Cognitive Assessment in a Computer-Based Coaching Environment in Higher Education: Diagnostic Assessment Of Development of Knowledge and Problem-Solving Skill in Statistics

Zhidong Zhang

Department of Educational and Counselling Psychology

McGill University, Montreal

April, 2007

A Thesis Submitted to McGill University in Partial Fulfillment of the Requirements of the Degree of Ph.D. in Educational Psychology

Your file Votre référence
ISBN: 978-0-494-32337-3
Our file Notre référence
ISBN: 978-0-494-32337-3

Acknowledgment

I wish to acknowledge and express my appreciation to these contributors and helpers to my studies and thesis completion.

I would like to convey the deepest thanks to my supervisor, Dr. Carl Frederiksen who contributed his insight, wisdom and dedication to my thesis project. His excellent ideas and suggestions are reflected in many aspects from the research proposal to the completion of last version of the thesis. Thanks also go to Dr. Susanne Lajoie and Dr. Carolyn Turner. As committee members, they proposed many suggestions and comments in improving the thesis from structure and design, to analysis.

Special thanks go to these professors, whose courses contributed to the development of my thesis: Dr. Mark Aulls, Dr. Robert Bracewell, Dr. Janet Donin, Dr. Jim Ramsay and Dr. Cynthia Weston.

I owe special thanks to Dr. Alenoush Saroyan for her valuable advice in helping me to successfully manage my thesis schedule.

I would like to convey thanks to my colleagues who shared academic dialogues and interests. Furthermore, they presented many valuable opinions. I would like to mention a few of them: Ernest Bauer, George Carani, Kevin Chin, Susan Jingyan-Lu, Tom Patrick, and Erika Vonhuene.

I am grateful to my family. My mother was a constant source of motivation in my studies; I received tons of encouragement from her. My wife Li Zhi has always contributed her effort, support and time. My daughter Alice Yang Zhang contributed her research expertise.

Finally, many thanks go to groups of graduate students who devoted their time responding to the stand-alone test, and to staff members at the executive office and library.

Abstract

Diagnostic cognitive assessment (DCA) was explored using Bayesian networks and evidence-centred design (ECD) in a statistics learning domain (ANOVA). The assessment environment simulates problem solving activities that occurred in a web-based statistics learning environment. The assessment model is composed of assessment constructs, and evidence models. Assessment constructs correspond to components of knowledge and procedural skill in a cognitive domain model and are represented as explanatory variables in the assessment model. Explanatory variables represent specific aspects of student's performance of assessment problems. Bayesian networks are used to connect the explanatory variables to the evidence variables. These links enable the network to propagate evidential information to explanatory model variables in the assessment model. The purpose of DCA is to infer cognitive components of knowledge and skill that have been mastered by a student. These inferences are realized probabilistically using the Bayesian network to estimate the likelihood that a student has mastered specific components of knowledge or skill based on observations of features of the student's performance of an assessment task.

The objective of this study was to develop a Bayesian assessment model that implements DCA in a specific domain of statistics, and evaluate it in relation to its potential to achieve the objectives of DCA. This study applied a method for model development to the ANOVA score model domain to attain the objectives of the study. The results documented: (a) the process of model development in a specific domain; (b) the properties of the Bayesian assessment model; (c) the

performance of the network in tracing students' progress towards mastery by using the model to successfully update the posterior probabilities; (d) the use of estimates of log odds ratios of likelihood of mastery as a measure of "progress toward mastery;" (e) the robustness of diagnostic inferences based on the network; and (f) the use of the Bayesian assessment model for diagnostic assessment with a sample of 20 students who completed the assessment tasks. The results indicated that the Bayesian assessment network provided valid diagnostic information about specific cognitive components, and was able to track development towards achieving mastery of learning goals.

Resumé

L'évaluation cognitive diagnostique (DCA) a été explorée en utilisant les réseaux bayésiens et le design centré sur l'évidence (EDC) dans un domaine d'apprentissage des statistiques (ANOVA). L'environnement de l'évaluation stimule des activités de la résolution des problèmes qui se sont produits dans un environnement en-ligne d'apprentissage des statistiques. Le modèle d'évaluation est composé des construits d'evaluations et des modèles d'évidence. Les construits d'évaluations correspondent à des composantes du savoir et des compétences procédurales dans un modèle à domaine cognitif et sont représentés comme étant des variables explicatives dans le modèle d'évaluation. Les variables explicatives représentent des aspects spécifiques de la performance des étudiants sur les problèmes d'évaluation. Les réseaux bayésiens sont utilisés pour joindre les variables explicatives aux variables d'évidence. Ces connexions permettent au réseau de propager l'information évidente aux variables du modèle explicatif dans le modèle d'évaluation. Le but de la DCA est de déduire les composantes cognitives du savoir et de la compétence qui ont été maîtrisées par l'étudiant. Ces déductions sont réalisées de façon probabilistique en utilisant le réseau bayésien afin d'estimer la probabilité qu'un étudiant a maîtrisé des composantes spécifiques du savoir ou compétence basé sur des observations de caractéristiques de la performance de l'étudiant d'une tâche d'évaluation.

L'objectif de cette étude était de developer un modèle d'évaluation bayésien qui peut implémenter la DCA dans un domaine spécifique des

statistiques et de l'évaluer en relation au potentiel d'accomplir les objectifs de la DCA. Cette étude a appliqué une méthode pour le développement du modèle au domaine de modèle du score ANOVA afin d'atteindre les objectifs de l'étude. Les résultats documentent: (a) le processus du développement du modèle dans un domaine spécifique; (b) les propriétés du modèle d'évaluation bayésien; (c) la performance du réseau pour tracer le progrès des étudiants vers la maîtrise en utilisant le modèle pour mettre à jour avec succès les probabilités postérieures; (d) l'utilisation des estimés du logarithme de l'odds ratio de la probabilité de la maîtrise comme une mesure du progrès vers la maîtrise; (e) la robustesse des déductions diagnostiques basée sur le réseau; et (f) l'usage du modèle d'évaluation bayésien pour l'évaluation diagnostique avec un échantion de 20 étudiants qui ont completé les tâches d'évaluations. Les résultats indiquent que le réseau d'évaluation bayésien fournit de l'information diagnostique valide à propos des composantes cognitives spécifiques et a été capable de tracer son développement vers l'atteinte de la maîtrise des buts d'apprentissage.

List of Tables

List of Figures

Table of Contents

CHAPTER ONE: INTRODUCTION

1.1 Identification of the Problem

With the rapid development of technology, networked computers are increasingly used in colleges, universities, and other training programs to support innovative problem-based learning through web-based learning tools and on-line learning environments. These new tools and environments can be used to challenge and improve learners' progress in knowledge acquisition, skill development, and problem-solving. However, current assessment tools, procedures and modern theories of testing do not provide effective and precise assessments of student cognitive processes and knowledge development in these learning environments. Therefore, instructors using web-based and other on-line learning systems to support student learning are critically concerned with the problems of identifying student learning strategies, with examining transitions in the development of student expertise (Alexander, 2003; Lajoie, 2003), and with developing cognitive assessments based on student learning processes. Moreover, there is a serious concern that conventional tests are not well suited to newer models of instruction and learning that emphasize the active construction of knowledge and that promote learning in dynamic problem-based environments. Fortunately, advances in cognitive and educational psychology have resulted in a better understanding of how people acquire, organize, and use knowledge, (Greeno, Collins, & Resnick, 1996). For instance, recent theories of evidence-centered assessment design (ECAD) hold promise for developing

effective assessment procedures and tools (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001).

The research topic explored here concerns design and implementation of assessment systems that are based on cognitive objectives using evidence-centered assessment (ECA) (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001). An effective assessment theory must be based on evidence of student conceptual knowledge, procedural knowledge, strategies, knowledge applications, expertise, and skills in task performances. By incorporating diverse aspects of knowledge acquisition and skill development such a theory can be used for diverse assessment purposes.

1.2 Theoretical Frameworks for Learning Assessment

Theories of assessment are very important in the design and development of assessment procedures. The beliefs of assessment researchers and practitioners typically guide the design and implementation of assessment activities. Learning assessment is influenced by its relationships to other areas of education (e.g. curriculum) which are reflected in its theoretical framework. Three theoretical frameworks have influenced my assessment design: (a) Shepard's (2000) historical and learning culture framework, (b) Pellegrino, Chudowsky and Glaser's (2001) cognitive framework, and (c) Mislevy, Steinberg, Almond, Haertel, and Penuel's (2001) evidence-centred framework.

1.2.1 Shepard's Historical and Cultural Framework

In her "learning culture" assessment framework, Shepard (2000) adopts a historical perspective in conceptualizing the interlocking tenets of a model of learning that encompassed not only theories of curriculum and instruction, but also cognitive and constructivist learning theories, and assessment. Shepard focuses on expertise and cognitive abilities as principled and coherent ways of thinking about and representing problems within current cognitive and constructivist learning theories. In Vygotsky's (1978) social-historical perspective, cognitive abilities develop through social interaction. New perspectives on how people learn provide a basis for redesigning and reorganizing curriculum and assessment theoretical foundations that are, thus, epistemologically robust assessments, which must reflect both current cognitive and constructivist theories, and parallel changes in curriculum development. Shepard particularly emphasizes formative assessment in examining student-learning processes and in assessing the step-by-step acquisition of competence. Consequently researchers and practitioners have adopted dynamic assessment, the use of feedback, and student self-assessment.

Shepard locates assessment in the relatively large context of education and culture. Cognitive and social constructivist learning theories are important both in the design of a constructivist curriculum and in the assessment of learning. Assessment must also meet the demands of challenging subject matter and must instantiate what it means to learn and to understand the content of different subject domains (Shepard, 2000). Shepard's learning culture framework

depicts situations in which researchers can identify the role of assessment in larger educational contexts, and its relation to other aspects of education. This framework usefully insists that the design and development of an assessment must be validated with reference to subject domains and classroom contexts of students learning.

1.2.2 Pellegrino's Cognitive Framework

Pellegrino, Chudowsky, and Glaser (2001) regard assessment as a process of reasoning from evidence. They postulate a triadic model for assessment in which the three dimensions of cognition, observation, and interpretation, must be coordinated and interrelated. Cognition refers to how students represent various domains of knowledge and develop competence in these domains. In assessment, observations are made in tasks or situations that allow one to observe students' performance. Interpretations pertain to different methods of making sense of assessment data.

Cognition is represented by means of a theory consisting of a set of beliefs about individuals' knowledge, performance and learning. This idea coincides with Shepard's cognitive and constructivist learning theories. A subtle difference is that in Shepard's framework, cognitive constructivist learning theories encompass constructivist, social, and cultural perspectives, while in Pellegrino et al's framework, cognitive theory encompasses a more abstracted perspective. The theory can be elaborated qualitatively and quantitatively, and in general or specific ways. Cognition can be modeled in various ways. For

example, in a general sense, cognition can be represented as sets of theories or models. In a specific sense, it can be characterized as expertise in a given domain. Qualitatively, it can be represented in terms of declarative and procedural knowledge. While quantitatively, it can be expressed probabilistically in the form of Bayesian networks.

In order to validate cognitive theories and models, observations are interpreted from the perspective of the triad model. This approach was borrowed from natural science. Once a hypothesis has been determined, a theory-driven design is adopted. Data-driven processes use observations as evidence to test hypotheses. However, observations in this model are complex processes and the steps in moving from data collection to establishing rules of evidence and to making inferences based on carefully assembled models of assessment.

Interpretation involves sets of methods for making sense of data. These methods bridge cognition theory and empirical observation (Pellegrino, Chudowsky, & Glaser, 2001). They specify how observations derived from assessment tasks constitute evidence of such cognitive variables as skills and expertise. Methods of interpretation also encompass processes for developing "measurable objects".

Pellegrino, Chudowsky, and Glaser (2001) explicitly characterize triadic assessment theory as determining how the three components are integrated into a coordinated whole to provide us with a theoretical framework for decomposing and analyzing assessment. The development of cognitive processes and expertise is situated (Clancey, 1997) and observations within specific domains

are distributed (Derry, DuRussel, & O'Donnell, 1998). Cognitive theories require specific statistical tools to support interpretations that bridge cognition and observation. Mislevy's model (Mislevy, Steinberg, & Almond, 2000) specifies a design framework that can be used to establish an effective assessment framework in order to deliver effective cognitive assessment.

1.2.3 Mislevy's Evidence-centred Assessment Framework

Evidence-centred assessment (ECA) design was initially developed at the Educational Testing Service by Mislevy, Steinberg, and Almond (2000). This framework provides an effective structure and process for designing, producing, and delivering assessments that can be used to enhance the validity of learning assessments. The statistical mechanism of Bayesian networks can be used effectively to connect cognitive processes and evidence from given task performances.

Mislevy's framework contains three logically connected models: student model, evidence model, and task model.

Student models represent student knowledge, skills and expertise. Although they cannot be directly observed, knowledge, skills and expertise can be indirectly inferred through what students say or do which provide evidence about assessment constructs, that is, student-model variables (Mislevy, Steinberg, & Almond, 2000).

Evidence models consist of two submodels: (a) the evaluative submodel, and (b) the statistical model. The evaluative submodel is composed of a set of

evidence rules by means of which features of student responses and performance are extracted. The statistical model is applied to make inferences about student model variables (assessment constructs) based on evidence variables. In current applications, the statistical model usually takes the form of an item response model, a latent class model, or a Bayesian network model. The actual statistical model adopted in an assessment framework depends on both the student model and the task model

Task models provide a framework for establishing the contexts and tasks which will be used to observe individual performances. They may be expressed in different specifications based on researcher beliefs and goals, and the research design. A task model is crucial to the assessment process because it determines what kinds of task model variables can be extracted from data.

ECA framework provides us with a level of generality that underpins many conventional and web-based assessment formats. Mislevy, Steinberg, Almond, Haertel, and Penuel (2001) provide three examples to illustrate how ECA framework works: (a) the GRE, (b) the Dental Interactive Simulation Corporation (DISC), and (c) the MashpeeQuest, although all three have the same assessment rationale and ECA design, they have different student variables, tasks variables, and optional statistical models.

Shepard's historical and cultural framework, Pellegrino's cognitive framework, and Mislevy's ECA framework describe the relation of assessment to other aspects of education, interpret relations among components of assessment systems, and provide theoretically-based tools for constituting effective

assessment systems. The three assessment frameworks specify the ecology, epistemology, and methodology of an effective assessment system. Therefore, these frameworks will be referred to at different levels when considered later in a detailed analysis of assessment structure.

1.3 Implications of Cognitive and Learning Theory for Assessment

The effective assessment of learning can be understood and evaluated with respect to several theoretical issues. (a) What theories and models of learning and the development of competency have emerged in modern cognitive research? What are their implications for the assessment of student knowledge, performance, and learning? (b) From a cognitive perspective, what deficiencies are there in currently implemented assessment methods? (c) What promising features have been implemented in alternative assessments as compared to conventional assessment procedures? (d) What are the main streams of research exploring various assessment systems or procedures, and how can they contribute to developing systematic cognitive assessment procedures?

1.3.1 Changes in Learning Theories Relevant to Assessment

Assessment is a critical component in Shepard's (2000) triadic framework. Fundamentally, assessment provides feedback to learners, educators and stakeholders about how well a given instructional strategy serves a specific learning process, and how well an assessment procedure promotes student learning. However, as theories of learning have developed, assessment theories

and methods have failed to keep pace (Mislevy, 1993). Assessment theories should be able to identify cognitive processes and results.

Theories of learning have changed a great deal since the beginning of the 20th century. Metaphors of learning are constantly shifting in the natural sciences, computer science, cognitive science, and educational epistemology. Basically, three paradigms have emerged: associationist, information-processing (IP), and situated-constructivist. Unfortunately, test theory, assessment, and assessment procedures have not developed in a parallel fashion.

The associationist paradigm views learning as changing the strength of stimulus-response associations (Mayer, 1996). The assumption is that external behaviors reflect mental processes. Associationists believe that general and precise laws of learning can be identified (Brown, 1994) and applied uniformly and universally across all kinds of learning and learning situations. Historically, researchers began by using the experimental method to observe animal and human mental activity, ultimately transferring research results on animal behavior to human mental processes.

