Intelligent Tutoring Systems and Learning Outcomes: Two Systematic Reviews

by

Wenting Ma

M.A. Simon Fraser University, 2008
M.Sc. Simon Fraser University, 2006

Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Educational Technology and Learning Design Program
Faculty of Education

SIMON FRASER UNIVERSITY

Spring 2017

Approval

Name:	**Wenting Ma**
Degree:	**Doctor of Philosophy**
Title:	***Intelligent Tutoring Systems and Learning Outcomes: Two Systematic Reviews***
Examining Committee:	**Chair:** Sen Campbell Associate Professor

Philip Winne
Senior Supervisor
Professor

John C. Nesbit
Co-Supervisor/Supervisor
Professor

Engida Gebre
Internal Examiner
Assistant Professor
Faculty of Education

Firstname Surname
External Examiner
Professor
Computer Science
University of Saskatchewan

Date Defended/Approved: April 12, 2017

Abstract

Intelligent Tutoring Systems (ITSs) are computer programs that dynamically model learners' psychological states to provide individualized instruction. ITSs have been developed for diverse subjects to help learners to acquire domain-specific, cognitive and metacognitive knowledge at all educational levels. In this thesis, I report on two studies conducted to examine the current state of the ITS field. The first study is a meta-analysis conducted on research that compared the outcomes from students learning from ITSs to those learning from non-ITS learning environments. It examines 107 studies, published prior to 2013, with a total of 14,321 participants. The results show that ITSs outperform teacher-led, large-group instruction (g = .42), non-ITS computer-based instruction (g = .57), and textbooks or workbooks (g = .35). However, no statistically significant difference was detected between learning from ITS and learning from individualized human tutoring (g = -.11) or small-group instruction (g = .05). The second study evaluates research on the relative effectiveness of Bayesian networks in constructing student models in ITSs, which involves 143 studies published between 1992 and 2014. The study explores how Bayesian network was adopted to support the development of student models, relative to its strengths and weaknesses in investigating learning constructs and their contributions to the effectiveness of BN student modeling. A number of implications are drawn with respect to the application of BN in ITS design. Both reviews provide evidence that ITSs are relatively effective tools for learning. Furthermore, ITS researchers are invited to reconsider three fundamental research questions that have been examined since the emergence of ITSs and how they contribute to and constrain advances in effective ITS design in light of developments in artificial intelligence research. Finally, recommendations for future research directions are provided to researchers in the ITS community.

Keywords: intelligent tutoring systems; student model; Bayesian network; effect size; meta-analysis;

Dedication

To my family including Lei, my parents and parents-in-law

Acknowledgements

I would like to thank Dr. Phil Winne, my senior supervisor, who provided me with opportunities to learn about educational psychology and technology. I deeply appreciate what he has taught me since I became his graduate student. Not only have I learned a lot from his wealth of knowledge, but I have also gained hands-on experience in conducting quality research from being a part of his research team. Over the years, I have received great support, encouragement and guidance from Phil. I feel fortunate to have a senior supervisor from whom I can always learn and seek assistance and with whom I can openly talk.

I would also like to thank my supervisor, Dr. John Nesbit, whose input was critical to the writing of this thesis. Like Phil, John has been my supervisor since I started graduate studies at Simon Fraser University. John is knowledgeable and insightful. He is always willing to help me whenever I have questions or inquiries about any research that I am currently carrying out years. I am very thankful for his continuous support, his challenging of my thinking, and his encouragement in guiding me through my years of graduate studies. I have also gained valuable insights from his extensive knowledge and wisdom. John is not only my supervisor, but he is also like a good friend with whom I can speak whenever I encounter any difficulty, whether it be academic work or life in general.

Apart from my committee, I would like to thank Dr. Sheryl Guloy for proofreading my thesis at various stages of its development. She has helped me to express my thoughts more clearly in writing. With her strong background in the educational technology, Sheryl has also commented on my work as a peer and provided me with feedback on areas that I have not yet fully developed. Her insights have challenged my thoughts and helped to improve the quality of my work. I am very thankful that she was available to take on this task given her already busy schedule on concurrent projects.

I also want to acknowledge Dr. Olusola Adesope and Qin Liu for their contributions to the meta-analysis study I conducted in this thesis. With his rich experience in meta-analysis methodology, Dr. Adesope helped me to process the coding form in the Comprehensive Meta-Analysis software and analyzed the results in both fixed-effect and random-effect models. Qin has also assisted me to search for and code the new studies published during the period when I was working on coding and data analysis. I am very grateful for their assistance as they have greatly smoothed the long path for me to conduct this meta-analysis.

Finally, I would like to thank my family, friends, professors, and colleagues who have encouraged me throughout the years. Without their supports, I could not have gone through this long and challenging journey.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AI	Artificial Intelligence
BN	Bayesian Network
CAI	Computer-assisted Instruction/Computer-aided Instruction
CBM	Constraint-based Modeling
ICAI	Intelligent Computer-assisted Instruction/Computer-aided Instruction
ITS	Intelligent Tutoring System
MT	Modeling Tracing

Chapter 1.

Introduction and Background

1.1. Introduction

One-to-one instruction has long been promoted as a more effective approach than classroom teaching (Desmarais & Baker, 2012). Influenced by research on Artificial Intelligence (AI), Intelligent Tutoring Systems (ITSs) emerged in the 1970s as a more adaptive and individualized paradigm for computer-based instruction than its predecessors (Martin, 1999). Research on ITSs is interdisciplinary, spanning artificial intelligence, cognitive science, psychology, learning science and instructional technology. The field draws implications from multiple disciplines that creates both “challenge” and “richness” in its landscape (Nkambou, Bourdeau, & Mizoguchi, 2010, p.5). Over the past few decades, ITSs have been widely integrated into a large number of subject domains to support various learning activities from basic reading to comprehensive hands-on training such as PHP language programming (Weragama & Reye, 2013), managing the equipment in a thermal power plant (Hernandez-Leal, Sucar, Gonzalez, Morales, & Ibarguengoytia, 2011), and job interview training (Anderson, et al., 2013).

Given the rapidly growing number of students who seek online resources to satisfy their own learning agenda or professional training, it is widely hoped that learning environments can become more responsive and intelligently adapted to individual needs (Ciloglugil & Inceoglu, 2010). Since they first appeared in the 1970s, ITSs have been viewed as “one of the most promising approaches to deliver individualized instruction” (Ahuja, & Sille, 2013, p.40). By maintaining a robust cognitive model of the learner, ITSs are able to dynamically assess the learner’s knowledge and calibrate tutorial strategies to facilitate meaningful learning (Everson, 1995). An ITS is designed to play a tutoring role by understanding what students learn, how they perform over time, and offering timely intervention to assist them. Like many prior technologies, the wide adoption of ITSs has great potential to enrich “the learning opportunities of students” and provide them a wider field for “intellectual exploration” (Duchastel, & Imbeau, 1988, p.104).

Researchers in the ITS community have worked extensively on exploring how to make the intelligent tutor “more flexible, autonomous and adaptive to the needs of each student” (Conati, 2009). Nevertheless, building an ITS is a non-trivial task. It requires enormous efforts and human resources to develop and successfully implement inferential and analytical capabilities derived from artificial intelligence techniques. Moreover, for accurate modeling of student cognition, an ITS requires more than just a model of the student’s knowledge structure and the capacity to make diagnostic assessment of the student’s knowledge. The challenge of building a student model to precisely and fully capture learners’ characteristics demands a strong theoretical foundation in the ITS field (Desmarais & Baker, 2012).

Although a few meta-analyses of research on the effects of ITSs have already quantified how well these systems promote learning, a clear picture of how and why ITSs are effective has yet to emerge. In particular, each analysis had its own research agenda, focusing on a single subpopulation, subject area, or a narrow set of moderating variables. Furthermore, as the ITS field spans a number of distinct disciplines, great variation exists in the relatively extensive research literature regarding definitions, terminology, intellectual frameworks and conceptual interpretations of research findings. To understand the current research state in the ITS field, in this dissertation, I investigated the overall effectiveness of ITSs in assisting students to achieve their learning goals and explored variables that moderated these effects on student learning at a fine grained level. Specifically, I conducted two reviews. One is a meta-analysis that compares the outcomes of students’ learning from ITSs to the outcomes of students learning in non-ITS environments. Unlike previous reviews, it includes research across subject domains and all educational levels. By synthesizing 104 effect sizes extracted from a large range of empirical studies, I investigated the moderating variables that affect learning. In addition, I discussed how the findings of this research enrich the theoretical foundation of the field and draw implications for ITS research. I also conducted a second review that examined 143 research studies in which a Bayesian network was used for student modeling. By aggregating the data collected on various moderator variables in response to the study’s seven research questions, I identified a set of constructs tracked in student models and that supported pedagogical strategies. This forms the basis for drawing implications to advance the field’s understanding of ITS design.

1.2. Background

When Pressey's teaching machine first appeared in 1926, educators' understanding of how technology could be used for teaching began to expand (Pacella, 2014). Although this very first machine was mechanically inelastic, offering only a set of fixed questions and answers, it was designed with attention to pedagogical strategies and learning theories (Shute & Psotka, 1996). A new pedagogy, programmed instruction, emerged in 1954 when B. F. Skinner built the teaching machine and presented it at a conference on practical applications of behavioral science (Benjamin, 1988). Programmed instruction is an instructional methodology centered on Skinner's principle of stimulus control and reinforcement to shape behavior (Skinner, 1954). Instruction follows a linear logical sequence and decomposes content into well-defined small curriculum units (Gagné, 1965). Learning is supported through a systematic, reinforcing approach in which students can advance incrementally, receive immediate feedback to their responses, and be rewarded in a self-paced manner (Skinner, 1968).

Computer-assisted instruction (CAI) is a type of instruction that uses computers to deliver course content. Its roots lie in behaviorism, stimulus-response associations, and psychometric traditions (Edwards, 1970). Extending programmed instruction to present interactive text, CAI programs gradually evolved to support more sophisticated learning tasks using a combination of texts, graphics, sound, and videos as well as enhanced user interactions in all areas of the curriculum (Ward, 2002). CAI has been widely adopted to facilitate students' self-paced learning, supplement classroom activities and assist in measuring learning (Ramani & Patadia, 2013).

With the advent of cheaper and more powerful personal computers, CAI gained popularity and evolved in a wide array of instructional techniques. These progressed developmentally from linear CAI to branching CAI to more mature CAI, and then gradually transitioned to Intelligent Tutoring Systems (ITSs) as the need for individualized instruction expanded (Wenger, 1987). In this chapter, I briefly review CAI's development history, impact on instructional technology, and its evolution into ITSs.

1.2.1. Computer-Aided Instruction (CAI) Systems

Computer-aided Instruction (CAI) systems began in 1950s. It is a type of instruction that utilizes computers to assist individual learners in educational practice (Anderson, 1986). Also known as computer-assisted instruction, CAI is viewed to be rooted in Pressey's multiple-choice machine and the punchboard device presented at the meeting of American Psychological Association in 1925. CAI allowed a learner to work on a question and receive immediate feedback about its correctness, and it also tracked all the learner's attempts (Mann, 2009).

B.F. Skinner (1959), sometimes called the father of programmed instruction, invented a teaching machine in mid-1950s based on the principle of operant conditioning (Saettler, 1990, p. 296). As Skinner described, operant conditioning is a method that facilitates learning through a pattern of reinforcement. This theory "provided the scientific basis of programmed instruction" (McDonald, Yanchar, & Osguthorpe, 2005, p. 85). Stemming from the science of behaviorism, learning was seen to result from "an immediate and systematized reinforcement" that rewarded correct responses in a "stimulus-response-feedback" flow until "a prescribed level of proficiency is reached" (Anderson, 1986, p. 164).

Early CAI programs were based on linearly programmed content. Students followed a step-by-step path, traversing learning mathematics content that was divided into small units of information called a frame. If the answer to a problem was correct, a new frame advanced the student to the next question or topic was presented. Students could not progress until a correct answer was offered. Such an approach allows the student to progress according to how he/she works on individual questions and makes it possible for more advanced students to progress at a faster pace (Anderson, 1986). With its potential, programmed instruction soon became a buzzword in education and sparked lots of interest as reflected in research on a number of CAI initiatives which explored its efficacy in supporting learners by using a fundamentally learner-centered approach (McDonald, et al., 2005). Those early CAI initiatives were considered to be the antecedent of modern CAI programs (Ward, 2002).

A more advanced form of programmed instruction was originated by N. Crowder in 1959. He designed a CAI program that trained US Air Force members to troubleshoot problems in electronic equipment. His system implemented a branching approach in which a correct answer or a new instruction was selectively presented according to the previously given response (Sackney & Mergel, 2007). Different from Skinner's view of focusing only on reinforcing correct

behaviors, Crowder (1959) argued that a student should not only learn from correct answers but also from mistakes. He believed that doing this would promote learning through cognitive reasoning (Owen & Aworuwa, 2005). Therefore, in his form of branching programming, errors are anticipated by the system and students' erroneous responses are placed on a remedial learning path (Lockee, Moore, & Burton, 2004). Although the errors were not based on "any sort of analysis of the error patterns or procedural bugs", Crowder's approach was "the first use of errors in a tutorial system" (Lockee, et al., 2004, p.547). Branching programming saved a large amount of instructional time compared to linear programs, although no significant differences on the effectiveness of learning were found between linear and branching CAI programs (Larson, Burton, & Moore, 2008).

Since the 1960s, with its great flexibility to deliver courses and potential to alleviate challenges in instructing an increasing number of students, many CAI systems have been developed in the past few decades and applied across a broad spectrum of subject domains (e.g., arithmetic, chemistry, medical health, computer programming) spanning educational levels from elementary to higher education (Anderson & Skwarecki, 1986; Grimes, 1977; Kulik, Kulik, & Bangert-Drowns, 1985; Poole,1995; Ramani & Patadia, 2013; Ranade, 2006). Compared to early CAI systems, modern CAI systems can present interactive materials using graphics, text, audio and video to foster more engaging learning experiences. Moreover, the advent of more powerful and affordable personal computers further promoted adoption of CAI in classrooms (Arnold, 1997). The proliferation of CAI systems signifies it has grown from "an add-on" to "a learn-from technology" (Mann, 2009, p. 7).

Computer-based learning can enhance students' mastery of content (Rosenberg, Grad, & Matear, 2003). With the "emphasis on active learning, enrichment of collaborative learning, encouragement of greater student independency", CAI provides a novel way to deliver a large number of interactive opportunities for learning and engage students as active participants to achieve better learning outcomes (Basturk, 2005, p. 170). With high user interactivity and engagement of the past few decades, CAI technology has been recognized as an effective instructional method that can support teaching endeavors and improve students' overall academic results (Ranade, 2006; Wang & Wong, 2008). It has profoundly reshaped the relationship among students, teachers and computers as well as how students learn.

Despite CAI's positive influence on student learning, CAI systems were often criticized for the lack of individualized instruction. Although students work individually with a computer, since

early CAI systems are not designed to collect information about individual students and programmed to flexibly adapt instruction, it was impossible to offer authentic tailored feedback to support individualized instruction (Nwana, 1990). To address this need, since the beginning of 1960s, CAI systems have gradually evolved into “intelligent” CAI systems with the capability of handling individualized differences among learners.

1.2.2. Emergence of Intelligent Tutoring Systems

As discussed above, efforts to develop an “Intelligent Computer-aided Instruction System (ICAI)” began in 1960s and continued in the 1970s (Nkambou, Bourdeau, & Mizoguchi, 2010). At that time, AI technology “was in full bloom” and being applied in both computer and cognitive science. At the same time, CAI had matured as an instructional technology that was striving to overcome its limitations to support a larger number of students (Nkambou, Bourdeau, & Mizoguchi, 2010, p. 2). With Bloom’s report of the 2-sigma effect on individual tutoring (1984), the situation proved favourable for developing intelligent tutoring systems through a multidisciplinary approach that combined both domains (Carbonell, 1970; Nwana, 1990).

While CAI was rooted in Skinner’s behaviorist theory, ICAI systems diverged. They shifted from a “logic focus” to a knowledge-based system that “could make intelligent decisions based on prior knowledge” (Salman, 2013, p. 157). ICAI was designed to simulate a human tutor by adaptively responding to students’ individual needs with effective tutoring strategies (Anderson & Skwarecki,1986). Salgado-Zapata (1989) used the two words “adaptive” and “dynamic” to delineate the key characteristics of ICAI. According to him, the ability to adapt to students’ changing tutorial needs was the core attribute distinguishing ICAI from traditional CAI. Being adaptive to students’ evolving knowledge meant the system has to be dynamic enough to support various kinds of adaptive activities, eliciting useful information about changes in students’ knowledge over time (Vassileva, 1990). ICAI was considered to be a promising instructional tool that could spontaneously generate adaptive instruction, mimicking one-on-one human tutoring.

For historical reasons, research on intelligent computer-assisted programs was initially referred to by the term ICAI. This was gradually replaced by ITS when Sleeman and Brown (1982) proposed the term Intelligent Tutoring Systems (ITSs), describing these systems as representing a new generation of computer-based instruction that emphasized learning by doing and representation of learners’ knowledge. The term student modeling also was introduced to describe the goal of ITSs to support an abstract representation of student knowledge

(Hategekimana, 2008). In 1988, the first ITS conference was held. It offered a venue to “share and consolidate ideas” and stimulate research funding dedicated to work on ITSs (Nkambou, et al., 2010, p. 2). Since then, with the integration of artificial intelligence (AI) techniques into education, there has been an enormous growth in the field of ITSs spanning “a large spectrum of incremental developments” (Ahuja & Sille, 2013, p. 39).

The period of the 1970s to 1990 witnessed the first generation of ITS development (Nkambou et al., 2010). During this period, a series of ITS systems were built involving methods of Socratic tutoring, buggy libraries, genetic graphs, case-based reasoning, natural language processing, authoring systems and so on (Ahuja, & Sille, 2013). Among those early ITSs, the Cognitive Tutor developed at Carnegie Mellon University in early 1980s (Anderson, Corbett, Koedinger, & Pelletier, 1995) has become one of the most widely deployed. The Cognitive Tutor is rooted in Anderson’s Adaptive Character of Thought (ACT-R) theory. It builds students’ cognitive competency by developing relevant declarative and procedural knowledge in the context of use (Anderson, 1993). Students’ progress is tracked in each step with a model tracing technique tailored to provide the most appropriate intervention as needed (elaborated in section 2.2.2). Most of those ITS efforts were designed to imitate a human tutor and effectively communicate targeted knowledge to students. In 1990, Self claimed that emulating a human tutor as a goal was overly emphasized in the ITS research community and misdirected its development. Shute and Psotka (1996) echoed this view suggesting that it was not necessary to expect an ITS to communicate with students in the same fashion as human tutors do. Instead, with the support of AI techniques, the aim was to develop ITSs to move beyond only duplicating human tutors to capture teachers’ underlying reasoning process and multiple facets of students’ learning (Woolf, 2008). Furthermore, to offer students personalized learning experiences, ITS researchers were recommended to consider their field as engineering design grounded on a framework of relevant learning theories, methodology and techniques (Nkambou et al., 2010).

From 1990 to present, the field of ITSs has encompassed a great number of technological and research innovations, sparked with great vitality and vibrancy (Ahuja & Sille, 2013). Striving to deliver effective adaptive tutorial services to individual students, ITS researchers focused not only on modeling students’ learning performances or skills, they also extended the role of the student model to integrate a wide range of new constructs including learners’ knowledge about how they learn (metacognitive skills), motivation, engagement behaviors and affective states (e.g., Arroyo, et al., 2014).

As more complex learning environment became available, ITSs could support not only learners who study alone, at their own pace, but also facilitate collaborative learning activities for small groups of students by modeling group knowledge and learning interaction (Desmarais & Baker, 2012). For instance, Comet is an ITS designed to enhance students' clinical reasoning skills as they work in small groups in a problem-based learning (PBL) session (Suebnukarn & Haddawy, 2007). Comet actively monitors group activities and could intervene in students' group discussion with specific hints when they became stuck or lost focus in terms of the discussion topic.

Another important trend in the ITSs field contributing to better student modeling was the development of educational data mining (EDM) methods and machine learning (ML) in AI. EDM is a knowledge discovery process that uncovers novel learning patterns and potentially useful information from a large amount of traced student data (Guruler & Istanbullu, 2014). It is "concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in" (Baker, 2010, p. 324). With EDM methods, it is feasible to build more sophisticated student models by exploiting large volumes of data collected across a broad range of learner constructs leading to more precise prediction of students' ongoing state changes (Buchheit, Garrett, Lee, & Brahme, 2000). Machine learning techniques make it possible for a system to learn from experience, improve and evolve based on large-scale observations (Shapiro, 1992). ITSs use ML techniques to improve the student model and forge new tutorial strategies to more effectively adapt to students (Woofe, 2008). With ML techniques, ITSs are able to reason with uncertainty in a context of incomplete learner data to enhance instructional interventions (Hämäläinen & Vinni, 2006).

1.3. Structure of this Thesis

This thesis explores the development of intelligent tutoring systems and extends the work of prior meta-analyses of ITS research. It expands understanding of how this field has grown since it first emerged. Through a synthesis of years of research findings, this thesis provides insights to guide future research in this field. Specifically, it is structured in five chapters as follows:

In Chapter 1, I introduce what will be explored in this thesis and build a rationale for exploring the development of ITSs as a research field. Then I discuss the historical development of ITSs as they evolved from earlier computer-aided instructional systems. Lastly, the structure of this thesis is presented with a brief summary of each chapter.

In Chapter 2, I review literature in the ITS field. I begin by discussing the ITS definition used within this thesis. I also examine how four key components of an ITS afford personalized student tutoring. Then, I review the five major student modelling techniques widely applied in ITSs and discuss their relative strengths and weaknesses in representing students' learning characteristics and in providing help. I review four quantitative reviews that examined the effectiveness of ITSs. These four reviews provide an excellent overview of the current state of the ITS field regarding effects on student learning as compared to non-ITS learning environments and afford useful insights regarding moderator variables explored in research. I point out known limitations of meta-analysis that make it difficult to draw a general conclusion. This sets a stage for conducting a comprehensive, systematic review of student modeling to further explore the landscape of the field and guide future development. I review ITSs modelled using a Bayesian network (BN), given its popularity and purported strength to infer student behaviors and identify factors that affect student learning. I define what a BN model is and introduce three common types of BNs. Lastly, I examine the strengths and limitations of a BN in student modeling.

In Chapter 3, I report on the procedures and findings of a published meta-analysis that examined the overall effectiveness of ITSs involving 107 effect sizes with 14,321 participants (Ma, Adesope, Nesbit & Liu, 2014). I first describe criteria for selecting studies and explain the strategies for study coding and effect size extraction. Next, I elaborate on the procedure of data analysis following standard guideline for meta-analysis and explain interpretations of fixed-effect and random-effects models. Then, I report the findings in relation to each of the four research questions. Lastly, the results of the meta-analysis are further discussed with reference to the prior quantitative reviews. A number of suggestions are offered to inform the design of ITSs.

To include research studies that would otherwise be excluded by strict selection criteria, I further examine 143 ITS studies using a Bayesian Network for student modeling in Chapter 4. Firstly, I describe the methodology used to conduct this review. I start with the selection criteria and explain the search process used to identify and collect relevant ITS studies using a BN. Next, I introduce the coding form and learning constructs used to capture key characteristics in BN studies. Then, I explain the process of aggregating data to address the seven research questions

and analyze how applying a BN in student modeling facilitates individualized learning. Finally, I summarize the results and discuss the implications of this review for the future research work in the ITS field. The limitations and constraints of the current review are also reported.

In Chapter 5, the results of both reviews are summarized with regard to common and distinctive features found in each of them. With a large body of literature reviewed and synthesized, general implications are derived to understand current trends and recommend refinements to existing practices in ITS development. Moreover, ITS researchers are invited to reconsider fundamental research questions, upon which ITSs were designed. Then, the limitations and constraints of the two reviews are discussed. I conclude by recommending future research directions for the ITS community based upon the findings of both reviews.

Chapter 2.

Literature Review: Intelligent Tutoring Systems

2.1. What is an Intelligent Tutoring System?

Individualized instruction has long been valued in the education community (Stellan & Mitrovic, 2006). Bloom (1984) reported a two-sigma effect whereby average students who received one-on-one tutoring from expert tutors achieved scores two standard deviations higher on standardized tests when compared to peers receiving traditional classroom instruction. This result inspired and sparked the interest of the research community to develop an intelligent software system that could simulate human tutors and reproduce a similar result on a one-on-one basis (Desmarais & Baker, 2012).

In 1982, Sleeman and Brown coined the term Intelligent Tutoring System. They reviewed the research on CAI and made a clear distinction between ITSs and CAI systems (Santhi, Priya, & Nandhini, 2013). CAI systems allow only limited adaptation because of the constraints imposed by pre-scripted instructions, feedback and branching. On the other hand, to deliver tailored instruction to students, ITSs are systems capable of dynamically maintaining a model of student knowledge and updating it as a student progresses.

Although the term intelligent tutoring system was not explicitly used in the report, SCHOLAR is often considered the first ITS (Corbett, Koedinger, & Anderson, 1997). In 1970, Carbonell (1970) reported on a computer program, SCHOLAR, built to tutor students on South American Geography. SCHOLAR was able to ask students questions and offer feedback to their answers using limited mixed-initiative instructional dialogues in which students and the system took turns to lead the conversation. What distinguished SCHOLAR from other CAI systems at the time was that its architecture domain knowledge was explicitly represented as an independent component separate from the natural language interface (Ma et al., 2014). Carbonell (1970) explained that such a separation makes it possible to model individual student knowledge based on its representation of a particular subject domain, which provides SCHOLAR with the foundation needed to diagnose students' cognitive state of knowledge and offer individualized instruction.

Over two decades ago, researchers expressed scepticism about the feasibility of building an effective student model to provide personalized instruction (Desmarais, & Baker, 2012). The development of artificial intelligence (AI) offered hope to do just that. Since 1980, there has been an increasing interest in ITS research and a rise in the number of ITSs. A number of evaluation studies were published that compared the effectiveness of ITSs with that of software systems which used other instructional approaches. These studies compared systems that provided instruction, across a wide range of subject domains, involving learners across all educational levels (Ma et al., 2014). The emergence of ITSs significantly changed the landscape of educational technology. In the next sections, I elaborate on the definition of an ITS and its key components.

2.1.1. Definition

Although the term *intelligent tutoring system* is widely used in the research literature, many research studies still use alternative terms. This variation in terminology could result in the neglect of relevant research contributions (Steenbergen-Hu, & Cooper, 2014). As well, lack of consensus about a unified definition of ITS may also lower readers receptivity to implications arising from this review. It is, therefore, sensible to develop a clear definition for ITSs.

Many efforts were made to distinguish an ITS from other computer-based instructional systems based on its capability to generate adapted instruction. In 1982, Sleeman and Brown defined ITSs as “adaptive systems which use intelligent technologies to personalize learning according to individual characteristics such as knowledge of the subject, mood and emotion” (p. 2). Shute and Psotka (1996) asserted that for an ITS, “the most critical element is real-time cognitive diagnosis (or student modeling) [and] the next most frequently cited feature is adaptive remediation” (p. 14). In 1999, Self claimed that “ITSs are computer-based learning systems which attempt to adapt to the needs of learners and are therefore the only such systems which attempt to ’care’ about learners in that sense” (p.350). More recently, Conati (2009, p.2) suggested that “ITS is the interdisciplinary field that investigates how to devise

educational systems that provide instruction tailored to the needs of individual learners, as many good teachers do". Similarly, Pacella (2014) also emphasized that, in addition to possessing expertise on the subject, an ITS must be able to store student information including prior knowledge and progress in addition to conducting meaningful assessment for each step and presenting tailored learning materials to students.

For the sake of this analysis, I adopt the ITS definition excerpted from Ma et al. (2014, p. 902) to guide the discussion of ITSs throughout this thesis.

"An ITS is a computer system that for each student:

1. Performs tutoring functions by (a) presenting information to be learned, (b) asking questions or assigning learning tasks, (c) providing feedback or hints, (d) answering questions posed by students, or (e) offering prompts to provoke cognitive, motivational or metacognitive change
2. By computing inferences from student responses constructs either a persistent multidimensional model of the student's psychological states (such as subject matter knowledge, learning strategies, motivations, or emotions) or locates the student's current psychological state in a multidimensional domain model
3. Uses the student modeling functions identified in point 2 to adapt one or more of the tutoring functions identified in point 1"

2.1.2. Key Components of an ITS

Although ITSs vary in features of the user interface, subject domains and learning variables modelled, some essential architectural components serve the core functionality of teaching students by adapting to changes in knowledge. In the first ITS, SCHOLAR, Carbonell (1970) implemented three core modules to support individualization: the expert module, the learner module and the instruction module. All three modules were connected together to generate limited mixed-initiative instructional dialogues to communicate with students.

Many early ITSs followed Carbonell's three-component model (e.g., Polson & Richardson, 1988). Later on, Dede (1986) extended this model by specifying the user interface as the fourth module in an ITS. While a user interface is not really a new module, per se, as any ITS relies on it to directly communicate and interact with students and to collect student data to be utilized in other modules in the system, recognizing the user interface as a stand-alone component in the ITS architecture highlights its importance in supporting a smooth learning experience. It also reflects the increasing complexity of user activities found in more advanced ITSs. In addition, with the same three modules that Carbonell (1970) proposed, Santhi, Priya and Nandhini (2013) defined a fourth module named Control Engine. In the proposed ITS, the control engine provided the visualization of the domain model and the student model to analyze various learning constructs of learners. Also, Sani and Aris (2014) proposed a similar four-module ITS architecture, which enables interaction with students through an interface module and the collection of data used for adapting and personalizing instructional assistance.

Based on the ITS architecture commonly reported in the literature, I summarize the following four components that orchestrate the delivery of tailored support to individual students (e.g. Ma et al., 2014; Sani, & Aris, 2014; Wenger, 1987), which is depicted in Figure 2.1.1:

1. A domain model

This model represents all knowledge that the designer intends to be learned by students. This includes declarative and/or procedural knowledge such as concepts, logical statements, topics, rules, and question banks etc. It is sometimes called the expert domain or expert knowledge module in some ITSs.

A domain model generally contains specific knowledge elicited in great detail from domain experts. A knowledge domain serves as a source of learning goals and as a standard to evaluate students' performance and knowledge during learning (Salman, 2013). Consequently, problem solutions must be generated in the same context experienced by students so their solution steps can be compared to the standard. Also, multiple solution paths and evaluation criteria must also be identified in a domain module to allow for evaluating variations in students' solutions (You, Liu, Long, & Pan, 2013)

2. A student model

This model reflects the student's most current knowledge state. It captures characteristics of the student's learning and dynamically assesses changes in the student's knowledge as instruction proceeds. Relevant aspects of a student's learning include but are not limited to the student's cognitive knowledge, exam scores, learning preferences, affect, metacognition and learning behaviors. It is the core component that enables an ITS to understand a student somewhat like a human tutor and to offer adapted instruction accordingly.

Similar to a domain model, a student model serves, in addition to its representational function, as an important source of student information, feeding other components in the ITS about how students' progress and mastery of knowledge compare to the expert domain knowledge. In some ITSs, the student model can also infer students' learning patterns and model students' misconceptions. Through maintaining a library of students' mistakes and suboptimal behaviors, a student model provides a diagnostic source for other ITS components to generate pedagogical supports for remediation (Kass, 1989).

3. A tutor model

A tutor model, also called a pedagogical model in some ITSs, offers tailored instruction and prepares suitable learning content for students. It plays a central role in pedagogical interventions and is responsible for "moment-by-moment adaptation" (Ohlsson, 1986, p. 293). From the data collected in the student model, the tutor model interprets student behaviors and synthesizes information about students to determine the pedagogical supports to provide in the next moment. In each instructional moment, it addresses the specific "changing cognitive needs of the individual learner" and intervenes in students' activities when necessary (Ohlsson, 1986, p. 293). Pedagogical actions should be selected to ensure students receive appropriate instructions in a timely manner and are progressing on the right track. Because varied pedagogical decisions relating to when and how to use learning materials leads to distinctive learning experiences and outcomes, choosing suitable content in the appropriate context at the right time requires great versatility (Martens & Uhrmacher, 2004).

4. A user Interface module

This module is the “front-end” component of an ITS. It facilitates communications and interactions between the system and students. It displays relevant domain information to students and collects their input through the interface. It also supports students as they navigate throughout the system and allows them to receive instructional responses and feedback generated from the tutor model. Depending on the nature of interactions, it enables students to work flexibly at their own paces and receive instructions for adapting their learning. As components of an ITS such as tutoring model are mediated by the interface to directly communicate to students, a good interface design in terms of user friendliness and presentation layout can greatly impact students’ acceptance of the system and can potentially influence the effectiveness of an ITS in promoting learning outcomes.

Not all ITSs have all four distinct components and they may vary in level of complexity (Conati, 2009). Some models may be packaged in one component in such a way that performs the functions but is technically integrated in one module. Nevertheless, tracking students’ progress and adapting instruction to their needs are the core features of ITSs that fundamentally distinguish them from the other types of CAI systems (Ma et al., 2014).

Representing a student is not an easy task. Students vary dramatically in their prior knowledge, cognitive ability and learning environments. Their knowledge also dynamically changes. Therefore, to accommodate individual students with the most appropriate instructions at each learning moment requires accurately reflecting the students’ current knowledge state, diagnosing the root causes of their errors and offering personalized assistance to those experiencing difficulty (Stellan & Mitrovic, 2006). These requirements demand a mechanism to systematically maintain and update students’ information – student modeling. In the next section, I discuss a few common student modeling techniques widely used in ITSs.

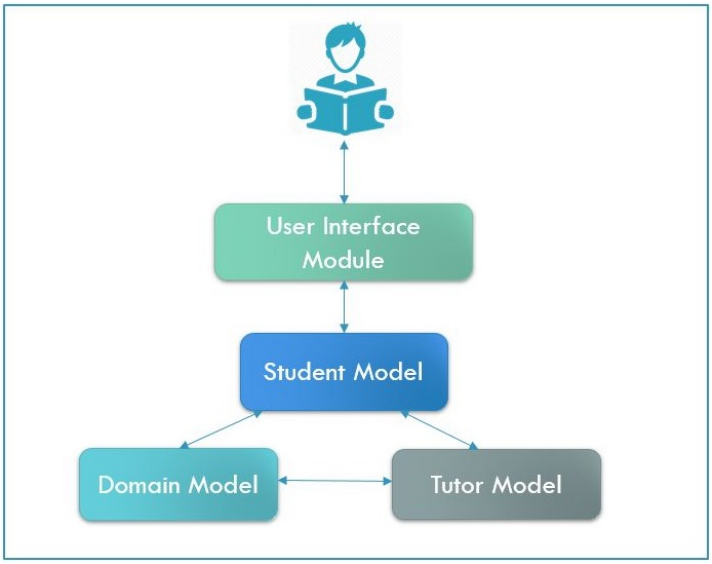

Figure 2.1-1. A Typical Architecture of an ITS

2.2. Types of Student Modeling in ITSs

As previously discussed, the most salient characteristic of an ITS lies in its capability to diagnose students' current state of knowledge on the fly and offer the most suitable instruction to support their consequent cognitive development. The diagnosing facility of an ITS is the student model. O'Shea and Self (1983) defined the student model as a program that contains specific information about the student being taught. This information could range from a simple count of how many incorrect answers have been given to some complicated data structure, which purports to represent a relevant part of the student's knowledge of the subject. Therefore, a typical student model functions to capture the student's most current state on relevant aspects of learning and provides the data needed for the system to tailor instruction towards the student's learning goals.

In this section, I review five major student modeling techniques that are most influential and widely applied in the research of the ITS community. The origin, development, strengths and constraints of the modelling techniques when adapting instruction to students are discussed.

2.2.1. Overlay Modeling

Overlay modeling was initially developed by Stansfield, Carr and Goldstein (1976) for a CAI gaming program to coach students on using logic and probability to play the game WUSOR. Overlay modeling has since been used in many user modeling systems (Virvou, 2003). This student modeling technique structures students' state of knowledge as a subset of expert knowledge (Nwana, 1990). Specifically, a student's knowledge is represented as "a network of tick-marks" that is "laid over the representation of the subject matter to show which part he/she has learned" (Ohlsson, 1986, p.297). In this modeling approach, knowledge is structured into learning components or elements such as concepts or topics. By representing students' conceptual knowledge independently, the overlay modeling technique facilitates the prediction of students' current state of knowledge by comparing the differences between what the system has collected about a particular students' level of knowledge and the system's model of the subject domain (Chrysafiadi & Virvou, 2015). Depending on the design of ITSs, the mastery of each learning element could be determined qualitatively with Boolean values such as learned *or* not learned or quantitatively based on the probability that a student has acquired it (Brusilovsky & Millán, 2007). Instruction is accordingly adapted to students' current level of knowledge.

With the overlay modeling technique, a student's knowledge becomes "a progressively more complete subset of the expert's knowledge units" (Ohlsson, 1986, p.297). This technique assumes that a novice becomes an expert only when the student has mastered the content defined in the expert domain. However, incorrect user behavior or knowledge may not necessarily be associated with incomplete knowledge but rather with misconceptions (Nwana, 1990). Moreover, expert knowledge is not attained by simply filling in the gaps in student knowledge. Rather, it involves a more complicated learning process in which students reflect upon what they learn and refine this along the way (Ciloglugil & Inceoglu, 2012). Therefore, overlay modeling is often criticized for being too simplistic in its assumptions about students' knowledge and as being unable to respond appropriately to their misconceptions (Ciloglugil & Inceoglu, 2012, p.555).

In response to this limitation, the bug model aims to capture both students' correct and incorrect/buggy knowledge (Brusilovsky & Millán, 2007). The most widely applied bug model is

perturbation modeling, which is used to diagnose students' incorrect knowledge and erroneous steps leading to mistakes, thereby creating a bug library. A bug library consists of a collection of incorrect perturbations generalized from the common errors students commit (Chrysafiadi & Virvou, 2015). This model allows a set of incorrect perturbations to be associated with individual elements of the domain knowledge (Brusilovsky & Millán, 2007). When a student commits an error, this error is identified as a perturbation in student knowledge, which is then used to diagnose the reason for the error to provide instructions needed to guide the student's learning. Compared to the overlay model, a perturbation model facilitates more effective remediation of students' incorrect knowledge and provides them with individualized guidance.

2.2.2. Model Tracing

Cognitive Tutors refer to a series of intelligent tutoring systems (e.g. Pump Algebra Tutor, LISP tutor) developed at Carnegie Mellon University since 1982. Cognitive tutors have been widely deployed in hundreds of classrooms at secondary schools, colleges and universities (Koedinger, Anderson, Hadley& Mark, 1997). Based upon extensive research on artificial intelligence and cognitive science, cognitive tutors are used in a variety of subject domains including algebra, geometry, and computer programming to help students develop problem-solving skills (Koedinger, 2001). Rooted in Anderson's Adaptive Character of Thought (ACT-R) theory (Anderson, 1993), cognitive tutors were implemented with the model tracing (MT) technique that keeps track of each step students take in the problem-solving process and offers support when students take unsuccessful courses of action.

The ACT-R theory provides a description of how human cognition works (Koedinger, 2001). According to this theory, "human cognition is complex" and complexity is viewed as the "complex composition" of "simple knowledge units" acquired through "relatively simple principles" (Anderson & Schunn, 2000, p. 2). In other words, it "arises from an interaction of procedural and declarative knowledge" (Anderson, 1996, p. 355).

The ACT-R theory makes a clear distinction between declarative and procedural knowledge (Koedinger et al., 1997). Declarative knowledge refers to factual knowledge that "can

be directly accessed including facts, concepts, pictures and stories" (Koedinger & Anderson, 1998, p.162). ACT-R theory defines declarative knowledge as "a network of small units of primitive knowledge called chunks" (Anderson & Schunn, 2000, p. 3), or "knowledge units" (Anderson, 1996, p.357). Procedural knowledge, on the other hand, consists of "a large number of rule-like units called productions" (Anderson & Schunn, 2000, p. 3). It is featured as "goal-oriented and mediates problem-solving behavior" (Corbett & Bhatnagar, 1997, p. 245). Essentially, declarative knowledge is regarded as the "direct encoding of things in our environment" whereas procedural knowledge is the "direct encoding of observed transformations" (Anderson, 1996, p. 364). The effectiveness of learning thus relies on the "amount of knowledge encoded and the effective deployment of the encoded knowledge" (Anderson, 1996, p. 355).

Grounded in ACT-R theory, a cognitive skill is defined to consist of both "goal-related domain knowledge" and "goal-independent procedural knowledge" (Steenbergen-Hu & Cooper, 2014, p.331). The acquisition of a cognitive skill is achieved through a process of applying respective declarative knowledge in the context of problem-solving activities (Koedinger & Anderson, 1998). When a cognitive skill is mapped to procedural knowledge, it is encoded as a series of independent production rules in an "if-then" construct and problem-solving goals (Anderson, 1996; Corbett & Bhatnagar, 1997). By matching these rules to the student model, a cognitive tutor is able to track the steps of students' cognitive processing, diagnose their misconceptions, and offer appropriate instructions to remediate their knowledge (Fournier-Viger, Nkambou & Nguifo, 2010). This process is called *model tracing*.

In MT, production rules are "condition-action units which respond to various problem-solving conditions with specific cognitive actions" (Anderson & Schunn, 2000, p. 3). Their "conditions and actions" are described "in terms of declarative structures" (Anderson, 1996, p. 356). A typical production rule in a cognitive tutor describes "how to retrieve declarative knowledge to solve problems" in an "if-then" format (Anderson, Matessa, & Lebiere, 1997, p. 439). It operates by matching "the problems' goals and current state" and by articulating "new sub-goals" (Sklavakis, & Refanidis, 2008).

Anderson (1993) provided an example of a production rule for programming recursion, which involves "responding to some goal, retrieving information from declarative memory, and possibly taking some action or setting a subgoal" (Anderson, Matessa, & Lebiere, 1997, p. 441). This production rule only actualizes when its conditions are "satisfied by the current knowledge in

declarative memory" and may lead to "creating new declarative structure" (Anderson, 1996, p. 356). Anderson's (1996) production rule follows:

> **IF** the goal is to identify the recursive relationship in a function with a number argument
>
> **THEN** set as subgoals to
>
> 1. Find the value of the function for some N
> 2. Find the value of the function for N- 1
> 3. Try to identify the relationship between the two answers. (p. 355)

Thus, for step 1, when N = 5, factorial (5) = 120; for step 2, N-1=4, factorial (4) = 24. With all its conditions satisfied, it could lead to a new declarative structure that factorial (N) = factorial (N-1) x N. By "the firing of such production rules" in specific problem-solving contexts in MT, knowledge acquisition occurs "step by step" in the activation process with regard to the deployment of encoded knowledge (Koedinger et al., 1997, p. 441). It involves generating a large number of production rules relating to task goals, task states and actions to achieve those goals (Koedinger et al., 1997).

The strength that MT has for student modeling is the large amount of traced performance and reasoning data that are captured while students are learning. Thus, MT is sensitive to students' progress, thereby enabling it to provide a wide variety of effective tutoring services. For instance, with the dynamic evaluation of a student's current knowledge state, it has the information needed to recommend appropriate steps for further problem-solving and adapting the pace of instruction to students' individual progress. When a weak knowledge area is identified, it offers "targeted, high-quality remediation" to fix the knowledge gap according to the growing complexity of goal structure for that student's learning (Kodaganallur, Weitz, & Rosenthal, 2005, p.117). MT enables students to learn in the context of problem-solving activities through the learning-by-doing paradigm (Koedinger et al., 1997).

One well-known limitation of the model tracing technique involves the step-by-step tracking of students' problem-solving behaviors. Thus, it requires a large number of production

rules and solution paths to support this effort. Developing just one production rule for a cognitive tutor requires 10 hours or more of development, depending on the complexity required (Anderson, et al., 1995). Therefore, it is time-consuming and expensive to develop such a cognitive model. Furthermore, MT can only be applied to those domains in which tasks related to problem-solving strategies or solutions are explicit and clear. For many ill-defined domains such as law, writing or language learning, when strategies for solutions cannot be easily defined, it is practically impossible to map the domain knowledge to prescribed rules or actions (Fournier-Viger, Nkambou & Nguifo, 2010).

In sum, MT offers a robust student modeling technique to trace performance accurately and to identify misconceptions. By representing students' competency in production sets, it situates students in problem-solving contexts where they can master targeted and complex cognitive skills (Anderson et al., 1995).

2.2.3. Expectation and Misconception Tailoring (EMT)

AutoTutor is an ITS that engages students in natural-language dialogues for tutorial instruction, and it has been widely applied in many subject domains (e.g. computer literacy, physics, behavioral research method, etc.) for over two decades of research development (Wolfe, et. al, 2013). Due to the inherent complexities of natural language processing and open-ended attributes of human conversations, using tutorial dialogues to exchange ideas has been challenging for ITS research. Technical breakthroughs in the fields of computational linguistics, informational retrieval, artificial intelligence, and discourse processes have made it feasible to capture dialogue patterns regarding domain knowledge, self-explanations, goals, questions, arguments and other forms of discursive dialogue during knowledge construction (Graesser, et al., 2004).

AutoTutor was built under the assumption that deep learning occurs when learners actively participate in information processing in which they self-explain what they learn and justify relevant causal relationships (VanLehn, Jones, & Chi, 1992; Wolfe et al., 2013). Natural language dialogues (NLP) can facilitate this process through "some conversational contexts" that involve "imprecise verbal content, a low to medium level of user knowledge about a topic, and earnest literal replies" (Graesser et al., 2004, p.181). Following what most human tutors do in tutoring,

AutoTutor adopted a tutorial NLP that simulates the kind of discursive exchanges between human tutors and students when scaffolding learning, namely *expectation and misconception tailored dialogue* (EMT). EMT-based tutors such as AutoTutor encourage students to articulate and explain their learning and reasoning when solving particular problems (e.g., Graesser, Person, Harter, & the Tutoring Research Group, 2001). EMT-based and human tutors are similar in that correct answers are expected for particular questions or appropriate steps are anticipated in a procedure; these are called *expectations* in EMT. *Misconceptions* are also detected when tutors track students' reasoning in problem-solving activities and are subsequently correct following the system feedback received. Typically, an EMT-based tutor has a list of expectations and misconceptions for corresponding key questions in the subject domain and coaches students through dynamic dialogue exchanges. Both *expectations* and *misconceptions* form the domain model in EMT tutors and are analogous to MT tutors in that expectations are similar to expert knowledge whereas misconceptions are similar to buggy libraries.

AutoTutor uses an animated conversational agent to deliver dialogues whereas students submit responses by using a keyboard (Graesser et al., 2004). Each expectation is associated with a set of pumps, hints and prompts, presented in order by AutoTutor, to encourage students to elaborate on their answers while advancing conversational moves (Graesser, Olney, Ventura, & Jackson, 2005). The agent follows the "pump → hint → prompt → assertion cycle" in scaffolding students to complete all expectations (D'Mello, & Graesser, 2012, p.10). While dialoguing with students, AutoTutor periodically verifies if an expectation is successfully covered by requiring students to articulate the answers. When misconceptions in reasoning are detected, relevant instructional supports such as short feedback, corrections, and summaries are presented to help the student revisit concepts and correct erroneous statements (Nye, Graesser, & Hu, 2014). If the student still fails to answer the question, an assertion with correct information is provided to the student and AutoTutor sets an appropriate expectation adapted to the knowledge state of the student. This process repeats until all expectations are met and all misconceptions are corrected for all requisite problems.

AutoTutor uses Latent Semantic Analysis (LSA), with its "conceptual pattern-matching algorithm", to evaluate the quality of students' discursive input as measured by expectations and misconceptions (Graesser et al., 2004, p.185). LSA is a statistical mechanism that measures the semantic similarity of texts such as words, paragraphs, and essays. Consequently, it is often used for pattern recognition, pattern matching, and pattern completion operations (D'Mello, & Graesser,

2012; Graesser et al., 2004). In EMT-based tutors, LSA is used to evaluate the probability that an approximate semantic match between the students' responses and the ideal answer has been achieved (Graesser et al., 2005; Wolfe et al., 2013). Accordingly, the tutor is able to adapt the instruction given to students based on the analysis derived from this comparison (Kopp, Britt, Millis, & Graesser, 2012; Wolfe et al., 2013).

In AutoTutor, the LSA algorithm was used to calculate "the proportion of expectations covered, using varying thresholds of cosine values on whether information in the learner essay matched each expectation" (Graesser et al., 2004, p.185). Specifically, it defines that each Expectation $\boldsymbol{E_i}$ is considered fully learned by a student when "the content of the student's cumulative set of turns meets or exceeds a threshold $\boldsymbol{T}$ in its LSA cosine value (which varies from near 0 to 1)", which means that "$\boldsymbol{E_i}$ is covered if the cosine match between $\boldsymbol{E_i}$ and the student input $\boldsymbol{I}$ (including turns 1 though N) is high enough: cosine ($\mathbf{E_i}$, $\mathbf{I}$) ≥ $\mathbf{T}$" (D'Mello, & Graesser, 2012, p. 12). Similarly, when a student input $\boldsymbol{I}$ matches a misconception $\boldsymbol{M}$ with a match score higher than the threshold $\boldsymbol{T}$, the student is diagnosed with the misconception. To make the model more accurate, the LSA settings can be fine-tuned to achieve the best match between input texts and expectation texts.

The quality of interactions between learners and tutors impacts the quantity and quality of learners' responses when answering questions through mixed-initiative dialogues, thereby affecting their learning performance (Wolfe et al., 2013). The EMT framework supports a discourse pattern that engages students in deep reasoning and effective knowledge construction and makes feasible the measuring of their learning outcomes at a fine-grained level.

2.2.4. Constraint-Based Modeling (CBM)

Constraint-based modeling (CBM) is another prominent technique for student modeling widely used in intelligent tutoring systems (e.g. Mitrovic, Martin& Suraweera, 2007; Mitrovic, 2012). Since the first CBM tutor SQL-Tutor was developed in 1995, CBM has been extended to support not only domain knowledge modeling but also student knowledge modeling, affect, metacognitive skills, and collaborative skills (Mitrovic, 2012). Grounded on Ohlsson's (1994) constraint-based modeling and the theory of learning from errors (1996), CBM is fundamentally different from the model tracing approach in that it represents knowledge "in the form of

constraints", specifying "abstract features of correct solutions" instead of "generating problem-solving paths" and "performing tasks in a domain" through a set of pre-defined production rules (Mitrovic, Martin, & Suraweera, 2007, p. 38).

The underlying assumption of CBM is that correct solutions will never violate any corresponding principles in a domain (Mitrovic, 2012). In other words, when a specific solution does not violate any pre-defined constraints, "it is deemed correct" (Mitrovic, Martin & Suraweera, 2007, p. 39). Unlike MT, CBM does not attempt to capture the exhaustive correct and incorrect actions that students take while learning. Rather, CBM fundamentally focuses on "the domain principles that *every* correct solution must follow" (Mitrovic, 2012, p. 64). Therefore, it is only necessary to define a set of constraints when mapping specific solutions to particular domain principles and to provide appropriate feedback to students' whenever their responses violate these principles. The sequence of students' actions that leads to errors is not central to the CBM model. Although CBM provides students with remediating feedback, this modeling technique is not dependent upon students' misconceptions. Therefore, the student model based on CBM does not "represent" a student's actions but the "effects" to which his/her actions have led (Mitrovic, 2012, p. 41).

Each constraint consists of three elements: a relevance condition, a satisfaction condition and a set of feedback messages (Martin, 1999). It is represented "in an ordered pair (C_r, C_s)" in which C_r is the relevance condition and C_s is the satisfaction condition" (Mitrovic, 2010, p. 41). Stellan & Mitrovic (2006) defined the general form of a constraint as follows:

> **IF** the properties C_r hold
> **THEN** the properties C_s have to hold also (or else something is wrong) (p.8).

The relevance condition describes the context in which the constraint is applicable. The satisfaction condition defines the condition that holds true for a correct solution to satisfy the relevance condition (Koedinger & Anderson, 1998). The feedback messages are what the tutor provides to students to repair their knowledge when a violation is committed (Martin, 1999). It is notable that the two conditions specified in the constraint form have a loose connection with each other, only suggesting that if C_r is true, then C_s "ought" to be true (Mitrovic, 2010, p.65).

In CBM, there are two types of constraints: syntactic and semantic (Martin, 1999). A syntactic constraint represents the "syntactic properties of the domain" and examines the syntax of students' queries (Stellan & Mitrovic, 2006, p. 8). A semantic constraint is defined by the domain

and used to match the meaning of a student's solution to the ideal one to evaluate its correctness and is generally more complex than a syntactic one, (Martin, 1999; Stellan & Mitrovic, 2006). Each constraint is a knowledge unit associated with "a very specific error" (Martin, 1999, p. 31). Therefore, constraints afford being "evaluative" and "making [a] judgement" about solutions and thus enable the identification of students' errors. In comparison, production rules describe the actions to take when the goals and conditions in the IF clause are met and are thus often regarded as being "generative" in nature and unable to explicitly diagnose students' errors (Mitrovic, 2010, p. 65). Sometimes, however, a constraint allows for multiple solutions and, in such cases, it can also be used to evaluate alternative solutions, similar to how MT tutors use production rules to generate all possible correct solutions (Mitrovic, 2012).

In the CBM paradigm, constraints are used to define domain knowledge as well as to represent students' knowledge. In a CBM student model, each constraint has three counters that record its relevance to students' answers. Specifically, it includes the number of times the constraint is "relevant for the student solution", the number of times it is "relevant to the ideal solution", and the number of violations students commit (Martin, 1999, p. 32). Such information is used to determine the appropriate set of new problems that students should be given in relation to their mastery level. A record of satisfied and violated constraints are stored in the short-term student model and used to reflect the cognitive state of the student as he/she progresses in the subject over time (Mitrovic, Martin & Suraweera, 2007).

Compared to several student modeling techniques, including MT, CBM is "computationally simple" and only requires "pattern matching" by using constraints to compare students' solutions to pre-defined correct solutions (Mitrovic, Martin & Suraweera, 2007, p. 39). Therefore, CBM is particularly useful in domains in which the order of the procedural steps when problem-solving is not critical to finding the solution; e.g., database design (Stellan & Mitrovic, 2006). Furthermore, CBM is flexible in that students can explore alternative solution paths as long as doing so does not violate the prescribed constraints in a given domain. CBM, thus, supports creativity among students (Mitrovic, 2010). Since CBM models a knowledge domain with all ideal solutions, "in terms of pedagogically significant state", at an abstract level, it greatly reduces the amount of authoring work required to design and implement a CBM tutor (Mitrovic, 2012).

One well-known limitation of CBM in student modeling is that it focuses primarily on the relevance and satisfaction conditions of constraints in a domain and does not "consider it important to know how the student arrived at a specific problem state" (Koedinger & Anderson,

1998, p.166). Therefore, with insufficient knowledge of students' procedural steps, CBM-tutors are incapable of determining whether a student solves the problem as result of true understanding or just by guessing. This limits CBM's ability to recommend the future steps that learners must take (Fournier-Viger, Nkambou & Nguifo, 2010). Nevertheless, CBM has developed, over the years, from a theoretical idea to a mature student modeling technique with unique strengths in offering adaptive tutoring support to learners in both well- and ill-defined domains and tasks (e.g. Suraweera & Mitrovic, 2004). CBM has also been combined with other modeling approaches to reach optimal learning outcomes (Mitrovic, 2012).

2.2.5. Bayesian Network Modeling

Bayesian networks (BNs) have been widely used in many intelligent tutoring systems as a powerful student modeling technique. For instance, Conati and Zhao (2004) used a BN to model grade 6 and 7 students' cognitive and affective states as they practice number factorization in the educational game Prime Climb. Andes is another intelligent tutoring system that uses a BN. Andes was developed to help college students in their homework to improve the learning of physics (VanLehn, Lynch & Schulze, 2005). Students' problem-solving steps were tracked by a static Bayesian model, wherein nodes and links represent the extent of students' mastery. Suebnukarn & Haddawy (2007) described a collaborative tutoring system Comet for medical problem-based learning (PBL). Comet also uses a BN and interaction log to model the hint strategies commonly used by human tutors in medical PBL to guide students to construct relevant case hypothesis; to track respective learning activities; and to develop clinical-reasoning skills (Suebnukarn & Haddawy, 2007).

As opposed to many student modeling techniques, a Bayesian network (BN) is also a "graphical and probabilistic modeling framework" that affords "high representative power" in the form of a networked structure, which can "be derived from data" and, thereby, reduce "the need for substantial knowledge engineering" (Desmarais & Baker, 2012, p. 16). BNs offer robust probabilistic computational power to handle uncertainty surrounding observations made as students' learn and to support the diagnostic analysis of their cognitive states during learning (Santhi, Priya, & Nandhini, 2013). A more detailed description of BN student modeling is elaborated in section 2.5 of this chapter.

2.3. Prior Quantitative Reviews of ITSs

In this section, I summarize and compare the results of four quantitative, analytical articles that have been recently published on ITSs. The reviews of these are organized chronologically in terms of publication date. I elaborate upon the primary findings of these reviews to provide insights on design science and the domain of ITSs. The effect sizes reported in these reviews were calculated and interpreted with regard to the standardized mean difference between groups.

2.3.1. A Review of the Effectiveness of Tutoring Systems

VanLehn (2011) conducted a quantitative analysis comparing the effectiveness of human tutoring, computer tutoring and no tutoring on STEM subjects. This review covered 95 comparison studies, published between 1975 and 2010, and assessed effects across conditions. To explore the effectiveness of interaction granularity, VanLehn categorized computer tutoring into three sub-types: substep-based tutoring, step-based tutoring and answer-based tutoring.

VanLehn (2011) found that human tutoring, when compared to the no tutoring condition, only yields an effect size of 0.79 standard deviation units. This result is surprising because it is far from the 2.0 sigma effect reported in the Bloom (1984) studies. After investigating Bloom's studies more closely, VanLehn explained that the observed effect could be attributed to the mastery learning expectations that tutors held for their students. After reviewing other studies that compared human tutoring to no tutoring, with the highest reported effect size being 0.82, he further asserted that the mean effect size of human tutoring on students should be closer to 0.79. In reference to the effect size of d = 0.76 in step-based tutoring, VanLehn concluded that ITSs could be nearly as effective as human tutors and posited that the effects of ITSs can be increased by improving "the [controlling] parameters [of]...its pedagogical decision-making" (p.213). The overall effectiveness of all tutoring types range from 0.31 to 0.79.

VanLehn (2011) also noted that when the interaction granularity in user interface decreases, the effect size increases accordingly; e.g., from answer-based tutoring to step-based tutoring, the effect size of ITSs increases from 0.31 to 0.75. However, there seems to be an effect ceiling to the interaction granularity; when it reaches the "peak" effect, further decreasing it

produces little increase in the effect size. This observation appears to be consistent with VanLehn's conclusion that the mean effect size of computer tutoring is close to 0.79.

In addition, Van Lehn's review also suggests that human tutoring, substep-based and step-based computer tutoring all yield comparable effects of ITSs on student achievements because all three types of tutoring support students in bridging their knowledge gaps and in scaffolding them toward correcting mistakes on their own (VanLehn, 2011). Therefore, VanLehn (2011) recommended that more step-based tutoring systems for STEM courses be developed to help students with doing their own homework.

2.3.2. A Review of the Effectiveness of ITSs on K-12 Students

Steenbergen-Hu and Cooper (2013) conducted a meta-analysis evaluating the effectiveness of ITSs in K-12 mathematical learning. This analysis involves 34 empirical studies, published between 1997 and 2010. The meta-analysis compared the effectiveness of ITSs with studies mainly conducted in traditional classrooms as well as some studies on human tutoring or homework practices. The mathematical learning under review includes basic math and algebra for a range of students, from elementary to high schools.