The information-processing paradigm is based on the metaphor that the mind is a symbolic digital computer (Kyllonen, 1996; Mayer, 1996). When psychologists realized that they had to abandon the associationist view of learning as the strengthening of stimulus-response associations, they believed that they could use symbolic data and the information-processing procedures of computers as a metaphor for human cognition. The human-computer metaphor is based on the premise that both computers and humans engage in cognitive

processes such as acquiring and retrieving knowledge. Computers perform cognitive tasks by processing information. They take symbols as inputs, apply operators to that input, and produce outputs. Psychologists argued that humans are also information processors. According to information-processing theory, learning is a process of knowledge acquisition in which information is transmitted from teachers to learners (Mayer, 1996). Learners are information processors, and learning is associated with the construction of mental representations. The strength of this metaphor lies in fact that it allows psychologists to analyze mental processes sequentially and to formulate cognitive models and structures. However, there are limitations to this metaphor in that it ignores the fact that learning is active, schematic, and effortful. Moreover, it does not take into account the emotional, affective, and motivational aspects of learning.

The situated-constructivist paradigm views learning as knowledge construction in the sense that it regards learners as sense makers (Hardy & Taylor, 1997; Mayer, 1996). Learners actively construct rather than passively receive knowledge (Brown, 1994). Human learning involves both knowledge and feelings. The quality of experiences depends on how they function (Confrey, 1995; Duit, 1995; Eisner, 1993; Ernest, 1993; Fosnot, 1993). Perception, conception, and physical action cannot be separated. Learners construct knowledge and meaning from their own experiences. Cognition is embedded in social and cultural contexts where emotions and cognitive activities are jointly situated in both brains and environments. According to this metaphor, learning is socially situated in groups that function as "communities of learners" and in such

socially-enacted activities as "reciprocal learning" (Brown, 1994). This model of learning assumes that learners learn by operating in zones of proximal development (Wertsch, 1985) which are defined as the distance between a learner's current level of knowledge and the level s/he can reach with the help of teachers and/or tools. The model stresses learning in the real world and is beneficial in providing a view of learners as active, strategic, self-conscious, self-motivated, and purposeful participant in learning environments.

1.3.2 An Integrated Learning Model, Information Processing, Constructivism and Situated Metaphors

Unified theories of cognition and learning as multi-faceted phenomena are beginning to emerge (Carroll, 1993, 1998; Horn, 1998; Scarr, 1998). It is necessary to develop more complex theoretical frameworks of learning and cognition.

Information-processing theory is limited by its atomistic view of information and its failure to deal with the fact that humans process information for specific purposes and in specific contexts. The constructivist metaphor stresses the purpose of cognition, and identifies differences between information and knowledge which is constructed out of information. Information is transformed into knowledge in constructivist environments. Knowledge is changed step-by-step into higher-level knowledge and meta-knowledge. And at a certain stage, it emerges as a new format of information which integrates information with other reprocessed information.

Yet, the opposition between IP and constructivist views of cognition and learning is simplistic. Integrating situated cognition (Brown, Collins, & Duguid, 1989; Clancey, 1997; Collin, Brown, & Newman, 1989) and IP can produce a cognitive theoretical framework that can be applied to complex learning processes. The situated perspective partly overlaps with constructivism. Situated cognition maintains that cognitive processes "stretch out" from internal cognitive processes to external social situations. Thus, cognitive processes occur in both the individual and through the interaction of individuals in social situations. Cognitive processes involve both information and knowledge. Thus, cognitive activities comprise strongly interrelated cognitive, social, and cultural aspects. An integrative cognitive theoretical framework must incorporate IP, constructive, and situated metaphors.

The development and application of learning theories inevitably requires corresponding theories of assessment in order to validate and interpret different aspects of learning in various learning environments. However, assessment and measurement theories are as yet not sufficiently well developed to measure and interpret learning in complex authentic environments and domains.

1.3.3 The Discordance between Methods of Learning and Assessment

Unfortunately, recent changes in modes of instruction have produced discrepancies between learning and assessment. Criticisms have been voiced

from researchers in the learning sciences, cognitive science, and educational measurement (Birenbaum, 1996; Embretson, 1993; Hambleton, Swaminathan, & Rogers, 1991; Horn, 1998; Snow, 1998; Snow & Lohman, 1989, 1993; Thissen, 1993). The problem is that theories of measurement and assessment have not adapted themselves to developments in theories of learning and cognition. Assessment has not responded to changes in the interpretation of processes and results in learning and instruction.

Changes in learning and cognitive theory have challenged assessment and test theory since the 1950s (Bechtel, Abrahamsen, & Graham, 1998) when IP theory in the cognitive sciences (Simon & Kaplan, 1989) began to influence perspectives on the measurement and assessment of cognitive processes. Cognitive theory interpreted the learning process as computing symbols within cognitive architectures (Newell, Rosenbloom, & Laired, 1989; Pylyshyn, 1989). Learning in cognitive theory is different from learning as response strengthening in traditional learning theory (Mayer, 1996). Cognitive theories seek to describe what happens in learning in much more detail entailing mental representation, memory, reasoning, and problem solving strategies. Traditional test theory based on the true score model cannot handle this complexity. Changes in learning theory require measurement and test theory to interpret cognitive processes in alternative ways.

1.3.3.1 Changing Conception of Learning Tasks and Processes

There have been important changes in conceptions of learning and instruction. Knowledge and skills are no longer conceived of as limited and static, but rather as extensive and dynamic. A competently functioning person has acquired new knowledge and can use it to solve new unforeseen problems. Learners must have not only declarative knowledge and procedure knowledge, but also application knowledge and strategies. In the information era, learners are seen as adaptable, self-regulated learners, capable of communicating and cooperating with others. Required competencies include: (a) cognitive competencies: problem solving and critical thinking; (b) meta-cognitive competencies: self-reflection; and (c) social competencies: communicating and cooperating (Birenbaum, 1996). Such competencies call for instructional strategies and alternative forms of learning which in turn require new strategies and procedure for measuring and testing in order to provide effective feedback and assessment of learning (Shepard, 2000).

1.3.3.2 Changing Learning Environments

Clearly, changes in learning environments should inform and be informed by changes in theories of cognition and learning. New learning tools such as videotapes, computers, and the World Wide Web can be seen in more and more classrooms. Research has focused on intelligent tutorial systems, computer-coached learning, virtual learning, and case-based learning (Arcos, Muller, Orue, Arroyo, Leaznibarrutia, & Santaner, 2000; Hmelo, 1998; Lajoie & Lesgold, 1989).

There is greater focus on problem-based and collaborative learning. Instructors and learners have reasons to believe that the objective of learning is not just to acquire declarative and procedural knowledge and skills, but also to develop effective strategies for understanding, creating and applying knowledge to new situations. They also recognize that virtual, web-based, simulated, and problem-oriented and collaborative learning environments can introduce different learning processes and outcomes. Learning activities in such complex learning environments can lead to experiences in which cognitive processes are highly distributed and socially situated (Greeno, 1998).

Such situations pose serious challenges for assessment. As learning environments change, effective assessment and testing procedures must respond by providing meaningful results pertinent to the knowledge and skills such environments afford. Testing procedures will require robust measurement and testing models to provide strong empirical support for inferences based on them. Learning in such environments requires explanations based on a new form of measurement and theoretical frameworks. Unfortunately, current measurement and test theory cannot yet respond to these changes.

1.3.3.3 Side Effects of Conventional Testing Procedures and Tools

The primary negative side effect of conventional testing practice is the tendency for testing to reduce teaching to the level of testing technology—away from learning and reasoning skills to more easily measurable skills. Another negative side effect is that present testing often encourages students to

memorize facts rather than to understand, which is what the construction of knowledge requires. These side effects are apparent in testing by multiple-choice items (Collins, 1990). When based on objective cognitive analysis, multiple-choice questions (MCQ) can indeed measure higher-order cognitive skills if the item can be validated as assessing higher-level mental processes. However, in practice, most MCQs measure the simple recall of information (Frederiksen, 1990). In short, test formats used on student achievement examinations lack the support of cognitive theory and construct validity.

1.4 Desirable Features of Cognitive Assessments

As a unifying concept of conventional assessment, standard test theory is a statistical model that encompasses classical true score theory and item response theory. Standard test theory appears to be largely incompatible with the implications and findings of contemporary psychological theories and research on assessment practices (Pellegrino, Baxter, & Glaser, 1999). Alternative objectives must be considered in order to establish new assessment frameworks and appropriate statistical and evidence models based on a construct-centered approach (Messick, 1992, 1994, 1995).

1.4.1 Cognitively Diagnostic Assessment

Cognitive science provides a theoretical basis for developing new methods of assessment that can improve instruction and learning (Frederiksen, 1990). Assessment is not only a procedure for measuring objects and reporting

scores statistically, it can also support inferences about what happens in the mind, what learners know, and how learners process information. Cognitive research typically emphasizes knowledge representation and organization, and problem-solving procedures and strategies which must be part of a cognitively-based assessment framework (Mislevy, 1993).

The substantive foundations of diagnostic assessment emerge from the connection between the theories of instruction, aptitude and the theory of cognition (Embretson, 1990). The purpose of diagnostic assessment is to explore observed facts and to make inferences about the nature of entities underlying those facts (Marshall, 1990). Diagnostic assessments are designed to make inference about the state of students' mastery of specific cognitive skills and knowledge on the basis of observations of their performance in task environments. Diagnostic information is based on observations that are influenced by both cognitive and psychometric models (Corter, 1995). Cognitive models inform observations that arise from task situations, and statistical models allow inferences about explanatory variables underlying such observations.

1.4.2 Dynamic Assessment Focusing on Learning Processes

Dynamic assessment is one of the more successful methodologies for assessing transition in learning (Lajoie, 2003). It can be defined as a moment-by-moment assessment of learners during problem solving so that feedback can be provided in the context of the activity (Lajoie & Lesgold, 1992).

Dynamic assessment has been increasingly emphasized in the last two decades. It focuses on assessing the learning processes by which knowledge

acquired and problem solving skills are developed. Two well known examples of dynamic assessment are currently being implemented in educational assessment: portfolio assessment and computer-based assessment (Lajoie & Lesgold, 1992).

Dynamic assessment is not necessarily cognitive, but cognitive assessment often demonstrates features of dynamic assessment especially in enriched learning environments because cognitive assessment often involves tracking the process by which knowledge and skills are acquired, and problem-solving strategies are formed. Lajoie (2003) postulates regarding the relations of dynamic assessment to expertise and new learning environments:

> Dynamic assessment implies that human or computer tutors can evaluate transitions in knowledge representations and performance while learners are in the process of solving problems, rather than after they have completed a problem. Immediate feedback can then be provided to learners during problem solving, when and where they need assistance. The purpose of assessment in these situations is to improve learning in the context of problem solving. (p. 22)

Lajoie (2003) cites ECA as an example of how to implement dynamic assessment. Conversely, ECA can be embedded in dynamic cognitive environments. It is not necessary but quite possible that knowledge, skills, and expertise can be assessed in dynamic assessment processes.

1.4.3 Performance Assessment of Declarative and Procedural Knowledge

Performance assessments are becoming increasingly popular because they promise authentic and direct appraisals of educational competence leading

to positive consequences for teaching and learning outcomes (Messick, 1994). In cognitively complex domains, learning often involves performance on complex tasks, although conventional assessments seldom use such performance tasks as measurable objects. This tendency has led to the failure to use authentic performance tasks in assessment. Performance assessment emphasizes monitoring the acquisition of both declarative and procedural knowledge thus increasing their construct validity. The characteristics of alternative assessments have become increasingly prominent due to increased demands for assessments stemming from advances in cognitive theory related changes in the goals and standards of instructional practices, and the increased use of multimedia and web-based learning systems. Therefore, current research on alternative assessment theories and practices increasingly emphasizes theory-based assessment frameworks that integrate complex cognitive task designs.

1.4.4 Summary

These theoretical frameworks provide a robust basis for developing alternative assessment designs. They allow researchers to explore effective assessment procedures in terms of cognitive theories combined with modern statistical models. Assessment procedures based on traditional learning theories are not appropriate for use in such new knowledge and problem focused learning environments as web-based learning. Modern learning theories require new assessment procedures and theories to measure learning.

Conversely, current assessment procedures including conventional tests used to measure achievement in universities have been confined to reporting on the acquisition of information or skills rather than to diagnosing relevant errors and demonstrating learner progress in acquiring and developing knowledge and skills.

Cognitive assessment focuses on both diagnosis and learning, and should assess development of components of competency in the performance of complex, authentic tasks. These characteristics are all necessary, and may be implemented using web-based and computer-based learning assessments.

A new theoretical framework for assessment, and the identification of current problems with current assessment practices informed the development of alternative assessments appropriate to new assessment purposes and learning environments.

CHAPTER TWO: OBJECTIVES AND RESEARCH ISSUES

2.1 Purpose and Objectives of the Study

This study aims to explore a cognitive diagnostic assessment procedure for student learning problem solving skills in the domain of statistics. The assessment procedure is diagnostic because it will identify mistakes and deficiencies. The assessment system will report dynamic cognitive processes and student learning processes at each step. The assessment system is cognitive-based involving two kinds of cognitive performances: students' problem-solving processes and semantic explanations. There are five research objectives:

1. To develop and explore a method for diagnostic cognitive assessment in a complex problem solving domain (statistics) based on the implementation of a cognitive-based Bayesian assessment model which is applicable to other complex problem solving domains.

2. To develop an assessment procedure that can be applied to task performance in various situations.

3. To develop a model that can potentially be implemented on web-based coached practice environments and to assess performance in well-understood cognitive domains.

4. To explore the potential of Bayesian Belief Network (BBN) models in cognitive assessment.

5. To evaluate the assessment model and design with data from students in statistics and simulated data.

2.2 Research Issues Investigated

In terms of these five objectives, three specific issues will be addressed.

1. How can an effective assessment model and environment using Bayesian belief networks (BBNs) be designed to provide valid diagnostic assessments of cognitive knowledge and performance skills on tasks in complex problem-solving domains?

2. How well can the assessment model diagnose the mastery or non-mastery of components of cognitive knowledge and competency on the basis of performance data of individuals performing appropriate tasks?

3. How robust are diagnostic assessments over variations in the conditional probability tables used in the network?

CHAPTER THREE: COGNITIVE MODELS IN ASSESSMENT

The representation and organization of knowledge is a priority in the design of cognitive assessment systems as they inform the validity of such systems. Two cognitive models will be used in complex assessment: (a) expert models (Frederiksen & Donin, 2005), and (b) student models (Mislevy, Steinberg, Almond, Breyer, & Johnson, 2001; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001). Although expert models act as a basis for inferring student models, both models can be used to describe student knowledge structure. The cognitive assessment literature has emphasized student models in order to explain progress in student learning. Models of expert knowledge and performance have been neglected and are usually discussed in research on expertise in cognitive science and in research on artificial intelligence (AI), expert systems, and intelligent tutoring systems. Nonetheless, cognitive assessments include expert models as they are closely tied to student models. Comparing student models to expert models is important in tracking the development of expertise. Examination of both models is important in analyzing the entire assessment process.

3.1 Expert Models, Expertise and Types of Knowledge Representation

Expert models and expert systems are often discussed together. Ignizio (1991) states that an expert system is a model within specific domains composed of procedures that exhibit a degree of expertise in problem solving that is comparable to that of a human expert. An expert system contains knowledge

derived from an analysis of human expertise in some domain and this knowledge is used to train individuals using the expert system to solve problems in that domain.

Ignizio (1991) merges the concepts of expert model and expert system. In fact, an expert model is an organized database of declarative and procedural knowledge for a particular domain. It represents the subject matter knowledge of an expert in that domain. Expert systems (rule-based systems) and semantic networks are two ways of modelling expert knowledge (Hay & McTaggart, 2003). An intelligent tutoring system is an example of an expert system, whose aim is to provide users with help in acquiring expert knowledge in some domains. (Hay & McTaggart, 2003). If the expert model incorporated in an intelligent tutor demonstrates high quality of expertise and knowledge structures, learners can quickly and effectively adapt themselves to the learning environment (i. e., the tutorial system). The expertise and the classification of knowledge are very important in describing trajectories of student learning and representation of knowledge.

3.1.1 Expertise and Cognitive Assessment

Expertise as an assessment model has been explored in such assessment paradigms as the web-based cognitive assessment of performance. The development of expertise from the "acclimation to proficiency" (Alexander, 2003) provides opportunities for giving feedback and diagnostic information in different ways. This perspective is highlighted by contrasting the Model of Domain Learning (MDL) with such traditional models of expertise as

expert/novice theory. Alexander (2003) and her colleagues believe that expert/novice theory has over-simplified the features of expertise. They claim that MDL focuses on learning in academic domains and describes the development of expertise in three stages: (a) acclimation, (b) competence, and (c) proficiency based on research investigations in such academic domains as social studies, biology, educational psychology, and special education (Alexander, Jetton, & Kulikowich, 1995; Alexander, Murphy, Woods, Duhon, & Parker, 1997; Alexander, Sperl, Buehl, Five, & Chiu, 2004; Murphy & Alexander, 2002). The "continuum" view of expertise as a multistage process of development (i.e. acclimation, early competence, mid-competence, and proficiency) indicates that models of expertise can be included in a cognitive model of evidence-based assessment and thus connected to statistical models of diagnostic assessment (Williamson, Steinberg, Mislevy, & Behrens, 2003).