Overall, the researchers concluded that ITSs show no statistically significant effect on K-12 students when compared to traditional classroom instruction, using a random effects model, with a Hedge's g that ranged from 0.01 to 0.09. For the few studies that compared ITSs with human tutoring or homework practices, a small to modest effect, ranging from 0.20 to 0.60, was found on the effectiveness of ITSs over these two conditions. Although these results were consistent with previous reviews in educational technology, the researchers reported that the effect size of ITSs in the current meta-analysis seems to be smaller than those found in other reviews. A plausible explanation for the small effect could relate to whether ITSs were used as the primary instruction or supplemented with other instructional methods. The researchers cautiously concluded that computer technology is more instructionally effective when used as a tool to support rather than to replace traditional teaching and practices.

Several moderator variables were explored to understand the impact of ITSs on student achievements over other instructional modes. By analyzing the *ITS duration* variable,

Steenbergen-Hu and Cooper (2013) found out that the influence of ITSs is greater for students with less than one school year as compared to those with one school year or more. Steenbergen-Hu and Cooper concluded that this difference may have resulted from the higher motivation levels that students experience when first using a novel ITS and that can "wear off" over time (p. 984). It is also possible that the researchers' difficulty in maintaining close involvement in the daily instructional activities, over the duration of the study, may have had an impact on the extent to which ITSs were used during learning. Steenbergen-Hu and Cooper (2013) also found that ITSs were found to be less effective for student groups identified as lower performers in comparison with general student groups. This finding raises concerns that the use of educational technology could potentially widen the learning gap experienced by students who are already disadvantaged relative to other students in terms of their performance, aptitude, or background. Similar findings have also been reported in other reviews (e.g. Ceci& Papiero, 2005).

Steenbergen-Hu and Cooper (2013) identified three salient issues in relation to the methodology of a study. First, the effectiveness of ITSs is related to the timing of the evaluation –student performance was found to be the highest when data were collected before the end of the school year. Second, Steenbergen-Hu and Cooper observed that ITSs were reported as more effective in relation to course-related outcomes rather than in relation to standardized tests. Third, studies with smaller sample sizes reported greater average effect sizes than those with larger sample sizes. As suggested by the researchers, these results echo similar findings in previous reviews of educational technology, which found that varying methodological features (e.g. experiment duration) could potentially impact the "magnitude" of effect sizes of study interventions (p. 984). In my view, this study has provided insight on the specific contexts and ways in which an ITS can be designed to promotion better learning outcomes.

2.3.3. A Review of the Effectiveness of ITSs on College Students

One year later, Steenbergen-Hu and Cooper (2014) published another meta-analysis on the effectiveness of ITSs for college students' academic learning. This analysis included 39 empirical studies, published between 1990 and 2011, with 22 different types of ITSs. A variety of

subject domains were covered by this study, including physics, statistics, computer science, mathematics, business (accounting and economics) in higher education.

Overall, this meta-analysis revealed that ITSs have a moderate positive effect on college students' achievement, with an effect size that ranges from *g* = .32 to .37. Although ITSs were found to be less effective when compared to human tutors, they produced higher learning outcomes than all other types of instructional modes; e.g., traditional classroom instructions, textbooks, computer-assisted instructions, lab, etc. The effects of ITSs were found not to be significantly distinguishable from other instructional modes on comparison variables such as different ITSs, subject domains, duration of ITS treatment, or degrees of use. These results confirm findings from previous reviews in that computer-assisted tools were generally more effective than traditional instruction in influencing achievement in learning within higher education.

The researchers also compared these results to findings from the previous meta-analysis they had conducted for K-12 students on mathematical learning. Steenbergen-Hu and Cooper (2014) suggested that ITSs may have been more effective for the older learners because they generally had better prior knowledge and learning skills in a computer-assisted learning environment than the younger learners had. Consequently, the older learners may have benefited more from having used ITSs than the younger ones. Thus, in addition to methodological design, duration of the intervention, and experimental settings, the grade level of learners is another variable that must be considered when investigating ITSs.

Based upon an analysis of moderator variables, Steenbergen-Hu and Cooper (2014) identified two key findings. One is that the ITS effects were significantly higher in studies conducted during the earlier years of schooling than those conducted later. However, the researchers cautioned that this result is not conclusive because grouping studies based on time is subject to arbitrary or subjective categorizations, thereby potentially leading to variable results. The researchers also discussed the relationship among the teacher, pedagogy, and ITSs on learning outcomes. Steenbergen-Hu and Cooper found that teachers and teaching pedagogy play critical roles in boosting the effectiveness of ITSs in computer-assisted learning environments, drawing attention to the need for the future exploration of this research stream.

Three areas of exploration for future ITS researchers were identified by Steenbergen-Hu and Cooper (2014). Based upon the observed effects of ITSs on learning, the first area is that ITS researchers pay more attention to less structured or ill-defined subject domains as studies have

mostly been focused on well-defined domains such as computer programming, physics, and algebra. Based upon the results of the moderator analysis, the second area involves the exploration of how the level of instructor involvement and the relevance of pedagogical activities facilitates the effective use of ITSs during learning to maximize the benefits of ITSs, such that optimal learning outcomes are achieved. Also, with data supporting the finding that an ITS is a good metacognitive tool to support self-regulation, the third area of study recommended by the researchers involves research questions that explore how ITSs can scaffold or facilitate the development of self-regulatory skills that foster learning.

2.3.4. A Review of ITS in Computer Science Education

Following the holistic analysis of the effects of ITSs comparing the learning outcomes of ITS and non-ITS instruction across all disciplines (Ma et al., 2014), Nesbit et al. (2014) conducted another meta-analysis focusing particularly on computer science education, as ITSs have been most commonly applied to this subject domain. This analysis covers 22 studies, published between 1998 and 2013, involving 1,447 participants. The topics include programming languages (C, C++, Java, etc.), computer literacy, database design, software design, and system security.

The researchers found that, overall, the use of ITSs in computer science education yields a moderate, positive effect size, g = .46, with a standard error of the mean .05. To fully understand why ITSs are more effective than non-ITS instruction, five key moderators (comparison instruction, student modeling, instructional use of the ITS, use of feedback, and misconceptions modeling) are identified to explore what may have contributed to the effectiveness of ITSs, at a fine-grained level, using a fixed-effects model. For comparison instruction, the result suggests that ITSs led to a significantly better learning outcome than two kinds of non-ITS instruction including teacher-led, group instruction (g = .67) and non-ITS, computer-based instruction (CBI) (g = .89). The researchers suggested that the effectiveness of ITSs over those two instructional modes could be associated with the student modeling component in ITS. It is possible that personalized instructions and feedback, optimized learning path and task assignments, which are supported in the ITS' student model, may have assisted students in learning content more deeply, at a finer level of granularity, and helped them in overcoming learning challenges in a more timely fashion than had the non-ITS instructional methods .

Based on this review, an insufficient number of studies necessary to conclude that ITSs lead to better achievement than one-to-one human tutoring or textbooks/workbooks exists. Still, ITSs have been found to greatly outperform other types of instructions, whether delivered as the primary instructional method (g = .45) or combined with other instructional activities (g = .55). This result indicates that, regardless of whether misconceptions were modeled (g = .41) or not modeled (g = .68), ITSs were more effective than the non-ITS modes included in this analysis. No statistically significant difference was found between the two effect sizes of misconceptions modeling in ITSs.

For student modeling, using constraint-based ITSs has been found to be more effective than non-ITS studies, with a small but statistically significant effect. For other commonly reported types of student models such as model tracing and Bayesian network, an insufficient number of studies exists to conclude that ITSs are more favorable. To evaluate the influence of the various student modeling methods, the researchers recommend that more theoretical work on defining and categorizing these heterogeneous methods be done. Among the 22 studies, only one study did not provide feedback. Instead, adapted tasks were assigned to students in this study. A non-significant effect size was found. This result is inconsistent with that reported in Ma et al. (2014), who found that ITSs with only adaptive task assignments produced better learning outcomes than non-ITS modes. To understand this discrepancy, the researchers recommended that more studies be conducted to explore the impact of adaptive task assignments on student performance. It was also suggested that more instructional functions could be implemented in ITSs to provide more adaptive instruction that support students' individualized learning experiences.

Overall, this review provides a great opportunity to tap into the most commonly taught subject domain, computer science, in ITS research. The power of ITSs over non-ITS systems on student performance was evident as a result of its adapting of tutoring strategies over time, based on an analysis of the individual ITS tutoring components. ITS researchers were also urged to work on a common framework to understand and evaluate different student modeling techniques and to pay more attention to the effects of individual instructional components when conducting future ITS studies.

2.4. Rationale: the Need for a Comprehensive Review in Student Modeling Techniques

The four quantitative reviews of ITSs evaluate the effectiveness of ITSs over other instructional methods and have brought many insights that can be applied to ITS research. They covered a wide variety of subject domains and included a large number of students who belong to a diversity of age and educational levels. They also examined a number of moderator variables and how these contributed to the effectiveness of ITSs and drew conclusions that ITS researchers can use to design more useful ITSs in the future. The insights gained from the recommendations and suggestions in the analyses in these reviews are valuable and well received.

In general, a meta-analysis is a statistically powerful means to conduct primary reviews on evaluation studies and identify patterns in the study domain. Its strength lies in its being able to synthesize all available evidence from multiple independent studies and extract useful information to test hypotheses (Greco, Zangrillo, Biondi-Zoccai, & Landoni, 2013). Thus, meta-analysis provides a rigorous and formal method by which to integrate effect sizes and provides results that can be used to derive useful insights from the data. However, a meta-analysis review is limited in its coverage of all literature that is published in a field because included studies must meet strict selection criteria. Among the four meta-analysis reviewed, researchers were only able to select those empirical studies that included at least one comparison group against which the ITS treatment group could be compared. Therefore, these meta-analyses excluded a large number of longitudinal empirical studies that evaluated the effectiveness of ITSs through pre- and post-tests within the same group of participants. Thus, only a small subset of all ITS evaluation studies were reviewed. Furthermore, researchers were often unable to generalize their findings because of the limited number of studies included in their analyses. Consequently, these researchers needed to interpret their results with caution. For instance, Steenbergen-Hu and Cooper (2013) emphasized that the influence of differences in the duration of interventions on the differential effectiveness of ITS should be interpreted with caution, given that only 31 independent studies were analyzed. Further research is required to confirm the conclusion.

In addition, three out of the four reviews reported the relative effectiveness of ITSs either in a specific subject domain (e.g., computer science in Nesbit, et al., 2014) or a specific student group (e.g., college students in Steenbergen-Hu and Cooper, 2013). The other two reviews focus on studies comparing ITSs and non-ITS conditions. None of the four analyses pay detailed attention to the student model techniques, which is a central concern of ITS research.

Considering this gap in the current reviews, I propose that a need exists for a comprehensive review of student modeling techniques in the field of ITSs. Bayesian networking is a powerful graphical and probabilistic modeling framework, which has been widely used in a wide range of applications (Desmarais & Baker, 2012). With its popularity in many ITSs for student modeling, BN provides a good starting point for further examining the contributions of student modeling methods to the effectiveness of ITSs.

In the next section, I introduce the origin, formalism, and development for a Bayesian network, and discuss its usefulness in handling uncertainty and predicting future events in complex domains. Then, I expand the discussion on three specific BN approaches to student modeling and their applications in ITS. Furthermore, I present the strengths and limitations of BNs in student modeling with the aim of depicting its characteristics to understand its potential in supporting future ITS developments.

2.5. Bayesian Network for Student Modeling

Bayesian networks (BNs), also known as *belief networks*, are an efficient approach to manage uncertainty in artificial intelligence (Santhi, Priya, & Nandhini, 2013). They belong to the family of "probabilistic *graphical models* (GMs)" (Ben-Gal, 2007, p. 307,) and are a "graphical description of a probability distribution that permits efficient probability propagation combined with a rigorous formalism" (Santhi, Priya, & Nandhini, 2013, p.3).

The theory of BNs was originally developed from the work of Thomas Bayes, an 18th century mathematician and theologian, who published 'An essay towards solving a problem in the doctrine of chances' in the *Philosophical Transactions of the Royal Society of London*, 53: 370-418;' in 1764. This essay "contains a special case of Bayes' Theorem…concerned with conditional probabilities" (Holmes& Jain, 2008, p.1). The conditional dependencies of variables in BNs are calculated using statistical and computational methods; therefore, BNs are considered to have a multidisciplinary origin in areas such as "graph theory, probability theory, computer science, and statistics" (Ben-Gal, 2007, p. 307).

Since the late 1980s, BNs have been widely applied to user modeling and student modeling (Conati, 2010). It is a powerful modeling technique that has been well received in areas such as data mining, machine learning, speech recognition, medical diagnosis, natural language processing and so on (Ben-Gal, 2007). BNs are capable of handling uncertainty in knowledge representation, reasoning and inferences, and diagnostic and pattern recognition (Ramírez-Noriega, Juárez-Ramírez& Martínez-Ramírez, 2016).

Student learning is often considered a source of unreliable information (Ben-Gal, 2007). The process of reasoning and diagnosing students' current level of knowledge and mental state involves making inferences based on uncertain observations, behaviors and measurements because of the dynamic nature of students' interaction with resources, instructors and learning environments (Chrysafiadi, & Virvou, 2015). The embrace of BNs lies in its capability to handle uncertainty in student modeling, "encoding expert knowledge, and performing automatic probability update in light of new evidence" (Ting, Khor & Sam, 2012, p. 576).

In the following sections, I elaborate on the details of a BN to provide background on its origin and development. I first introduce the definition of BNs. Following that, I discuss three primary types of BNs and their applications in ITS. I then explore the strengths and constraints of BNs in student modeling and provide examples of how they are employed to handle uncertainty efficiently when addressing real-life problems.

2.5.1. What is Bayesian Network (BN)?

A typical BN provides a theoretical framework, based on probability theory, for handling uncertainty in artificial intelligence (Santhi, Priya, & Nandhini, 2013). It is a graphical representation of a probability distribution over a set of random variables in a given domain (Gamboa & Fred, 2002; Holmes& Jain, 2008). The graphical structure of variables in a BN is generally called a *directed acyclic graph* (DAG), which consists of a set of nodes representing random variables and a set of directed edges indicating direct dependence among those variables (Ben-Gal, 2007). Thus, it combines "graph theory and Bayesian inference" to collect evidence and update the current belief in the given network (Nguyen & Do, 2009, p. 42).

The DAG structure of a BN is often considered the “qualitative” aspect of a BN model whereas the network parameters are generally described as the “quantitative” aspect of it (p.307, Ben-Gal, 2009). The parameters are defined as “a set of local distributions combined with a set of conditional independence assertions”, representing a *joint probability distribution* (JPD) (Santhi, Priya, & Nandhini, 2013, p.453), with each node’s conditional probability distribution depending only on its parents (Ben-Gal, 2009).

A formal definition of a BN is given below by Conati (2002, p. 282):

“a Bayesian network is a directed acyclic graph where nodes represent random variables and links represent direct dependencies among these variables. If we associate to each node X_i in the network a conditional probability table (CPT) that specifies the probability distribution of the associated random variable given its immediate parent node’s parents (X_i), then the Bayesian network provides a compact representation of the Joint Probability Distribution (JPD) over all the variables in the network.”

$$\mathbf{P}(X_1, \ldots, X_n) = \Pi^{n}_{i=1} \boldsymbol{P}\,(X_i \mid Parents(X_i)) \quad (1)$$

The joint probability $P(X_n)$ is calculated in Formula 1, with the node X_i conditionally independent from each other. As Gamboa & Fred (2002, p.453) suggested,

“The BN structure encodes the assertions of conditional independence as a directed acyclic graph such that: (a) each node corresponds to a variable; (b) the parents of the node corresponding to X_i are the nodes associated to the variables in Π_i. The pair formed by the structure (graph) and the collection of local distributions, $P(X_i \mid \Pi_i)$, for each node in the domain, constitutes the Bayesian Network for that domain”. Therefore, a BN can be fully defined by its graphical structure and associated variables (Heusch, 2007).

The structure of a BN can be viewed as a knowledge base, representing the beliefs of variables in a system and the relationships between them. When an event occurs in the system, BN is able to collect them as evidence, infer probabilities given the changes and the existing dependencies among the variables, and propagate beliefs throughout the network (Mihajlovic & Petkovic, 2001). The structure of the acyclic graph and semantics of a BN determine that each variable/node in the graph is “conditionally independent from its non-descendent given its parents” (Ghahramani, 1998, p.169). It ensures that no node can be its own ancestor or descendent. Such

an attribute serves to reduce the number of network parameters required to “characterize the JPD of the variables” and to provide a more efficient approach to “compute the posterior probabilities given the evidence” (Ben-Gal, 2009, p. 307). Although the links between variables in BNs represent causal relationships that are unidirectional, the propagation triggered by the reasoning process of BNs can be performed *in any direction* (Ben-Gal, 2009).

2.5.2. Types of BNs

In this section, three primary types of BNs are discussed in detail. For each type, one or two examples of how the BN is applied to student modeling in an ITS are given. The strengths and constraints of each type of BN are then explored to determine the way a model for future event predictions can be optimized.

Static BNs

A typical BN provides an intuitive way to perform JPD over a set of random variables in a directed acyclic graph, wherein links represent the casual relationship between those variables (Heusch, 2007). A *static* Bayesian network is one type of BN that is characterized by its running of probabilistic inference over variables with values that do not evolve over time (Conati, 2010). What is changed in a static BN is “the belief over the state of these variables” when new evidence of state changes in the existing variables is added (Conati, 2010, p. 283).

A static BN is one type of graphical model that represents events and objects in the real-world and defines them as a set of random variables with different states (Mihajlovic & Petkovic, 2001). The links between those variables are represented in directed edges and reflect the causal relationship between the observed evidence represented by those variables (Heusch, 2007). Each variable in the BN is associated with a probability function, representing the conditional independence defined by “the edges leading into the variable” (Mihajlovic & Petkovic, 2001, p. 4).

A noteworthy property of a static BN is that it is generally easier to develop and maintain, when compared to other types of BNs, e.g. Dynamic BN, as it generally contains a smaller set of parameters and a simpler network structure. Therefore, static BNs are computationally more efficient and faster than a DBN. Because of this attribute, a static BN is often used to initiate network parameters and to expand over time with the increasing complexity required for dynamic processes (Ricks & Mengshoel, 2010).

In student modeling, this attribute of a static BN is often used as a tool to evaluate students' knowledge (Conati, 2010). For instance, the student model in Andes, an ITS for Newtonian physics, was built in a static BN, representing task-specific knowledge and making inferences on the probability that mastery has been achieved by a student (Conati, Gertner, & VanLehn, 2002). When a student starts working on a specific exercise, all related inferential information for this problem is loaded from the pre-defined problem solution graph to the static BN structure including its physics rules, relevant solution steps, and tutoring strategies. Based on the last exercise they solved and their previous solution steps, students are evaluated on their capability to apply specific physics rules and solution elements and the probabilities of the mastery of corresponding knowledge (Conati, Gertner, & VanLehn, 2002).

A static BN is not capable of dealing with temporal information. It requires external resources to process time-related information to be encoded into the collected evidence (Ricks & Mengshoel, 2010). To overcome this challenge, BNs are extended to deal directly with temporal and sequential variables and model dynamic causal influences over time in complex systems.

Dynamic BNs

As previously discussed, traditional BNs cannot handle temporal information (Hernandez-leal, et al., 2011). The dynamic Bayesian network (DBN) is a modeling technique used to represent dynamic domains with temporal or sequential data (Vlasselaer, Meert, Van Den Broeck, & De Raedt, 2016). Since DBN was proposed by Dean and Kanazawa (1989), it has been widely applied to represent and make inferences about sequential events in discrete time (Kwon, & Suh, 2012). As time proceeds, dynamic transitions require effective temporal updates over the network in complex domains. A DBN is capable of capturing such "dynamic causal influences between

covariates" and supports the changing nature of a dynamic domain as the system evolves over time (Song, Kolar, & Xing, 2009, p. 1732).

Typically, a DBN consists of "a series of time slices" that capture the values and states of all variables at a given time (Sucar, 2015, p. 161). For each time slice, it defines "a dependency structure" between the variables as the "base network" and repeats this structure for all time slices (Sucar, 2015, p. 161). DBN can be viewed as a set of "slices of a static BN over time", which connect with one another via links (Hernandez-leal, et al., 2011, p.39). Specifically, a static BN is created for a specific instance. This structure is then repeated over time. The individual static BN variables are connected via arcs across different temporal time slices (Ziani & Motamed, 2007).

One drawback of DBN is that it can grow very complex. It is unnecessary to use it in models that require only a few changes in most variables (Hernandez-Leal et al., 2013). Moreover, DBN only allows for a fixed number of time intervals between stages; therefore, it is not able to handle models requiring varying levels of time granularity (Hernandez-Leal et al., 2013). In the next section, a temporal node Bayesian network (TNBN) designed to address challenges associated with DBN is introduced to explain how it can be used to manage processing and reasoning of temporal information with a less complex computational requirement.

Temporal BNs

The temporal node Bayesian network is another type of BN that extends the traditional BN to build less complex BNs for temporal reasoning in dynamic domains (Arroyo-Figueroa, & Sucar, 1999). It is an event-based modeling technique and consists of a set of temporal nodes that offers a "compact graphical representation" of domains and defines "time intervals in which events can occur" (Fiedler, Sucar, & Morales, 2015, p. 578). Unlike DBN, A TNBN represents changes of states at different times instead of the changes of state values. It is capable of handling multiple levels of "time granularity" and is used to "manage uncertainty and temporal reasoning" (Hernandez-Leal et al., 2013, p. 956). Therefore, it is powerful technique to be applied to domains to handle uncertainty in measuring data over time, e.g., it has been used for medical diagnosis, which involved analyzing "large amounts of longitudinal data" (Orphanou, Stassopoulou, & Keravnou, 2014, p. 134).

In a TNBN, there are two types of events: instantaneous and temporal. An instantaneous event is one that has no time delay before it occurs. This means its child event occurs right after

its parent one takes place. A temporal event refers to one that occurs with possible time delays. A temporal node models state changes and time duration with a set of temporal intervals as well as "state changes of a variable" (Hernandez-Leal et al., 2013, p. 956), with each node defined by an ordered pair of the value of the variable and the time interval regarding the value change for that variable (Arroyo-Figueroa, & Sucar, 1999). The intervals can vary in their number and size which allows for multiple levels of time granularity. The temporal nodes are connected by edges representing the probabilistic causal-temporal relationship between them (Arroyo-Figueroa, & Sucar, 1999). In each node, there is no absolute temporal timing reference because its time intervals are relative to its parent node.

Formally, Fiedler, Sucar, and Morales (2015, p. 579) defined a temporal node as follows:

Definition 1 A temporal node is a random variable defined by a set of states each characterized by an ordered pair (λ, τ) where λ is the value taken by the random variable and τ is the temporal interval $[t_i – t_f]$ (t_i and t_f are the start and end times of the interval, respectively) in which the state change occurred. A default state of no change that corresponds to the event "not occurring" is also associated to every temporal node. There is at most one state change for each temporal node in the temporal range of interest.

Definition 2 Let V be a set of instantaneous and temporal nodes, and E a set of edges between those nodes. A TNBN is a pair $B = (G, \theta)$ where $G = (V, E)$ is a directed acyclic graph (DAG), and θ is a set of conditional probability distributions that quantify the network.

In general, TNBN is an effective probabilistic graphical model that represents "temporal relationship between events and their state changes" and predicts the evolution of dynamic temporal processes in discrete time.

2.5.3. Strengths of BNs in Student Modeling

BNs are a remarkably powerful modeling technique for uncertainty management (Ben-Gal, 2007). In general, BN offers several advantages for modeling reasoning under uncertainty

and for inference processing. Conati (2010, p.287) summarized the following strengths a BN possesses for student modeling when compared to other modeling techniques:

- They provide a more compact representation of the joint probability distribution (JPD) over the variables of interest.

- Algorithms have been developed that exploit the network's structure for computing the posterior probability of a variable given the available evidence on any other variable in the network. While the worst case complexity of probabilistic inference in Bayesian networks is still exponential in the number of nodes, in practice it is often possible to obtain performances that are suitable for real-world applications.

- The intuitive nature of the graphical representation facilitates knowledge engineering. It helps developers focus on identifying and characterizing the dependencies that are important to represent in the target domain. Even when dependencies are left out to reduce computational complexity, these decisions are easy to track, record and revise based on network structure, facilitating an iterative design-and-evaluation approach to model construction.

- Similarly, the underlying network structure facilitates the process of generating automatic explanations of the results of probabilistic inference, making Bayesian networks very well suited for applications in which it is important that the user understands the rationale underlying the system behavior, as it is often the case for Intelligent Tutoring systems (e.g., Zapata-Rivera and Greer, 2004).

- Finally, Bayesian networks lend themselves well to support decision making approaches that rely on the sound foundations of decision theory. This means that selection of tutorial actions can be formalized as finding the action with maximum expected utility given probability distributions over the outcomes of each possible action and a function describing the utility (desirability) of these outcomes (e.g., Murray et al., 2004; Mayo &Mitrovic, 2001).

In addition, in the form of a network graph, a BN's nodes "can represent not only random variables, but also hypotheses, beliefs, and latent variables" (Ben-Gal, 2007, p. 5). Therefore, it is easy for a BN to represent "both causal and probabilistic semantics" in its structure. Given this flexibility, nodes can be used to represent a variety of student characteristics including "knowledge, misconceptions, emotions, learning styles, motivation, goals" in a student model (Chrysafiadi, & Virvou, 2015, p. 9). This attribute makes it possible to integrate both "prior/expert knowledge" and observed data "seamlessly …within a single network" (Mayo & Zealand, 2001, p. 127).

Overall, BNs are a powerful modeling technique to explore the "undetermined relationships among variables" and describe "these variables upon discovery" (Niedermayer, 1998, p. 126).

2.5.4. Limitations of BNs in Student Modeling

While BNs powerfully address challenges in the inferential process and in managing uncertainty, there are still some inherent difficulties in applying BNs as an approach for student modeling. Firstly, it is computationally complex to explore an unknown network in BNs and discovering network involves an "NP-hard task" (Murray, 1998, p. 425), which may lead to an extremely high computational cost, depending on the number and combinations of variables that need to be calculated and inferred (Niedermayer, 1998). For instance, efforts were made to simplify the network structure and CPT of Andes, an ITS based on BNs to teach students how to solve physical problems, to develop a simpler BN. However, the result of the performance testing on simulated students still showed the infeasibility of the modified models because they still entailed a high computational inference on the model (Conati, Gertner & VanLehn, 2002). Even after optimizing the algorithm, it still created delayed response times on the performance of larger networks. Therefore, it is critical to verify, on occasion, whether the probabilistic update in the network is empirically acceptable and, thus, feasible in practice (Conati, 2010).

In addition, the selection of a proper distribution model to describe the data and the setting required for network parameters has a notable impact on the efforts needed to build a BN for

student modeling in an ITS (Murray, 1998). Insufficient or inaccurate data, required to set prior probabilities of a BN and define the values of network parameters (conditional probabilities), could lead to an unreliable network or even distort the entire network, resulting in an unreliable inference and incorrect evidence for a probabilistic update (Niedermayer, 1998). In addition to populating parameters, when new evidence is observed, it takes effort to implement algorithms to propagate probabilities over the network, which increases the knowledge engineering efforts required to build an effective student model (Millán, Pérez-de-la-Cruz, & García, 2003).

Furthermore, network parameters are either updated from nodes and variables or are estimated by domain experts. The approach to enriching network parameters from student data is preferred over acquiring expert judgement, which could be costly and subject to human errors. Although much research on techniques to conduct probability elicitation has been conducted, an elicitation process is fairly time-consuming and, thus, sometimes practically infeasible (Conati, 2010). Also, while many nodes and variables in a BN, developed within the context of student modeling, might seem observable, it could be difficult to collect them in practice. For instance, a variable such as the emotional state of a student would make defining network parameters difficult, especially with a new ITS (Conati, 2010).

These concerns aside, BNs are still regarded as a remarkably powerful technique to support complex inference modeling, causality analysis and statistical induction in a wide range of areas (Niedermayer,1998) and, thus, provide insights on considerations for future ITS developments on how student modeling adaptations can be optimized.

Chapter 3.

Overview of the Effectiveness of ITSs: A Meta-analysis

Prior ITS reviews show ITSs promote learning and assist students in achieving better learning outcomes than other modes of instructions (see Sections 2.3.1-2.3.3). Yet, those previous meta-analyses are limited to only the subsets of ITS research that cover selected subjects and educational levels. Having a comprehensive meta-analysis that examines a larger number of studies has a greater statistical power to detect the overall mean effect size or the effect size at different levels of each moderator variable, when compared to a set of smaller sized analyses for the same collection of research studies. By including more evaluation studies in the collection pool, a meta-analysis' results could also be more conclusive. In addition, to compare the effect sizes of two or more categories across different studies requires that two or more levels of a moderator variable exist in the same meta-analysis. Only in this way can we conclude whether differences between levels are statistically significant. Therefore, to attain a holistic view of effects of ITSs on student learning, a comprehensive meta-analysis of ITSs is conducted to compare student performances in ITSs to non-ITS learning environments in all subject domains across all educational levels.

In the following sections, I describe the purpose and analysis of a comprehensive meta-analysis, including: the procedure of selecting a pool of evaluation studies in ITS, conducting analyses of moderator variables and examining ITS effects at a fine granular level. This comprehensive study was published in Ma et al. (2014).

3.1. Purpose of the Study and Research Questions

This review incorporates and synthesizes research studies on the relative effectiveness of ITSs and examines the respective moderator variables to examine how these contribute to the effects of ITSs. Specifically, I address the following research questions:

1. Do students using ITSs have different learning outcomes from students using other modes of instruction?

2. Do the effects associated with ITSs vary with characteristics of the ITSs?
3. Do the effects associated with ITSs vary with characteristics of the students, outcome assessments, and research setting?
4. Do the effects associated with ITSs vary with the methodological features of the research?

3.2. Method

3.2.1. Selection Criteria

Selection criteria were applied include studies in the meta-analysis if they:

(a) reported original data;

(b) assessed learning outcomes after students interacted with software that matched the definition of an ITS presented in the introductory section of this review;

(c) compared learning outcomes from the ITS with outcomes from a non-ITS mode of instruction;[1]

(d) were publicly available, either online or in library archives;

(e) reported sufficient data to calculate effect size; and

(f) reported measurable cognitive outcomes such as recall, transfer, or a mix of both.

[1] Studies that compared group learning from ITS with a control group that received no instructional treatment were retained, but these studies were meta-analyzed separately to provide interpretive context for the results of the principal analysis that bore more directly on the research questions.

3.2.2. Search, Retrieval, and Selection of Studies

A comprehensive search for relevant research was conducted in ERIC, PsycINFO, Springer Link, and Web of Science. The search returned 26,613 titles using these key terms the fields: intellige* tutor*, intellige*, agent, cognit* tutor*, adapt* tutor*, cognit* virtual companion, and intellige* coaching system*. The reference sections of review articles on ITSs were manually searched for studies to add to the selection pool (Arnott, Hastings, & Allbritton., 2008; Conati, 2009; VanLehn, 2011; Wang et al., 2008; Steenbergen-Hu & Cooper, 2013; Steenbergen-Hu & Cooper, 2014).

During initial screening phase, abstracts of articles were read to identify studies that fit criteria a, b, c and d. 362 articles were then identified and respective full-text copies were further evaluated against all six inclusion criteria. Finally, a total of 107 studies, published prior to 2013, were found to match the inclusion criteria with a total of 14,321 participants. These were coded following a predefined coding form and coding instructions developed for this meta-analysis. All effect sizes were calculated with Hedges' correction for bias due to small sample sizes (Lipsey & Wilson, 2001).

3.2.3. Study Coding and Effect Sizes Extraction

The coding form includes 44 fixed-choice items and 37 comment items to capture detailed information about the studies including:

- author,
- year published,
- source of the study,
- research questions,
- type of ITS,
- control treatment,

- grade level of participants,
- research settings,
- duration of the study,
- reliability reporting, and
- statistics for computing the effect size of each study.

For studies that compared the learning outcomes of more than two groups, the coding strategy avoided repetitively counting the control group, which can lead to statistical dependence among contrasts and inflate the overall weight in the study (Borenstein, Hedges, Higgins & Rothstein, 2009; Lipsey & Wilson, 2001). Specifically, when there is more than one control group in a study, the control group receiving no instructional treatment is dropped from the main meta-analysis according to selection criterion c. Other comparison groups are combined by calculating their mean weighted by sample size. Likewise, when there is more than one treatment group that learns from an ITS in a study, these groups are also combined by calculating their weighted mean.

3.2.4. Data Analysis and Interpretation

In this meta-analysis, standard guidelines were followed (Adesope & Nesbit, 2012; Adesope, Lavin, Thompson, & Ungerleider, 2010; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001; Nesbit & Adesope, 2006). After all studies were coded, the corresponding spreadsheet was imported to IBM® SPSS® Statistics software (version 21) and later to Comprehensive Meta-analysis 2.2.048 for further analysis (Borenstein, et al., 2009). The Comprehensive Meta-Analysis software was used to generate the unbiased mean effect size (Hedges' *g*), the standard error of *g*, 95% lower and upper confidence interval around *g*, and values for the test of heterogeneity including *Q*, *p* and *I*-squared.

When the confidence intervals span a range above zero, they are interpreted as signifying a statistically significant result, favoring learning from ITSs over other instructional modes. Moreover, the upper and lower 95% confidence intervals are used to detect between-levels

differences among different categories of analyses. Specifically, when the confidence intervals of two or more categories are not overlapping, the effect sizes are considered to be statistically significantly different from one another.

In meta-analysis, the observed effect sizes of individual studies are averaged into a mean effect size. These are tested for the assumption of homogeneity of effects by the *Q* statistic to determine whether each effect size estimates the same population effect size. When all findings are drawn from the same population, *Q* has an approximate chi-square distribution with *k*-1 degrees of freedom, where *k* is the number of studies that account for a particular subset of analysis. When *Q* exceeds the critical value of the chi-square distribution, (i.e., $p < .05$), the mean effect size declared statistically significantly heterogeneous, which suggests `individual effect sizes do not estimate a common population mean (Borenstein, et al., 2009; Lipsey & Wilson, 2001).

Primary effect sizes were identified as outliers if standardized scores were extreme, $-3.3 \geq Z \geq 3.3$; $p < .001$. Two studies were identified as outliners. One yielded an effect size $g = 2.25$, whereas the other produced an effect size $g = -1.10$. No methodological flaws were identified found in either of these studies. Comprehensive Meta-Analysis was run to determine whether a homogeneous distribution existed after excluding the two outliers (Hedges & Olkin, 1985). The forest plot of all 107 effect sizes was first examined and then the two potential outliers were removed one at a time. The recalculated results showed that the removal of potential outliers did not improve the fit of the remaining effect sizes to a simple model of homogeneity. As recommended by Tabachnick and Fidell (2013), each outlying effect sizes was adjusted toward the next nearest effect size in the distribution to $g = 1.5$ and $-.5$, respectively.

In this review, both fixed-effect and random-effect models were calculated in all data analyses. A fixed-effect model operates under the assumption that all the studies included in the meta-analysis share one true effect size whereas a random-effects model assumes that there is more than one true effect and effect sizes could vary from one study to the other (Borenstein et al., 2009; Lipsey & Wilson, 2001). Given the diversified research interventions implemented and the variability from aggregating across a multiplicity of conditions, a random-effects model is usually considered a more accurate model than a fixed-effect model (Borenstein et al., 2009; Denson, 2009; Hedges & Vevea, 1998; National Research Council, 1992). Therefore, I report detailed results for random-effects model and add summary results for the fixed-effect model. This affords comparisons to fixed-effect results reported in previous ITS meta-analyses.

Moderator variables and levels showing significant differences under a fixed-effect model are also reported to provide additional research insights. Any mean effect size reported without specifying the type of model was generated by a random-effects model.

3.3. Data Analysis and Results

As previously discussed, nine studies that involved 784 participants, in which control group received no instructional treatment were excluded from the final meta-analysis. Although these studies are not directly related to the four research questions, they are examined separately to extract potentially useful context for the main analysis. In Table 3.1, the studies produced a statistically significant, weighted mean effect size of $g = 1.23$, under a random-effects model.

Table 3.1. Characteristics of "No Treatment" Control Studies

Study	Domain	Grade	Student Model	Study Setting	Effect Size (g)	95% CI	
						LCI	UCI
Arroyo et al. (2011)	Math II	PS	Other	Classroom	-0.17	-0.58	0.24
Beal et al. (2010) (2)	Arithmetic and fraction	6	Other	Classroom	-0.26	-1.07	0.56
Beal et al. (2010) (3)	Arithmetic and fraction	6	Other	Classroom	0.71	-0.17	1.59
Chen (2011)	Programming	PS	Not Reported	Laboratory	0.66*	0.33	1.00
Halpern et al. (2012) (1)	Research methods and scientific reasoning	PS	Not Reported	Not Reported	9.61*	8.42	10.80
Halpern et al. (2012) (2)	Research methods and scientific reasoning	PS	Not Reported	Laboratory	0.69*	0.15	1.23
Shute et al. (2007)	Algebra	12	Other	Classroom	0.38*	0.08	0.67
Wang et al. (2011) (1)	Earth Science	10	Other	Laboratory	0.17	-0.81	1.15
Wang et al. (2011) (2)	Earth Science	10	Other	Laboratory	0.06	-0.79	0.90

* $p < .05$

Figure 3.1 depicts the distribution of effect sizes for the main meta-analysis after adjusting for the two outliers. The effect sizes range between -.25 and .75 standard deviations. A positive effect size suggests the ITS group performed better than groups who received other modes of instruction, whereas a negative effect size indicates the opposite. The overall distribution, in Figure 1, shows students in ITS groups outperformed their counterparts in respective control groups in a majority of the studies.

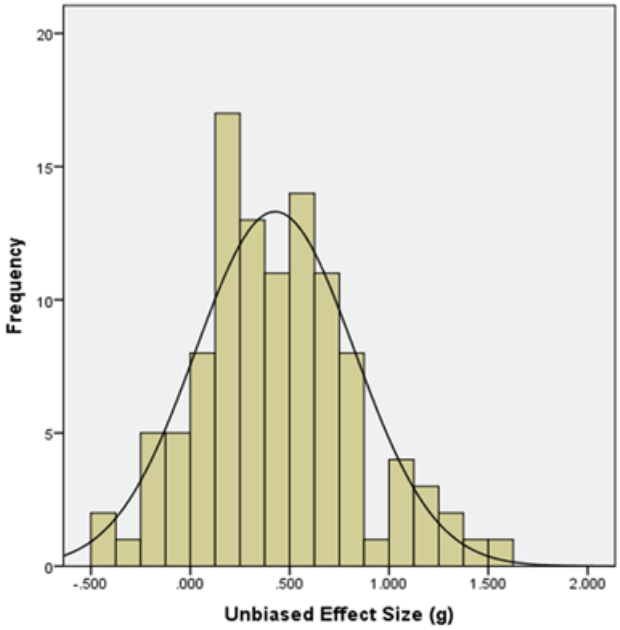

Figure 3.3-1. Distribution of 107 effect sizes (M = .43; SD = .40)

Table 3.2 includes all 107 studies that met the inclusion criteria. It presents characteristics of each study including the author(s), subject domain, grade level of participants, type of ITS, comparison treatment, study setting, the unbiased effect size, Hedges' *g*, and 95% lower and upper confidence intervals around each unbiased effect size.

Table 3.2. Characteristics of Coded Studies and Concomitant Effect Sizes

Study	Domain	Grade Range	Type of ITS	Control	Setting	Hedges g	95% CI Upper	95% CI Lower
Abu-Naser (2009)	Computer Science	PS	Other	LGHI	Classroom	0.55*	0.05	1.05
Aist et al. (2001) (1)	Language & Literacy	K-5	Other	LGHI	Classroom	0.35	-0.21	0.92
Aist et al. (2001) (2)	Language & Literacy	K-5	Other	LGHI	Classroom	-0.07	-0.64	0.50
Albacete & VanLehn (2000)	Physics	PS	Bayesian Network	CBI	Laboratory	0.62*	0.01	1.22
Arbuckle (2005)	Mathematics & Accounting	9-12	Model Tracing	LGHI	Classroom	0.64*	0.20	1.07
Argotte et al. (2011) (1)	Mathematics & Accounting	PS	Other	LGHI	Classroom	0.44	-0.26	1.15
Argotte et al. (2011) (2)	Physics	PS	Other	LGHI	Classroom	0.28	-0.61	1.16
Argotte et al. (2011) (3)	Physics	PS	Other	LGHI	Classroom	0.72	-0.44	1.88
Arnott et al. (2008)	Humanities & Social Science	PS	Other	LGHI	Classroom	0.75*	0.39	1.12
Beal et al. (2010)	Mathematics & Accounting	6-8	Other	SGHI	Classroom	-0.33	-1.09	0.44
Beal et al. (2007)	Mathematics & Accounting	9-12	NR	LGHI	Classroom	0.44*	0.11	0.76
Cabalo et al. (2007)	Mathematics & Accounting	Mixed	Model Tracing	LGHI	Classroom	0.18*	0.00	0.36
Carnegie Learning (2001)	Mathematics & Accounting	PS	Model Tracing	CBI	Laboratory	0.11	-0.13	0.36
Chambers et al. (2008)	Language & Literacy	K-5	Other	LGHI	Classroom	0.12	-0.08	0.31
Chen (2008)	Mathematics & Accounting	K-5	Other	TWS	Classroom	0.14	-0.15	0.44
Chien et al. (2008)	Mathematics & Accounting	9-12	Model Tracing	CBI	Classroom	1.08*	0.56	1.61
Chin et al. (2010) (1)	Biology & Physiology	6-8	NR	LGHI	Classroom	0.52*	0.00	1.03
Chin et al. (2010) (2)	Biology & Physiology	K-5	NR	LGHI	Classroom	0.22	-0.16	0.61
Conati & VanLehn (1999)	Physics	PS	Bayesian Network	TWS	Laboratory	0.15	-0.36	0.67
Conati & Zhao (2004)	Mathematics & Accounting	6-8	Bayesian Network	CBI	Classroom	0.65	-0.31	1.61
Corbett (2001)	Computer Science	PS	Model Tracing	CBI	Laboratory	0.70	-0.17	1.56
Fossati et al. (2008)	Computer Science	PS	Constraint-Based Model	IHI	Classroom	-0.19	-0.62	0.24
Fossati et al. (2009)	Computer Science	PS	Constraint-Based Model	IHI	Classroom	-0.16	-0.47	0.16
Graesser et al. (2003)	Physics	PS	EMT	TWS	Classroom	0.57	-0.24	1.37
Graff et al. (2008)	Mathematics & Accounting	Mixed	Other	LGHI	Classroom	1.19*	0.86	1.52
Hagerty & Smith (2005)	Mathematics & Accounting	PS	NR	LGHI	Classroom	0.49*	0.24	0.74

Han et al. (2010)	Computer Science	9-12	Bayesian Network	CBI	Classroom	1.08*	0.69	1.46
Hayes-Roth et al. (2010)	Others & Not Reported	PS	NR	TWS	Classroom	2.25*	1.21	3.29
Heffernan (2003)	Mathematics & Accounting	9-12	Model Tracing	CBI	Classroom	0.40	-0.11	0.92
Hu et al. (2007) (1)	Mathematics & Accounting	PS	NR	LGHI	Classroom	0.53*	0.15	0.91
Hu et al. (2007) (2)	Mathematics & Accounting	PS	NR	LGHI	Classroom	0.22	-0.05	0.50
Hwang et al. (2008)	Mathematics & Accounting	9-12	Other	LGHI	Classroom	0.69*	0.35	1.02
Jeremic et al. (2009)	Computer Science	PS	Other	LGHI	Classroom	0.51	-0.13	1.14
Johnson et al. (2009)	Mathematics & Accounting	PS	Model Tracing	TWS	Classroom	0.56*	0.03	1.10
Kinshuk et al. (2000) (1)	Mathematics & Accounting	PS	Other	LGHI	Laboratory	0.14	-0.29	0.58
Kinshuk et al. (2000) (2)	Mathematics & Accounting	PS	Other	LGHI	Laboratory	0.20	-0.24	0.63
Kinshuk et al. (2000) (3)	Mathematics & Accounting	PS	Other	LGHI	Classroom	0.14	-0.30	0.57
Koedinger (2002)	Mathematics & Accounting	6-8	Model Tracing	LGHI	Classroom	0.53*	0.35	0.72
Koedinger et al. (1997)	Mathematics & Accounting	9-12	Model Tracing	TWS	Classroom	0.32*	0.07	0.57
Kozierkiewicz et al. (2011)	Others & Not Reported	NR	Bayesian Network	LGHI	Laboratory	0.33	-0.02	0.69
Kumar (2002)	Computer Science	PS	Other	TWS	Classroom	0.08	-0.40	0.55
Lane & VanLehn (2005)	Computer Science	PS	Other	TWS	Classroom	0.04	-0.72	0.79
Lanzilotti & Roselli (2007)	Mathematics & Accounting	K-5	Other	LGHI	Classroom	0.09	-0.51	0.70
Lesta & Yacef (2002)	Mathematics & Accounting	PS	Other	LGHI	Classroom	0.43*	0.29	0.57
McLaren & Isotani (2011)	Chemistry	9-12	Model Tracing	LGHI	Classroom	0.15	-0.24	0.54
McNamara et al. (2006) (1)	Language & Literacy	6-8	Other	CBI	Classroom	1.32*	0.39	2.26
McNamara et al. (2006) (2)	Language & Literacy	6-8	Other	CBI	Classroom	1.36*	0.36	2.37
Mills-Tettey et al. (2010) (1)	Language & Literacy	K-5	Other	LGHI	Laboratory	-1.10	-1.93	-0.26
Mills-Tettey et al. (2010) (2)	Language & Literacy	K-5	Other	LGHI	Laboratory	0.14	-0.56	0.84
Mills-Tettey et al. (2010) (3)	Language & Literacy	K-5	Other	LGHI	Laboratory	1.11*	0.36	1.86
Mitrovic (2003)	Computer Science	PS	Constraint-Based Model	LGHI	Classroom	0.50*	0.15	0.84
Mitrovic et al. (2009)	Mathematics & Accounting	PS	Constraint-Based Model	SGHI	Classroom	-0.09	-0.96	0.78
Mitrovic & Ohlsson (1999)	Computer Science	PS	Constraint-Based Model	LGHI	Laboratory	0.75*	0.17	1.33
Morgan & Ritter (2002)	Mathematics & Accounting	9-12	Model Tracing	LGHI	Classroom	0.29*	0.10	0.47
Mostow et al. (2002) (1)	Language & Literacy	K-5	Other	TWS	Classroom	0.42	-0.09	0.93
Mostow et al. (2002) (2)	Language & Literacy	K-5	Other	TWS	Classroom	0.28	-0.22	0.79
Mostow et al. (2002) (3)	Language & Literacy	K-5	Other	TWS	Classroom	0.54*	0.02	1.05

Pane et al. (2010)	Mathematics & Accounting	9-12	Model Tracing	LGHI	Classroom	-0.19	-0.34	-0.04
Person et al. (2001)	Computer Science	PS	EMT	LGHI	Laboratory	0.78*	0.15	1.42
Phillips & Johnson (2011)	Mathematics & Accounting	PS	Model Tracing	CBI	Classroom	0.32	-0.01	0.65
Pinkwart et al. (2009) (1)	Humanities & Social Science	PS	NR	TWS	Laboratory	0.41	-0.35	1.18
Pinkwart et al. (2009) (2)	Humanities & Social Science	PS	NR	TWS	Classroom	-0.11	-0.57	0.35
Poulsen (2004)	Language & Literacy	K-5	Other	LGHI	Laboratory	0.15	-0.50	0.81
Radwan (1997)	Mathematics & Accounting	K-5	NR	LGHI	Laboratory	0.58*	0.03	1.13
Ramadhan (2000)	Computer Science	PS	Other	CBI	Laboratory	0.52	-0.42	1.46
Reif & Scott (1999)	Physics	PS	NR	IHI	Classroom	-0.42	-1.12	0.29
Ritter et al. (2007)	Mathematics & Accounting	9-12	Model Tracing	LGHI	Classroom	0.30*	0.08	0.51
Rosé & Bhembe (2003)	Physics	PS	Model Tracing	LGHI	Laboratory	0.13	-0.52	0.78
Rowe & Schiavo (1998)	Computer Science	PS	Other	LGHI	Classroom	0.91*	0.25	1.56
Schulze et al. (2000)	Physics	PS	Bayesian Network	LGHI	Classroom	0.23*	0.02	0.44
Shneyderman (2001)	Mathematics & Accounting	9-12	Model Tracing	LGHI	Classroom	0.30*	0.16	0.44
Shute & Glasser (1990)	Humanities & Social Science	PS	NR	LGHI	Laboratory	-0.10	-0.94	0.74
Smith (2001)	Mathematics & Accounting	9-12	Model Tracing	LGHI	Classroom	-0.12	-0.31	0.06
Stankov et al. (2004)	Computer Science	PS	Other	LGHI	Classroom	0.71	-0.12	1.54
Stankov et al. (2008) (1)	Computer Science	PS	Other	LGHI	Classroom	0.85*	0.40	1.30
Stankov et al. (2008) (2)	Chemistry	6-8	Other	LGHI	Classroom	0.17	-0.43	0.77
Stankov et al. (2008) (3)	Physics	6-8	Other	LGHI	Classroom	0.08	-0.35	0.52
Stankov et al. (2008) (4)	Humanities & Social Science	K-5	Other	LGHI	Classroom	0.54	-0.03	1.10
Stankov et al. (2008) (5)	Humanities & Social Science	K-5	Other	LGHI	Classroom	1.07*	0.47	1.67
Stankov et al. (2008) (6)	Humanities & Social Science	K-5	Other	LGHI	Classroom	0.80*	0.16	1.43
Stankov et al. (2008) (7)	Computer Science	PS	Other	LGHI	Classroom	1.18*	0.51	1.85
Stankov et al. (2008) (8)	Computer Science	PS	Other	LGHI	Classroom	0.36	-0.26	0.98
Stankov et al. (2008) (9)	Mathematics & Accounting	6-8	Other	LGHI	Classroom	0.28	-0.60	1.17

Stankov et al. (2008) (10)	Mathematics & Accounting	6-8	Other	LGHI	Classroom	0.09	-0.79	0.97
Stankov et al. (2008) (11)	Mathematics & Accounting	K-5	Other	LGHI	Classroom	0.31	-0.25	0.87
Suraweera & Mitrovic (2002)	Computer Science	PS	Constraint-Based Model	CBI	Laboratory	0.68*	0.17	1.19
Suraweera & Mitrovic (2004)	Computer Science	PS	Constraint-Based Model	CBI	Laboratory	0.17	-0.55	0.89
Tsiriga & Virvou (2004)	Language & Literacy	Mixed	Other	CBI	Classroom	0.50*	0.11	0.89
VanLehn et al. (2007) (1)	Physics	PS	EMT	SGHI	Laboratory	-0.24	-0.74	0.26
VanLehn et al. (2007) (2)	Physics	PS	Other	TWS	Laboratory	0.70*	0.09	1.31
VanLehn et al. (2007) (3)	Physics	PS	Other	TWS	Laboratory	0.20	-0.30	0.69
VanLehn et al. (2007) (4)	Physics	PS	Other	SGHI	Laboratory	0.75*	0.22	1.28
VanLehn et al. (2007) (5)	Physics	PS	Other	TWS	Laboratory	0.04	-0.26	0.35
VanLehn et al. (2005) (1)	Physics	PS	Bayesian Network	LGHI	Classroom	0.78*	0.54	1.03
VanLehn et al. (2005) (2)	Physics	PS	Bayesian Network	LGHI	Classroom	0.53*	0.18	0.87
VanLehn et al. (2005) (3)	Physics	PS	Bayesian Network	LGHI	Classroom	0.45*	0.11	0.79
VanLehn et al. (2005) (4)	Physics	PS	Bayesian Network	LGHI	Classroom	0.65*	0.29	1.02
VanLehn et al. (2010) (1)	Physics	PS	Model Tracing	LGHI	Classroom	0.23*	0.02	0.44
VanLehn et al. (2010) (2)	Physics	PS	Model Tracing	LGHI	Classroom	0.78*	0.54	1.03
VanLehn et al. (2010) (3)	Physics	PS	Model Tracing	LGHI	Classroom	0.53*	0.18	0.87
VanLehn et al. (2010) (4)	Physics	PS	Model Tracing	LGHI	Classroom	0.45*	0.11	0.79
VanLehn et al. (2010) (5)	Physics	PS	Model Tracing	LGHI	Classroom	0.65*	0.29	1.02
Veermans, et al. (2000)	Physics	PS	Other	CBI	Laboratory	-0.15	-0.73	0.44
Wheeler & Regian (1999)	Mathematics & Accounting	9-12	NR	LGHI	Classroom	0.70*	0.46	0.94
Wijekumar et al. (2012)	Language & Literacy	K-5	Other	LGHI	Classroom	0.32	-0.02	0.66
Wisher et al. (2001)	Humanities & Social Science	PS	Other	LGHI	Classroom	1.41*	1.14	1.69
Woo et al. (2006)	Biology & Physiology	PS	Other	TWS	Classroom	1.15*	0.56	1.75

Note. PS = postsecondary; EMT = Expectation and Misconception Tailoring; LGHI = Large-Group Human Instruction; SGHI = Small-Group Human Instruction; IHI = Individual Human Instruction (human tutoring); CBI = Individual Non-ITS Computer-Based Instruction; TWS = Individual Textbook or Workbook Studying; NR = Not Reported; * $p < .05$

Tables 3.3 through 3.8 present results for the fixed- and random-effects models, including: number of participants (*N*) in each category, number of studies (*k*), weighted mean effect size (*g*+) and its standard error (*SE*), 95% confidence interval around the mean, and a test of heterogeneity (*Q*). Each weighted mean effect size was obtained through the weighting of independent effect sizes by inverse variances. In the following sections, I elaborate on these results, organized by research question.

3.3.1. Research Question 1: Do Students Using ITS Have Different Learning Outcomes Than Students Using Other Modes of Instruction?

Table 3.3 shows the overall weighted mean of all statistically independent effect sizes. Under a fixed-effect model, it shows a moderate statistically significant effect of learning with intelligent tutors (g = .36; p <.001) with significant heterogeneity [$Q(106) = 390.52$, $p < .001$, I^2 = .73]. Under a random-effects model, the overall weighted mean effect size is also statistically significant and moderate (g = .41; p <.001).

For the breakdown of the comparison treatment instruction in all studies, the majority of studies compared the use of intelligent tutors with large-group human instruction (k = 66). Under both the fixed- and random-effects models, the use of ITS yields moderate, statistically significant mean effect sizes when compared with large-group human instruction, which included but was not limited to: traditional classroom instruction (g = .44), individual computer-based instruction (CBI, g = .57) and the individual use of textbooks or workbooks (g = .36). Furthermore, there are no statistically significant differences in the mean effect sizes comparing the use of ITS and small-group human instruction, defined as any form of synchronous instruction in groups of up to 8 students conducted with the presence of a human tutor.

Since the between-levels variance was statistically significant under both the fixed- and random-effects models (p < .001), post-hoc analyses were conducted to explore the variance. The analyses revealed that 66 studies, which compared the use of ITSs to large-group human instruction (g = .44), produced similar effect sizes to studies comparing ITSs to individual

computer-based instruction (CBI, *g* = .57) and to individual use of textbooks or workbooks (*g* = .36). However, all of these methods has a statistically significantly higher weighted mean effect sizes than studies that compared the use of ITSs to human tutoring. Altogether, these results show students who used ITSs learned more than students who used other modes of instruction except for small-group and individual human tutoring.

The statistically significant heterogeneity in the overall result suggests there is unattributed variability among the individual effect sizes that comprise the overall result. Therefore, moderator analyses were conducted on ITS characteristics, sample characteristics and methodological features of the studies to further explore for factors that may contribute to the variability in effect sizes.

Table 3.3. Overall Effect and Weighted Mean Effect Sizes for Comparison Treatments

			Effect size		95% CI		Test of heterogeneity			
	N	*k*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*df*	*p*	*I2 (%)*
Overall: Fixed-Effect Model	14,321	107	0.36*	0.02	0.32	0.39	390.52	106	<.001	0.73
Overall: Random-Effects Model	14,321	107	0.41*	0.04	0.34	0.48				

			Random-Effects Model						Fixed-Effect Model					
			Effect size		95% CI				Effect size		95% CI			
	N	*k*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*
Comparison Instruction							27.54	<.001					27.35	<.001
Large-group human instruction	11,296	66	0.44*	0.05	0.35	0.53			0.37*	0.02	0.33	0.41		
Small-group human instruction	184	4	0.05	0.28	-0.50	0.61			0.10	0.16	-0.21	0.41		
Individual human instruction	404	5	-0.11	0.10	-0.31	0.10			-0.11	0.10	-0.31	0.10		
Individual CBI	1,034	15	0.57*	0.11	0.34	0.79			0.47*	0.06	0.34	0.59		
Individual textbook or workbook	1,403	17	0.36*	0.09	0.18	0.53			0.30*	0.06	0.19	0.41		

* $p < .05$

3.3.2. Research Question 2: Do the Effects Associated with ITS Vary with Characteristics of the ITS?

In Table 3.4, the results show how different features and characteristics of ITSs contribute to the overall effect of learning with these systems. The effects on learning are examined in relation to different characteristics of ITSs including: type of ITSs, the nature of intervention provided by the ITS, whether the ITS modeled misconceptions, and the provision of feedback by the ITS. Most commonly in these studies, ITSs were the principal means of instruction (k = 35), provided feedback to students (k = 86), and modeled student misconceptions (k = 58). Under a random-effects model, two types of ITSs, constraint-based modeling and expectation and misconception tailoring, did not produce significant effects. However, ITSs with model tracing, Bayesian network modeling and other types of student modeling produced statistically significant effect sizes. Although ITSs with Bayesian network modeling have a higher weighted mean effect size (g = .54) than model tracing (g = .35), constraint-based modeling (g = .24), and expectation and misconception tailoring (g = .34), the between-levels difference is not statistically significant under a random-effects model. Conversely, statistically significant differences are detected under a fixed-effect model (Q_B [5] = 36.37, p < .001). Post-hoc analyses show ITSs that used Bayesian network modeling have a statistically significantly higher weighted mean effect size than those that use model tracing and constraint-based modeling.

In this study, I adopted the categories used by Steenberger-Hu and Cooper (2014) and coded the instructional roles of ITSs as follows:

- principal instruction: the ITS is the principal means of instruction;
- integrated class instruction: the ITS is an integral part of regular classroom instruction;
- separate in-class activities: the ITS is used for separate laboratory or other exercises that take place during class time;
- supplementary after-class instruction and homework: the ITS is used as part of out-of-class assignments.

Table 3.4. Weighted Mean Effect Sizes for Characteristics of Intelligent Tutoring Systems

			Random-Effects Model						Fixed-Effect Model					
			Effect size		95% CI				Effect size		95% CI			
	N	*k*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*
Type of ITS							4.18	0.52					36.37	<.001
Model Tracing	5970	21	0.35*	0.07	0.22	0.47			0.25*	0.03	0.20	0.31		
Constraint-Based Modeling	569	7	0.24	0.16	-0.08	0.56			0.20*	0.09	0.03	0.37		
Bayesian Network Modeling	1,417	10	0.54*	0.10	0.35	0.73			0.52*	0.06	0.41	0.63		
Expectation and Misconception Tailoring	142	3	0.34	0.35	-0.35	1.02			0.24	0.18	-0.12	0.59		
Other	4,425	53	0.44*	0.06	0.32	0.56			0.44*	0.03	0.38	0.50		
Not reported	1,798	13	0.40*	0.10	0.20	0.59			0.43*	0.05	0.32	0.54		
ITS Intervention							2.41	0.79					32.38	<.001
Principal instruction	4,505	35	0.37*	0.07	0.23	0.51			0.32*	0.03	0.26	0.38		
Integrated class instruction	4,045	15	0.33*	0.08	0.17	0.49			0.25*	0.03	0.18	0.31		
Separate in-class activities	1,939	24	0.47*	0.10	0.27	0.67			0.53*	0.05	0.43	0.62		
Supplementary after-class instr.	933	8	0.43*	0.11	0.22	0.64			0.36*	0.07	0.23	0.48		
Homework	2,480	15	0.45*	0.07	0.32	0.59			0.46*	0.04	0.38	0.54		
Not reported	419	10	0.48*	0.13	0.23	0.74			0.47*	0.10	0.27	0.66		
Feedback Provided?							4.55	0.10					13.53	<.001
No	1,411	10	0.54*	0.15	0.25	0.83			0.40*	0.05	0.30	0.51		
Yes	11,728	86	0.42*	0.04	0.34	0.50			0.37*	0.02	0.33	0.41		
Not reported	1,182	11	0.21*	0.10	0.02	0.41			0.15*	0.06	0.04	0.27		
Model Misconception?							0.02	0.99					5.14	0.08
No	1,508	21	0.40*	0.07	0.27	0.54			0.39*	0.05	0.29	0.49		
Yes	9,911	58	0.40*	0.05	0.31	0.49			0.33*	0.02	0.29	0.37		
Not reported	2,902	28	0.42*	0.10	0.23	0.61			0.43*	0.04	0.35	0.51		

* $p < .05$

Findings reported in Table 3.4 indicate ITSs are effective in all the instructional roles in which they were evaluated. Under both fixed- and random-effects models, the use of ITSs is associated with statistically significant effect sizes in all categories. The between-levels difference is not statistically significant under a random-effects model. However, the between-levels variance is statistically significant under a fixed-effect model, (Q_B [5] = 32.38, $p < .001$). Post-hoc analyses show studies that used ITS for separate, in-class activities and for homework have statistically significantly higher weighted mean effect sizes than those that used ITS for other purposes such as principal instruction.