3.1.2 Expertise and Problem Solving Strategies

Since the 1970s, theories of expertise have established a base by exploring expert-novice problem-solving performance in different domains, especially medicine (Alexander, 2003; Arocha & Patel, 1995; Joseph & Patel, 1990). It has been found that experts share several cognitive characteristics (Chi, Farr, & Glaser, 1988). According to Lajoie (2003):

> Experts seemed to share the following characteristics: superior memory for information in their domain, better awareness of what they know and do not know, greater pattern recognition, faster and more accurate solutions (although they tend to spend more time initially analyzing problems prior to solving them) and deeper, more highly structured

> knowledge. Despite commonalities, a key constraint to developing expertise is that it is domain specific. Experts are experts at something, be it chess or avionics. This is important because it demonstrates that expertise is more than general intelligence. (p. 21)

In research on problem solving, Patel, Evans, and Kaufman (1990) and Patel and Groen (1986) identified directionality of reasoning in clinical diagnostic domains: Forward on data-driven and backward on hypothesis-driven reasoning. Medical experts employ forward reasoning while novices and intermediate medical practitioners are more inclined to use backward reasoning developing and testing hypotheses against available data (Arocha, 1990; Patel & Groen, 1993; Patel, Groen, & Arocha, 1990). In the literature on expertise and problem-solving strategies, reasoning and problem-solving strategies are popular topics. Consequently problem-solving strategies and expertise are useful in designing cognitive models for assessing learning in complex domains.

3.1.3 Trajectories of Expertise Development, and Dynamic Assessment

Lajoie (2003) initially argued for the view that the development of expertise can be fostered along cognitive trajectories. She proposed that the development of expertise has two goals: determining what experts know, and how to help novices enhance their competence.

> Identifying what experts know can help determine the trajectory towards competence for the task. This trajectory, or path, is not necessarily linear and it can have several signposts where learning transitions can take place. Once such trajectories are mapped out assessments can be designed that assess learning transitions along the road to competence. Research must specify how to promote transitions or changes in

> competence in different learning situations. Models of expertise that include different trajectories to competence can be used to design instruction and assessment for both in and out-of-school contexts. (p.21)

Thus, expertise develops along a non-linear, cognitive trajectory. Knowing a given trajectory of expertise can help in identifying changes in learner domain competencies. However, a number of investigations have shown that cognitive trajectories can follow different directions. In analyses of avionics experts, a trajectory may consist of problem solving plans, actions, and the use of mental models (Lesgold, Lajoie, Logan, & Eggan, 1990). In a real-world study of expert surgical nurses (Lajoie, Azevedo, & Fleiszer, 1998), a trajectory of expertise is composed of the following components in the following order: hypothesis generation, planning of medical intervention, action performed, results of evidence gathering, interpretation of results, heuristics, and the overall solution paths. These multi-signpost trajectories provide assessment possibilities for revealing diagnostic information for learners.

Lajoie argues that the acquisition of expertise is a transitional process that can be enhanced through dynamic assessment which provides ways to evaluate transitions in the organization and representation of knowledge and performance. Dynamic assessment is a moment-by-moment assessment of learner problem solving (Lajoie & Lesgold, 1992). It focuses on actual learning processes (Lidz, Jepsen, & Miller, 1997) and shares many of the functions and features with diagnostic assessment which emphasizes such cognitive aspects as learner errors expressed as misrepresentation, novice problem-solving strategies, and

reasoning (Embretson, 1993; Feltovich, Spiro, & Coulson, 1993; Lajoie & Lesgold, 1992;).

Lajoie and Lesgold (1992) describe dynamic assessment and cognitive diagnostic assessment as follows:

> Dynamic assessment implies diagnostic assessment in that it is used to both monitor and improve the learning situation. The diagnostic monitoring of skill and knowledge acquisition implies that information relevant to the process of learning in a domain can be recorded and preserved to provide a continuous record of changes in knowledge, skill, and understanding as students encounter problems of increasing complexity. (p. 366)

Dynamic and diagnostic attributes of assessment will work together in complex learning environments such as web-based tutorial systems. Dynamic assessment is an effective method for assessing transitions in expertise and for tracking trajectories of expertise development.

3.1.4 Types of Knowledge as Possible Cognitive Models in Cognitive Assessment

According to Lajoie (2003) and Alexander (2003), transitions in the development of cognitive processes can be characterized in different ways. For instance, they can be characterized in terms of plans, goals, actions, and outcomes, or in terms of types of knowledge such as declarative and procedural (Anderson, 1982; Bitan, Karni, & Bitan, 2004; Byrnes & Wasik, 1991; Corbett & Anderson,1995; Rittle-Johnson & Alibali, 1999; ten Berge & van Hezewijk, 1999).

Shavelson and Ruiz-Primo (1998) proposed that a cognitive assessment framework must include three types of knowledge: declarative, procedural, and

strategic. Declarative knowledge involves the knowledge of facts and concepts within some domains such as force, mass, acceleration in physics. Procedural knowledge involves knowing how to do something ("hands on"). Strategic knowledge involves knowing where, why and how to apply specific knowledge, that is, when a complex task should be completed and what problem-solving plans are required.

Shavelson and colleagues have conducted several studies based on different hypotheses about different cognitive dimensions. For instance, Shavelson (2000), and Ayala, Ayala, and Shavelson (2000) characterize reasoning in terms of three cognitive "dimensions": basic knowledge and reasoning, spatial mechanical reasoning, and quantitative science reasoning, which they used to examine student reasoning on scientific problems in laboratory learning environments. Yin, Ayala, and Shavelson (2001) explored student problem solving in a science program and identified strategies of problem solving strategies: attending, processing information, reading and planning, observing, and conjecturing.

Most of this research is based on think-aloud protocol and analyses of data collected on student learning processes. Shavelson and colleagues claim that student learning should emphasize processes of both thinking and

performing and propose a "hands on and minds on" perspective. These cognitive hypotheses allow theoretical interpretations of results in terms of data collected, and analysis. However, these cognitive dimensions have not yet been clarified as well-articulated expert and student models.

3.1.5 Expertise Contained in Semantic Networks Distributively and in Procedure Structures Hierarchically

A semantic network is a graphical notation for representing knowledge in patterns of interconnected nodes and arcs. What is common to all semantic networks is a declarative graphic representation that can be used to represent knowledge and/or support reasoning (Sowa, 1987, 1991, 2000). Sowa (2005) classified semantic networks into six types of networks: (a) definitional, (b) assertional, (c) implicational, (d) executable, (e) learning, and (f) hybrid. Sowa (2005) characterizes these networks as follows:

> Definitional networks emphasize the subtype or a relation between a concept type and a newly defined subtype. The resulting network, also called a generalization or subsumption hierarchy, supports the rule of inheritance for copying properties defined for a supertype to all of its subtypes. Since definitions are true by definition, the information in these networks is often assumed to be necessarily true. Assertional networks are designed to assert propositions. Unlike definitional networks, the information in an assertional network is assumed to be contingently true, unless it is explicitly marked with a modal operator. Some assertional networks have been proposed as models of the conceptual structures underlying natural language semantics. Implicational networks use

> implication as the primary relation for connecting nodes. They may be used to represent patterns of beliefs, causality, or inferences. Executable networks include some mechanism, such as marker passing or attached procedures, which can perform inferences, pass messages, or search for patterns and associations. Learning networks build or extend their representations by acquiring knowledge from examples. The new knowledge may change the old network by adding and deleting nodes and arcs or by modifying numerical values, called weights, associated with the nodes and arcs. Lastly, hybrid networks combine two or more of the previous techniques, either in a single network or in separate, but closely interacting networks. (p.1-2)

Different semantic networks provide mechanisms for describing relations among different categories of knowledge, and can be used to describe different levels of expertise. The semantic networks of an expert are richer, more intricate and more interconnected than those of novices (Derry, 1990). The expert semantic networks are more internally coherent than those of a novice (Roth, 1990; Tweney & Walker, 1990). Experts recognize and store more patterns, and organize information into larger chunks (Chase & Simon, 1973; Perkins, 1981). These findings are closely associated with different focus of analysis and methods of data-collection such as cognitive analysis (Annett, 2000) and discourse analysis (Schiffrin, 1994). Discourse analysis, especially propositional analysis, can represent expert declarative and schematic knowledge in semantic networks at different levels (Chi, 1997).

3.1.5.1 Application of Semantic Networks to Monitoring and Assessing Declarative Knowledge and Skills

Semantic networks can be used to monitor cognitive processing (Frederiksen & Breuleux, 1990), and student acquisition of knowledge and skills in many domains such as biology, chemistry, and medicine. They have also been used in assessing the capacity for discourse and comprehension in neuropathology (Frederiksen, 1999; Frederiksen & Breuleux, 1990; Frederiksen & Stemmer, 1993). Discourse analysis is one way of developing semantic networks. Discourse is viewed as a sequence of natural language expressions produced by speakers or writers to represent and communicate conceptual knowledge to listeners or readers in various contexts (Frederiksen & Stemmer, 1993). Propositional analysis is used to represent knowledge at a micro-level (Frederiksen, 1975) and can be used to develop semantic networks. Semantic networks produced through propositional analyses can reflect the development of knowledge and skills. Frederiksen and Breuleux (1990) proposed two approaches for defining cognitive models representing semantic networks: canonical frames and semantic grammars.

> Models of semantic representation and methods for analyzing the processes involved in generating and manipulating semantic structures ought to provide a basis for cognitive monitoring or diagnosis of learners' knowledge and performance in semantically complex tasks. Cognitive diagnosis in semantically complex domains involves the evaluation of: (a) an individual's state of knowledge in a domain, (b) the semantic representations an individual generates in the performance of a task, and (c) the processes that are employed in retrieving, generating, applying, modifying, or in other way manipulating knowledge representation. (p. 356)

Frederiksen and Breuleux characterize canonical frames and semantic grammar as follows:

> A canonical representation (or frame) is a particular network structure or pattern that contains variables. Variables are symbols in a pattern that can be replaced by specific values... A canonical frame defined in this way thus is capable of representing a large number of structures or "instantiations." The canonical frame approach to defining a propositional representation consists of defining an exhaustive set of such patterns, each of which represents a particular type of structural possibility...
>
> A semantic grammar adopts a generative approach to definition, specifying a model by means of rules that generate all acceptable patterns within the grammar. It is well known within the theory of generative grammar that a relatively small set of recursive rules can be much more powerful than a large of canonical frames. (p. 359-360)

Frederiksen (1986) developed a semantic BNF grammar for analyzing proposition to monitor cognitive processing in such content domains as medicine (Frederiksen, 1999; Patel & Arocha, 1995; Patel & Groen 1986). Propositional models can represent relations in semantic processes as cognitive processes both qualitatively and quantitatively. Propositional models typically contain several types of propositions: events, systems, states, propositional relations, identities, algebraic relations, functions, binary dependency relations, and conjoint dependency relations (Frederiksen & Breuleux, 1990). These proposition types plus the BNF grammar constitute a system for developing semantic networks.

Based on propositional analysis, Frederiksen proposes a general model of cognitive evaluation in which relations between expert and student models have been described (Frederiksen & Breuleux, 1990). In expert models, structures are

organized into three submodels: (a) expert knowledge model where experts demonstrate their knowledge base; (b) expert processing models which represents sources of information received such as formal language, experienced events, graphic information, numeric data; and (c) expert data/task models divide the expert processing model into two parts: rules applied to specific task information and methods for applying rules. In short, although there are many forms of data analysis and semantic frame building, semantic networks offer greater opportunities for monitoring and assessing declarative and schematic knowledge.

3.1.5.2 Expertise in Procedure and Strategy Knowledge, and Assessment

In such cognitively complex domains as science education, medicine, or nursing (Ayala, Ayala, & Shavelson, 2000; Frederiksen, 1999; Lajoie, Azevedo, & Fleiszer, 1998), learner progress is based on judgments as to whether they have acquired declarative, procedural, schematic, and strategic knowledge, and whether they can use that knowledge to solve problems competently. In general, these kinds of knowledge are implicitly contained in such learning contexts as problem-based learning (Frederiksen & Breuleux 1990). To develop expertise in a professional domain, students must not only acquire and apply a rich body of declarative and procedural knowledge for solving authentic problems, but they must also learn to function in the various social contexts in which professionals typically collaborate to solve problems. Thus both declarative and procedural knowledge are indispensable in the development of expertise. For example,

problem solving varies according to different occasions and organizational formats such as small group collaborations. Although researchers have been aware of procedural and strategic knowledge in the last three decades (Schacter, 1989), they have yet to be successfully assessed and monitored (i.e. Ayala, 2003; Baxter, Shavelson, Goldman, & Pine, 1992; Hunt, 1995; Marshall, 1995). As computer-based coaching systems and web-based learning systems become more common in various content domains and complex learning environments, assessment of learning becomes an increasing concern (Lajoie, 1993; Schwartz, Biswas, Bransford, Bhuva, Tamara, & Brophy, 2000; Sugrue, 2000).

Expertise can be expressed in different types of knowledge, skills, reasoning and problem solving strategies. All of these components of expertise demonstrate the common cognitive characteristics indicating progress from novice to expert. In recent assessment research on computer-based learning environments, hierarchical models (Frederiksen & Donin, 1999; Frederiksen, Donin, Bracewell, Mercier, & Zhang, 2002), network models (Heffernan, 2001), and mixture models (Lajoie, 1993) have been employed to monitor the acquisition of procedural knowledge and problem solving strategies. Although other models have been used in the design of such tutorial systems as Multi Agent Architecture for Adaptive Learning Environment (MAGALE) (McCalla, Vassileva, Greer, & Bull, 2000), and Web-based Authoring Tools for Algebra Related Domain (WEAR) (Virvou & Moundridou, 2000), hierarchical models are potentially useful in cognitive assessment and are well adopted to Bayesian nets as their statistical models.

Minor references to the use of hierarchical models in student models and statistical models can be found in the assessment literature. However, little has been done with respect to expert models. Heffernan (2001) used a network model to design a tutorial system for algebra. This model provided possible paths for students to choose. If students encountered problems at some stages, the tutorial system provided helpful feedback.

Another tutorial system, Bio-world (Lajoie, 1993) provides a computerized coaching environment in which secondary school students learn to diagnose medical problems. Bio-world is a mixed or "semi-hierarchical" network. Students can choose such keywords as "AIDS" and then select from several available patients. To help students to uncover more diagnostic information, Bio-world provides a notebook in which diagnostic reasoning structures can be hierarchically developed. For example, in moving from hypothesis to the disease in question, students move through a hierarchical space.

The McGill Statistics Tutorial System (Frederiksen & Donin, 1999) adopts a cognitively complete hierarchical design and is organized on the basis of studies of how problem-solving procedures are structured in the memory of experts. The procedure frame represents complex procedures by decomposing them into hierarchies of actions and goals:

> At the top level, solving a data analysis problem involves six component procedures: (a) defining the research problem, (b) specifying the data, (c) carrying out a descriptive analysis of the data, (d) performing an ANOVA with the data, (e) conducting any post-hoc analysis, and (f) drawing conclusions based on the results obtained from previous steps. Each of these main procedures is composed of subprocedures. For example, procedure

> (d), performing the ANOVA, is composed of eight main subprocedures to be performed: (a) specifying the research design, (b) specifying a linear model for scores on the dependent variable, (c) obtaining least squares estimates of the grand mean and all effects in the linear model, (d) partitioning the total sum of squares according to the ANOVA model, (e) preparing an ANOVA table for organizing results, (f) computing ANOVA statistics, and (g) conducting F tests. (p. 399)

This hierarchical model contains conceptual, theoretical and procedural knowledge and records student performance while solving such statistics problems as two-way ANOVA problems by representing statistical models as Bayesian networks.

3.2 Student Models in Cognitive Assessment

Student models, as related to cognition and assessment, have been proposed and applied in intelligent tutorial systems (ITSs) (Hay & McTaggart, 2003), and also discussed as tools (Reusser, 1993) in cognitive assessment (Ohlsson, 1990) and diagnostic assessment (Corbett & Anderson, 1995). Reusser postulated that "the student model, which encompasses both the learner's knowledge and behavior as he or she interacts with the ITS, acts as a guidance system that helps lead the student through the domain's knowledge base" (p. 6). Therefore, the student model can be seen as a process through which assessors can assess student performance during the development of knowledge and skills.