Table 3.4 also shows that, under both fixed and random-effects models, the use of ITSs is associated with statistically significant effect sizes. The overlap in confidence intervals indicates that effect sizes are not moderated by whether the ITS provides feedback. Also, the use of ITS produces moderate, statistically significant effect sizes regardless of whether the ITS modeled misconceptions. The between-levels difference was not statistically significant under either the fixed and random-effects models.

3.3.3. Research Question 3: Do the Effects Associated with ITSs Vary with Characteristics of the Students, Outcome Assessments, and Research Setting?

To address the third research question, Tables 3.5, 3.6 and 3.7 show results of moderator analyses based on student and study characteristics, outcome assessments and research settings respectively. Specifically, Table 3.5 shows the effects of using ITSs across different grade levels, subject domains, and levels of prior knowledge. For grade levels, studies are binned according to school year in the following categories:

- elementary school: students from kindergarten through grade 5;
- middle school: students from grades 6 through 8;
- high school: students from grades 9 through 12;

- postsecondary: students at universities and colleges;
- mixed grades: three studies spanning grade bands.

In Table 3.5, the result show that the use of ITSs produces moderate statistically significant mean effect sizes at all grade levels under both the fixed and random-effects models. The between-levels difference is not statistically significant under a random-effects model but statistically significant under a fixed-effect model, (Q_B [5] = 25.74, p < .001). Post-hoc analyses reveal that students in middle school and postsecondary levels achieved statistically significantly higher weighted mean effect sizes than those who used ITS in elementary and high schools.

Table 3.5 also shows ITSs produce positive and moderate to large effect sizes across different subject domains. Notably, under a random-effects model, the use of ITS produces a large effect size in the humanities (g = .63). For domains such as biology and physiology (g =.59), computer science (g =.51), physics (g =.38) and mathematics and accounting (g =.35), ITS produced moderate effect sizes. For chemistry as well as literacy and language learning, the use of ITS produced small and moderate mean effect sizes (g =.16 and g = .34), respectively. The between-levels variance is not statistically significant under a random-effects model but is under a fixed-effect model (Q_B [7] = 50.67, p < .001), indicating statistically significant differences across subject domains. Post-hoc analyses yield that studies that used ITS in the humanities and social sciences have a statistically significantly higher weighted mean effect size than those that used it in mathematics and accounting, physics, computer science, language and literacy, and chemistry.

Table 3.5 also shows that many participants have low prior domain knowledge (k = 32). Except for high prior domain knowledge, all other categories of prior domain knowledge are associated with statistically significant effect sizes. Specifically, the results suggest that students with low and medium prior domain knowledge learned more with ITSs than those with high prior domain knowledge. However, the certainty of this interpretation is significantly limited by at least three factors: (a) the large number of studies that did not report the prior domain knowledge of participants (k = 34); (b) the small number of studies having participants with high prior knowledge (k = 2); and (c) the significant heterogeneity of the effect size distributions.

Table 3.5. Weighted Mean Effect Sizes for Student and Study Characteristics

			Random-Effects Model						Fixed-Effect Model					
			Effect size		95% CI				Effect size		95% CI			
	N	*k*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*
Grade Levels							2.2	0.82					25.74	<.001
Elementary school	1,496	19	0.31*	0.08	0.16	0.47			0.26*	0.05	0.16	0.37		
Middle school	810	10	0.41*	0.13	0.15	0.66			0.45*	0.07	0.31	0.59		
High school	4,355	14	0.40*	0.10	0.21	0.59			0.25*	0.03	0.18	0.31		
Postsecondary	6,767	60	0.43*	0.05	0.33	0.53			0.43*	0.03	0.38	0.48		
Mixed grades	771	3	0.61	0.32	-0.02	1.25			0.42*	0.07	0.28	0.57		
Not reported	122	1	0.33	0.18	-0.02	0.69			0.33	0.18	-0.02	0.69		
Subject Domains							6.53	0.48					50.67	<.001
Mathematics & Accounting	8,038	35	0.35*	0.05	0.24	0.45			0.29*	0.02	0.25	0.34		
Physics	2,890	24	0.38*	0.07	0.26	0.51			0.41*	0.04	0.33	0.49		
Computer Science	1,152	19	0.51*	0.11	0.30	0.72			0.46*	0.06	0.34	0.58		
Language & Literacy	1,075	14	0.34*	0.11	0.12	0.56			0.27*	0.06	0.15	0.39		
Chemistry	141	2	0.16	0.17	-0.17	0.48			0.16	0.17	-0.17	0.48		
Biology & Physiology	210	3	0.59*	0.27	0.07	1.11			0.51*	0.14	0.23	0.78		
Humanities & Social Science	671	8	0.63*	0.22	0.20	1.06			0.84*	0.08	0.68	1.01		
Others & Not Reported	144	2	1.23	0.96	-0.65	3.10			0.53	0.17	0.20	0.87		
Prior Domain Knowledge							3.45	0.49					11.87	0.02
Low	5,265	32	0.38*	0.06	0.27	0.49			0.37*	0.03	0.31	0.43		
Medium	1,356	17	0.28*	0.08	0.12	0.45			0.27*	0.06	0.16	0.38		
High	77	2	0.51	0.29	-0.06	1.07			0.53*	0.23	0.07	0.98		
Varied	2,699	22	0.48*	0.12	0.25	0.71			0.27*	0.04	0.19	0.34		
Not reported	4,924	34	0.46*	0.06	0.34	0.58			0.41*	0.03	0.35	0.47		

* $p < .05$

Table 3.6 reports the mean effect sizes for different outcome assessments, test formats, knowledge types and test sources. The learning outcomes are coded as retention, transfer, and mixed retention and transfer; test formats are coded as objective format (e.g., multiple choice items), short answer, and mixed format (e.g., combinations of multiple choice and short answer). Knowledge type is coded as procedural, declarative, and mixed procedural and declarative while test source is coded as researcher-developed, standardized or both. Under both the fixed- and random-effects models, the use of ITSs is associated with statistically significant mean effect sizes, regardless of the learning outcome. The between-levels variance is not statistically significant under either model.

ITSs yield statistically significant effect sizes across all test formats, except for mixed item formats. The between-levels variance is not statistically significant under the random-effects model but is significant under a fixed-effect model, (Q_B [4] = 108.17, $p < .001$). Table 3.6 also shows ITSs are effective for learning all types of knowledge. The between-levels variance is not statistically significant under a random-effects model but is under a fixed-effect model (Q_B [3] = 30.33, $p < .001$), suggesting there are statistically significant differences between knowledge types. Post-hoc analyses reveal studies that used ITSs to promote mixed procedural and declarative knowledge produced a statistically significantly higher weighted mean effect size than those that used ITS to acquire only procedural or only declarative knowledge. Finally, it also shows moderate, statistically significant effect sizes are obtained with all test types under both the fixed- and random-effects models. However, the between-levels variance is only statistically significant under the fixed-effect model (Q_B [3] = 14.92, $p < .001$). Post-hoc analyses show researcher-developed tests produced a statistically significantly higher weighted mean effect size than standardized tests.

Table 3.6. Weighted Mean Effect Sizes for Outcome Constructs, Test Format, Knowledge Type and Measuring Tool

			Random-Effects Model						Fixed-Effect Model					
			Effect size		95% CI				Effect size		95% CI			
	N	*k*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*
Outcome Constructs							1.12	0.77					3.56	0.31
Retention	3,922	33	0.35*	0.07	0.22	0.48			0.35*	0.03	0.28	0.41		
Transfer	1,683	18	0.44*	0.09	0.27	0.62			0.43*	0.05	0.33	0.52		
Mixed retention and transfer	6,371	32	0.43*	0.08	0.28	0.58			0.33*	0.03	0.28	0.38		
Not reported	2,345	24	0.42*	0.06	0.31	0.54			0.39*	0.04	0.30	0.47		
Test Formats							8.79	0.07					108.17	<.001
Multiple choice	1,777	18	0.26*	0.05	0.16	0.36			0.26*	0.05	0.16	0.36		
Short answer	1,170	11	0.25*	0.06	0.13	0.36			0.25*	0.06	0.13	0.36		
Mixed items	1,701	10	0.06	0.05	-0.03	0.16			0.06	0.05	-0.03	0.16		
Other	972	10	0.91*	0.07	0.77	1.05			0.91*	0.07	0.77	1.05		
Not reported	8,701	58	0.40*	0.02	0.35	0.44			0.40*	0.02	0.35	0.44		
Knowledge Type							1.18	0.76					30.33	<.001
Procedural	6,143	46	0.39*	0.05	0.28	0.49			0.36*	0.03	0.31	0.42		
Declarative	4,318	31	0.37*	0.07	0.23	0.51			0.26*	0.03	0.20	0.32		
Mixed procedural and declarative	777	6	0.65*	0.29	0.08	1.21			0.70*	0.08	0.55	0.86		
Not reported	3,083	24	0.43*	0.06	0.32	0.54			0.40*	0.04	0.33	0.48		
Test Source							0.71	0.87					14.92	<.001
Researcher developed	7,279	62	0.41*	0.05	0.32	0.50			0.40*	0.02	0.36	0.45		
Standardized	4,597	19	0.42*	0.10	0.21	0.62			0.27*	0.03	0.21	0.33		
Both	1,095	5	0.46*	0.07	0.33	0.59			0.46*	0.07	0.33	0.59		
Not reported	1,350	21	0.38*	0.07	0.24	0.52			0.34*	0.06	0.23	0.45		

* $p < .05$

Table 3.7 shows the results of analyses of contextual moderator variables: the setting where the research was conducted (laboratory or classroom), the continents where the study was conducted, the treatment duration, and the entire study duration. For research setting, 'classroom' refers to studies that have learning activities reported as part of an academic course of study or conducted in a classroom under the supervision of an instructor. Conversely, when learning activities were conducted solely for the purpose of research and learning was not assessed for academic credit, the setting is coded as *laboratory*. Approximate median splits on duration of treatment (split at one hour) and duration of study (split at one month) were used to create two categories for each of these variables. Table 3.7 shows most of the studies are conducted in the classroom (k = 81). Both classroom and laboratory studies produced moderate statistically significant effect sizes, under both the fixed and random-effects models. The between-levels variance was not statistically significant under the random-effects model but significant under the fixed-effect model, showing that classroom-based studies produced a higher weighted mean effect size than laboratory studies.

Table 3.7 shows the effectiveness of ITSs is evident regardless of the region where studies are conducted although a majority of the studies are conducted in North America (k = 75). Under the random-effects model, the use of ITSs is associated with moderate weighted mean effect sizes in North America (g = .38), Europe (g = .51), and Oceania (g = .36). The effect size in Asia is larger (g = .67). The between-levels variance is only statistically significant under the fixed-effect model (Q_B [3] = 12.80, p = .01). Post-hoc analyses reveal that the weighted mean effect size associated with studies conducted in Europe is larger than studies conducted in North America.

Table 3.7, indicates a tendency for higher mean effect sizes to be associated with longer treatment and study durations. Results of a random-effects model suggest treatments which are less than or equal to one hour in length produce a statistically significant mean effect size (g = .30), as do those greater than one hour (g = .39). However, neither the fixed- nor random-effects model show statistically significant, between-levels variance. Under the random-effects model, studies conducted for a month or less and those conducted for over a month are also associated with statistically significant, weighted mean effect sizes (g = .34 and g = .38). These results should be interpreted with caution because a large number of studies did not report the treatment and study durations.

Table 3.7. Weighted Mean Effect Sizes for Contextual Features

			Random-Effects Model						Fixed-Effect Model					
			Effect size		95% CI				Effect size		95% CI			
	N	*k*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*
Setting							3.30	0.07					4.29	0.04
Laboratory	1,596	26	0.29*	0.07	0.15	0.43			0.26*	0.05	0.16	0.36		
Classroom	12,725	81	0.44*	0.04	0.36	0.52			0.37*	0.02	0.33	0.41		
Continent							3.59	0.31					12.80	0.01
North America	11,065	75	0.38*	0.04	0.29	0.46			0.33*	0.02	0.29	0.37		
Europe	1,083	18	0.51*	0.10	0.32	0.71			0.55*	0.06	0.43	0.67		
Asia	962	6	0.67*	0.20	0.28	1.06			0.42*	0.07	0.29	0.55		
Oceania	1,211	8	0.36*	0.07	0.22	0.51			0.38*	0.06	0.27	0.50		
Treatment Duration							1.43	0.49					4.67	0.10
One hour or less	587	9	0.30	0.16	-0.01	0.62			0.18*	0.09	0.02	0.35		
Greater than one hour	7,589	59	0.39*	0.05	0.30	0.48			0.35*	0.02	0.30	0.40		
Not reported	6,145	39	0.47*	0.07	0.34	0.60			0.38*	0.03	0.33	0.43		
Study Duration							3.78	0.15					32.36	<.001
One month or less	2,044	32	0.34*	0.07	0.22	0.47			0.31*	0.05	0.22	0.40		
Greater than one month	9,577	53	0.38*	0.05	0.29	0.47			0.31*	0.02	0.27	0.35		
Not reported	2,700	22	0.57*	0.10	0.37	0.76			0.57*	0.04	0.49	0.66		

* $p < .05$

3.3.4. Research Question 4: Do the Effects Associated with ITS Vary with the Methodological Features of the Research?

Table 3.8 show how effect sizes are associated with varied methodological features of studies in this meta-analysis. The studies were categorized according to research design, source of publication and attrition. Under both fixed- and random-effects models, learning with ITSs was associated with moderate, statistically significant effect sizes regardless of research design. The between-levels variance was statistically significant only under the fixed-effect model (Q_B [3] = 65.68, $p < .001$). Post-hoc analyses indicate a larger mean effect size is associated with quasi-experimental designs in which prior differences were not controlled. The interpretation of this result should be taken with caution because a high number of studies did not explicitly report sufficient details in their research designs.

Studies published in journals often have higher methodological quality than those presented at conferences or as dissertations. In Table 3.8, under a random-effects model, mean effect sizes were statistically significant for studies published in journals, conference proceedings, as well as technical reports. However, dissertation studies did not yield a statistically significant effect size. The between-levels variance was not statistically significant under a random-effects model but it was under a fixed-effect model (Q_B [3] = 19.73, $p <.001$). Studies published in journals had a moderate mean effect size that is statistically significantly different from studies reported in conference proceedings and dissertations or theses. Finally, under both the fixed- and random-effects models, studies without attrition of participants and those with some attrition yielded moderate, statistically significant effect sizes.

Table 3.8. Weighted Mean Effect Sizes for Different Methodological Features

			Random-Effects Model						Fixed-Effect Model					
			Effect size		95% CI				Effect size		95% CI			
	N	*k*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*	*g+*	*SE*	*Lower*	*Upper*	Q_B	*p*
Random Assignment							7.39	0.06					65.58	<.001
Yes	5,588	34	0.31*	0.06	0.19	0.43			0.22*	0.03	0.16	0.27		
No -- prior difference controlled	3,075	23	0.38*	0.07	0.25	0.50			0.35*	0.04	0.27	0.42		
No -- prior difference not controlled	4,724	34	0.54*	0.06	0.42	0.67			0.55*	0.03	0.49	0.61		
Not reported	934	16	0.37*	0.11	0.15	0.58			0.30*	0.07	0.17	0.43		
Source							2.36	0.50					19.73	<.001
Journal	7,171	72	0.44*	0.04	0.36	0.53			0.42*	0.02	0.37	0.47		
Conference proceeding	4,045	23	0.33*	0.08	0.18	0.49			0.29*	0.03	0.23	0.36		
Dissertation/Thesis	1,419	5	0.27	0.15	-0.02	0.57			0.19*	0.05	0.08	0.30		
Technical report	1,686	7	0.46*	0.18	0.11	0.81			0.39*	0.05	0.29	0.49		
Attrition of Participants							4.08	0.13					30.29	<.001
None	3,191	36	0.39*	0.07	0.24	0.53			0.26*	0.04	0.19	0.33		
Some	4,075	23	0.29*	0.09	0.11	0.47			0.27*	0.03	0.20	0.33		
Not reported	7,055	48	0.48*	0.04	0.40	0.56			0.46*	0.03	0.41	0.51		

* $p < .05$

3.3.5. Are These Results Valid?

For this meta-analysis, I examined the potential impact of publication bias to determine whether results can be considered valid. Publication bias, the "file-drawer" effect, is a plausible threat to the validity of meta-analyses due to the fact that statistically significant results are more likely to be published and accessible for inclusion in meta-analyses than non-statistically significant results, which either may not be reported or reported in less accessible outlets (Orwin, 1983; Rosenthal, 1979). In Table 3.8, the between-levels variance of source of publication was found to be statistically significant under a fixed-effect model. Studies published in journals had a moderate mean effect size that is significantly greater than studies in conference proceedings and dissertations or theses. Therefore, it is crucial to conduct a further analysis of publication bias to validate the findings in this meta-analysis.

Two statistical tests are computed with Comprehensive Meta-Analysis to examine the potential for publication bias. First, a *Classic Fail-Safe N* test was computed to determine the number of null-effect studies needed to raise the *p*-value associated with the average effect above an arbitrary alpha level (set at $\alpha = .05$). This test result revealed that an additional 871 studies would be required to inflate the overall effect reported in this meta-analysis. Orwin's Fail-Safe *N*, a more stringent publication bias test, revealed that 656 missing null-effect studies would be required to bring the mean effect size identified in this meta-analysis to a trivial level of .05. Both test results show that, with 107 analyzed studies, the number of null-effect studies required to invalidate the overall effect size reported in this meta-analysis is larger than the $5k + 10$ limit suggested by Rosenthal (1995). Hence, although there is potential for publication bias in this meta-analysis, the results of these two tests suggest publication bias does not pose a significant threat to the validity of the findings.

3.4. Discussion

3.4.1. Summary of the Results

The overall result of this meta-analysis is that ITSs outperform the other modes of instruction in evaluation studies. Moderator analyses suggest using ITSs was associated with statistically significantly higher achievement outcomes than using each of the other modes of instruction except small-group human tutoring and individual human tutoring. There was no statistically significant difference in learning outcomes between ITS and these two forms of human tutoring. ITSs were also associated with greater achievement regardless of whether it was the principal means of instruction, an integral part of classroom instruction, a means to support in-class activities such as laboratory exercises, supplementary after-class instruction, or as part of assigned homework. In analyzing 18 other moderator variables related to characteristics of the ITS, students, research setting, outcome assessments, and research methods, no statistically significant differences were identified among levels of the moderators under a random-effects model.

3.4.2. Comparison with Previous Quantitative Reviews

In general, these results agree with previous reviews by VanLehn (2011), who examined the relative effectiveness of ITS in STEM subjects, and Steenberger-Hu and Cooper (2014), who investigated ITS use in all college-level subjects. The mean effect size for postsecondary education (g = .43) observed here was only slightly greater than the mean effect size (g = .37) reported by Steenberger-Hu and Cooper (2014). However, the mean effect sizes in the current meta-analysis for levels of K-12 education all were markedly greater than those reported in the meta-analysis of ITS effects in K-12 mathematics by Steenberger-Hu and Cooper (2013). Steenberger-Hu and Cooper reported ITSs showed no statistically significant effect on K-12 students' achievement compared to traditional classroom instruction. This discrepancy may be due to their smaller sample of empirical studies and their focus on only mathematical learning compared to the broader scope in my meta-analysis. Furthermore, in discussing their

finding, Steenberger-Hu and Cooper (2014) speculated “ITS may function better for more mature students who have sufficient prior knowledge, self-regulation skills, learning motivation, and experiences with computers” (p. 33). The current analysis, which directly compared ITS effect sizes at four levels of schooling, yielded no evidence to support that hypothesis.

The comparison of ITS with one-to-one human tutoring in this study produces a statistically undetectable mean effect size (g = -.11) similar to effect sizes reported by VanLehn (2011) and Steenberger-Hu and Cooper (2014) for human tutoring as a control condition. Unlike the previous reviews, small-group human tutoring was treated here as a separate category of control condition for which a statistically undetectable mean effect size of g = .05 was found under a random-effects model.

Unlike Steenberger-Hu and Cooper (2014), in this study, the no-treatment control conditions were analyzed separately from the main analysis. Whereas the current study report *g* = 1.23 for no-treatment controls, Steenberger-Hu and Cooper found *g* = .90, and VanLehn (2011) found .40 and .76 for differing levels of interaction granularity. Although these discrepancies could be cause for concern, the important fact that serves as a reality check on the ITS evaluation enterprise is that, in every review, the mean effect sizes for no-treatment controls are greater or equal to the largest mean effect sizes for comparisons between ITS and an alternate form of instruction.

3.4.3. Quality of Reporting

Based on findings in this meta-analysis, I suggest there is considerable room for improvement in how fundamental features of the primary research are reported. Basic statistics such as means and standard deviations were not reported in about a third of the studies, and the reliability of outcome measures was reported in only a few cases. In many studies, reporting was also insufficient for methodological features such as attrition, whether participants were randomly assigned to treatments, format and provenance of

achievement tests, and duration of treatment. The challenge in raising the quality of research reporting in an interdisciplinary field such as ITS can be attributed to the lack of shared understanding of methodological standards among researchers and editors, and the dissemination of ITS evaluation research across a remarkably eclectic set of journals including *Issues in Accounting Education*, *Thinking Skills and Creativity*, and *Methods of Information in Medicine*. This challenge, inherent to interdisciplinary research reporting, is reflected in the statement of scope and standards of the *International Journal of Artificial Intelligence in Education*, which remarks "if a paper presents a behavioural study of students using some system to support claims about improved learning, then it must conform to the standards developed in behavioural science" but also that "it is not reasonable to expect that authors will meet all the standards of all disciplines outside their main focus." I advocate journal editors specify requirements for reporting research design, sample size, attrition, reliability of measures, means and standard deviations for quantitative educational research, and also publish articles that inform their readership on how contemporary methodological practices such as "the new statistics" (Cumming, 2012) relate to their discipline.

One difficulty I found in synthesizing all studies in this meta-analysis was a lack of common terminology for describing and reporting about the design of an ITS as well as inconsistent practices in selecting which ITS features should be described in an evaluation report. For example, some researchers reported their ITS used model tracing but did not indicate whether misconceptions were modeled, whether knowledge tracing was used, or whether the ITS adaptively selected problems. Often, the method used for student modeling was not described in relation to other ITS, and important features of a system's design and behavior were not reported. I speculate that developing a taxonomy of ITS design that could underpin a reporting standard that would accelerate advances in ITS research. Certainly, more precise reporting about ITS design that draws from a common conceptual framework and terminology would greatly assist meta-analysts in identifying specific design features to compare their effects on learning outcomes.

3.4.4. Can Evaluation Research Contribute to a Theory of ITS Design?

In this study, each study evaluated a single ITS that consisted of a complex set of interrelated features, many of which were no reported by or even known to the primary authors. In most cases, the evaluations were performed to investigate whether to deploy the ITS or to evaluate a software engineering project holistically. Rarely were the studies conceived as research into a theoretical question about the relationship between ITS and learning outcomes. Nonetheless, a meta-analysis of evaluation studies can categorize primary studies according to theoretically significant features and enable observation of relationships between the features and learning outcomes that were not considered in the primary work. Although none of the primary studies in this analysis compared a version of their ITS that models student misconceptions with another version that did not, I was able to code for misconception modeling as a moderator variable and assess its influence on effect size. As it turns out, none of the ITS characteristics coded, including type of ITS and misconception modeling and feedback, were found to reliably influence effect size under a random-effects model (although a fixed-effect model found that one type of ITS, Bayesian network modeling, had a significantly greater influence on effect size). It is notable that when I reread the studies in which the ITS did not provide response feedback, I observed the primary adaptive feature, in each case, was individualized task selection. This suggests individualized task selection may offer benefits comparable to the well-established, positive effects of feedback on learning (Hattie & Timperley, 2007).

In addition, most research studies reported independent and dependent variables but not intervening process variables (i.e., measures observed during the learning process) that might help to explain observed effects or lack thereof. Although recent work in educational data mining indicates process variables are gaining a more prominent position in the study of ITS (Winne & Baker, 2013), such variables are rarely reported in empirical evaluations of ITS effectiveness. When a process variable was reported in the studies, it was often only meaningful in the context of the particular learning task of the study that reported it. As a consequence, when researchers find that ITSs outperform other methods of computer-based instruction, there is little researchers can do in a meta-analysis to account for the effect at the level of computer-student interactions. Similarly, if we want to seek an explanation for how Bayesian network modeling might outperform

other ITS designs, there are no common, interface-level data that could show whether such technical distinctions among the major types of ITSs are manifested across the research base in consistent, differentiated patterns of interaction with students.

3.4.5. What Meta-analysis Can Tell Us About ITS

Despite the great variety of conditions under which ITSs were found more effective than other modes of instruction, these results do not support the direct inference that ITS should, in any way, replace other modes currently in use. First, in all the studies analyzed in this review, there was the rational tendency to evaluate an ITS relative to the goals and scope of the project that created it. Because a project's goals are determined by what might feasibly be accomplished by an ITS, the studies have an inbuilt bias toward instructional conditions and roles in which ITS could compete favorably with other modes of instruction. Another possible source of bias is that the development of an ITS, like any major instructional design project, typically involves more detailed attention to learning goals, materials and activities than the more typical instructional practices represented by the control conditions in the analyzed evaluations. It may be that the significant effect sizes are due as much to the kind of intensive instructional planning and analysis that can enhance any instructional approach than to the particular features of ITS.

What the results of this meta-analysis provide is strong evidence that, in some situations ITS can successfully complement and substitute for other instructional modes, and that these situations exist at all educational levels and in many common academic subjects. The results do not support any particular explanations for the effectiveness of ITS, but they are consistent with an attribution to the most frequently implemented ITS features enabled by student modeling, namely highly individualized task selection, prompting and response feedback. That ITSs were found to be relatively effective whether or not they model frequent student misconceptions suggests the need for comparative research on the conditions under which misconception modeling adds value to individualized instruction.

This meta-analysis and previous reviews by Steenberger-Hu and Cooper (2013, 2014) examined evaluation research in which the use of ITSs was compared to a variety of other modes of instruction. While reviews of this type are useful in marking general progress in the field of ITSs, a more powerful use of meta-analysis to drive those capabilities forward may be to review comparisons between ITSs. This strategy would be especially informative when analyzing studies that compare two or more versions of the same ITS such that each version represents a theoretically informed design variation. VanLehn (2011) adopted elements of this strategy to investigate the effects of interaction granularity on learning outcomes, and full meta-analyses comparing different versions of the same systems could be used to investigate many other potentially effective ITS features such as animated pedagogical agents (Baker et al., 2006), misconception modeling (Myneni, Narayanan, Rebello, Rouinfar, & Pumtambekar, 2013) and metacognitive prompts (Wu & Looi, 2012). I believe such strategic application of meta-analysis to the end products of ITS research, development, and evaluation can inform and advance the design science of ITS.

Chapter 4.

Overview of Student Modeling in Bayesian Network

In Chapter 3, I reported the ITS meta-analysis that investigated the effectiveness of ITS compared to non-ITS environments. It is a statistically powerful way to detect the overall mean effect size of ITSs compared to other modes of instruction from a comprehensive set of research studies. However, to conduct the required statistical analysis, studies retained for the meta-analysis have to assess cognitive outcomes, and report statistics needed to calculate an effect size. These are strict screening and inclusion criteria. As a result, a good number of experimental studies that may bring useful insights to the landscape of the ITS field were filtered out.

To supplement this meta-analysis, I conducted a second meta-analysis to focus on studies that fundamentally address issues of student modeling, specifically, a Bayesian Network. This is a mature student modeling technique with power to handle uncertainty and afford inferences based on uncertain observations, behaviors and measurements (Chrysafiadi, & Virvou, 2015).

4.1. Purpose of the Study and Research Questions

This review synthesizes research on the relative effectiveness of Bayesian networks in constructing student models in intelligent tutoring systems (referred to as “ITS BN studies” in all instances), and it addresses the following research questions:

1. What research questions were investigated in ITS BN studies?
    a. Evaluation of BN student models?
    b. Evaluation of learning outcomes?
2. What types of BNs have been applied in ITS BN studies?
3. What are the contextual settings of ITS BN studies (e.g., subject tutored, grade level etc.)?
4. What constructs are modeled in BN student modeling? (e.g., level of knowledge, affect, motivation, etc.)?

5. What pedagogical approaches are applied in ITS BN studies?
6. What instructional strategies are applied in ITS BN studies?
7. What are the characteristics of BN student models?

4.2. Method

4.2.1. Selection Criteria

Studies were eligible for inclusion in the review when they shed light on research questions. In particular, studies must have been characterized by the following:

(a) applied at least one type of BN in student modelling,
(b) reported sufficient details of how BN student modeling was designed, and
(c) were made publicly available, online or in library archives.

According to criterion (a), studies that applied either one type of BN or one type of BN plus other algorithm(s) were included in the selection pool. However, studies that applied BN to areas other than student modeling were excluded.

For criterion (b), studies were excluded when they reported only very high-level design features and, therefore, could not provide sufficient data for extraction, based upon a pre-defined coding form that had been developed for this study (See Section 4.2.3 and Appendix B) . Studies that reported about the BN design for student modeling but that did not assess the effectiveness of the BN design or the learning outcomes were still included for analysis if they met criterion b.

4.2.2. Search, Retrieval, and Selection of Studies

A comprehensive search for relevant research was conducted in five major bibliographic databases: ERIC, PsycINFO, Springer Link, Science Direct and Web of Science. The key term Bayesian network was combined with the following key terms in the search: intellige* tutor*, intellige* agent*, cognit* tutor*, adapt* tutor*, cognit* virtual companion*, intellige* coaching system*, and pedagog* agent*. In total, the result returned

26,547 articles, which were individually screened. Table 4.1 shows results from the search using these key terms across the bibliographic databases.

Table 4.1. Summary of Literature Search Parameters

Search Key Terms	Bibliographic Databases	# of Documents Retrieved
intellige* tutor*, or intellige* agent*, or cognit* tutor*, or adapt* tutor*, or pedagog* agent*, or cognit* virtual companion*, or intellige* coaching system*	ERIC	229
	PsycINFO	35
	Springer Link	18,817
	Science Direct	7,173
	Web of Science	320
And		
Bayesian Network	**Total Hits**	26,574

During the screening phase, the abstracts of the articles were reviewed and compared with criteria a and b to exclude irrelevant studies. The 2,182 articles that passed the initial screening were retrieved, and full-text copies were further evaluated against all three inclusion criteria. During this screening phase, BN studies that were referenced in the reviewed articles were manually retrieved and added to the selection pool. Finally, the 143 studies that met the inclusion criteria were coded using a pre-defined coding form and coding instructions, which were developed specifically for this review. These include studies published between 1992 and 2014.

4.2.3. Coding Study Characteristics

The coding form included 32 fixed-choice items and 14 comment items that detailed information about the studies. Appendix B lists the complete coding form. Items

recorded about each study included: year published, research questions, subject domains, type of BN, major student modeling parameters, educational level of participants, research settings, duration of the study, type of research study, reliability reporting and descriptive statistics reported for the reported quantitative study.

During the coding process, three categories of BN studies were identified and coded. The first category included studies evaluating the learning outcomes of students who worked with an ITS, which had been built with a proposed Bayesian student model. The effectiveness of the specific BN model under review was considered to be associated with students' performances. The second category of studies generally evaluated the capability of the Bayesian student model to accurately predict the current state of students. Their approach involved comparing the learning outcomes of a group of students, working with the proposed BN model, with another group of students, working with another type of student model. The discrepancy between the groups' respective learning outcomes was examined. The third category included studies that provided the design of the Bayesian student model but did not perform assessments with real users. Some of these studies used simulated students to test the effectiveness of the BN model to perform as predicted. In such cases, the simulated students' prior knowledge was pre-defined and adjusted according to the test requirements as were other relevant defining variables. Since all parameters about simulated students were manipulated at the will of the researchers, the associated learning performances were not considered reflective of the validity needed for the purposes of this review. Consequently, these studies were not coded.

4.2.4. Data Analysis and Interpretation

In this review of Bayesian student modeling, data of relevant learning constructs were collected and entered into the coding form. The data were aggregated to identify patterns that could provide insights relevant to the seven research questions. In terms of Research Question 1, which addresses the research questions explored in the BN studies, I investigated the categories of research types reported. I also explored the studies in reference to a set of characteristics such as research approach (qualitative or

quantitative), research design, data collection instrument, type of publication, learning outcome, etc. These characteristics were examined with the intention of providing a comprehensive understanding about the nature of the included studies and to provide insights on how these studies addressed their respective research questions. Similarly, for Research Question 3 about the contextual settings of BN studies, a series of variables were explored to examine the contextual factors that could have contributed to a study's results. These include country, subject domain, educational level, type of knowledge, and targeted level of student knowledge. For Research Question 4, to understand what learning constructs were traced in the Bayesian student model, I listed all the learning constructs extracted from the BN studies and then calculated the frequency of each to identify the constructs most commonly examined in the existing pool of studies. This examination also facilitated my gaining insights on the array of constructs that can be tracked and analyzed using BN. For the Research Questions 2, 5 and 6, specific variables were analyzed to understand the types of BNs utilized in all studies and to identify the respective pedagogical and instructional strategies adopted to facilitate learning. For the last research question, with reference to the insights gained on the six previous questions, the characteristics of BN to support student modeling were holistically examined to understand the extent to which it predicts student performances.

4.3. Results

In this chapter, I present the results of the data analysis conducted on 143 studies, retrieved during a search of the literature, from 1992 to 2014. The results are presented in relation to this study's seven research questions organized in the list below:

4.3.1. Research Question 1: What research questions were investigated in ITS BN studies?

During this analysis, I examined the kinds of research questions that were investigated in ITS BN studies. In addition to research type, I also present data according to nine variables associated with research design characteristics to provide further insights on Research Question 1.

Research Type

In this study, research type refers to the research questions that a particular experiment was designed to investigate. Table 4.2 presents three research types used to classify BN studies. Design proposal refers to research studies that only provide the design of a BN student model with or without the details of the ITS. Instruction assessment includes studies which assess specific instructions that guide students during learning in BN studies. System assessment refers to studies that design and evaluate the accuracy of the Bayesian student model or the ITSs with real users. A BN study evaluated with simulated students was not classified in this category because its experimental conditions are predefined and manipulated to suit varied testing needs; thus, they are instead categorized as a design proposal. Overall, the proportion of studies (50.3%) implemented as a system assessment is similar to those (49%) categorized as a design proposal. No studies were categorized as instruction assessments.

Table 4.2. Research Type in BN Studies

Research Type	Number of Studies	Percentage
Design Proposal	71	49.7%
Instruction Assessment	0	0.0%
System Assessment	72	50.3%

Type of Publications

Table 4.3 presents the types of publications for the included BN studies. It suggests that conference proceedings (69.9%) account for the highest percentage among all ITS research studies, followed by journal articles (22.4%) and book chapters (7%). It is worth noting that some studies were published as dissertations as well as one of the three aforementioned types of publications. After examining the publications derived from a particular study, dissertations were not included in data calculation because later publications reported the same findings and were of higher quality. Instead, dissertations were added to an archival folder for reference. One research report, summarizing the design and development of an adaptive math tutoring prototype, was also included in the current analysis (Shute, Graf, & Hansen, 2006).

Table 4.3. Publication Type in BN Studies

Publication Type	Number of Studies	Percentage
Conference Proceedings	100	69.9%
Journal Article	32	22.4%
Book Chapter	10	7.0%
Dissertation	0	0.0%
Others-Report	1	0.7%

Qualitative Studies Conducted

Table 4.4 presents the number of qualitative BN studies. It suggests that, excluding studies that were classified as not reported or applicable (49%), most BN studies (50.3%) did not adopt a qualitative approach when evaluating the proposed ITS design. Only one study (0.7%) was found to use a case study approach when investigating the effect of an adaptive collaborative assistant on the improvement of students' helping behaviors in a peer tutoring activity (Walker, Rummel, & Koedinger, 2011).

Table 4.4. Number of Qualitative BN Studies

Qualitative Study Conducted	Number of Studies	Percentage
Yes	1	0.7%
No	72	50.3%
Not Reported or Applicable	70	49.0%

Research Design

Table 4.5 presents the research design of quantitative BN studies. Excluding studies classified as not reported or applicable (57.3%), a majority of studies used quasi-experimental (14.7%) and experimental research design (11.2%) approaches. Ten studies were conducted as within-subject experiments. The remaining 14 studies fall into the *others* category, with the following subcategories: 1) studies that collect and analyze user logs or data to evaluate the accuracy of the student model to predict learner performance/affects (9 studies); 2) studies that collect users' self-reports and questionnaires regarding their learning experience with the BN ITS (2 studies); 3) a combination of both 1) and 2) (3 studies).

Table 4.5. Type of Research Design in BN Studies

Research Design	Number of Studies	Percentage
Experimental	16	11.2%
Quasi-experimental	21	14.7%
Within-subject experiment	10	7.0%
Others	14	9.8%
Not Reported or Applicable	82	57.3%

Categories of Learning Outcome

Table 4.6 presents five common types of learning outcomes used to categorize the 143 BN studies. Among the five types, verbal information is defined as "declarative knowledge" (Gagné & Driscoll, 1988, p. 44). It is the ability to articulate previously learned

facts, concepts and procedures (Driscoll, 2000). Intellectual skills refers to the understanding of how to execute an action, which is also known as "procedural knowledge" (Gagné & Driscoll, 1988, p. 47). Cognitive strategies are the skills with which "learners regulate their own internal processes of attending, learning, remembering, and thinking" (Gagné, 1985, p. 55). Thinking creatively and problem solving are also part of cognitive strategies (Driscoll, 2005). Attitudes are defined as an "acquired internal state that influences the choice of personal action" (Gagné & Driscoll, 1988, p. 58). Motor skills emphasize smooth and accurate performance involving muscular coordination, which is an indication of performance but not a primary educational goal (Gagné & Driscoll, 1988).

Overall, excluding studies classified as not reported or applicable (28.7%), intellectual skill (46.2%) is the most studied skill in ITSs among all BN studies. The second most studied skill is verbal information (14%) which includes 22 studies. Seven studies (4.9%) targeted both of those skills in ITSs. Nine studies were found to support the development of cognitive strategies in ITS, e.g., scientific inquiry skill. In the current analysis, no studies were found to develop attitudes or motor skills.

Table 4.6. Categories of Learning Outcome in BN Studies

Categories of Learning Outcome	Number of Studies	Percentage
Verbal Information	20	14.0%
Intellectual Skill	66	46.2%
Cognitive Strategy	9	6.3%
Attitude	0	0.0%
Motor Skill	0	0.0%
Verbal Information + Intellectual Skill	7	4.9%
Not Reported or Applicable	41	28.7%

Positive Experimental Outcome in ITSs

To explore the characteristics of ITSs regarding student models, I not only include empirical studies that provide sufficient empirical data and experimental details but also literature that reports the design of student models in building an ITS, which is run without an experiment to evaluate its accuracy and effectiveness. For the included empirical literature, Table 4.7 shows the number of studies that were found to have a positive outcome based upon an evaluation of the ITS' effects. In the current analysis, the phrase *positive outcome* is defined broadly to include not only studies with positive ITS effects in empirical evaluation but also to include studies that have reported positive learning experiences associated with ITS. For instance, positive experiences were self-reported by students, which suggests that they found the learning environment to be enjoyable. Excluding studies classified as not reported or applicable (51.7%), a majority of studies (41.3%) reported a positive outcome. Only nine studies were found to have no positive outcome or no effect, indicating that most studies resulted in some degree of positive outcome using BN student modeling in ITSs. One study, which reported a mixed result, involved adding eye-tracking data to increase bandwidth in student modeling. The data provides support for using an eye-tracker to predict students' ability to apply self-explanation skills.

Table 4.7. Positive Experiment Outcome in ITSs in BN Studies

Positive Outcome in ITS	Number of Studies	Percentage
Positive Outcome	59	41.3%
No Positive Outcome or No Effect	9	6.3%
Others - Mixed Results	1	0.7%
Not Reported or Applicable	74	51.7%

Positive Learning Outcome in ITS Table 4.8 shows the number of studies that yielded positive learning outcomes for students when they worked with ITSs. Overall, excluding studies classified as not reported or applicable (71.3%), a majority of studies (25.2%) reported a positive learning outcome. Only four studies reported no positive learning outcome or no effect. One study reported a mixed learning outcome, suggesting some students demonstrated significant progress or a certain degree of improvement in

their learning while others showed no improvement at all, after working with a math ITS on decimal numbers.

Table 4.8. Positive Learning Outcome in BN Studies

Positive Learning Outcome in ITS	Number of Studies	Percentage
Positive Outcome	36	25.2%
No Positive Outcome or No Effect	4	2.8%
Others - Mixed Results	1	0.7%
Not Reported or Applicable	102	71.3%

Other Dependent Variable

In addition to student performance, Table 4.9 shows the additional dependent variables used during experiments for BN research. Overall, most BN studies do not use other dependent variables (94.4%).

Table 4.9. Other Dependent Variables in BN Studies

Other Dependent Variable	Number of Studies	Percentage
Others	6	4.2%
Not Reported or Applicable	137	95.8%

As listed in Table 4.10, only six studies use additional dependent variables in their evaluation:

Table 4.10. Other Dependent Variables in Other Category

Other Dependent Variable	Number of Studies
Number of mountains students in both groups climbed in the game and whether students used help or not	1
Students' confidence, confusion and effort estimates	1
The quality of the peer tutor's help	1

Frequency of constraint violation	1
Number of problems attempted in each condition	1
Pupil size, affect and reasoning styles	1

Independent Variable

Table 4.11 identifies the independent variables for experiments conducted in BN research. Overall, excluding the studies classified as not reported or applicable (65%), a majority of studies (32.9%) were found to have *use of ITSs* as the independent variable. Only three studies were found to use other variables as the independent variable in their evaluation, as listed in Table 4.12.

Table 4.11. Independent Variables for BN Studies

Independent Variable	Number of Studies	Percentage
Use the tool/system	47	32.9%
Others	3	2.1%
Not Reported or Applicable	93	65.0%

Table 4.12. Independent Variables in Other Category

Independent Variable Conditions in Other Category	Number of Studies
Use version 1 or version 2 of the proposed ITS	1
Use the game-based ITS, or the non-game-based ITS, or no ITS	1
Relevance of computer support and peer tutor's noticing of support	1

Instrument and Procedure

Table 4.13 shows the instrument or procedure used in experiments to collect additional student data in the BN studies. Overall, excluding the studies classified as not reported or applicable (78.3%), 22 studies have used surveys or questionnaires to collect students' attitudes or opinions about how they interact with ITSs during learning. Five studies (3.5%) collected students' self-reports on their experiences working with ITSs. One study (0.7%) conducted an in-depth group interview with students of the treatment group to collect similar students' data on their experience with the ITS during learning (Han & Lee, 2010).

Table 4.13. Instruments and Procedures in BN Studies

Instrument and Procedure	Number of Studies	Percentage
Survey/Questionnaire	22	15.4%
Test	0	0.0%
Observation	0	0.0%
Self-report	6	4.2%
Others	1	0.7%
Survey/Questionnaire + others	2	1.4%
Not Reported or Applicable	112	78.3%

4.3.2. Research Question 2: What types of BNs have been applied in ITS BN studies?

This research question explores the types of BNs applied in ITS BN studies. In the current analysis, when a study reports the use of BN in the student model without specifying a type, it is coded as a general BN in the category of *Bayesian Network*. Table 4.14 reflects that a majority of studies have adopted general BN student modeling (77.6%). Twenty-six studies applied dynamic BN in the student model. Two studies (1.4%) applied temporal BN and one study (0.7%) used both static and temporal BN. Two studies used a customized BN approach. Specifically, BNT-SM and atomic BN were used.

Table 4.14. Types of BN applied to Student Modeling in BN Studies

Types of BN applied in Student Modeling	Number of Studies	Percentage
Bayesian Network	112	78.3%
Dynamic Bayesian Network	26	18.2%
Static + Temporal Bayesian Network	1	0.7%
Temporal Bayesian Network	2	1.4%
BNT-SM	1	0.7%
Atomic Bayesian Network	1	0.7%

Knowledge Domain Model Built in BN

In addition to student models, BN application in knowledge domain models were also examined. Table 4.15 shows that, excluding the studies that were classified as not reported or applicable (46.2%), a majority of studies did not adopt BN in building their knowledge domain (38.5%).

Table 4.15. Knowledge Domain Built using BN in BN Studies

Knowledge Domain Built using BN	Number of Studies	Percentage
Yes	22	15.4%
No	55	38.5%
Not reported or applicable	66	46.2%

Tutor Model Built using BN

Table 4.16 presents the number of ITS studies that built their tutoring models using BN. Similar to knowledge domain modeling, excluding the studies not reported or

applicable (59.4%), a majority of studies did not adopt BN in building their tutoring model in ITSs (30.1%).

Table 4.16. Tutor Model Built by BN in BN Studies

Tutor Model Built using BN	Number of Studies	Percentage
Yes	15	15.4%
No	43	38.5%
Not reported or applicable	85	46.2%

4.3.3. Research Question 3: What are the contextual settings of ITS BN studies?

Contextual settings provide environmental information about how a specific study is conducted and may yield additional information that could influence the study results; however, contextual settings are not considered when designing a research study. For this research question, there are five contextual categories, regarding the demographics and characteristics of ITS studies that adopted BN in their student modeling.

Country

Table 4.17 shows the distribution of countries in which the included research studies were implemented. The research studies were conducted across 32 countries. USA, Canada and China comprise 46.2% of the total distribution and are the countries that have conducted the most studies in this area. For the remaining 29 countries, the number of studies per country ranges from one to seven. The data reflects the popularity of using BN for student modeling in the ITS communities across a variety of regions globally.

Table 4.17. Country Distribution in BN Studies

Country/Region	Number of Studies	Percentage
USA	36	25.2%
Canada	21	14.7%
China	9	6.3%
Spain	7	4.9%
Malaysia	6	4.2%
Mexico	6	4.2%
Italy	5	3.5%
Greece	4	2.8%
Brazil	4	2.8%
Argentina	3	2.1%
Taiwan	3	2.1%
Egypt	3	2.1%
France	3	2.1%
India	3	2.1%
Singapore	3	2.1%
UK	3	2.1%
Korea	2	1.4%
Netherlands	2	1.4%
Australia	2	1.4%
New Zealand	2	1.4%
Columbia	2	1.4%
Portugal	2	1.4%
Iran	2	1.4%
Thailand	2	1.4%
Cuba	1	0.7%
Philippines	1	0.7%
Germany	1	0.7%
Russia	1	0.7%
Vietnam	1	0.7%
Morocco	1	0.7%
Indonesia	1	0.7%
Turkey	1	0.7%

Subject Domain

Table 4.18 presents the number of BN studies by subject domain. The top three subject domains are computer science, math, and physics, representing 60.2% of the overall subject domains in study. Excluding the studies classified as not reported or applicable (16.8%), the remaining studies in other subject domains account for 23% of the overall studies. This result is consistent with that of my meta-analysis that found the dominant subjects for ITS research, since its emergence, continue to be science, math, and physics.

Table 4.18. Subject Domains in BN Studies

Subject Domain	Number of Studies	Percentage
Computer science	38	26.6%
Math	31	21.7%
Physics	17	11.9%
Language Learning	8	5.6%
Medical Education	8	5.6%
Engineering	6	4.2%
Biology	4	2.8%
Aeronautics	3	2.1%
Other Domains -Human Development	1	0.7%
Other Domains - Job Interview Skills	1	0.7%
Other Domains - Teacher Education	1	0.7%
Other Domains- Analytical Skill Training	1	0.7%
Not reported or applicable	24	16.8%

Educational Level

Table 4.19 presents the distribution of BN studies by educational levels. It suggests that, excluding studies classified as not reported or applicable (36.4%), the majority of the studies are implemented within post-secondary settings (39.2%). The total number of studies conducted at elementary and secondary school levels is only half of the number of those studies that have been conducted at post-secondary institutions. Only a small percentage of studies was conducted for professional training (1.4%). ITS researchers' tendency to run studies in post-secondary settings probably stem from their being able to recruit university students more conveniently than it would be to recruit those in elementary and secondary schools, which requires greater effort to coordinate with schools administrators and curricular schedules.

Table 4.19. Educational Level in BN Studies

Educational Level	Number of Studies	Percentage
Elementary	16	11.2%
Secondary	13	9.1%
Post-secondary	56	39.2%
Others - Professionals training (astronaut training, orthopedic surgery training)	2	1.4%
Others - mixed groups of various educational levels	4	2.8%
Not reported or applicable	52	36.4%

Knowledge Type

Table 4.20 presents the distribution of BN studies by knowledge type. Excluding studies classified as not reported or applicable (21%), the majority of the studies, across

all domain areas, investigated procedural knowledge (57.3%). This result is supported by data showing that the top three subject domains, in Table 17, are computing science, math, and physics because the type of knowledge associated with these three subject domains are procedural in nature.

Table 4.20. Knowledge Type in BN Studies

Knowledge Type	Number of Studies	Percentage
Declarative	22	15.4%
Procedural	82	57.3%
Both	9	6.3%
Not reported or applicable	30	21.0%

Targeted Level of Knowledge

Table 4.21 presents the distribution of BN studies by the participants' level of knowledge that was targeted by the ITS. Excluding studies classified as not reported or applicable (53.8%), the majority of the remaining studies were aimed towards novice students (35%). There are only nine studies (6.3%) conducted for students at intermediate or advanced levels and two studies (1.4%) intended for both novice and intermediate levels. This result may indicate that implementing ITSs to support more advanced learners be practically difficult and costly because this would require a large amount of domain knowledge to be modelled and a high degree of computational power to keep track of student performance and internal cognitive processing in student modeling.

Table 4.21. Targeted Level of Knowledge in BN Studies

Targeted Level of Knowledge	Number of Studies	Percentage
Novice	50	35.0%
Intermediate	8	5.6%
Advanced	1	0.7%
All Levels	4	2.8%
Mixture of Novice and Intermediate	2	1.4%
Not reported or applicable	77	53.8%

4.3.4. Research Question 4: What constructs are modeled in BN student modeling (e.g., level of knowledge, affect, motivation, etc.)?

Table 4.22 presents all constructs captured in student models examined in BN studies. The full list of constructs in student modeling are:

a. time spent

b. number of attempts

c. motivation

d. prior knowledge

e. knowledge level or performance score/level

f. affective state

g. learning style

h. demographic

i. others

Table 4.22 reveals that 49 studies (34.3%) out of all 143 studies just modeled students' knowledge or performance in ITSs. An additional 55 studies (38.5%) modeled students' knowledge plus one more variable.

Table 4.22. Constructs Modeled in BN Student Models

Constructs Modeled in BN Student Modeling	Number of Studies	Percentage
e	49	34.3%
ei	40	28.0%
eg	5	3.5%
de	6	4.2%
i	3	2.1%
ef	3	2.1%
edi	1	0.7%
efi	2	1.4%
aei	2	1.4%
f	2	1.4%
aeg	2	1.4%
g	2	1.4%
gi	2	1.4%
fi	2	1.4%
adei	2	1.4%
cefi	1	0.7%
cegi	1	0.7%
cei	1	0.7%
cdef	1	0.7%
abehi	1	0.7%
aeghi	1	0.7%
adgi	1	0.7%
aefi	1	0.7%
ai	1	0.7%
dehi	1	0.7%
def	1	0.7%
efgi	1	0.7%

Constructs Modeled in BN Student Modeling	Number of Studies	Percentage
egdi	1	0.7%
eghi	1	0.7%
egi	1	0.7%
eiag	1	0.7%
fe	1	0.7%
fh	1	0.7%
gf	1	0.7%
hegi	1	0.7%

Note: Given the number of variables, the constructs are presented in the abbreviated format in English letters.

To understand more clearly which constructs are most captured in student models, the frequency with which each construct appears in BN studies is calculated and listed in Table 4.23. It shows that knowledge level has the highest frequency among all constructs in BN studies. Excluding the category *others*, learning style, affective state, and prior knowledge are the three most modeled constructs in BN studies. Student motivation is the least modeled construct in BN studies. Only one included study explicitly captured the number of attempts on correct answers.

Table 4.23. Frequency of Constructs Modeled in BN Student Models

Frequency of Constructs Modeled in BN Student Modeling	Number of Studies
A. Time Spent	11
B. Number of Attempts	1
C. Motivation	4
D. Prior Knowledge	14
E. Knowledge Level or Performance Score/Level	127
F. Affective State	16
G. Learning Style	21
H. Demographic	6
I. Others	69

Table 4.24 presents a variety of constructs modeled in the category *others* in all BN studies. The data suggests that students' self-regulatory skills, learning activities and behaviors when interacting with ITSs, and misconceptions are the top constructs modeled in the BN studies.

Table 4.24. Constructs Modeled in Student Models in Others Category

Constructs Modeled in BN Student Modeling	Number of Studies
Self-regulatory/metacognitive skills/states	12
Students' learning activities and behaviors	11
Students' misconceptions	8
Scientific inquiry skill	5
Students' errors/mistakes	3
Students' interests, personality and preferences	3
Scaffolds, hints and help students used and ignored	3
Scaffolds, hints and help students used and ignored	3
Students' collaborative skills	2
State of problem-solving	2
Students' confidence in their work	2
Students' carelessness	2
Parameters of students' learning interaction	2
Whether the student elicits help or not	1
Test items related to knowledge nodes	1
Social and psychological signals/cues	1
The students' learning paths identified by indexes that highlight students' cognitive actions and navigation behaviors	1
Students' learning rate	1
Students' computer expertise and connection speed	1
Learner's belief about his cognitive achievement	1
Students' behavior in playing games	1

Constructs Modeled in BN Student Modeling	Number of Studies
The randomized position of each representation icon in trial	1
Type and amount of help asked	1
Kind of help received	1
Students' disengagement	1
Students' learning problems	1
Students' plan/goals and associated actions	1
Students' attention and gaze data	1
Similarity between the problem and a candidate example students worked on	1

4.3.5. Research Question 5: What pedagogical approaches are applied in ITS BN studies?

Table 4.25 presents the distribution of the pedagogical approaches with reference to learning theories taken in the BN studies. Overall, there are 11 studies (7.7%) that embedded collaborative and social learning theories in their respective learning design, supporting students' social and interactive learning in a collaborative manner. There are only two studies (1.4%) that used a modified cognitive tutor following the ACT-R design. One study reported to have adopted both cognitive load theory and problem-based learning to guide design for developing conceptual, practical and strategic knowledge for students in Indian rural area (Toshniwal & Yammiyavar, 2013). Another study integrated collaborative, social learning theory with community of practice to build a collaborative learning environment to support students' situated, multi-step problem-solving skills in algebra (Singley, Singh, Fairweather, Farrell, & Swerling, 2000).

Excluding almost half of all studies classified as not reported or applicable (49.7%), a majority of researchers in the remaining 72 studies have followed some type of pedagogical framework to design and guide their research studies (39.9%, in *other* category). For instance, to promote teaching practice for apprentice teachers, Chieu and Herbst (2011) adopted Piagetian epistemology to design an intelligent learning

environment that fosters learning by having students' adapt their prior knowledge to the feedback they receive. The Andes ITS study followed the coached problem-solving approach to provide students with relevant assistance in overcoming impasses while solving a physics problem (Gertner, Conati, & VanLehn, 1998). Moreover, Weragama and Reye (2013) used Vygotsky's Zone of Proximal Development (ZPD) to scaffold students with exercises for the PHP language programming in the ITS. Such efforts indicate that ITS researchers have actively worked to integrate pedagogical elements embodied in various theoretical framework to obtain effective learning results during the learning design phase.

Table 4.25. Pedagogical Approach with Theoretical References in BN Studies

Pedagogical Approach with Theoretical Reference	Number of Studies	Percentage
Cognitive Learning Theory	2	1.4%
Collaborative and Social Learning Theory	11	7.7%
Instructional Design Theory	0	0.0%
Others	57	39.9%
Cognitive Learning Theory + Others	1	0.7%
Collaborative and Social Learning Theory + Others	1	0.7%
Not reported or applicable	71	49.7%

4.3.6. Research Question 6: What instructional strategies are applied in ITS BN studies?

Table 4.26 presents the distribution of the respective instructional strategies applied in BN studies. Excluding 78 studies classified as not reported or applicable (54.5%), a majority of researchers in the remaining 65 studies have integrated some type of instructional strategies into their ITS design (27.3%). There are 13 studies (9.1%) categorized as problem-oriented, which probably correspond to the procedural, problem-solving learning activities required in the top three subject domains based on the previous discussion. Five studies adopted collaborative instructional strategies to promote students' interaction during learning. Only two studies were classified as either example-oriented, or natural language dialogue, or program visualization.

Table 4.26. Instructional Strategies applied in BN Studies

Instructional Strategies	Number of Studies	Percentage
Example-oriented	2	1.4%
Program Visualization	2	1.4%
Program Analysis	0	0.0%
Natural Language Dialogue	2	1.4%
Collaborative	5	3.5%
Problem-oriented	13	9.1%
Natural Language Dialogue + Problem-oriented	1	0.7%
Collaborative + Problem-oriented	1	0.7%
Others	39	27.3%
Not reported or applicable	78	54.5%

The learning pace a student is able to take in an ITS reflects the instructional intention of enabling students the flexibility to control how they learn. A learner-paced ITS allows students to explore the learning content at their own will and follow their own strategies and speed. Students are not restricted to study in a pre-defined curricular order and complete learning tasks within a given period of time. System-paced learning in ITS,

on the other hand, involves students following a structured approach to learn. As such, students have to complete work as instructed with little control over what to learn next. Table 4.27 presents the number of studies in each category in the BN studies. Excluding studies classified as not reported or applicable (49.7%), among the remaining 57 studies, there are more ITSs designed in a self-paced manner (23.8%) than a system-paced one (16.1%).

Table 4.27. System-paced or Learner-paced in ITS Design in BN Studies

System-paced or Learner-paced in ITS Design?	Number of Studies	Percentage
Self-paced	34	23.8%
System-paced	23	16.1%
Not reported or applicable	86	60.1%

Table 4.28 presents the number of pedagogical agents in the BN studies. Typically, a pedagogical agent is an animated character that plays a virtual instructor or peer role in an ITS to guide or provide feedback to students on their performance during learning. The data suggests that, excluding the studies classified as not reported or applicable (18.9%), a majority of studies (47.6%) do not have a pedagogical agent in the design. 37 studies (25.9%) reported having one pedagogical agent in their ITSs. Eleven studies (7.7%) reported to have included multiple pedagogical agents in their ITS design.