In their research on teaching avionics troubleshooting skills in computer-based learning environments (CBLE), Lajoie and Lesgold (1992) characterize their student model dynamically as follows:

> Student modeling refers to the programming techniques that enable an instructional system to develop and update an understanding of the learner's performance on the system. More broadly defined, student modeling includes the processes that utilize the system's knowledge about the student as a basis for diagnosing student problems and selecting instructional approaches that best address the diagnoses. (p. 375)

In a later study, Derry and Lajoie (1993) expanded the definition of student modeling.

> Narrowly speaking, student modeling refers to the programming techniques and reasoning strategies that enable an instructional system to develop and update an understanding of the student and her performance on the system. More broadly defined, student modeling also includes the processes that actually utilize the system's knowledge about the student as a basis for diagnosing student problems and for selecting instructional approaches that best fit current diagnoses. (p. 2)

Thus, the definitions of student modeling are nearly identical. The only major difference lies in the addition of "reasoning strategies" Furthermore, "understanding of the student and his/ her performance," is stressed as opposed to simply focusing on "performance", per se. The last sentence of Derry and Lajoie (1993) refers to "approaches to the best fit current diagnoses" rather than "best address the diagnoses." In short, Lajoie and Lesgold, and Derry and Lajoie have elaborated and expanded the definition of student modeling. They emphasize student performance, the diagnosis of student problem solving, and the effects of student modeling on the selection of instructional approaches.

Frederiksen and Breuleux (1990) present another point of view on the study of relationship between expert and student model:

> A student model is defined in terms of an expert system in which a learner is described in terms of his or her knowledge of production rules in the system. The student model is determined by inferring the rules the learner has applied on the basis of his or her response. (p. 355)

Based on the three components of the expert model: knowledge, processing, and data representation, Frederiksen and Breuleux (1990) proposed a four-step procedure for developing student models. Frederiksen and Breuleux's (1990) research on monitoring cognitive processing in semantically complex domains establishes the principle that student and expert models can be developed in parallel.

Such concepts and definitions certainly help to characterize student models carefully. However, considering the purposes and functions of student models are important to better understanding of how they can be used in cognitive assessment.

3.2.1 Purposes and Functions of Student Modeling

The purposes of student modeling are diverse and closely associated with the functions of student models. Zhou (2000) suggests that a student model is useful for guiding pedagogical decision-making in ITSs. For example, in medicine, an author may intend to help first-year medical students solve medical problems. Obviously, the design of this type of tutorial system is pedagogically oriented. Because of the importance of medical diagnostic skills, decision-making

is often a focus for research. When Ganeshan, Johnson, Shaw, and Wood (2000) explored the causal relationships between symptoms and disease states they described the purpose of the student model in the following manner:

> It is for capturing "all of the knowledge the student is expected to bring to bear on the diagnostic process including the steps and their associated properties, the findings associated with the steps, the hypothesis, the hierarchical relationships between hypotheses, the causal relationship between the findings and hypothesis, and the strengths associated with these relationships." (p. 36)

Student models are also relevant to the design of tutorial systems. For example, Online Assessment of Expertise (OLAE) (VanLehn, 2001) and Probabilistic Online Assessment (POLA) (Conati & VanLehn, 1996) are physics learning tools. OLAE is mainly for assessment, while POLA is mainly for probabilistic online assessment. Thus, student models have multiple uses in computer-based and web-based learning.

In summary, student models are important. In ITS, student models fundamentally fulfill three functions (Gitomer, Steinberg, & Mislevy, 1995). First, to help determine a set of instructional options, which can tailor appropriate pedagogical suggestions for an individual; second, to predict student actions, from which their validity can be inferred; third, to enable "the ITS to make claims about the competency of an individual with respect to various problem-solving abilities" (p. 74). Further identification of the purposes and functions of student modeling will be beneficial to the design of cognitive assessments based on ECA (Mislevy, Almond, & Lukas, 2004).

3.2.2 Examples of Application of Student Models in Tutorial Systems

Student models are also relevant to the design of tutorial systems. Tutorial systems in physics and medicine have used student models. Several examples illustrating applications of student models in physics and medicine follow.

3.2.2.1 A Student Model in Web-based Tutoring System for Problem Solving of Digital Logic Circuits

Kassim, Ahmed, and Ranganath (2001) use a student model to trace student progress in a web-based problem solving learning environment for digital logic circuits. Records of student progress are kept in a database, where they are monitored and instructional options are selected based on student models.

Based on their research, Kassim et al (2001) regard student models as dynamic representations of the knowledge and skills that students demonstrate in solving problems in digital logic circuits. Kassim et al. (2001) wrote that "the student's inputs to the system provide evidence of learning and are used to update the student model" (p. 26).

Kassim et al (2001) attempt to establish a relation between the expert model and student model which they call an overlay model in which student knowledge is regarded as "a subset of the expert's knowledge and the goal of tutoring is to enlarge this subset toward the expert's knowledge" (p. 27).

3.2.2.2 Student Models in ANDES and OLAE: Physics Learning Tutorial and Assessment System

ANDES is an intelligent tutoring system for students learning Newtonian physics in an introductory physics course (Vanlehn & Niu, 2001; Vanlehn, Niu, Siler, & Gertner, 1998). Students receive visualizations, immediate feedback, and procedural and conceptual help. The ANDES student model emphasizes the development of declarative and procedural knowledge. ANDES has two models: A homework Assignment Editor and a Tutor. The homework module involves Bayesian reasoning to maintain a long-term model of the students' mastery of physics concepts and preferred problem solving techniques. The tutor has four components: workbench, helper, assessor, and author's toolbox. The workbench, which includes a calculator and algebraic equation solver, allows learners to choose activities and complete series of tasks. Learners receive feedback on final answers or intermediate results. The helper module provides information on plans and goals, and helps learners to solve physics problems. It also explains workbench feedback. The assessor is a relatively independent module on Online Assessment of Expertise (OLAE) and Probabilistic Online Assessment (POLA) which are associated with ANDES. The Author's Tool Box is used to modify the physics knowledge base, and to create and modify individual homework activities.

Assessment functions were included in early versions of ANDES. The OLAE was developed and associated with each version of ANDES (Martin & VanLehn, 1995a). OLAE provides detailed reports of student performance,

students' abilities to solve physics problems. OLAE adopts three purposes in assessing student competence in solving physics problems (VanLehn & Martin, 1998):

> (1) to collect detailed performance data on student actions as they performed tasks in a web-based learning environment ,
>
> (2) to analyze student competencies in detail, and
>
> (3) to ensure data analysis can be sound and computationally feasible. (p.181)

Based on this, a detailed student model of OLAE is expressed as student knowledge representations containing sets of rules. The student model is designed to cover correct and incorrect physics rules. OLAE also uses a three-level model of mastery. In OLAE, each rule is assumed to be in one of three states:

> (1) Non-mastery: the student never applies the rule.
>
> (2) Partial mastery: the student applies the rule when using paper and pencil, but does not use the rule when mentally planning a solution.
>
> (3) Full mastery: the student applies the rule whenever it is applicable. (p. 184)

This student model is potentially associated with a theory of expertise in which trajectories and development are represented as three level scales.

3.2.2.3 A Student Model in the CIRCSIM Tutor system for Physiology Learning

The CIRCSIM-Tutor project is building a language-based ITS to assess medical students in learning to solve medical problems in physiology. Hence, the student model helps students learn to solve problems using qualitative-causal reasoning and highlights student initiative (Farhana, Evens, Michael, & Rovick, 2002). Student initiative occurs when a student temporally takes control of a

session, by saying something that forces the tutor to change its course of action and respond to a new situation. Student questions serve as initiatives.

Students communicate through writing with CIRCSIM which controls conversations, alternatively teaching concepts and asking questions. Such interactions can be problematic as conversations are still constrained by the system. The researchers are currently exploring different ways for students to interact with the tutorial system through natural language.

3.3 Summary of Expert Models and Student Models

Expert and student models are referred to as cognitive models in the design of DCA. Expert models are organized databases and frameworks of knowledge and expertise in given domains. Student models are systems designed to record student knowledge and behaviour, then are defined in terms of expert systems in which domain expertise is applied. The designers of student models collect the products of student learning to understand the students' learning trajectories.

This chapter has reviewed relations between expert systems and expertise, and representation of expertise in semantic networks. Student models and their functions, along with some exemplars have been examined. Student models have various uses in research and tutoring applications based on the purposes and characteristics of the learning environments.

From a cognitive assessment perspective, expert and student models are cognitive models which can be used to build assessment models. The application

of these models is an issue of great concern to assessment researchers.

CHAPTER FOUR: STATISTICAL AND TASK MODELS IN COGNITIVE ASSESSMENT

4.1 Statistical Models Applied in Achievement Assessment

A statistical model in Mislevy's framework (Williamson, Bauer, Steinberg, Mislevy, Behrens, & DeMark, 2004) is embedded in an evidence model, which links observations of evidence variables to theoretical assessment constructs (components of the student model). An evidence model consists of two submodels: an evaluative model and a statistical model. The evaluative model is a set of rules for extracting components of student's knowledge and skills from student scores on evidence variables (reflecting performance). Evidence variables are developed based on learning tasks. Statistical models are critical in transferring information from evidence variables to assessment constructs (theoretical variables). A specific statistical model used in an assessment system depends on the assessment purposes, the student model, and the task model.

In order to focus on cognitive assessment and evidence-centred design (ECD) (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004) this section will review three mainstream theories that are likely to continue to play powerful roles in the future development of assessment: item response theories (IRT), latent class models (LCM) and Bayesian networks (BN).

4.1.1 IRT: Assumptions and Models

Item response theories (IRT) have been developing for more than four decades and still dominate achievement measurement as a psychometric paradigm (Embretson, 1984; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Roger, 1991; Junker, 1999). Initial contributions to IRT can be traced to Richardson (1936) and Tucker (1946). Richardson established the connection between IRT model parameters and classical item parameters and Tucker postulated the concept of the item characteristic curve (ICC), a monotonic increasing curve specifying that as the level of proficiency increases, the probability of a correct response to an item increases. ICC is a critical element in IRT because it describes a non-linear relationship between the proportion of correct responses to an item and a criterion variable (Baker, 1992). No matter how IRT models change, ICC is still a fundamental characteristic of them.

IRT is a model-based measurement because it specifies how trait levels and item properties are related to a person's item responses (Embretson & Reise, 2000). IRT models describe the relations between item response scores and proficiency levels where item responses are usually expressed in terms of their probabilities.

Two important assumptions, unidimensionality and local independence, provide a theoretical foundation for IRT models and their extensions. In addition, the unitary items assumption will also be discussed.

4.1.1.1 Two Fundamental Assumptions for Item Response Theories

According to unidimensionality, one trait or proficiency is necessary to account for the performance of an examinee on a task. However, in practice, this assumption cannot be strictly met because other cognitive, personality-related, and test-taking factors impact test performance. A dominant component or factor must influence test performance for a set of test data to adequately meet this assumption. This dominant component is referred to as the ability measured by the test (Hambleton & Swaminathan, 1985). Two methodologies for assessing unidimensionality are usually used: test of essential unidimensionality (DIMTEST) (Stout, 1987, 1990), and factor analysis.

DIMTEST is a statistical procedure for testing the hypothesis that an essentially unidimensional IRT model fits observed binary item response data on a given test. Here a set of items are considered unidimensional when the average between-item residual covariance after fitting a one-factor model approaches zero as test length increases.

Factor analysis methods may be used to determine whether responses are consistent with unidimensionality. The non-linear factor analytic model (Nandakumar, 1994) and the accompanying principle of weak local independence have been shown to be useful for identifying the number of dimensions underlying a set of binary item responses. Other factor analysis methods such as full information factor analysis (Zwick, 1987) may also be used to evaluate the unidimensionality of an IRT model.

Unidimensionality is reflected in a theoretical, single item parameter, the estimation of which is achieved through mathematical estimates. A major functional consideration in IRT unidimensionality is to enhance the interpretability of a set of test scores for a given performance.

Local independence is the assumption that the response to any item is unrelated to the response to any other items when the trait level is controlled. Items may be highly correlated in the whole sample. However, if the trait level is controlled, local independence implies that all test items are unrelated (Embretson & Reise, 2000). Thus, after taking examinee abilities into account, examinee responses to different test items will be unrelated (Hambleton, Swaminathan, & Roger, 1991).

Because the assumptions of local independence and the unidimensionality of the latent space are equivalent, factor analytic techniques can be used in estimating local independence (Hambleton & Swaminathan, 1985). The Q3 method (Yen, 1984, 1993) is another effective approach for investigating the local dependence issue. The Q3 index represents correlations between items after isolating latent trait variables. Based on differences between raw scores and expected scores in IRT models, a residual is obtained and correlations are obtained for residual scores among item pairs. If a Q3 value is relatively large, the relevant item may share other factors with the other items.

4.1.1.2 The Assumption of Unitary Items

The basic assumption in most IRT models is that items correspond to unitary tasks. Some newer IRT models assume that items can be decomposed into subtasks or attributes, each of which becomes a component of item parameters that are still subject to unidimensionality.

4.1.1.3 Classification of IRT Models in Achievement Assessment

IRT models can be categorized in terms of: (a) response scales (b) trait dimensionality, and (c) number of item parameters. Responses can be dichotomous or polytomous; trait dimensionality can be unidimensional or multidimensional; the number of parameters can be one, two, or three. Any combination of these features constitutes a particular model category.

4.1.1.3.1 Dichotomous unidimensional models.

Dichotomous unidimensional models can often be found in assessments of achievement using Multiple Choice Questions (MCQs). Basic IRT models usually assume that traits are one-dimensional. Rasch models (Rasch, 1960) combine dichotomous and unidimensional features. Rasch models explain the occurrence of a data matrix containing dichotomously scored answers of a sample of persons on a fixed set of items by assuming that the items measure the same latent trait. A similar IRT variation is in Lord (1953), Lord and Norvick (1968) and Birnbaum (1968) except that Lord (1953) uses a normal Ogive model

and Birnbaum uses the logistic model. However, both models predict similar probabilities. For practical purposes, the Logistic model is more commonly applied in current achievement assessment. The normal Ogive model can be transformed into the Logistic model by means of a scaling factor (D=1.7) which is used to multiply the power of exponent terms in the Logistic model equation. The logistic model will be used in discussions of all models in this section.

IRT models are parametric models which involve an item difficulty index **β**, and may involve a discrimination index **α**, and a guessing index **γ**. Models with only **β** are referred to as one-parameter models. Models with **α** and **β** are referred to as two-parameter models, and models with **α**, **β**, **γ** are referred to as three-parameter models. Models become more complicated as parameters are decomposed into sub-components. The linear logistic latent trait model (LLTM, Embretson, & Reise, 2000) is such a case.

(1) One-parameter logistic (1PL) models

1PL models are Rasch models. They depict the relationships between examinee scores on a single trait and an item parameter (Rasch, 1960). A single latent trait is assumed to be sufficient to characterize differences between examinee trait scores and item parameters (Embretson & Reise, 2000). Examinee proficiency can be conveyed in the exponent format of trait and item parameters.

$$P(X_{is}=1|\theta_s,\beta_i)=\frac{exp(\theta_s-\beta_i)}{1+exp(\theta_s-\beta_i)} \qquad \text{(Formula 4.1)}$$

Where $X_{i\,s}$ is the response of examinee s to item i (takes 0 or 1)

θ_s is the trait level for examinee s.

β_i is the difficulty of item i.

Formula 4.1 cannot be directly understood. It describes a non-linear relationship assessing the proficiency of examinees based on responses to a set of items, between latent trait and item difficulty. It can be more easily understood in a logit format.

$$\text{Ln}\ \frac{P_{is}}{1-P_{is}} = \theta_s - \beta_i \qquad \text{(Formula 4.2)}$$

Formula 4.2 is the natural logarithm of the odds ratio which is modeled by the simple difference between an examinee trait score θ_s and item difficulty β_i. The model specifies that the logarithm of the ratio of the probability of examinee success on item i, P_{is}, to the probability of examinee failure on item i, 1- P_{is}, is a function of the item and ability parameters. It is the difference between trait level and item difficulty level.

(2) Two-parameter logistic (2PL) models

The only difference between 1PL and 2PL models is that 2PL models include a second parameter, the item discrimination parameter. In 1PL models, the default assumption is that all items demonstrate the same discrimination power and in 2PL models the assumption is that all items demonstrate different discrimination levels.

2PL models specify the relationship of examinee proficiency to examinee probability for a correct response to an item. This relationship involves the latent trait, an item difficulty parameter, and an item discrimination parameter:

$$P(X_{is}=1|\theta_s,\beta_i,\alpha_i)=\frac{exp[\alpha_i(\theta_s-\beta_i)]}{1+exp[\alpha_i(\theta_s-\beta_i)]} \qquad \text{(Formula 4.3)}$$

The item discrimination parameter α_i is proportional to the slopes of the ICC at point b_i on the trait scale (Hambleton, Swaminathan, & Rogers, 1991) and can theoretically be defined in the range of ($^{-}\infty$, $+\infty$). However, in practice it is recommended that items with negative discrimination indexes or with values larger than 2 be removed. (Hambleton & Swaminathan, 1985). Thus, the range of a good discrimination index is between 0 and 2. The discrimination index for unidimensional models is a single value and for multidimensional models it is a vector. In the multidimensional model, the discrimination index becomes a summation of multidimensional components (Reckase, 1997; Reckase & McKinley, 1991).

(3) Three-parameter logistic (3PL) models

3PL models have become popular since multiple-choice questions have become common response formats in secondary and tertiary educational settings. Birnbaum (1968) explicated the 3PL model. 3PL models add a new parameter **γ**, a lower-asymptote parameter, also called the pseudo-chance-level parameter, which represents the probability of low proficiency examinees correctly endorsing an item. **γ** represents a binomial floor on the probability of

getting an item correct (Sireci, Wainer, & Braun, 1998). Essentially 3PL models correct an estimating bias when examinees with low proficiency levels respond to multiple-choice items.

$$P(X_{is}=1 \mid \theta_s, \beta_i, \alpha_i) = \gamma_i + (1-\gamma_i)\frac{\exp[\alpha_i(\theta_s-\beta_i)]}{1+\exp[\alpha_i(\theta_s-\beta_i)]} \qquad \text{(Formula 4.4)}$$

Here, γ ranges between 0 and 1.

(4) LLTMs allowing for Item Complexity

In the Linear Logistic Latent Trait Models (LLTMs) (Embretson, 1990), item parameter b_i (Formula 4.1), the item difficulty index, has been modified and expanded. In 1PL models, the difficulty index is one variable—β_i; but in LLTM, the difficulty index is a linear combination of three sub-parameters η_m, q_{im}, and d.

$$\beta_i = \sum_{m=1}^{M} \eta_m q_{im} + d \qquad \text{(Formula 4.5)}$$

Where β_i is the item difficulty index for item i; q_{im} is the complexity score of item i on factor m (m=1, …, M factors); η_m is the difficulty weight of factor m; and d is a normalization constant. LLTMs model item difficulty as a linear combination of values of these m parameters.

4.1.1.3.2 Applications of dichotomous, unidimensional models to cognitive assessment.

In the model discussed above, dichotomy and unidimensionality have confined the application of LLTM to achievement assessment and diagnostic assessment of knowledge and performance skill development. With changes in

learning environments, learning tasks have become increasingly complicated. Many studies focus on parameter estimation (Baker, 1992; Bock, 1972), model fit (Andrich, 1978), and response categories (Bennett, Morley, & Quardt, 1998). These IRT models are used in assessment in many different learning domains. Verguts and de Boeck (2000) applied a Rasch model to detect learning on an intelligence test. Their initial idea was to use a basic IRT model, but it was obvious that 1PL models cannot satisfy more complex tasks. Assessment research based on ECD (Mislevy, Almond, & Lukas, 2004; Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001) has addressed dichotomous and 1PL models used in different assessment designs. For example, the GRE which tests verbal, quantitative, and analytical reasoning skills has been used to explore relations between student models and evidence models. Variable θ represents proficiency in a specified task domain. ECD assumes that single latent traits are appropriate only when students are solving unitary tasks.

Dichotomous and unidimensional models have been used in task-based language assessments (Mislevy, Steinberg, & Almond, 2002) to establish evidence for claims about whether students have mastered some skills or pieces of knowledge. The study examined different relationships between a single claim based on a single continuous student assessment variable, and multiple claims based on multiple continuous student assessment variables. The assumption being examined is a single IRT variable θ accounts well for performance across tasks in particular domains.

In measurement and assessment practice, most tasks are very complicated, and dichotomous unidimensional models cannot satisfy the objectives of these achievement tests. Therefore, multidimensional models have been applied to assessment in various content domains (Embretson, 1993; Kelderman & Rijkes, 1994).

4.1.1.3.3 Multidimensional latent trait IRT models applied in cognitive assessment.

Multidimensional IRT models (Embretson, 1993; Embretson & Reise, 2000) were developed from Rasch models. In these models, the trait θ_s is no longer a variable consisting of a single dimension as in unidimensional models. θ_s is replaced by a set of components: θ_{s1}, θ_{s2}, ..., θ_{sk}. This section introduces two typical multidimensional latent trait IRT models used in cognitive assessment: the multidimensional Rasch model, and the multidimensional linear logistic model (MLLM).

(1) Multidimensional Rasch models

McKinley and Reckase (1982) describe multidimensional Rasch models as follows.

$$P(X_{is}=1|\underline{\theta}_s,\delta_i)=\frac{\exp(\sum_{m=1}^{M}\theta_{sm}+\delta_i)}{1+\exp(\sum_{m=1}^{M}\theta_{sm}+\delta_i)} \qquad \text{(Formula 4.6)}$$

where X_{is} is the response of examinee s on item i (0 or 1)

θ_{sm} is trait level for examinee s on dimension m

δ_i is easiness intercept for item i

M is the number of dimension

This model assumes that multiple dimensions are equally weighted for each item. Thus different trait levels cannot be separately estimated. This problem has been solved by multidimensional extensions of two parameter logistic models (Embretson, & Reise, 2000).

(2) Multidimensional Linear Logistic Models (MLLMs)

MLLMs aim to abandon the unidimensionality requirement in favour of multidimensionality which cognitive assessment requires in some domains. Fischer and Seliger (1997) proposed a model in which no assumption about the latent dimensionality of an item is necessary and they applied the model to assess children's intellectual development. In MLLMs, examinees are characterized by parameter vectors $\theta_j = (\theta_{1j},...,\theta_{nj})$. The components of θ_j are associated with items I_1, ..., I_n. There is no assumption about the mutual dependence or independence of latent dimensions. Consequently, this model is more flexible for many areas of research.

MLLMs are usually applied in repeated measurement designs for two or more time points. Given two time points, two model equations are as follows.

$$P(+|S_j I_i T_1) = \frac{\exp(\theta_{ij})}{1+\exp(\theta_{ij})} \qquad \text{(Formula 4.7)}$$

$$P(+|S_j I_i T_2) = \frac{\exp(\theta_{ij} - \delta_j)}{1+\exp(\theta_{ij} - \delta_j)} \qquad \text{(Formula 4.8)}$$

Where θ_{ij} is examinee $S_j's$ position on latent dimension D_i measured by item I_i at time point T_1.

δ_j is the amount of change in S_j between T_1 and T_2

MLLMs are widely used to measure change, and the expanding of unidimensionality to multidimensionality is useful for heterogeneous cognitive tasks.

To solve different cognitive tasks, MLLMs can be classified into compensatory and non-compensatory models. Embretson (1990) provides a clear description of these two types of models.

> A noncompensatory model that is appropriate for processing components is the multicomponent latent trait model. This model contains both person and item parameters for each component. In this case, non-compensatory implies that each component is necessary for item solving, so that the model is multiplicative. Thus, a low ability in one component implies a low probability of item solving regardless of the ability levels on the other components.
>
> In a compensatory latent trait model, a person's response potential for a given item depends on the item's threshold and a weighted combination of several abilities. (p. 423)

Thus, the two models are used to assess cognitive tasks in terms of task features and assessment design where different subtasks require different knowledge structures.

4.1.1.3.4 Tatsuoka's rule space model (RSM) and its application in cognitive assessment.

Tatsuoka's RSM represents a combination of cognitive and psychometric models. It begins with the design of an incidence Q matrix which identifies cognitive components of performance ("attributes") that are required to complete tasks (items) by which an individual's proficiency is assessed (Junker & Sijtsma, 2001; Tatsuoka, 1990, 1995). Items that require the same attributes for their successful performance are identified (using the Q matrix), and constitute an item type. For each item type, a Rule Space Model R is constructed in which an IRT model is used to assess rule mastery (i.e., proficiency θ). The model includes a slip probability for each item representing the occurrence of a statistically random error component ("slip") in solving the item. The rule space model is used to estimate examinee proficiency and diagnose examinee performance. An example of the application of Tatsuoka's RSM will be described, a summary of how to develop a RSM is presented, and finally a description of how the RSM is used to provide diagnostic assessment will be given.

(1) An example of the application of Tatsuoka's RSM

Tatsuoka's RSM has been applied to assess algebra problem solving (Tatsuoka, 1990, 1995). The RSM represents cognitive attributes which specify the declarative knowledge, cognitive processes and solution strategies that are involved in solving algebra problems. A classification space formulated in terms of relationship between these attributes and items is cognitively modelled by a Q

matrix consisting of two vectors: attributes A_1, A_2,... A_k, and items I_1, I_2, ...I_k. Five attributes have been identified for fraction addition problems (Tatsuoka, 1995):

A1 Convert the first mixed number to a simple fraction

A_2 Convert the second mixed number to a simple fraction

A_3 Take the common denominator and make equivalent fractions

A_4 Add the two numbers

A_5 Answer to be simplified to the simplest term. (p.337)

Solving given fraction addition problems requires one or more of these attributes. There may be a variety of combinations among the attributes A_1—A_5 for different addition fraction items. If the response space is dichotomous, examinee responses to items would be combinations of 0's and 1's in various patterns. Based on the cognitive attributes, different combination patterns reveal different diagnostic information and progress information as examinees solve fraction problems.

(2) Developing the RSM

Gierl, Leighton, and Hunka (2000) summarize Tatsuoka's RSM in test development and analysis. They propose that a task domain such as fraction addition must be defined by clarifying the items and attributes required to perform the tasks. First, dependency relations linking attributes are represented in order to specify the ordered or hierarchical relationships constituting an attribute model. Second, a potential pool of item types reflecting all possible patterns of attributes must be listed in a Q matrix. Third, items, inconsistent with the attribute model, are eliminated in order to obtain the effective Q matrix. Further, an IRT model is

applied to estimate a latent proficiency θ_m for each attribute. Finally patterns of scores on these attribute mastery scores are used to classify subjects into knowledge state categories using discriminant analysis techniques. Attribute patterns provide diagnostic information about errors or misconceptions, and about knowledge states when compared to ideal patterns.

(3) Approach to Diagnosis in Tatsuoka's RSM

The Rule Space Model for an item type class is used to calculate two measures based on an examinee performance patterns on sets of items of particular item type (Tatsuoka,1983, 1985). If $\underline{X}=(x_1, x_2, \ldots, x_m)$ is a vector of item score (0 or 1 on each of n items), the Rule Space Model estimates two "parameters" for each subject, based on the vector of scores for that subject: (a) θ_R, the subject level of mastery (proficiency) of the knowledge components needed to solve items of the particular item type; and (b) ζ_R, a measure of the discrepancy of subject performance $\underline{X}$ from an ideal performance that would be predicted for a subject at a particular proficiency level θ. This discrepancy index reflects the occurrence of "slips" (errors) in the subject's performance on the items.

Rule Space Models are developed for all the different item types (Katz, Martinez, Sheehan, & Tatsuoka, 1998). Bug distributions are introduced into the models as slip parameters. Students' response patterns are used to compute the values of θ_R and ζ_R, in the rule space for each item type, and students are matched to the closest rule centroid (i.e., mean values of θ_R and ζ_R) in the rule

space. Discriminant analysis is used to calculate posterior probabilities for each item type. In this way, diagnoses of bugs in specific rules are obtained for each student.

Psychometrically, this is a complex model. Its application is confined to cases where all attributes have been clearly identified. Items must represent at least one attribute. The attributes and items constitute a complex space in a complex task domain.

4.1.1.3.5 IRT models that introduce multiple responses into the model.

As task or item levels become more complex, polytomous IRT models can be considered. Polytomous IRT models are needed to describe relations between participant trait levels and their probability of responding in particular categories to an item (Andersen, 1995; Embretson & Reise, 2000).

Polytomous models are classified as direct and indirect in terms of steps needed to determine the conditional probabilities of test subject responses in given categories. Indirect models require two steps for estimating parameters and proficiency. Typical indirect models are graded response models (GRMs; Samejima, 1969, 1996) and modified graded response models (M-GRMs, Muraki, 1990). Direct models require one step for estimating parameters and proficiency. Typical direct models are partial credit models (PCMs) initially developed by Masters (1982) and general partial credit models (G-PCMs; Muraki, 1992).

GRMs are appropriate when task data can be characterized in a format of ordered categorical responses. These ordered categories are used to assess student performance such as in rating scales. The mathematical model is:

$$P_{ix}^{*}(\theta) = \frac{\exp[\alpha_i(\theta - \beta_{ij})]}{1 + \exp[\alpha_i(\theta - \beta_{ij})]} \qquad \text{(Formula 4.10)}$$

where x is an examinee's raw item response

i is the number of an item

$P_{ix}^{*}(\theta)$ is the "operating characteristic curve" used to estimate examinees' trait level θ based on category thresholds. Thus, in a graded response item with four categories, three β_{ij} parameters are involved. Based on each $P_{ix}^{*}(\theta)$, the item operating characteristic for each category $P_{ix}(\theta)'s$ is defined by the difference between two consecutive $P_{ix}(\theta)'s$. GRM item parameters determine the shape and location of item category response curves. The between category threshold parameter β_{ij} dictates the location of the operating characteristic curve.

M-GRM (Muraki, 1990) is used to analyze questionnaire data in which items correspond to equally-spaced categories along a scale. Here β_{ij} is the difference between a location parameter b_i and a threshold parameter c_j. According to Embretson and Reise (2000):

> The difference between the GRM and M-GRM is that in the GRM one set of category threshold parameter (β_{ij}) is estimated for each scale item, whereas in the M-GRM one set of category threshold parameters (c_j) is estimated for the entire scale, and one location parameter (b_i) is estimated for each item. (p. 103)

PCMs (Masters, 1982; Master & Mislevy, 1993) are used to analyze test items requiring multiple steps to complete, as in solving mathematical problems. PCMs are direct models so the probability of getting a given category response is expressed in the exponential model (Embretson & Reise, 2000):

$$P_{ix}(\theta) = \frac{\exp[\sum_{j=0}^{x}(\theta - \delta_{ij})]}{\sum_{r=0}^{m_i}[\exp\sum_{j=0}^{r}(\theta - \delta_{ij})]} \qquad \text{(Formula 4.11)}$$

where $\sum_{j=0}^{0}\theta - \delta_{ij} \equiv 0$

In Formula 4.11, δ_{ij} is the "item step difficulty" for step j. For item i, θ is the subject's score on the latent dimension where there are $m_i + 1$ steps from 0 to m_i on item i.

PCMs were developed from Rasch models, and assume that all items share the same discrimination parameter. Having considered different discrimination parameters, or slopes between different items, Muraki (1992) modified PCMs by adding a parameter α_i to each exponent term. The resulting G-PCM provides more information for learning assessment:

$\exp\sum_{j=0}^{x}\alpha_i(\theta - \delta_{ij})$. Polytomous models have been developed for a variety of different tasks or item responses, and designs.

4.1.1.3.6 Application of multicomponent IRT models in cognitive assessment.

Multicomponent IRT models have been applied to assess multiple latent proficiency components (Embretson, 1997; Whitely, 1980). Although cognitive tasks are often assumed to require multiple processing stages and strategies, it may not be appropriate to view these cognitive aspects as different trait dimensions. Rather, in multicomponent IRT models, they are seen as distinct cognitive components. Multicomponent IRT models have been applied to spatial tasks (Pellegrino, Mumaw, & Shute, 1985) and to mathematical problems (Embretson, 1995). Three multicomponent IRT models will be considered: (1) LLTM models which incorporate task components into IRT models, (2) MLTM models which strictly require mastery of multiple traits m that correspond to corresponding task components m, and (3) GLTM models.

Embretson (1993), and Embretson and Reise (2000) merge Linear Logistic Latent Trait Models (LLTMs) and multicomponent latent trait models (MLTMs).