Table 4.28. Number of Pedagogical Agents in BN Studies

Number of Pedagogical Agents	Number of Studies	Percentage
Zero	68	47.6%
One	37	25.9%
Multiple	11	7.7%
Not reported or applicable	27	18.9%

4.3.7. Research Question 7: What are the characteristics of BN student models?

BN is a powerful technique to model process under uncertainty with its "rigorous probabilistic formalism with graphical representation and efficient mechanisms" (Pek, & Poh, 2004, p. 282). In this section, the following characteristics of BN student modeling were identified that may contribute to the effectiveness of ITSs to promote learning.

Ability to Handle Inherent Uncertainty in Student Modeling

When an ITS is devised to model and predict students' knowledge, it involves a high level of inherent uncertainty because students' mental states are unlikely to be explicitly revealed and can implicitly change without being easily noticed from overt behaviors. In this study, it was found that, the wide adoption of BN in building student models lies in its power to handle uncertainty of predicting student's current states (knowledge, emotion, performance etc.). BN modeling maintains a graphical probabilistic/belief network to represent the interdependency and interrelationship of nodes of interests, e.g., knowledge. In a typical BN student model, knowledge and skills assessments on students and structures are performed on a corresponding belief network. By eliciting evidence of change from the most recent representation of a student's knowledge, BN algorithm runs automated reasoning to predict the current knowledge state and then propagates updates to all related nodes in the entire belief network (Mayo & Mitrovic, 2000). This process is conducted through a rigorous computational mechanism based on a sound theoretical foundation in statistics and weighs all corresponding evidence in the network, instead of directly relying upon students' subjective responses and actions, which may include guesses or slips of the tongue (Reye, 2004).

Moreover, BN student modeling can capture and represent changes in students' mental states over time. For instance, DBN, a specific type of BN, is used to model temporal evolution of a variable node. In a DBN student model, each node is tracked with multiple values, reflect changes across time slices and indicate their temporal interdependencies (Conati & Zhao, 2004). Each time slice represents a moment depicting

the occurrence of a student activity when evidence is collected for a particular node in the network (Seffrin, Rubi, & Jaques, 2014). Therefore, for a specific time-sliced node, representing the current moment, it has one more time slice for a previous state and another one for a future state (Ting & Zadeh, 2007). These time-sliced nodes are interconnected by the temporal arcs that define their influence on each other as a result of temporal changes, which can be further analyzed to understand students' progress over time. Therefore, BN student modeling performs inferences more efficiently in a highly structured belief network than other types of student modeling that operate using a set of diagnostic rules. Therefore, BN student modeling can obtain a more accurate prediction of student knowledge, enabling the ITS to make effective pedagogical decisions to accommodate ongoing study needs for students in multidimensional learning constructs.

Ability to Model a Wide Range of Constructs in Student Modeling

In this analysis, thirty-seven different constructs were found to have been modeled in BN student models, when combining variables listed in Table 4.22 and 4.23 (see Research Question #4). Among the 37 constructs, many are related to student learning including cognitive, metacognitive, behavioral, aptitudinal, motivational and affective dimensions. This diversity in the aspects of learning considered by a BN student model is believed to provide the model with the capability to reason and make inferences from across a variety of learning dimensions, thereby allowing ITS researchers greater flexibility in selecting a subset of constructs most relevant to their respective research questions.

BN student modeling can also support a wide range of dimensions upon which an ITS can adapt to students because of the large amount of student data captured by the model. Other types of student models are restricted to a limited number of constructs. For instance, in addition to students' current mastery of knowledge, the BN student model also captures information on students' help-seeking behaviors. Therefore, an ITS can adapt instruction to provide hints to students when they are stuck but fail to elicit any system help. In addition to the adapted learning content, the ITS can also guide students to learn

to monitor their progress and seek help when necessary. The adapted instruction is generated by integrating evidence collected from students' behaviors as they progress, instead of relying solely on students' current state of knowledge. Through leveraging multi-dimensional learning constructs, BN student modeling increases the user modeling bandwidth with a greater quantity and quality of user information to assess and represent students' latest states. In this way, BN student modeling can provide more targeted adaptation for personalized learning to individual students.

Ability to Build Short-term and Long-term Student Models in an ITS

BN student models can be used to track and assess students' ongoing state changes, over the short-term and long-term, in targeted modeling constructs such as knowledge, affect or metacognitive skills. Short-term student modeling refers to the tracking of the student learning process while a student works on the current learning task and moves to a new one upon completion. During this dynamic process, new evidence of the student's progress is derived and used to propagate to each of the relevant nodes in the entire BN network as the posterior probability. In this way, the updated BN network maintains a long-term student model with all known nodes representing individual learning tasks that the student has completed so far. When the student starts a new learning task, the posterior probability of those existing nodes in the long-term model are used as the prior probabilities to initiate the related nodes in the new short-term model. New evidence will be further collected along with the student's progress on the new task. This loop will continue until all tasks are completed. In other words, the short-term student model represents a snapshot of the student's most recent behaviors or actions and is kept temporarily during BN modeling. The long-term student model depicts the full view of the students' current states on the targeted constructs and is kept permanently as part of BN modeling.

Some ITS researchers have made both the short-term and long-term models explicit in their design, whereas others have kept them as part of one BN model. For instance, Conati and Zhao (2004) explicitly built both short-term and long-term models using DBN for their Prime Climb educational game on number factorization for elementary

school students. The short-term model was used to track a student's current action on solving a specific number factorization, contextualized within the act of climbing a particular mountain in the game. Once the student completes climbing this mountain, the student's progress is saved in the short-term model and used to update the probabilities of the corresponding knowledge nodes in the long-term model. Once the update is complete, the short-term model is disposed while a new short-term model is simultaneously created when the student starts to climb a new mountain. In the physics ITS Andes (Conati, Gertner, & Vanlehn, 2002), students' current problem-solving behavior is captured in a task-specific network as part of the overall Bayesian student model. The network makes inferences based on the student's most current behaviors and actions on solving a problem. Once the problem is solved, the short-term network saves the posterior probability of the related physics rule nodes in the long-term student model and is discarded afterwards. Similarly, a new task-specific network is created when the student starts working on a new problem. The process of tracking closely the student's progress allows for the constant updates needed to support long-term student models.

4.4. Discussion

4.4.1. Summary of the Results

In this study, I address seven research questions to uncover the characteristics of ITSs with BN student modeling from 143 studies, conducted in 32 countries, between the years of 1992 to 2014. For almost half of these studies, only the framework for designing an ITS using BN student modeling was provided. The remaining studies implemented ITSs with BN student models, which were evaluated for effectiveness with participants; these

studies employed experimental designs or collected feedback from participants using self-report instruments such as questionnaires. Also, the analysis reveals that BN ITSs have been widely used in a great variety of subjects across a range of educational levels, including professional training. Excluding studies that were classified as not reported or applicable, the majority of the remaining studies reported positive experimental and learning outcomes.

For Bayesian student modeling, a total of 37 constructs were tracked and modelled in student models. These constructs spanned multiple dimensions at the cognitive, metacognitive, behavioral, aptitudinal, motivational and affective levels. Capturing such a variety of user data in real-time makes it possible for researchers to identify specific student attributes during the course of learning, thereby enabling the provision of appropriate instructional support to intervene and remediate knowledge as deemed necessary. Excluding those studies classified as *not reported*, the results indicate that the studies integrated one or multiple types of instructional strategies into the ITS design to facilitate different levels of learning activities and interactions. Similarly, among the 143 studies, researchers adopted one or multiple pedagogical frameworks to guide the design of ITSs and build authentic, interactive learning environments to promote meaningful learning among students.

4.4.2. Quality of Reporting

To include more ITS studies using BN student modeling in the review, the selection criteria allowed including studies without evaluation experiments but having ITS design proposals. This wider spectrum of studies confirmed that reporting on fundamental features of ITS research was insufficient, an issue that had been already been identified in the meta-analysis. In some empirical studies, reporting omitted key methodological features such as attrition, types of test items in assessment tests, descriptions of comparison groups, and reliability of outcome measures. Insufficient reporting about details pertaining to pedagogical approaches and instructional strategies also existed in

both empirical and design studies. In many studies, especially those relating to BN design, not enough explanation was given about how underlying pedagogical frameworks guided researchers' designs for the modeling components of the ITS. The rationale for selecting and implementing corresponding instructional strategies and learning scaffolds are not often provided; this gap in reporting makes it challenging for readers to understand how these strategies can be adopted within their particular learning contexts. Without presenting the ways in which an ITS design embodies pedagogical and learning theories, the reporting of these studies could be perceived as more of a technical exercise than as being one that aims to forward research within the learning sciences.

4.4.3. What This Review Can Tell Us About ITS

This review confirms that Bayesian networks have been effectively used to model a great variety of learning constructs spanning multiple facets of learning in student models. Among these constructs, a set of contextual and process variables are modeled in addition to common variables describing student performance. For instance, a student's prior knowledge and gender are profiled and analyzed as contextual characteristics that may influence performance (Stevens & Thadani, 2006). To understand the course of learning per se, process variables, such as the number of attempts to answer an exercise correctly or student actions logged over a period of time, are captured to gain insights on learning as it occurs (e.g. Bedor, Mohamed, & Shedeed, 2004). This diverse set of learning constructs operationalized at a fine level of granularity through student modeling makes it possible to elicit evidence of learning in its immediate context and facilitates the examination of how learning occurs incrementally and gradually over time. The data provide a rich resource for discerning the interplay of those constructs as they shape students' motivational and participatory behaviors during learning, thereby providing insights on how to assess the impact of interactions of varied levels of complexity on student learning. The captured data can increase predictive accuracy of student states by facilitating the transformation of user performance data into learner parameters leveraged through BN inference.

Despite the great potential of utilizing these variables to seek evidence of learning, it should be noted that almost half of the 143 studies were still at the ITS design phase at the time they were published rather than at the prototype implementation or end product stage of ITS development. Needless to say, the instantiation of the design involves a process of validating related assumptions and hypotheses as well as discarding unrealistic requirements. The infeasibility of capturing all or only a portion of the constructs in a student model, as proposed in many studies, becomes apparent during the implementation phase and is inevitably constrained by a number of unforeseen factors. Therefore, the operationalization in modeling certain learning constructs is still in doubt and should not be taken for granted as being readily applied within a typical ITS.

Another observation emerging from this review is that, among the 143 studies, the research questions primarily center around whether a proposed BN student model accurately predicts students' performance or whether the new ITS can effectively promote learning gains. In many cases, when a positive learning outcome had been detected, a conclusion was often broadly drawn and the effect attributed to the overall individualized treatment in the ITS. Further investigations to uncover the actual underlying learning mechanism with regard to the associated contributing variables were not conducted. Although some BN studies captured a broad set of variables in the student model, the data were not further analyzed to understand how the variables influenced learning. Therefore, despite perceived ITS benefits, it would be challenging for ITS researchers to generalize about whether a set of similar attributes could be captured and that learning effects could be replicated across a broader audience and different learning contexts. Thus, I recommend that ITS researchers not only direct attention to the general evidence of learning gains but also explore critical underlying parameters and corresponding contexts surrounding positive learning outcomes. Hypotheses derived from the set of contextual, process and student performance variables should be developed and then tested. Such investigations would yield insights into how student behavior during learning is influenced by the interplay of these variables.

With its high predictive power, Bayesian student modeling is capable of diagnosing the student's current state accurately and efficiently. Yet, diagnosing students properly is just the very first step in facilitating individualized learning through an ITS. A series of

sequent steps have to take place in the respective domain and tutoring modules in order to produce highly individualized tasks. Yet, many studies did not articulate sufficiently how new evidence of learning, detected in a student model, was interpreted and utilized in these two modules. The procedure of feeding the evidence nodes and causal links from the Bayesian student models into these two modules often remains a black box to readers. The justification for offering certain response feedback and differential instructional strategies to support students is also not well documented. Therefore, I suggest that the entire flow of user data, its processing and utilization across the critical components in an ITS should be transparent and clearly presented to readers so relevant research can be properly evaluated and drawn upon by the ITS community.

Chapter 5. General Summary

5.1. Summary of the Results of Two Reviews

In this thesis, I conducted a meta-analysis and a review of Bayesian student modeling to understand the effect and development of ITSs over the past decades. The meta-analysis examined empirical studies that evaluated the effects of ITSs prior to 2013. The overall result suggests ITSs outperform other modes of instruction except one-to-one human tutoring and small-group instruction. The review of Bayesian student modeling investigated the strength of BN and summarized studies between 1992 and 2014 regarding characteristics of Bayesian student modeling and its capacity to handle uncertainty in predicting student performance. The meta-analysis focused on detecting the effects of ITSs with regard to a set of moderator variables describing characteristics of ITSs, and followed a statistical procedure that combines data from multiple evaluation studies. The BN review, on the other hand, explored the research questions investigated and the respective modeling parameters captured in the student models in individual studies. The review complements the meta-analysis, which required the exclusion of a great number of studies in its analysis because of the need for strict inclusion criteria. By broadening the kind of studies that could be analyzed, the review provided in greater insight into the characteristics of current research practice in the ITS community, and facilitated the exploration of more facets of the student capabilities modeled in ITSs at a granular level.

Through the synthesis of results from 107 studies, the meta-analysis provides strong evidence that ITSs can be effectively used to complement and substitute for other modes of instruction, in a variety of academic subjects, across a range of educational levels. In the BN review, a majority of studies found ITSs generated positive learning outcomes in comparison to other instructional modes. In a way, both classes of studies underline the effectiveness of ITSs where students receive individualized instruction and assistance to reach their own learning goals.

In addition, these two studies found ITSs have been widely applied to a number of common academic subjects. In both sets of studies, mathematics, physics and computer

science are the most popular subjects for which ITS researchers strive develop students' declarative and procedural knowledge. Furthermore, both kinds of studies show that ITSs have been adopted to support pedagogical instruction across a range of educational levels. More ITSs tend to be built for and used in post-secondary institutions rather than elementary, secondary or other educational settings. Similarly, both studies found that a majority of ITSs have been designed to target novice students or participants with low prior domain knowledge rather than more advanced students or professionals.

In addition to these findings, each study identified a specific set of characteristics that bear on the research questions addressed by ITS studies. While the previous chapters have elaborated upon these results, the following section will discuss their implications for the overall design of ITSs.

5.2. Implications for the Design of ITSs

A Fundamental Question for the Design of an ITS

Both reviews in this study revealed that ITS studies share the assumption that one-on-one tutoring is the most effective instructional method, outperforming all others. The researchers aimed to reproduce or improve on the beneficial effects of one-to-one human tutoring. This effect is thought to be primarily attributed to the individualized instruction made available to students during learning with an ongoing accurate snapshot of a student's current state of knowledge or skillset. The more accurately an ITS can diagnose a student's current state of learning, the greater is the likelihood the ITS can provide productive assistance. Thus, the ITS needs to simulate a human tutor to reproduce results achieved through the one-on-one interaction between a human tutor and a student. The limited access students have to teachers for one-on-one help provides a rationale and impetus for developing ITSs. Consequently, much time, resources and effort have been invested to pursue this research goal. Since the emergence of the term *intelligent tutoring*

system, in 1982, a variety of ITS research groups have worked on numerous projects and their stance on this underlying assumption is now firmly entrenched within the educational community. Based on results of the reviews that I conducted, I believe it is time to revisit assumptions underlying ITS design. We should begin by asking this fundamental question: What do we expect an ITS to do in terms of supporting student learning? Alpha Go's victory in beating a professional human Go player reflects the great advances made in machine learning (Knight, 2016). Given the evolution of artificial intelligence, which had been highly influential in the development of ITS research, should providing individualized instruction to mimic the efficacy of a human tutor still be the only goal targeted by the design an ITS? Can we not go beyond what a human tutor commonly does to support students with revolutionary practices beyond the limitations of human beings? For instance, could an ITS model a student's misconceptions and visualize how they occur by dynamically tracing the reasoning processes in the student's mind? Artificial intelligence has evolved greatly over the past decade. However, our fundamental understanding of designing an ITS seems not to have evolved apace. The technical advancement of AI alone does not bear pedagogical fruit until researchers identify approaches that handle uncertainties to predict key constructs in learning activities and integrate these within instructional and pedagogical requirements. Therefore, I urge ITS researchers to review the affordances of new techniques in artificial intelligence as well as in other related fields, and to reflect on how these may provide new directions to ITS design, potentially leading to breakthroughs in the ITS field.

Are We Asking the Right Research Questions?

Both analyses reported here show a majority of empirical studies have focused on evaluating the overall effectiveness of an ITS used within a particular project. There is nothing wrong with this research approach. However, we, as ITS researchers, must move beyond this evaluative research question, which is generally investigated as part of a preliminary assessment of computer-assisted applications. For many research initiatives,

we have not ventured further to investigate how underlying features in the ITS contribute to the effects of ITSs. In this light, I propose three new directions for future research.

Firstly, the BN review identified that learning constructs were used by student models as the foundation for tracing changes over time and those constructs served as reference points for understanding student learning. Constructs were used to diagnose the state of student learning so the tutoring module could assign appropriate learning tasks. While studies used a set of these learning constructs for tracing, many studies did not further investigate how these constructs influenced particular aspects of student learning. Therefore, I recommend the influence of such constructs be further examined at a finer granularity in relation to relevant cognitive and affective variables and student performance. With a better understanding of how specific constructs relate to and mediate student learning, researchers can better identify which variables are needed to develop more accurate student models.

Secondly, further to conducting general ITS evaluations by comparing an ITS to other modes of instructions, I suggest comparing two or more versions of the same ITS with each version operationalizing a theoretically informed design variation. For instance, a comparison could evaluate one version of an ITS providing immediate feedback and another with delayed feedback. Another study might investigate the influence on learning progress when students are provided varying degrees of control over pace in an ITS. Such comparative research could be very informative regarding effects of particular ITS features and could, in turn, provide insight for designing ITSs tailored to a wider range of students according to their varied learning attributes and needs.

Thirdly, both reviews revealed much attention has been given to exploring the development of ITS student models. While accurate student modeling is critical to the design of an ITS that provides adapted learning support, pedagogical decisions that influence the design of appropriate student support is of equivalent weight. Without the capability to offer the most appropriate assistance when it is required by a student, an ITS would simply be a traditional CAI application. Therefore, I recommend pedagogical strategies be explicitly considered when designing ITS assistance to provide kinds of interventions needed to achieve desired learning outcomes. Specifically, a series of

related pedagogical questions should be explored with reference to particular learning theories in the ITS field to inform the design of the tutoring/teaching module in an ITS. These questions include:

- How should we map the associated learning tasks, activities and resources to a specific student state?
- How can we ensure that students are always on the right track?
- What strategies can help students close the gap in their knowledge?

Do We Have an Ecological Environment to Build the Right ITS for Students?

In the early 1980s, ITSs emerged as a new field to research how an instructional system could intelligently predict student progress and offer the most effective assistance to guide students' problem solving and learning. Since then, the field of ITSs has embraced a series of technological and research innovations. It has attracted numerous ITS groups and researchers and has grown rapidly. In spite of much research in this area, the meta-analysis revealed a lack of common terminology in this field. For instance, different ITS systems present the same domain knowledge using different terminology. Furthermore, a lack of commonly defined research practices exists within the field. Various researchers have built their own system architectures, have run distinctive computing and reasoning algorithms, and have applied varied strategies to represent domain knowledge (Glavinić, Stankov, Zelić, & Rosić, 2008). Reusability and interoperability of intelligent systems were rarely considered in the research. In this regard, a lack of common ground leads to misunderstanding and confusion as well as the inefficient investment of time and money on similar research projects across regions instead of building upon existing research. Therefore, I propose the ITS field begin to unify research practices. Instead of working in isolation, I recommend research efforts draw upon a common conceptual framework, taxonomy of terminology and design practices. Only when ITS researchers

begin to work collectively will research advance through a research environment that fosters collaborative contributions to ITS development (Brawner, Goodwin, & Sottilare, 2016).

5.3. Quality of Reporting

As previously discussed, both the meta-analysis and BN review study revealed there is insufficient structure to organize reports about fundamental features in primary ITS research. In many empirical studies, methodological features were not reported, such as means and standard deviations, student attrition rates, study duration, the reliability of outcome measures, and the random assignment of participants, In addition, pedagogical and instructional theories which informed the design of an ITS were often omitted from research reports. Two possible reasons can be offered to account for insufficient reporting of research details. The primary reason relates to challenges imposed by page or word limits for publications, which requires researchers to omit information in favour of accommodating more study findings. As more researchers adopt the same approach to handle this challenge, the omission of methodological details gradually becomes a common practice that is accepted over time. The other reason is the lack of a common standard for reporting on ITS design principles and pedagogical considerations, primary features and key components, experiment details, and research findings and implications. Interdisciplinary research demands a shared body of knowledge and research insight to support consistent practices in the ITS community. Despite the variation of research goals, differentiation of pedagogical and learning theories, heterogeneity of methodologies, diversity of student background, I advocate for a common conceptual framework and terminology to be developed to promote shared understanding and to shed light on the research efforts of the discipline.

5.4. Limitations and Constraints

For this dissertation, I conducted two studies to examine the effectiveness of ITSs and their characteristics that support individualized instruction. One limitation was limited primary studies that were available for coding. For the meta-analysis, I searched and coded most relevant articles independently. I received some help from another PhD student who assisted me with the search and the coding of newly published studies while I was reviewing, compiling and coding the existing pool of publications. Although we discussed and reached agreement on codes for specific moderator variables within particular studies retrieved during her search, we did not work on the same set of research studies in parallel. This process limited our ability to determine our level of inter-rater agreement and to validate the accuracy of our coding. However, an author of the published manuscript of this meta-analysis conducted a random verification of the coded spreadsheet and the results found codes were consistent and accurate. He also questioned the coders' understanding of a number of moderator variables including research setting, comparison instruction and ITS intervention. These variables were revisited and re-coded by me after all four authors reached an agreement on how they should be interpreted and coded.

For the Bayesian student modeling study, I independently searched for relevant articles in the bibliographic databases and developed the coding book in reference to those used in other meta-analyses. The variables were revised to align with the research questions of this study. Based on experience gained in the meta-analysis study, I clearly defined specific moderator variables by listing the possible selection items and creating the *Other* category to avoid potential changes in the conceptualization of a specific moderator variable, which sometimes occurs over the lengthy course of coding work. Then, I performed all screening, reviewing and coding activities. No additional coder was involved during this phase. During coding, to ensure an overall consistency, I noted incidents involving subjective judgement. After all coding was completed, subsequently I randomly revisited 40 coded items to verify coding accuracy. During this verification process, I modified three items; therefore, the overall accuracy was judged consistent and acceptable.

In this study, all data came from a pool of empirical literature, which was analyzed, following secondary data analysis procedures, to discover new findings by synthesizing years of research efforts. Needless to say, the data sets across studies varied greatly in terms of data collection, research methodology, and research participants and experimental setting. Although not a limitation from the perspective of this study, concerns about secondary data analysis can arise because data sources vary in quality and questions may be raised about the appropriateness of the original data synthesis, especially with regard to the research questions. These issues may raise concerns about the kinds of inferences that can be made beyond the research studies investigated (Mueller & Hart, 2011).

Inbuilt biases affecting the design and implementation of heterogeneous experimental studies also pose constraints on inferences made and the way in which the synthesized results should be interpreted. A potential bias is the sponsorship of ITS research projects, possibly favoring ITSs over other instructional modes. In the studies reviewed, there was a tendency to set project goals and experimental conditions towards what may be more feasibly achieved by an ITS, making it easier to produce better learning outcomes for ITSs over other modes of instruction. Another potential inbuilt bias pertains to the possibility that more detailed instructional planning and attention were given to the design of the ITS condition over the other instructional practices found in the control conditions. Although the study's results do present strong evidence for the effectiveness of ITSs, one cannot infer ITSs should replace other modes of instruction because the effect may have resulted from supplementary well-planned and executed instructional strategies rather than from particular features of ITS.

Another limitation associated with both analyses conducted for this research study is the lengthy process of searching, coding and analyzing vast amounts of research literature. New publications, which were not available during the initial search efforts, were not part of the analysis. For the meta-analysis study, the 107 studies only included those that were available until 2012. Subsequent publications, between 2013 and 2016, have not yet been reviewed. For the study of Bayesian student modeling, the 143 research articles included in the review were published between 1994 and 2014. New trends and potential shifts in research interests for articles published between 2014 and 2016 have

not yet been investigated. Integrating recent publications into both studies would be an appropriate approach for a follow-up study to increase confidence in the conclusions drawn from the two studies. This may also bring about new insights that reshape our understanding of and the implications for the design of ITSs.

5.5. Conclusion

For this study, I conducted two ITS reviews to examine the effect of ITSs and to bring new insights to the ITS community. This research makes two major contributions to the ITS literature. First, it builds upon the work of prior ITS reviews, which only investigated the effectiveness of ITSs within particular subsets of ITS literature (by subject or level of education). The current meta-analysis instead provides a comprehensive evaluation that combines a larger number of studies from all subject domains across all educational levels in the ITS literature. It provides a holistic view of the effects of ITSs on student learning. By synthesizing a more comprehensive set of studies, the meta-analysis was able to attain the greater statistical power needed to detect the overall mean effect sizes. Its results are more conclusive than other prior reviews and provides evidence on the effectiveness of ITSs over other types of non-ITS learning environments.

The second contribution is a further exploration of findings discovered in the comprehensive meta-analysis. By taking a closer look at a particular type of student modeling technique deployed in ITSs, this review integrates a great number of research studies on Bayesian student modeling. It explores the specific types of research questions investigated in studies, the set of learning constructs captured in the student model, and associated instructional and pedagogical approaches that are used to identify which ITSs

characteristics could have contributed to student learning. The results of the second review not only support the findings of the first review, but they also reveal the collective effort of ITS researchers modeling the student modeling using a diverse set of variables at a fine level of granularity. The identification of a wide range of learning constructs in the current ITS BN literature provides ITS researchers with new directions for research. These include further investigating a specific subset of these constructs to determine how these mediate student learning. The study also revealed that researchers should trace incremental changes on variables during the learning process as well as observe how learning occurs in this context. Furthermore, findings of this review provide insight about using BNs to increase predictive power when faced with uncertainty in student modeling and more accurately diagnosing students' most current state of learning to elicit evidence for learning.

Overall, this study reviewed a large quantity of ITS literature and synthesized years of research efforts to increase understanding of the current state of ITS research. It soughts insights for refining existing practices and presented researchers with a new set of principles to advance the development of ITSs. It also provided the entire ITS community with guidance for future directions in research. This study was conducted with the view that an ITS will eventually be developed that will give students authentic personalized learning support that surpasses what even a human tutor can provide, thereby enabling students to construct their own understandings when freely inquiring into and exploring the world.

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis on the cognitive correlates of bilingualism. *Review of Educational Research, 80,* 207-245.

Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology, 104,* 250-263.

Ahuja, N. J., & Sille, R. (2013). A Critical Review of Development of Intelligent Tutoring Systems : Retrospect , Present and Prospect. *IJCSI International Journal of Computer Science Issues*, *10*(4), 39-48.

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

Anderson, J. (1996). A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365. http://doi.org/10.1037//0003-066X.51.4.355

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *The Journal of Learning Sciences*, *4*(2), 167–207. http://doi.org/10.1207/s15327809jls0402_2

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A Theory of Higher Level Cognition and its Relation to Visual Attention. *Human-Computer Interaction*, *12*, 439–462. http://doi.org/10.1207/s15327051hci1204_5

Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. *Advances in Instructional Psychology*, *5*, 1-34.

Anderson, J. R., & Skwarecki, E. (1986). The Automated Tutoring of Introductory Computer Programming. *Communications of the ACM*, *29*(9), 842–849. http://doi.org/10.1145/6592.6593

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssafidou, E., … Sabouret, N. (2013). The TARDIS framework: Intelligent virtual agents for social coaching in job interviews. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8253 LNCS*, 476–491. http://doi.org/10.1007/978-3-319-03161-3_35

Arroyo-Figueroa, G., & Sucar, L. E. (1999). A temporal Bayesian network for diagnosis and prediction. *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, 13-20.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. *International Journal of Artificial Intelligence in Education*, *24*(4), 387–426. http://doi.org/10.1007/s40593-014-0023-y

*Arnott, E., Hastings, P., & Allbritton, D. (2008). Research methods tutor: Evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods, 40*(3), 694-698.

Baker, R. S. J. d. (2010). Mining Data for Student Models. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems* (pp. 323–337). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-14363-2_16

Basturk, R. (2005). The Effectiveness of Computer-Assisted Instruction in Teaching Introductory Statistics. *Educational Technology & Society*, *8*(2), 170-178. http://doi.org/1176-3647

*Bedor, H. S., Mohamed, H. K., & Shedeed, R. A. (2004). A general architecture of student modelto assess the learning performance in intelligent tutoring systems. In *Electrical, Electronic and Computer Engineering, 2004. ICEEC '04. 2004 International Conference on* (pp. 173–178). in proceedings. http://doi.org/10.1109/ICEEC.2004.1374413

Ben-Gal, I. (2007). Bayesian Networks. *Encyclopedia of Statistics in Quality & Reliability*, *1*(1), 4. http://doi.org/10.1002/wics.48

Ben-Gal, I. (2009). Bayesian networks. In R. Ruggeri, F. Faltin, & F. Kenett (Eds.), *Wiley Interdisciplinary Reviews: Computational Statistics* (Vol. 1, pp. 307-315). http://doi.org/10.1002/wics.48

Benjamin, L. T. (1988). A history of teaching machines. *American Psychologist, 43*(9), 703–712. http://doi.org/10.1037/0003-066X.45.4.551.b

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chippenham, Wiltshire: Wiley.

Brawner, K., Goodwin, G., & Sottilare, R. (2016). Agent-Based Practices for an Intelligent Tutoring System Architecture. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience: 10th International Conference, AC 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part II* (pp. 3–12). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-39952-2_1

Brusilovsky, P., & Millán, E. (2007). User Models for Adaptive Hypermedia and Adaptive Educational Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization* (pp. 3-53). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-72079-9_1

Buchheit, R. B., Garrett, J. H., Lee, S. R., & Brahme, R. (2000). A Knowledge Discovery Framework for Civil Infrastructure: A Case Study of the Intelligent Workplace. *Engineering with Computers*, *16*(3), 264–274. http://doi.org/10.1007/s003660070009

Carbonell, J. (1970). AI in CAI: An artificial intelligence approach to computer aided instruction. *Science, 167,* 190-202.