(1) LLTMs (see section 4.1.1.3.1 (4) above) were developed to incorporate task components into predictions of task success by specifying relations between content factors and specific component tasks:

$$P(X_{ij}=1|\theta_j,\tau_k)=\frac{\exp(\theta_j-\sum_{k=1}^{K}\tau_k q_{ik})}{1+\exp(\theta_j-\sum_{k=1}^{K}\tau_k q_{ik})} \quad \text{(Formula 4.12)}$$

Where q_{ik} is the value of stimulus factor k (k=1, ..., K) in item i;

τ_k is the weight of stimulus factor k in item difficulty

Embretson (1993) applied LLTMs to a spatial folding task which involved spatial reasoning. Based on previous research (Shepard & Feng, 1972; Shepard & Metzler, 1971), it was suggested that the angle of rotation is linearly related to response time, and the number of surfaces carried also influences processing difficulty. Based on these studies, Embretson (1993) proposed an attached folding model for the spatial folding task that comprises four major task components: encoding, attaching, folding, and confirming. Having compared different cognitive models based on these components, Embretson (1993) established a linear equation of item difficulty index consisting of four q_{im}'s as their coefficients. The LLTM made progress by expanding the difficulty index, but it does not involve a multicomponent and ability parameter.

(2) MLTMs were developed to measure multiple processing components in which both multi-trait levels and multiple difficulties are estimated for each component m. Item success is assumed to depend on success of several components. MLTM is a strict non-compensatory model.

$$P(X_{ijT}=1|\underline{\theta}_j,\underline{\beta}_i)=\prod_{m=1}^{M}\frac{\exp(\theta_{jm}-\beta_{im})}{1+\exp(\theta_{jm}-\beta_{im})} \qquad \text{(Formula 4.13)}$$

Where θ_j is the trait levels of examinee j on M components (m=1, 2, ..., M)

$\underline{\beta}_i$ is the vector of item difficulties i's on the M components (m=1, 2, ..., M)

θ_{jm} is the trait level of person j on component m (m=1, 2, ..., M)

β_{im} is the difficulty of item i on component m (m=1, 2, ..., M)

Embretson and Reise's (2000) MLTM dealt with relations between abilities and item difficulties in spatial folding reasoning tasks. Abilities θ_{sm} and item β_{im} difficulties were estimated for each subtask.

(3) In order to build a comprehensive model, Embretson (1984) postulated the general component latent trait model (GLTM) which used a multiplicative relationship between item success probabilities for subtasks.

$$P(X_{ijT}=1|\underline{\theta}_j,\underline{\beta}_i)=\prod_{m=1}^{M}\frac{\exp(\theta_{jm}-\sum_{k=1}^{K}\tau_{km}q_{ikm})}{1+\exp(\theta_{jm}-\sum_{k=1}^{K}\tau_{km}q_{ikm})} \qquad \text{(Formula 4.14)}$$

Where τ_{km} is the weight of stimulus factor k on component m (m=1, 2, ..., M)

q_{ikm} is the value of stimulus factor k (k=1, ..., K) on component m for item i

GLTMs include Rasch models (Formula 4.1) for each subtask including trait scores θ_{im} and item parameters β_{im} ($\beta_{im}=\sum_{k=1}^{K}\tau_{km}q_{ikm}$, the item difficulty parameter for component m). GLTMs are extensions of MLTMs. The element which distinguishes Formula 4.13 from Formula 4.14 is the β term. While in MLTMs (see formula 4.13) the β term is a single β_{im}, in the GLTM (see formula 4.14), the β term is a summation of the products of τ_{km} and q_{ikm}.

The basic assumption is that because the completion of a task requires at least two different ability dimensions, MLTM emphasizes the trait dimensions,

which can distinguish initial ability from modifiability, the change between successive measurements. Maris (1995) applied MLTM to examine two components involved in success on synonym items: (a) generation of a potential synonym, and (b) evaluation of a potential synonym. The assumption is that these two cognitive aspects are viewed as different cognitive components rather than two dimensions of one single trait.

4.1.2 Latent Class Models (LCMs) Potentially Applicable to Cognitive Assessment

LCMs were developed to examine qualitative differences in knowledge structures and problem solving strategies in describing different traits and proficiencies of examinees in solving problems. LCMs are relatively independent of statistical models and have recently been used in cognitive assessment (Pellegrino, Chudowsky, & Glaser, 2001).

In LCMs, latent constructs have been characterized as discrete classes in either ordered or unordered ways (Pellegrino, Chudowsky, & Glaser, 2001; Rost, 1990). Selecting an ordered or an unordered model depends on the task to be completed and its problem solving features. The determining factor is how useful the particular measurement model is in reflecting the nature of the cognitive task and providing effective assessment information.

In a basic LCM, it is assumed that there are W latent classes (Pellegrino, Chudowsky, & Glaser, 2001) which correspond to theoretical attributes. These latent classes are directly connected to observed variables. If the design is

theoretically driven, the latent classes will explain the data and predict its distribution. However, if the design is data-driven, the latent classes will be quantitatively updated. The "weights" among different class members will be modified (i.e., the conditional probability of responses given each class membership).

This section introduces three LCMs used in cognitive assessment: (1) the restricted LCM, (2) the hybrid LCM, and (3) the unified LCM. Latent class models emphasize specific ability structures in developing parameters and characterizing relations between traits and tasks.

(1) Restricted LCMs

Haertel (1989) and Haertel and Wiley (1993) present a restricted LCM, referred to as the "binary skills model," to determine the skills required by sets of test items. The model was applied to reading achievement data from a large sample of 4^{th}-grade students and offers useful perspectives on test structure and examinee ability. Restricted LCMs are slightly different from basic LCMs. Mathematically, a set of latent attributes have been added to the between layer of the model. In the first layer, there are W latent classes which determine student traits; students in the same latent class share the same knowledge and problem solving skills. Thus, it is assumed that students possess the same array of attributes. Attribute variables and latent response variables conjunctively describe the stochastic version of restricted LCMs (Maris, 1995; Pellegrino,

Chudowsky, & Glaser, 2001). The "between" layer specifies classes of test items each of which are linked to individual test items in a specific item class.

(2) Hybrid LCMs

Yamamoto and Gitomer (1993) developed a hybrid model to assess cognitive skill representation, which combines a traditional IRT model with the latent class approach which is used to give diagnostic assessment information and to model qualitative aspects of performance. If an assessment is intended to provide diagnostic information about the acquisition of knowledge and skills, standard IRT models are inadequate for describing trajectories on these changeable aspects. Hybrid LCMs directly emphasize relationships between responses from data and a categorical theoretical structure. Yamamoto and Gitomer (1993) characterize hybrid models thus:

> The hybrid model was developed to cope with the need for models to represent qualitative aspects of performance, while at the same time recognizing that performance of some individuals may best be captured by continuous models. The HYBRID model is a hybrid of IRT and latent classes. Examinees are characterized either on an IRT scale or as belonging to one of several latent classes that represent key, qualitatively meaningful cognitive states. As with many measurement models, conditional independence is assumed to hold for both IRT and latent class groups. (p. 277-278)

This statement indicates how the two models are combined into a new hybrid model, extending its function based on separate features of the model. The hybrid model consists of a latent class model, and a two-parameter logistic IRT model which includes two item parameters α_i and β_i. In short, the hybrid model has three sets of parameters: (a) the item parameters α_i and β_i for each

item, and a trait score θ_j reflecting the proficiency of examinee j; (b) weights of individuals for the IRT model and the latent classes, and (c) a set of conditional probabilities for each latent class.

If a two-parameter logistic IRT model is applied, proficiency with conditional probability of a correct response for item i with parameter values $\zeta_i = (\alpha_i, \beta_i)$ given examinee j's score on trait θ_j:

$$P(x_i = 1 | \theta_j, \zeta_i) = \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]} \qquad \text{(Formula 4.15)}$$

where α_i is the item discrimination parameter

β_i is the item difficulty parameter.

Based on the assumption of conditional independence in IRT models and latent class groups, the joint conditional probabilities of a response vector X under (a) the IRT class, and (b) the latent class can be expressed respectively (for a test consisting of i items) as:

$$P(\mathrm{X} | \theta, \zeta) = \prod_{i=1}^{I} P(x_i = 1 | \theta, \zeta_i)^{x_i} [1 - P(x_i = 1 | \theta, \zeta_i)]^{1 - x_i} \qquad \text{(Formula 4.16)}$$

$$P(\mathrm{X} | \gamma = \kappa) = \prod_{i=1}^{I} P(x_i = 1 | \gamma = \kappa)^{x_i} [1 - P(x_i = 1 | \gamma = \kappa)]^{1 - x_i} \qquad \text{(Formula 4.17)}$$

Where $\gamma = 1$ represents the IRT group, and $\gamma > 1$ represents the κ latent classes.

Formulas 4.16 and 4.17 can be combined to provide a basis for modeling the proficiency of a response pattern vector X given the IRT parameter ζ and the latent classes γ in a merging model:

$$P(\mathrm{X}\mid\zeta)=\sum_{\kappa=1}^{K}P(\mathrm{X}\mid\zeta,\gamma)P(\gamma=\kappa) \qquad \text{(Formula 4.18)}$$

Formula 4.18 represents the hybrid model combining the IRT and the latent class models.

(3) Unified latent class model

DiBello, Stout, and Roussos (1995) present a unified format for LCMs. This multi-strategy model is a unified model that "brings together the discrete, deterministic aspect of cognition and the continuous, stochastic aspect of test response behavior that underlie item response theory" (p. 361). In unified model development, DiBello, Stout, and Roussos (1995) stress the response variations featured in multiple strategies, completeness, positivity, and slips. DiBello, Stout, and Roussos (1995) used the unified model in the domain of solving algebra problems to provide better diagnostic information for students' learning to solve algebra problems. However, the unified model is complex and has many parameters.

As cognitive components become more complicated and learning environments change, increasing numbers of parameters and the complexity of assumptions does not enhance the validity of cognitive constructs. Especially for web-based learning, more effective statistical models should be developed and applied to give stakeholders richer diagnostic information and assessments of learning processes.

4.1.3 Bayesian Networks in Evidence-centred Performance Assessment

Bayesian network models are relatively independent in their development and are rooted in AI. However, over the last two decades they have been applied to assess physics and mathematics problem solving, and to assess troubleshooting skills in aeronautical hydraulics in a tutoring system (Gitomer, Steinberg, & Mislevy, 1995; Martin & VanLehn, 1995a; Mislevy, 1995).

Bayesian networks can be used to assess complicated learning tasks where learners and instructors are more concerned with learning processes, the problems that occur, and the troubles that develop as learning progresses. The tasks usually have multiple steps which are often ordered or conditionally dependent. These features coincide nicely with the functions of Bayesian nets (Jensen, 2001).

4.1.3.1 Fundamental Representation of Bayes Theorem and Bayesian Networks

Bayesian networks are used to make predictions in situations where data or observations are limited. The basic theory is based on Bayes theorem and its variations (Pearl, 1988, 2000) which can be written mathematically as the Bayesian inversion formula:

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(A)} \qquad \text{(Formula 4.19)}$$

Where P(B|A) is the conditional probability of the hypothesis B given evidence A;

P(B) is the prior probability of hypothesis B;

P(*A*|*B*) is the conditional probability of evidence A given model B;

P(A) is the probability of evidence A (Jensen, 2001).

Formula 4.19 can be expanded into more than one hypothesis. For example, if a vector $\underline{B} = B_1,...,B_i,...,B_m$, then the probability of hypothesis B_i given evidence A is:

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)} \qquad \text{(Formula 4.20)}$$

Where $P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$

The Bayesian theorem describes relations between prior and posterior probabilities, and provides a theoretical basis for Bayesian network theory, on which Bayesian networks are defined (Jensen, 2001). A Bayesian network is represented as a directed acyclic graph (DAG):

> A set of variables and set of directed edges between variables;
>
> Each variable has a finite set of mutually exclusive states;
>
> The variables together with the directed edges form a directed acyclic graph (DAG) from B_1 to B_n;
>
> To each variable B with parents A_1, ..., A_n, there is an attached potential table P(B | A_1, ..., A_n). (p.18-19)

The definition identifies the variable children that are (e.g., B_i) of variables representing parent nodes (e.g., A_i).

A more general expression of the joint probability distribution of observations (x_1, x_2, ..., x_n) can be expressed as the product of the conditional

distribution of each variable x_i given only its parents pa_i: $P(x_i|x_1,\ldots, x_{j-1})=P(x_i|Pa_j)$. The joint probability of the variables in a DAG network of a parent-child relations (Pearl, 2000) can be written as:

$$P(x_1,\cdots,x_n)=\prod_{i=1}^{n}P(x_i\mid pa_i) \qquad \text{(Formula 4.21)}$$

We assume that when parent probability pa_i's have been given, the number of parameters grow as the size of the network increases. However, the joint distribution increases move rapidly. The D-separation rule and the chain rule for Bayesian networks (Jensen, 2001) have been effectively applied to solve the problem computationally enabling the efficient calculation of node probabilities (posterior probabilities) conditional on other network nodes.

4.1.3.2 A Basic Rule of Bayesian Network: D-Separation

The D-separation rule states that if variables A and B are separated by V which has also been instantiated, then A and B are D-separated. Information from A does not transfer to B. Thus, variables A and B are independent.

The chain rule for Bayesian networks states that over a set of nodes U= $\{A_1, \ldots, A_n\}$, the joint probability distribution P (U) is the product of all potential variables specified in a Bayesian network (Jensen, 2001):

$$P(U)=\prod_{i=1}^{n}P(A_i\mid pa\,(A_i)) \qquad \text{(Formula 4.22)}$$

Where $pa(A_i)$ is the parent set of A_i.

The D-separation and chain rule provide an elegant method for dealing with the huge amount of joint distribution information. When D-separation exists

between two variables, information does not transfer. The chain rule indicates that as long as we know the probability of the adjacent parent, the children's probabilities can be inferred. It is not necessary to trace back to remote parent variables. Therefore, the D-separation and the chain rule achieve compactness in calculating the joint probabilities of nodes by factoring joint distributions into local, conditional distributions for each variable given its parents.

4.1.3.3 Example of a Bayesian Net Applied in Assessment to Physics Problem Solving in College

Bayesian networks have been applied in college physics problem solving (Martin & VanLehn, 1995b). Martin and VanLehn built a Bayesian network model in OLAE, a computer-based physics learning environment. To model problem solving behaviour (applications of sequences of rules) by a sample of students, their assessment emphasized the problem solving process rather than only the results and diagnosis of the presence or absence of components of student knowledge and skills. In complicated problem-solving tasks such as physics problem solving, diagnostic assessment inevitably involves uncertainty. As students solve problems, typing errors (slips) occur, correct answers are guessed, or there are multiple solution paths. To tackle assessment uncertainties, OLAE uses Bayesian nets, since this approach allows both the ranking of hypotheses and considerations of the impact of prior knowledge.

OLAE Bayesian nets use four types of nodes to represent learning trajectories (Martin & VanLehn, 1995b).

> (a) the student knows a rule from the student model of elementary physics (rule node),
> (b) the student actually used a rule during solution of a given problem (rule application node),
> (c) the student believes a particular fact about the given problem (fact node), and
> (d) the student has performed a particular action (action node). Fact nodes include equations that the student might write. These nodes are connected by directed edges (arrows) in the net. (p.579)

The four types of nodes provide sufficient options for developing a large set of Bayesian nets for recording and assessing student problem-solving activities. Probability distributions of student knowledge and skills in problem solving have thus been characterized and modeled in Bayesian networks.

OLAE Bayesian nets assume prior probabilities to be uniform under the condition of limited sample size. However, the expectation maximization (EM) parameter estimation technique was used to estimate conditional probability parameters, as large amounts of data become available. The Bayesian nets were composed of two kinds of node (variable) relation: leaky-AND gates (to determine the probability that a Boolean function was computed incorrectly), and leaky-XOR gates (to satisfy the assumption that learners rarely infer the same fact twice). The XOR rule guarantees that only one input is true (Martin & VanLehn, 1995a; Martin & VanLehn, 1995b).

4.1.3.4 Example of a Bayesian Network in Assessment of Mathematics Problem Solving

Another application of Bayesian networks is to assess mathematical problem solving in secondary school (Mislevy, 1995). The main task was to build

an inference network, based on an analysis of mathematical problem solving. The study focussed on developing the consecutive steps students used in solving fraction problems. Mislevy (1995) summarized the method used to construct the Bayesian assessment model in seven steps.

Step 1. Recursive representation of the joint distribution of variables.

Step 2. Directed graph representation of step 1.

Step 3. Undirected, triangulated graph.

Step 4. Determination of cliques and clique intersections.

Step 5. Join tree representation.

Step 6. Potential tables.

Step 7. Updating scheme. (47-56)

In step 1, a representation of the joint distribution of the defined variables is described based on a directed acyclic graph (DAG). In step 2, a DAG is created to represent a sequential task. The DAG specifies the Bayesian inference rules needed to indicate conditional dependency relationships among variables used to complete joint probability distributions. Step 3 is used to confirm that the graph is composed of singly connected networks of these variables, and the information in the network is not allowed to loop. Step 4 is about determining cliques and clique intersections. Cliques define local structure analysis, and serve to "recognize" variable patterns. For instance, for two sets of variables, usually, groups of overlapping variables are represented by clique intersections. There are multiple ways to define cliques in terms of analysis angles (relations) and purposes. Step 5 is to join the cliques into a tree representation. As soon as the cliques and clique intersections are defined, possible connected structures among cliques and clique intersections can be

determined. Step 6 is used to make a local calculation of probabilities within cliques and their intersections. These probability distributions are depicted in conditional probability tables associated with each link in the clique. Potential values are prepared for use in step 7, local updating. In step 7, new evidence is acquired from marginal probability distributions of variables, and is used to update conditional probability tables.

Bayesian networks have been used in cognitive and diagnostic assessment in expert systems in medicine, avionics and aeronautical hydraulics ITS and social science (Andreassen, Jensen, & Olesen, 1990), the dental interactive simulation corporation (DISC) project in stomatology and dental hygiene (Mislevy, Steinberg, Almond, Breyer, & Johnson, 2001), MashpeeQuest, an on-line history project (Mislevy, Steinberg, Almond, Haertel, & Penuel, 2001), and HYDRIVE, an operational computer-based intelligent tutoring system built to help Air Force technicians develop skills for troubleshooting hydraulics system in aeronautical hydraulics (Gitomer, Steinberg, & Mislevy, 1995; Mislevy & Gitomer, 1996). These applications show how to build domain models, how to develop Bayesian networks based on domain models and how to compare models according to test performance. The different research aspects of Bayesian networks provide a solid foundation for exploring cognitive assessment.

4.2 Task Models and Task Environments in Cognitive Assessment

Task models play essential roles in the design of construct-based assessments and influence structures of assessment design from assessment

construct to evidential variables and tasks. In presenting the evidence-centred design approach to cognitive assessments, Mislevy, Steinberg and Almond (1999) stated that:

> A task model provides a framework for describing the situations in which examinees act. In particular, this includes specifications for the stimulus material, conditions, and affordances, or the environment in which the student will say, do, or produce something. It includes rules for determining the values of task-model variables for particular tasks. And it also includes specifications for the work product, or the form in which what the student says, does, or produces will be captured. Altogether, task-model variables describe features of tasks that encompass task construction, management, and presentation. (p. 19)

Mislevy, Steinberg and Almond (1999) provide a general description of the aspects of a task model that is of concern in the design of cognitive assessment. However, learning objectives and processes can have many formats and trajectories. In particular learning situations and task models may be developed in distinct ways. Fundamental considerations in developing a task model and establishing its relevance are: (a) the specification of cognitive tasks, (b) the establishment of a task model for these tasks, and (c) the task environments. These aspects determine the features and functions of task models, and relations between a task model and other models such as the student model, and a model of how the components of knowledge and competency are connected to the evidence based on student performance.

4.2.1 Cognitive Tasks and Measurable Objects

Defining cognitive tasks for cognitive assessment is a complex process because it requires consideration of assessment purposes, theoretical models of domain knowledge competency, and the design of statistical models that connects components of theoretical models to evidence based on task performance. From various cognitive task designs perspectives, Hollnagel (2003) summarized several steps in the development of cognitive tasks:

- The first step is observation for familiarisation. This should identify cognitive tasks and supporting functions from a functional analysis.
- The second step is to identify criteria for where to locate such tasks in the organization with respect to issues such as the mission requirements, manpower and support equipment, command structures, or operational effectiveness. This step requires a task synthesis to derive what is required to be done by the system and how it must be done...
- The third step is to look at task loading resulting from alternative implementations in terms of acceptable workload and then to consider how to redress what is unacceptable and reiterate the assessment with the revised implementation. This step derives preferred locations for the cognitive tasks.
- The forth step is to identify the necessary supporting functions, or to redress overload.
- The fifth step is to check the viability of the options, amend the task synthesis, and re-evaluate task-loading implications.
- The sixth step is to review the viable options. (p. 236-237)

These steps may help task designers define different tasks to fit alternative assessments motivated by theoretical models of cognitive assessment. Even for a single assessment purpose, different aspects of

knowledge or skills may be highlighted, and the cognitive tasks might have different formats or be organized in different ways. Therefore, even though some task design processes may reflect the steps listed above designers need not follow the same steps for cognitive tasks.

In the case of a simple mathematics problem or a multiple choice question in medicine or engineering, cognitive tasks are often relatively well defined and have clear problem spaces. Other tasks may not be so very well defined. For example, cognitive tasks in medical problem solving can be ill defined (VanLehn, 1989), so that it may be difficult to lay out a complete problem space. An unclear problem space is usually apparent in two ways: (a) problem statements may be unclear and may represent incomplete or ambiguous problem spaces; and (b) cognitive processes required to solve problems may be complicated even though problem spaces are relatively well defined. The latter case recently occurred in OLAE (Martin & VanLehn, 1995b). Although it is theoretically possible to clarify the task space, in practice, there are so many ways of doing a task that determining a complete task space can be very difficult.

One critical task attribute is object measurability. Cognitively, when an assessment task is limited to observable behaviors, it is measurable, and can be referred as to a measurable object. In a sense, learning tasks are relatively unconfined. However, they can certainly be limited by time, feasibility, and assessment. Therefore, measurability is important and very relevant to the design of statistical and theoretical models.

4.2.2 Task Models and Structures

Task models are structures for arranging and connecting cognitive tasks and measurable objects to statistical and theoretical models. They involve processes and networks, especially in complex cognitive task environments.

4.2.2.1 Compensatory and Non-compensatory Task Models

Task design in cognitive assessment can be traced back at least three decades even though theoretical models often were not very well defined. Embretson (1990) designed cognitive tasks for a diagnostic test to measure cognitive processes in mathematical reasoning. In her research, the design of a theoretical model was oriented toward diagnosing certain errors or problems that occur when students are learning to reason mathematically. Theoretical models specify the constructs to be assessed through a dynamic assessment process. She used an additive factor model in her design to provide a basis for developing cognitive tasks.

Two different task models were applied to assess mathematical reasoning and to determine the statistical models: compensatory and non-compensatory models, which correspond to different task models. The compensatory model used a multidimensional latent trait model in which response proficiency depends on the task's threshold and a weighted combination of several aspects of ability. The non-compensatory model employed a multicomponent latent trait model in which the ability to complete each subtask was assessed.

In compensatory task models the four task components were nested bottom-up so that each item covers the task listed for itself, and the tasks below it. In the theoretical model, the knowledge structure consists of four categories: factual, schematic, strategic, and algorithmic. Whereas Task 4 targets algorithmic knowledge, Task 3 covers strategic and algorithmic knowledge. In the same fashion, Task 2 requires schematic, strategic and algorithmic knowledge, and Task 1 covers all categories of knowledge. The advantage of the compensatory task model is that different knowledge components can be isolated in steps. For example, if an examinee can complete Task 3, but not Task 2, a simple subtraction will indicate that the examinee failed to master schematic knowledge. The contrast between any two combinations of tasks can lead to different response patterns.

The theoretical structure of the non-compensatory task model incorporates the same categories, but uses different subtasks. In this design, the task covers all four knowledge categories; sub-tasks 1, 2, 3 and 4 emphasize assessing each knowledge category. The advantage of this model is that each kind of knowledge can be tested without contrasting different subtasks.

4.2.2.2 Schema-based and Network Task Models as a Basis for Developing Measurable Objects

A cognitive task can be implemented as measurable objects using different formats. It can be a well-defined item, an ill-structured problem, a physical problem, or a collection of patient symptoms to be diagnosed (Marshall, 1990). Network task models can be used in such complex cognitive tasks. This section introduces two different network task models: A schema-based task model and a hierarchical task model.

Marshall (1990) in her research on ability structure analysis proposed a schema-based task model. She believed that cognitive complexity is relevant to models of schema knowledge, based on which a learner's competence can be assessed. The fundamental assumption in schema assessment is that knowledge structures are organized into networks. Marshall (1990) states that:

> There are at least three distinct sets of elements: declarative facts, preconditions, and subsequent procedures or rules. Any test item for a schema would call for subsets of these three sets. The diagnostic problem is to test various subsets and thereby to estimate efficiently the completeness of an individual's knowledge of the schema. (p.441-442)

In graph theory, a schema consists of a set of nodes and arcs. Concept knowledge is represented by nodes, and relations are expressed by arcs. The ability of learners to carry out cognitive tasks is closely related to the extent to which nodes and arcs can be applied in complex cognitive tasks (Marshall, 1990):

> Construct a set of items, I, to test the sampled nodes of S. In the extreme case, one might sample all N nodes, and one might evaluate each node by one test item, so that

> N=S=I. For the moment, assume that N>S=I. That is, we sample a subset of nodes S and construct one item to test each node. Present the item to the individual, and score the response as 1 or 0 according to the individual's success or failure in responding. Denote the response as X_i, with i=1, ..., S. the overall success of the individual can be expressed by $P= 1/S X_i$. (p. 444)

This statement indicates that different learners assume different patterns on task containing N nodes. Different patterns represent different task models and demonstrate a particular distribution of successful and unsuccessful performance. In addition, the number of nodes indicates the difficulty of a cognitive task.

A hierarchical task model is a special case of a network task model (Annett & Cunningham, 2000; Essens, Post, & Rasker, 2000; Frederiksen & Donin, 2005). The former two research groups focused on a cognitive task analysis of Marine Corps command posts in which the cognitive skills of a team are modeled using hierarchical task analysis. These analyses do not emphasize aspects of cognitive assessment.

The latter research group focussed on a statistics (ANOVA) learning system characterized as a web-based tutoring process in which students can select any problem from a web-based problem bank. The learning process is decomposed into several cognitive tasks. The stimuli are questions based on the learning and assessment characteristics. In each task, sub-tasks can be hierarchically arranged. The hierarchical model decomposes the task into basic components determined by learning goals and assessment. This model

combined with an appropriate statistical model, can extract information that has been learned and can determine how much is known about given subtasks.

According to Mislevy, Steinberg, and Almond (1999), task models cover such topics as specifications of stimulus material, conditions, affordances and the environment. Having specified a task model, the next issue is to investigate the task environment.

4.2.3 Task Environments in Cognitive Assessment

Cognitive task environments can be physical, social and informational environments in which cognitive tasks are produced, organized and supported. Mislevy, Steinberg and Almond (1999) specify relevance conditions, affordances and environments in their definition of task models. The scope of cognitive task environments in this study may be considered to be somewhat larger than in Mislevy's framework. Environments include everything except cognitive tasks and task models. Those that support and are involved in cognitive tasks belong to cognitive task environments.

Cognitive tasks are measurable objects that have long been applied in traditional testing and assessment situations, especially with such objective questions as multiple-choice questions (Haladyna, 1999) which are usually constructed from the domain knowledge of instructors or test developers. Test items may be based on textbook definitions of concepts or readings. As cognitive environments, such item-based tests are often simple and far from authentic,

although they provide easy affordances for cognitive task developers to produce measurable objects.

As criticism increases concerning the use of objective items in achievement assessment, authentic cognitive tasks are receiving much more attention (Birenbaum, 1996; Segers, 1996). Cognitive tasks used in performance assessment (Sackett, 1998) and tasks involving the assessment of complex cognitive skills (Mumford, Baughman, Supinski, & Anderson, 1998) are more complex than textbook tasks. Researchers have begun to realize that dynamic assessment is more valid and useful than static assessment, and that diagnostic information is more meaningful than simple test scores. This trend has led to the design of cognitive tasks emphasizing complexity, authenticity and dynamism.

4.2.3.1 Authentic Problems and Simulated Problem Situations as Complex Task Environments

The application of authentic tasks to cognitive assessment is flourishing. As Segers claims (1996):

> The assessment of students' level of competence in problem-solving is a determinant stimulus for the focusing of students' learning activities on problem-solving. [It demands that] the assessment system is based on authentic problems: contextualized assessment. In PBL it is essential that students learn by the analysis and solving of problems which are representative of the problems to which students will have to apply their knowledge in future. Consequently, a valid assessment system should evaluate students' competences with an instrument based on real life problems. (p. 204)

The selection and creation of authentic problems are important for cognitive task design and is feasible in such fields as mathematics, physics, and

chemistry. Segers (1996) provides an example from economics. The need for authentic tasks is important in such high-risk, high stakes fields as the military and medicine. Lesgold, Lajoie, Logan, and Eggan (1990) developed cognitive tasks to improve the performance of Air Force technical specialists. The task environment was a complex operational system. The researchers developed tasks to simulate avionics troubleshooting procedures based on protocol analysis and task extraction processes embedded in avionic systems.

Gadd and Pople (1990) simulated cognitive tasks for internal medicine teaching rounds on a computer-based system. Because interviewing patients is often difficult and deficient, simulated cognitive tasks are effective learning resources and can be used again and again. Human problem solving discourse has been used to model human-machine interaction in developing computer-based simulations. Clearly, such cognitive task environments are becoming more and more complex. For instance, cognitive tasks for medical students can involve at least two kinds of task environments: (a) diagnostic tasks based on the relevant experiences of medical experts, and (b) computer-based systems which are developed by examining expert cognitive processes, and are used to diagnose student errors and misconceptions. Therefore, information provided by the system, and authentic expertise are relatively different and independent.

4.2.3.2 Web-based Learning and Assessment Systems as Cognitive Task Environments

Web-based learning and assessment systems provide alternative cognitive task environments, and can use dynamic and instantaneous hypertexts. Lajoie, Azevedo, and Fleiszer (1998) developed a simulation-based intelligent tutoring system (ITS) for nurses working in a Surgical Intensive Care Unit (SICU). The rationale was to model expertise for nurses learning in complex decision-making environments. The system mimics real world tasks on web-based cognitive environments from which the cognitive tasks can be developed. The cognitive tasks demonstrate various characteristics:

> (1) ill-structured problems, (2) incomplete, ambiguous, and changing information, (3) shifting, ill-defined, and competing goals, (4) decisions occur in multiple event-feedback loops, (5) time constraints, (6)stakes are high, (7) multiple participants contribute to the decision making process(es), and (8) the decision maker must balance personal choice with organizational norms and goals. (p. 208)

These characteristics inform cognitive tasks that can be developed in a variety of formats to satisfy the needs of cognitive diagnostic assessment, though these characteristics were contextualized in decision-making situations.

Cognitive task environments can also emphasize different aspects of tasks. For example, in a drill-and-practice tutoring system for reading and writing Chinese characters (Almond, Steinberg, & Mislevy, 2002), cognitive tasks can be expressed as different task models. To explore different task features, a cognitive task can be categorized as reading, phonetic transcription, writing, and character identification. Relations between morphemes and phonemes have been

investigated in reading task environments. In phonetic task environments, relations between short strings of characters and phonetic pronunciations have been studied. In writing task environments, the pronunciation of characters used in sentences has been explored. Character identification task environments can involve morphemically similar characters.

Web-based task environments usually provide the possibility of task variables returning different diagnostic information. In a statistics tutoring system (Frederiksen & Donin, 2005), task variables corresponding to help categories can be used to assist learners as they attempt to solve ANOVA problems. Task component response times are another task variable.

Web-based task environments can express different aspects of conditions, affordances, and relevance. The creation and development of new cognitive tasks in such environments can potentially produce valid and valuable assessment information.

4.3 Summary and Conclusion

A statistical model is an engine that connects a theory model to a task model. The kind of statistical model that can be used in cognitive assessment design depends on the theory model, task model, and assessment purposes. Because an appropriate task model is based on a statistical model we began by reviewing statistical assessment models.

Statistical models cover many topics including IRT models, latent class models (LCM), and Bayesian network models. Developers of IRT models began

by analyzing dichotomous responses and expanded to polytomous models. Item parameters included in IRT models ranged from one to two or three parameters. Assumptions were expanded from unidimensional to multidimensional examinee traits. These aspects and others inform many item response models that are applicable to different cognitive tasks, measurable objects, and proficiency and ability assumptions.

In the assumptions of LCM, latent constructs have been characterized as discrete classes that may be either ordered or unordered. Latent classes are often directly connected to observed variables and they may be used in conjunction with IRTs to estimate examinee proficiency.

Bayesian network (BN) models consider statistical models in a different way. In a BN model, examinee abilities are usually expressed probabilistically. Information about variables from different network layers can be transferred based on D-separation, chain and other rules. Transferring information from top layer (theory) variables to bottom layer (observable) variables depends on networks of parent-child relations. Further the direction of information can be reversed to infer mastery of theory variables on the basis of observed (evidence) variables.

Task environments may be distinguished from traditional task models used in item-based tests. Task models, task environments, and their relations were discussed. Cognitive task and measurable objects were examined. Authentic and simulated cognitive task environments were also discussed.