Ceci, S. J., & Papiero, P. B. (2005). The rhetoric and reality of gap closing: When the "have-nots" gain but the "haves" gain even more. *American Psychologist, 60,* 149-160. doi:10.1037/0003-066X.60.2.149

Chrysafiadi, K., & Virvou, M. (2015). A Novel Hybrid Student Model for Personalized Education. In *Advances in Personalized Web-Based Education* (pp. 61–90). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-12895-5_3

Chieu, V.M. and Herbst, P. (2011). Designing an Intelligent Teaching Simulator for Learning by Practicing in the Practice of Mathematics Teaching. *ZDM-The International Journal of Mathematics Education, 43*(1), 105-117.

Ciloglugil, B., & Inceoglu, M. (2010). Exploring the state of the art in adaptive distributed learning environments. *Computational Science and Its Applications–ICCSA 2010*, 556–569. Retrieved from http://www.springerlink.com/index/P20N71157R023017.pdf

Ciloglugil, B., & Inceoglu, M. M. (2012). User Modeling for Adaptive E-Learning Systems. In B. Murgante, O. Gervasi, S. Misra, N. Nedjah, A. M. A. C. Rocha, D. Taniar, & B. O. Apduhan (Eds.), *Computational Science and Its Applications -- ICCSA 2012: 12th International Conference, Salvador de Bahia, Brazil, June 18-21, 2012, Proceedings, Part III* (pp. 550–561). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-31137-6_42

Conati, C. (2009). Intelligent tutoring systems: new challenges and directions. *Proceedings of the 21st International Jont Conference on Artificial Intelligence*, 2-7. Retrieved from http://videolectures.net/ijcai09_conati_its/

Conati, C. (2010). Bayesian student modeling. *Studies in Computational Intelligence*, *308*, 281-299. http://doi.org/10.1007/978-3-642-14363-2_14

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian Networks to Manage Uncertainty in Student Modeling. *User Modeling and User-Adapted Interaction*, *12*(4), 371-417. http://doi.org/10.1023/A:1021258506583

*Conati, C., & Zhao, X. (2004). Building and evaluating an intelligent pedagogical agent to improve the effectiveness of an educational game. *Proceedings of the 9th International Conference on Intelligent User Interfaces*, 6–13. http://doi.org/10.1145/964442.964446

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

Corbett, A. T., & Bhatnagar, A. (1997). Student modeling in the ACT programming tutor: Adjusting a procedural learning model with declarative knowledge. *Courses and Lecturesinternational Centre for Mechanical Sciences*, 243-254. http://doi.org/10.1007/978-3-7091-2670-7_25

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent Tutoring Systems. *Science*, *37*(1986), 849-874. doi:10.1007/978-3-319-07221-0

Crowder, N. A. (1959). Automatic Tutoring By Means of Intrinsic Programming. In E. Galanter (Ed.), *Automatic Teaching: The State of the Art* (pp. 109-116). New York: John Wiley & Sons.

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Dean, T., & Kanazawa, K. (1989). A Model for Reasoning About Persistence and Causation. *Computational Intelligence*, *5*(3), 142-150. http://doi.org/10.1111/j.1467-8640.1989.tb00324.x

Dede, C. (1986). A review and synthesis of recent research in intelligent computer-assisted instruction. *International Journal of Man-Machine Studies*, *24*(4), 329-353.

Denson, N. (2009). Do curricular and co-curricular diversity activities influence racial bias? A meta-analysis. *Review of Educational Research, 79,* 805-838.

Desmarais, M. C., & Baker, R. S. J. D. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9-38. http://doi.org/10.1007/s11257-011-9106-8

D'Mello, S. K., & Graesser, A. C. (2012). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems, 2*(4), 2-29.

Duchastel, P., & Imbeau, J. (1988). Intelligent Computer-assisted Instruction (ICAI): Flexible Learning Through Better Student-Computer Interaction. *Journal of Information Technology*, *3*(2), 102–105. http://doi.org/10.1057/jit.1988.18

Driscoll, M. (2000). Ten Things We Know About Teaching Online. In *Teaching online. Technology for Learning Newsletter, Lakewood Publications, 2*(1), 4-5.

Edwards, T. O. (1970). Optimizing computer assisted instruction by applying principles of learning theory. Report retrieved on June 14, 2016 from ERIC. http://eric.ed.gov/?id=ED105899

Everson, H. T. (1995). Modeling the student in intelligent tutoring sytems: The promise of a new psychometrics. *Instructional Science*, *23*(5-6), 433–452. http://doi.org/10.1007/BF00896881

Fiedler, L. J., Sucar, L. E., & Morales, E. F. (2015). Transfer learning for temporal nodes Bayesian networks. *Applied Intelligence*, *43*(3), 578–597. http://doi.org/10.1007/s10489-015-0662-1

Fournier-Viger, P., Nkambou, R., & Nguifo, E. M. (2010). Building intelligent tutoring systems for ill-defined domains. *Studies in Computational Intelligence*, *308*, 81-101. doi:10.1007/978-3-642-14363-2_5

Gamboa, H., & Fred, A. (2002). Designing intelligent tutoring systems: a bayesian approach. *Enterprise Information Systems III*, (Fred 1994), 146. Retrieved from http://books.google.com/books?hl=en&lr=&id=AvtEClDmY3cC&oi=fnd&pg=PA146&dq=Designing+intelligent+tutoring+systems:+a+bayesian+approach&ots=wsV9ozE2vC&sig=6di_UhAtD6NfybwiOTaQktfv6zc

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

Gagné, R. & Driscoll, M. (1988). Essentials of Learning for Instruction (2nd Ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gertner, A., Conati, C. and VanLehn, K. (1998). Procedural help in Andes: Generating hints using a Bayesian network student model. *Proceedings of the 15th National Conference on Artificial Intelligence*, Madison, Wisconsin, U.S.A., pp. 106-111.

Ghahramani, Z. (1998). Learning Dynamic {Bayesian} Networks. *Lecture Notes in Computer Science*, *1387*, 168-197. http://doi.org/10.1007/BFb0053992

Glavinić, V., Stankov, S., Zelić, M., & Rosić, M. (2008). Intelligent Tutoring in the Semantic Web and Web 2.0 Environments. In M. D. Lytras, J. M. Carroll, E. Damiani, R. D. Tennyson, D. Avison, G. Vossen, & P. Ordonez De Pablos (Eds.), *The Open Knowlege Society. A Computer Science and Information Systems Manifesto: First World Summit on the Knowledge Society, WSKS 2008, Athens, Greece, September 24-26, 2008. Proceedings* (pp. 172-177). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-87783-7_21

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 180–192. http://doi.org/10.3758/BF03195563

Graesser, A., Olney, A., Ventura, M., & Jackson, G. (2005). AutoTutor's Coverage of Expectations during Tutorial Dialogue. *FLAIRS Conference*. Retrieved from http://www.aaai.org/Papers/FLAIRS/2005/Flairs05-085.pdf

Graesser, A. C., Person, N. K., Harter, D., & the Tutoring Research Group (2001). Teaching tactics and dialogue in AutoTutor. *International Journal of Artificial Intelligence in Education*, *12*, 257-279.

Greco, T., Zangrillo, A., Biondi-Zoccai, G., & Landoni, G. (2013). Meta-analysis: pitfalls and hints. *Heart, Lung and Vessels*, *5*(4), 219-225.

Grimes, D.M. (1977). Computers for learning: The uses of computer assisted instruction (CAI) in California public schools. Sacramento, CA: California State Department of Education.

Guruler, H., & Istanbullu, A. (2014). Modeling Student Performance in Higher Education Using Data Mining. In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends* (pp. 105–124). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-02738-8_4

Hategekimana, C.P. (2008). *Cognition and Technology: Effectiveness of intelligent tutoring systems for software training* (Unpublished doctor's thesis). Iowa State University. Ames, Iowa, USA. Retrieved on August 14, 2016 from http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=2386&context=etd

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hämäläinen, W., & Vinni, M. (2006). Comparison of Machine Learning Methods for Intelligent Tutoring Systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 525–534). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/11774303_52

Han, K. W., Lee, E., & Lee, Y. (2010). The impact of a peer-learning agent based on pair programming in a programming course. *IEEE Transactions on Education*, *53*(2), 318–327. http://doi.org/10.1109/TE.2009.2019121

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random-effects models in meta-analysis. *Psychological Methods, 3,* 486-504.

Hernandez-Leal, P., Sucar, L. E., Gonzalez, J. A., Morales, E. F., & Ibarguengoytia, P. H. (2011). Learning Temporal Bayesian Networks for Power Plant Diagnosis. In K. G. Mehrotra, C. K. Mohan, J. C. Oh, P. K. Varshney, & M. Ali (Eds.), *Modern Approaches in Applied Intelligence: 24th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2011, Syracuse, NY, USA, June 28 -- July 1, 2011, Proceedings, Part I* (pp. 39-48). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-21822-4_5

Heusch, Guillaume Marcel, S. (2007). Face Authentication with Salient Local Features and Static Bayesian Network. *International Conference on Biometrics (ICB)*, 878-887.

Holmes, D. E., & Jain, L. C. (2008). Introduction to Bayesian Networks. *Innovations in Bayesian Networks*, 1-5. doi:10.1007/978-3-540-85066-3_1

Kass, R. (1989). Student Modeling in Intelligent Tutoring Systems-Implications for User Modeling. In A. Kobsa & W. Wahlster (Eds.), *User Models in Dialog Systems* (pp. 386–410). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-83230-7_14

Knight, W. (2016, March 18). Five Lessons from AlphaGo's Historic Victory. Retrieved Jan 20,2017, from https://www.technologyreview.com/s/601072/five-lessons-from-alphagos-historic-victory/

Kodaganallur, V., Weitz, R. R., & Rosenthal, D. (2005). A Comparison of Model-Tracing and Constraint-Based Intelligent Tutoring Paradigms. *International Journal of Artificial Intelligence in Education*, *15*(2), 117-144. Retrieved from http://dl.acm.org/citation.cfm?id=1434925.1434928

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent Tutoring Goes To School in the Big City. *International Journal of Artificial Intelligence in Education*, *8*, 30-43.

Koedinger, K. R., & Anderson, J. R. (1998). Illustrating Principled Design: The Early Evolution of a Cognitive Tutor for Algebra Symbolization. *Interactive Learning Environments*, *5*(1), 161–179. http://doi.org/10.1080/1049482980050111

Koedinger, K. R. (2001). Cognitive Tutors as Modeling Tool and Instructional Model. *Smart Machines in Education: The Coming Revolution in Educational Technology*, 145-168.

Kopp, K. J., Britt, M. A., Millis, K., & Graesser, A. C. (2012). Improving the efficiency of dialogue in tutoring. *Learning and Instruction, 22*, 320-330.

Kulik, J. A., Kulik, C. L. C., & Bangert-Drowns, R. L. (1985). Effectiveness of computer-based education in elementary schools. *Computers in Human Behavior*, *1*(1), 59–74. http://doi.org/10.1016/0747-5632(85)90007-X

Kwon, W. Y., & Suh, I. H. (2012). A temporal Bayesian network with application to design of a proactive robotic assistant. *Proceedings - IEEE International Conference on Robotics and Automation*, 3685-3690. http://doi.org/10.1109/ICRA.2012.6224673

Larson, M. B., Burton, J.K., & Moore, D.M. (2008). Programmed Technologies. In J. M. Spector, M. D. Merrill, J. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology: A Project* (pp. 187-197). Routledge.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

Lockee, B. B., Moore, D. M., & Burton, J. K. (2004). Foundations of programmed instruction. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 545-569). Mahwah, NJ: Lawrence Erlbaum.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, *106*(4), 901-918. doi:10.1037/a0037123

Mann, B. L. (2009). Computer-Aided Instruction. *Wiley Encyclopedia of Computer Science and Engineering*. 583-592.

Martin, B. (1999). Constraint-based modelling: Representing student knowledge. *New Zealand Journal of Computing*, *7*(2), 30–38. Retrieved from http://www.cosc.canterbury.ac.nz/~bim22/paper_nzjc.pdf

Martens, A., & Uhrmacher, A. M. (2004). Modeling of Tutoring Processes in Intelligent Tutoring Systems. In S. Biundo, T. Frühwirth, & G. Palm (Eds.), *KI 2004: Advances in Artificial Intelligence: 27th Annual German Conference on AI*, KI 2004, Ulm, Germany, September 20-24, 2004. Proceedings (pp. 396–409). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-30221-6_30

Mayo, M. & Mitrovic, A.(2001). Optimising ITS behaviour with Bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education, 12*, 124-153.

Mayo M., Mitrovic A. and McKenzie J. (2000). CAPIT: An Intelligent Tutoring System for Capitalisation and Punctuation. In Kinshuk, Jesshope C. and Okamoto T. (Eds.) *Advanced Learning Technology: Design and Development Issues*, Los Alamitos, CA: IEEE Computer Society (ISBN 0-7695-0653-4), pp. 151-154.

Mayo, M., & Zealand, N. (2001). Optimising ITS behaviour with Bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, *12*, 124–153. http://doi.org/http://www.ijaied.org/iaied/ijaied/

McDonald, J. K., Yanchar, S. C., & Osguthorpe, R. T. (2005). Learning from programmed instruction: Examining implications for modern instructional technology. *Educational Technology Research and Development*, *53*(2), 84–98. http://doi.org/10.1007/BF02504867

Mihajlovic, V., & Petkovic, M. (2001). Dynamic bayesian networks: A state of the art. *CTIT Technical Reports Series*, *34*, 1–37.

Millán, E., Pérez-de-la-Cruz, J. L., & García, F. (2003). Dynamic versus Static Student Models Based on Bayesian Networks: An Empirical Study. In V. Palade, R. J. Howlett, & L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003, Oxford, UK, September 2003. Proceedings, Part II* (pp. 1337–1344). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-45226-3_181

Mitrovic, A. (2010). Modeling domains and students with constraint-based modeling. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems* (Vol. 308, pp. 63–80). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-14363-2_4

Mitrovic, A. (2012). Fifteen years of constraint-based tutors: What we have achieved and where we are going. *User Modeling and User-Adapted Interaction*, *22*(1-2), 39–72. http://doi.org/10.1007/s11257-011-9105-9

Mitrovic, A., Martin, B., & Suraweera, P. (2007). Intelligent tutors for all: Constraint-based modeling methodology, systems and authoring. *IEEE Intelligent Systems*, *22*, 38–45. http://doi.org/10.1109/MIS.2007.74

Mueller, C. E., & Hart, C. O. (2011). Effective use of secondary data analysis in gifted education research: Opportunities and challenges*. Gifted Children, 4*(2), 3.

Murray, R. C., Vanlehn, K., & Mostow, J. (2004). Looking Ahead to Select Tutorial Actions: A Decision-Theoretic Approach. *International Journal of Artificial Intelligence Education*, *14*(3,4), 235-278. Retrieved from http://dl.acm.org/citation.cfm?id=1434913.1434915

Murray, W. R. (1998). A practical approach to Bayesian student modeling. *Fourth International Conference on Intelligent Tutoring Systems*, 424-433. http://doi.org/10.1007/3-540-68716-5

Myneni, L. S., Narayanan, N. H., Rebello, S., Rouinfar, A., & Pumtambekar, S. (2013). An interactive and intelligent learning system for physics education. *IEEE Transactions on Learning Technologies, 6*(3), 228-239.

National Research Council (1992). *Combining Information: Statistical Issues and Opportunities for Research*. Washington, DC: National Academy Press.

Nesbit, J. C., Adesope, O. O., Liu, Q., & Ma, W. (2014). How effective are intelligent tutoring systems in computer science education? In *2014 IEEE 14th International Conference on Advanced Learning Technologies* (pp. 99-103). doi:10.1109/ICALT.2014.38

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, *76,* 413-448.

Nguyen, L., & Do, P. (2009). Combination of bayesian network and overlay model in user modeling. *International Journal of Emerging Technologies in Learning*, *4*(4), 41–45. http://doi.org/10.3991/ijet.v4i4.684

Niedermayer, D. (1998). An Introduction to Bayesian Networks and their Contemporary Applications. *Innovations in Bayesian Networks*, 1-12. http://doi.org/10.1007/978-3-540-85066-3_5

Nkambou, R., Bourdeau, J. & Mizoguchi, R. (Eds.), (2010). *Advances in Intelligent Tutoring Systems*. Heidelberg : Springer Verlag, Studies in Computational Intelligence, Volume 308, 510p.

Nwana, H. S. (1990). Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, *4*(4), 251-277. http://doi.org/10.1007/BF00168958

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, *24*(4), 427–469. http://doi.org/10.1007/s40593-014-0029-5

Ohlsson, S. (1986). Some principles of intelligent tutoring. *Instructional Science*, *14*(3-4), 293–326. http://doi.org/10.1007/BF00051825

Ohlsson, S. (1994). Constraint-Based Student Modeling. In J. E. Greer & G. I. McCalla (Eds.), *Student Modelling: The Key to Individualized Knowledge-Based Instruction* (pp. 167-189). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-662-03037-0_7

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review*, *103*(2), 241-262. http://doi.org/10.1037/0033-295X.103.2.241

Orphanou, K., Stassopoulou, A., & Keravnou, E. (2014). Temporal abstraction and temporal Bayesian networks in clinical domains: A survey. *Artificial Intelligence in Medicine*, *60*(3), 133-149. http://doi.org/10.1016/j.artmed.2013.12.007

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157-159.

O'Shea, T. & Self, J. (1983). *Learning and teaching with computers: Artificial intelligence in education*. Sussex: Harvester Press.

Owen, R. S., & Aworuwa, B. (2005). Progammed Instruction, Programmed Branching, and Learning Outcomes. *Encyclopedia of Information Science and Technology (IV),* 2326-2329.

Poole, B. J. (1995). Education for an Information Age: Teaching in the computerized classroom, Madison: WCB Brown & Benchmark.

Pacella, D. (2014). Report on artificial cognitive modelling and intelligent tutor: a Literature Review, 1-22.

Pek, P. K., & Poh, K. L. (2004). a Bayesian Tutoring System for Newtonian Mechanics: Can It Adapt To Different Learners? *Journal of Educational Computing Research*, *31*(3), 281–307. http://doi.org/10.2190/VDAP-K5BX-EJX1-D8QY

Polson, M. C., & Richardson, J. J. (Eds.). (1988). *Foundations of Intelligent Tutoring Systems*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc.

Ranade, M. D. (2006). Development of CAI presentations for science teaching and overview of research findings. *International Journal of Science and Mathematics Education*, *4*(4), 763–789. http://doi.org/10.1007/s10763-005-9022-7

Ramani, P., & Patadia, H. (2013). Reaction of Students on Developed Computer Assisted Instruction for Teaching Arithmetic, *Scientific & Academic Publishing, 3*(1), 37-42. http://doi.org/10.5923/j.edu.20130301.06

Ramírez-Noriega, A., Juárez-Ramírez, R., & Martínez-Ramírez, Y. (2016). Evaluation module based on Bayesian networks to Intelligent Tutoring Systems. *International Journal of Information Management*. http://doi.org/10.1016/j.ijinfomgt.2016.05.007

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*, 1-33.

Ricks, B. W., & Mengshoel, O. J. (2010). Diagnosing Intermittent and Persistent Faults using Static Bayesian Networks. *Time*, 1-8.

Rosenberg, H., Grad, H.A., & Matear, D. W. (2003). The effectiveness of computer-aided, self-instructional programs in dental education: A systematic review of the literature. *Journal of dental education, 67*(5), 524-532.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638- 641.

Sackney, L., & Mergel, B. (2007). Contemporary Learning Theories, Instructional Design and Leadership. In J. M. Burger, C. F. Webber, & P. Klinck (Eds.), *Intelligent Leadership: Constructs for Thinking Education Leaders* (pp. 67–98). Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-1-4020-6022-9_5

Salman, A. R. (2013). The Use of Intelligent Tutoring System for Developing Web-based Learning Communities. *IJCSI International Journal of Computer Science*, *6*(1), 157-161.

Saettler, P. (1990). *The evolution of American educational technology*. Englewood, CO: Libraries Unlimited, Inc.

Salgado-Zapata, P.J. (1989). An Intelligent Computer-Assisted Instruction system for Underway Replenishment (Unpublished master's thesis). Naval Postgraduate School, USA. Retrieved on June 14, 2016 from http://calhoun.nps.edu/bitstream/handle/10945/27103/intelligentcompu00salg.pdf?sequence=1

Sani, S., & Aris, T. (2014). Computational Intelligence Approaches for Student / Tutor Modelling : A Review. *Fifth International Conference on Intelligent Systems, Modelling and Simulation*, 72-76. doi:10.1109/ISMS.2014.21

Santhi, R., Priya, B., & Nandhini, J. (2013). Review of intelligent tutoring systems using bayesian approach. *arXiv Preprint arXiv:1302.7081*. Retrieved from http://arxiv.org/abs/1302.7081

Seffrin, H. M., Rubi, G. L., & Jaques, P. A. (2014). A Dynamic Bayesian Network for Inference of Learners' Algebraic Knowledge. *In Proceedings of the 29th Annual ACM Symposium on Applied Computing* (pp. 235–240). New York, NY, USA: ACM. http://doi.org/10.1145/2554850.2555062

Self, J. (1990). Theoretical Foundations for Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education, 1*(4), 3-14.

Self, J. (1999). The defining characteristics of intelligent tutoring systems research : ITSs care, precisely. *International Journal of Artificial Intelligence in Education*, *10*(1998), 350-364.

Shapiro, G. (1992). Special issue, Knowledge Discovery in Data and Knowledge Bases. *International Journal of Intelligent Systems, 7*(7).

*Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. PytlikZillig, R. Bruning, and M. Bodvarsson (Eds.). *Technology-based education: Bringing researchers and practitioners together* (pp. 169-202). Greenwich, CT: Information Age Publishing.

Shute, V. J., & Psotka, J. (1996). Intelligent Tutoring Systems: Past, present, and future. In D. Jonassen (Eds.), *Handbook of research for educational communications and technology* (pp. 570-600). New York, NY: Macmillan.

Singley, M. K., Singh, M., Fairweather, P., Farrell, R., & Swerling, S. (2000). Algebra Jam: Supporting Teamwork and Managing Roles in a Collaborative Learning Environment. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (pp. 145–154). New York, NY, USA: ACM. http://doi.org/10.1145/358916.358985

Skinner, B. F. (1959). The programming of verbal knowledge. In E. Galanter (Ed.), *Automatic teaching: The state of the art* (63–68). New York: John Wiley & Sons.

Skinner, B.F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24*(2), 86-97.

Skinner, B.F. (1968). *The Technology of Teaching*. New York: Appleton-Century-Crofts.

Sklavakis, D., & Refanidis, I. (2008). An Individualized Web-Based Algebra Tutor Based on Dynamic Deep Model Tracing. In J. Darzentas, G. A. Vouros, S. Vosinakis, & A. Arnellos (Eds.), *Artificial Intelligence: Theories, Models and Applications: 5th Hellenic Conference on AI, SETN 2008, Syros, Greece, October 2-4, 2008. Proceedings* (pp. 389-394). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-87881-0_38

Sleeman, D. H., & Brown, J. S. (1982). Intelligent tutoring systems: an overview. In D.H. Sleeman & J.S. Brown (Eds.), *Intelligent Tutoring Systems,* 1-11. London: Academic Press.

Song, L., Kolar, M., & Xing, E. P. (2009). Time-Varying Dynamic Bayesian Networks. *Advances in Neural Information Processing Systems*, 1732-1740. Retrieved from http://papers.nips.cc/paper/3716-time-varying-dynamic-bayesian-networks

Stansfield, J.C., Carr, B., & Goldstein, I.P. (1976). *Wumpus advisor I: a first implementation of a program that tutors logical and probabilistic reasoning skills*. Unpublished AI Memo 381. Cambridge, MA: Massachusetts Institute of Technology, AI Laboratory.

Stellan, O., & Mitrovic, A. (2006). Constraint-based knowledge representation for individualized instruction. *Computer Science and Information Systems*, *3*(1), 1-22. http://doi.org/10.2298/CSIS0601001S

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K-12 students' mathematical learning. *Journal of Educational Psychology, 105*(4), 970-987. http://dx.doi.org.proxy.lib.sfu.ca/10.1037/a0032447

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology*, *106*(2), 331-347.

Stevens, R. H., & Thadani, V. (2006). A Bayesian network approach for modeling the influence of contextual variables on scientific problem solving. *International Conference on Intelligent Tutoring Systems, ITS*, *4053 LNCS*, 71-84. http://doi.org/10.1007/11774303_8

Sucar, L. E. (2015). Dynamic and Temporal Bayesian Networks. In *Probabilistic Graphical Models: Principles and Applications* (pp. 161-177). London: Springer London. http://doi.org/10.1007/978-1-4471-6699-3_9

Suebnukarn, S., & Haddawy, P. (2007). COMET: A Collaborative Tutoring System for Medical Problem-Based Learning. *IEEE Intelligent Systems*, *22*(4), 70-77. http://doi.org/10.1109/MIS.2007.66

Suraweera, P., & Mitrovic, A. (2004). An intelligent tutoring system for entity relationship modelling. *International Journal of Artificial Intelligence in Education*, *14*, 375-417. Retrieved from http://ir.canterbury.ac.nz/handle/10092/318\nhttp://hdl.handle.net/10092/318

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics (6th ed.)*. Boston: Allyn and Bacon.

Ting, C. Y., Khor, K. C., & Sam, Y. C. (2012). Evidence conflict analysis approach to obtain an optimal feature set for Bayesian tutoring systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7473 LNCS*, 576-583. http://doi.org/10.1007/978-3-642-34062-8_75

Ting, C.-Y., & Zadeh, M. R. B. (2007). Assessing Learner's Scientific Inquiry Skills Across Time: A Dynamic Bayesian Network Approach. In C. Conati, K. McCoy, & G. Paliouras (Eds.), *User Modeling 2007: 11th International Conference, UM 2007, Corfu, Greece, July 25-29, 2007. Proceedings* (pp. 207–216). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-73078-1_24

Toshniwal, O., & Yammiyavar, P. (2013). Intelligent Interactive Tutor for Rural Indian Education System. In A. Agrawal, R. C. Tripathi, E. Y.-L. Do, & M. D. Tiwari (Eds.), *Intelligent Interactive Technologies and Multimedia: Second International Conference, IITM 2013, Allahabad, India, March 9-11, 2013. Proceedings* (pp. 186–199). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-37463-0_17

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences, 2*, 1-59.

Vanlehn, K., Lynch, C., & Schulze, K. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of …*, *15*(3), 1–51. http://doi.org/citeulike-article-id:7925171

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197-221. doi:10.1080/00461520.2011.611369

Vassileva, J. (1990). A classification and synthesis of student modelling techniques in intelligent computer-assisted instruction. In D. H. Norrie & H.-W. Six (Eds.), *Computer Assisted Learning: 3rd International Conference, ICCAL '90 Hagen, FRG, June 11--13, 1990 Proceedings* (pp. 202–213). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/BFb0020881

Virvou, M. (2003). Modelling the Knowledge and Reasoning of Users in a Knowledge-Based Authoring Tool. *International Journal of Continuing Engineering Education and Life-Long Learning*, *13*(3-4), 399-412.

Vlasselaer, J., Meert, W., Van Den Broeck, G., & De Raedt, L. (2016). Exploiting local and repeated structure in Dynamic Bayesian Networks. *Artificial Intelligence*, *232*, 43–53. http://doi.org/10.1016/j.artint.2015.12.001

Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning*, *6*(2), 279-306.

Wang, F. L., & Wong, T. L. (2008). Designing Programming Exercises with Computer Assisted Instruction. In J. Fong, R. Kwan, & F. L. Wang (Eds.), *Hybrid Learning and Education: First International Conference, ICHL 2008 Hong Kong, China, August 13-15, 2008 Proceedings* (pp. 283–293). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-85170-7_25

Ward, M. (2002). *A Template for CALL Programs for Endangered Languages* (Unpublished master's thesis). Dublin City University, Dublin, Ireland.

Wenger, E. (1987). *Artificial Intelligence Tutoring Systems.* Los Altos: Morgan Kaufmann Publishers.

Weragama, D., & Reye, J. (2013). The PHP Intelligent Tutoring System. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 9-13, 2013. Proceedings* (pp. 583–586). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-39112-5_64

Winne, P. H., & Baker, R., S., J., D. (2013). The potentials of educational data mining for researching metacognition, motivation and self-regulated learning. *Journal of Educational Data Mining, 5*, 1-8.

Wolfe, C. R., Widmer, C. L., Reyna, V. F., Hu, X., Cedillos, E. M., Fisher, C. R., … Weil, A. M. (2013). The development and analysis of tutorial dialogues in AutoTutor Lite. *Behavior Research Methods*, *45*(3), 623-636. http://doi.org/10.3758/s13428-013-0352-z

Woolf, B. P. (2008). *Building Intelligent Interactive Tutors: Student-centered Strategies for Revolutionizing e-Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. http://doi:10.1007/BF02680460

Wu, A. K. ., & Lee, M. (1998). Intelligent tutoring systems as design. *Computers in Human Behavior*, *14*(2), 209-220. doi:10.1016/S0747-5632(98)00002-8

You, X., Liu, G., Long, W., & Pan, Z. (2013). Based on the Teaching Module in Intelligent Tutoring System. In J. Xu, M. Yasinzai, & B. Lev (Eds.), *Proceedings of the Sixth International Conference on Management Science and Engineering Management: Focused on Electrical and Information Technology* (pp. 277-282). London: Springer London. http://doi.org/10.1007/978-1-4471-4600-1_24

Zapata-Rivera, J.-D., & Greer, J. E. (2004). Interacting with inspectable Bayesian student models. *International Journal of Artificial Intelligence in Education*, *14*(2), 127-163.

Ziani, A., & Motamed, C. (2007). Temporal Bayesian Networks for Scenario Recognition. *Scia*, *4522*, 689–698.

Appendix A.

Coding Form Used for the Meta-analysis

CodingBook-Met
aAnalysis-Wentin

Appendix B.

Coding Form Used for the Review of Bayesian Network

BN-Review_Codebook_Thesis.xlsx