CHAPTER FIVE: METHOD

The fundamental task of cognitive diagnostic assessment for statistics learning with a tutoring system is to establish a model-based assessment framework which provides an effective approach to assessing the details of knowledge and skill acquisition. Based on Mislevy, Almond, and Lukas' (2004) evidence-centred assessment (ECA) framework, evidence comes directly from a performance task which is embedded in an authentic or simulated learning environment. The McGill Statistics Tutoring Project (MSTP) is an environment which serves as a platform for presenting cognitive tasks and for coaching students as they carry them out. First, the ANOVA tutoring system will be described as a task environment. See Frederiksen and Donin (2005) for a more complete description. Second, a stand-alone performance assessment test based on tasks developed in the ANOVA tutor environment will be described. Third, the data collection method will be described, including participants, and data. Fourth, assessment rubrics for scoring student task performance, evidence rules and variables are examined. Fifth, fundamental features of Bayesian networks will be tested on simulated and collected data. Finally, the assessment methodology will be summarized in terms of a model-based assessment framework.

5.1 The Statistics Tutorial System as a Task Environment

A research environment is a platform for implementing a research framework. In this project, the tutorial system functions as a hypertext research

environment. Tutorial systems have been used in assessment for at least a decade (VanLehn, 2001). Two types of tutoring systems have been used in the cognitive assessment of student problem solving in physics and statistics (Martin & VanLehn, 1995a; Frederisken & Donin, 2005): ITSs and knowledge/coaching-based tutoring systems.

5.1.1. The Features of Intelligent Tutoring Systems Function in Cognitive Assessment

ITSs (du Boulay, 2000; Wenger, 1987) have been designed to individualize the educational experience of students according to their level of knowledge and skills. ITSs provide students with individualized, dedicated tutoring based on AI analysis of the procedures. ITSs provide users with feedback, assistance, guidance, use simulations and other highly interactive learning environments requiring learners to use their knowledge and skills. Such learning environments help students apply their knowledge and skills more effectively.

ITSs rely on three knowledge models: expert models, student models, and instructor models. While expert models represent subject matter expertise and specify teaching contents and strategies, student models represent learner knowledge spaces and possible problem solving patterns. Instructor models encode instructional strategies.

ITSs are student-model-centred systems. They adapt students into learning environments and attempt to control student learning processes. Many of ITSs

have been used in training staff and to assess their knowledge and skills in applied physics and mathematics.

5.1.2. The Features of Knowledge-Based Tutoring Systems (KBTS) Functions in Cognitive Assessment

Knowledge-based coaching systems (Frederiksen & Donin, 2005) represent an alternative stream to tutoring systems. They can support learning in ways that are consistent with such cognitive theories as situated learning (Brown, Collins, & Duguid, 1989) and social constructivist theories (Confrey, 1995). A knowledge-based coaching system emphasizes the importance of student acquisition of self-monitoring competence and their need to function as a self-directed learner, providing coaching resources to support student learning and problem-solving strategies. KBTSs are based on models of domain knowledge so that the database includes learning tasks, coaching guidance and assistance in various problem-solving components. Conceptually organized problem-solving knowledge provides cognitive models for use in developing the student model component of a Bayesian network assessment model (Mayo & Mitrovic, 2000). Knowledge-based coaching system can support interaction between tutorial systems and learners, and dynamic assessment.

Tutorial systems present problems for students to practice but not automatically. Such systems contain learning problems at different levels of difficulty, which students working alone or in group can select. They also include hierarchically organized hints and information based on learner errors. Systems

help learners identify and connect their mistakes through guided self-evaluation. They provide diagnostic information in self-evaluation contexts and can assess student learning in an ECA framework.

They emphasize the ability of students to control their learning, and provide tutorial help and coaching. The knowledge assessment can take a Bayesian inference with evidence variables in evidence model.

5.1.3 A Selected Domain and a Tutorial System in Statistics Learning

In this research project, the tutorial system for learning statistics leans towards a knowledge-based coaching system. As cognitive tools, such systems assist students in developing knowledge and skills in various domains. The ANOVA tutor (Frederiksen & Donin, 2005) helps students learn to solve ANOVA data analysis problems in the context of coursework or independently (individually or collaboratively).

The ANOVA tutorial system has two-phases and a multi-component hierarchical design. ANOVA problems are classified as one-way, two-way and others, and each problem is organized into eleven tasks. After selecting a problem, students work through the component tasks (see Figure 5.1). In order to write a relatively complete analytical report, students must complete each task. When students begin working on a task, they can get help by consulting the hierarchically structured on-line tutorial system.

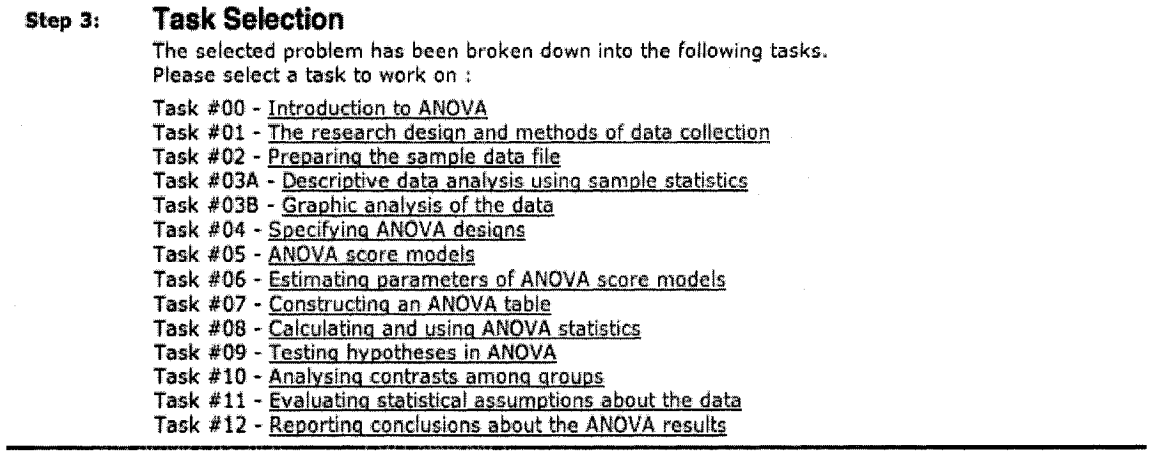

Fig 5.1. Task selection list in McGill Statistics Tutoring Project

Consider task 9, from a two-way ANOVA problem: "effects of cognitive organizers on students learning." After clicking task 9, "Testing of Hypotheses in ANOVA" appears with six options:

New Tutoring System,

Edit Personal Profile,

Get Solution Template,

Show Task Help,

Submit Your Answer, and

Show Previous Self-evaluation.

Students click "show task help" to get help. "Testing Hypothesis in ANOVA" under "Task Help" is expanded by clicking ⊕ to view a list of sub-help indices. "Ask tutor" or "Coaching" can be selected for each row of information. Help indices are organized into three dimensions. One dimension is task help (listed hierarchically) and the other two are "ask tutor" and "coaching."

"Asking tutor" and "Coaching" are parallel while "Task help" and "Asking tutor, and "Task help" and "Coaching" are crossed (see *Figure* 5.2).

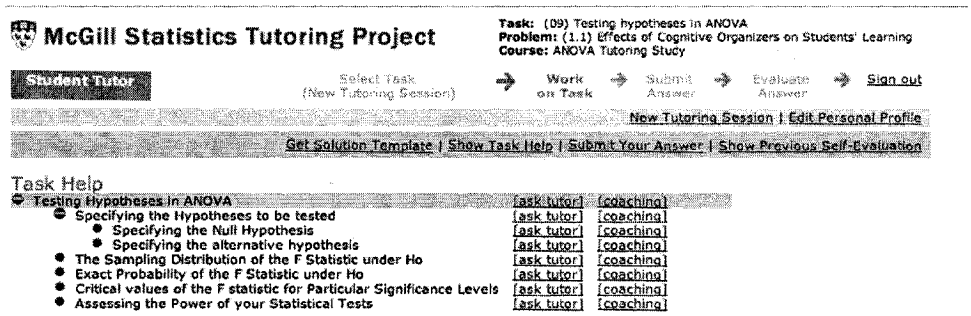

Figure 5.2. Tutor index segment of task 9: Testing Hypotheses in ANOVA

This feature helps students since topics are listed in an orderly and logical fashion. "Help Index" is presented as an outline with indented subtopics. For example, it models a task: Writing the ANOVA Score Model, with a two-way ANOVA model help index which has six sub-indexes on three levels: "Writing ANOVA Score Model", "Two-way (Two-factor) model" and "Grand Mean." (see *Figure* 5.3).

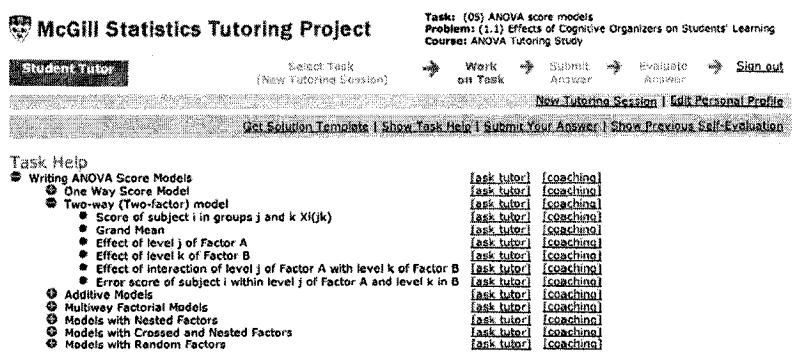

Figure 5.3. A path of help indices of task five: Writing ANOVA Score Model

This design allows learners to select any route in collecting problem solving information. To get help on the effect of level j of factor A, students can go from the "Writing ANOVA Score Model" to "Two-way (Two-Factor) Model" to reach the "Effect of level j of Factor A."

In each help indexed item, there are two phases directing students to two aspects of the information: asking the tutor and coaching.

There are six components in the "ask tutor" phase.

1. Goal: What is the goal of this component procedure?
2. Condition: What must be done before applying the current procedure?
3. Result: What will the results be?
4. Problem State: What is the current situation in which you will be applying the procedure?
5. Theory: What is the conceptual and theoretical background for the procedure?
6. Action: What operations are to be performed?

Two kinds of help are available for each procedural step.

Tool help for:

1. SAS: How to use the SAS tools to perform actions.
2. Graphic: How to use any graphic displays.

Figure 5.4 shows test of hypothesis trajectory and can be described as: test of hypotheses → specifying the hypotheses to be tested → ask tutor → goal. Under "Goal," the goal hypothesis is described in a short paragraph. This

information can help students think about the purposes of hypotheses in their ANOVA problem.

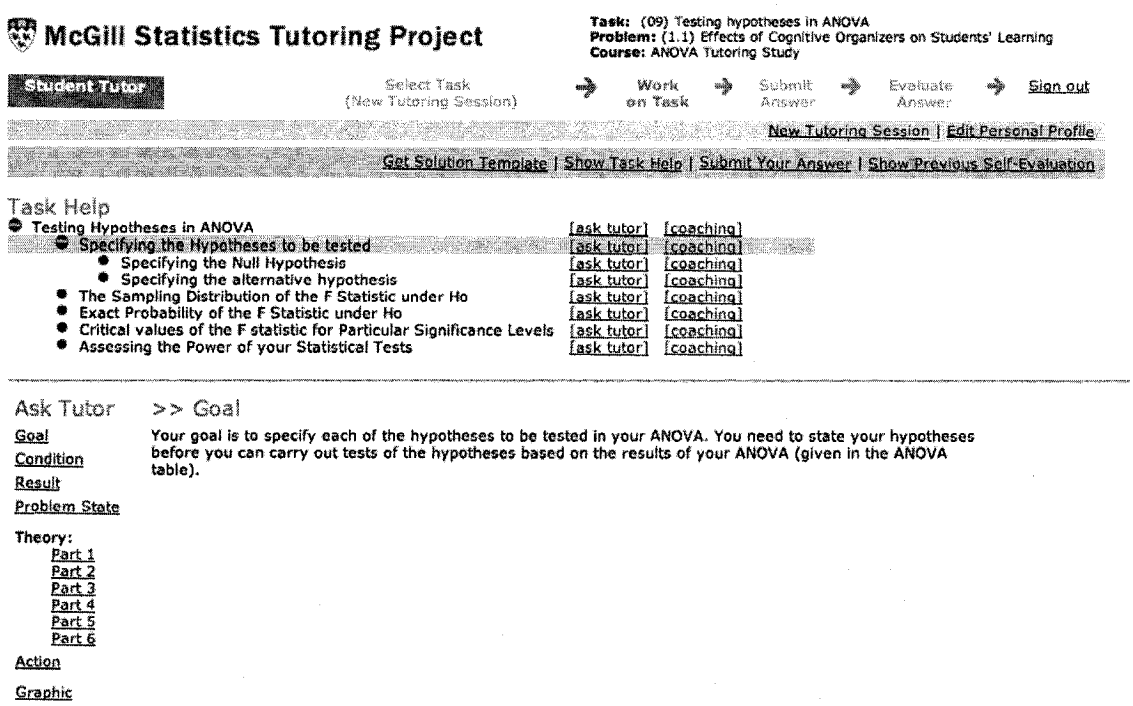

Figure 5.4. Tutor index segment of Task 9: Test of hypotheses trajectory of asking tutor and further goal.

There are four components in the coaching phase: "Questions", "Deep Questions", "Clarification of the Questions", and "Hints."

1. Question: The tutor asks a prompt question.
2. Deep Question: Questions requiring conceptual or theoretical answers.
3. Clarification: The tutor clarifies questions by giving more information.
4. Hint # 1

   Hint # 2

Hint # 3

Hint # 4: Hints give suggestions to prompt various actions required to solve this component of the problem.

Figure 5.5 shows another test of hypothesis trajectories: test of hypotheses → specifying the hypotheses to be tested → coaching → question. Under the terms of the question, learners can continue thinking and problem-solving as prompted by questions.

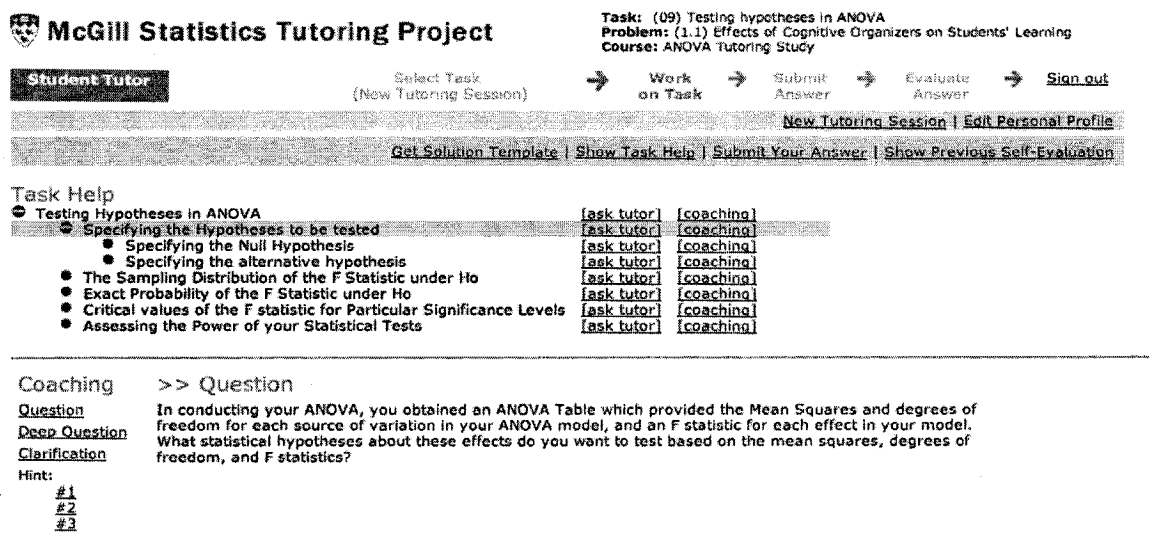

Figure 5.5. Tutor index segment of Task 9: test of hypotheses trajectory of coach and further question.

The "Ask tutor" and "Coaching" features emphasize different cognitive aspects to help students solve problems. In the "Ask tutor" phase, knowledge is organized hierarchically according to a cognitive model of expert procedural and knowledge structures. Learners can choose any topic for help. For example,

learners solving a two-way ANOVA problem may need help with certain concepts and procedures. They can choose an item such as "Theory, part 1." Note that Task 9 and a top-level topic have been selected as an illustration. The scenario can be seen in *Figure* 5.6:

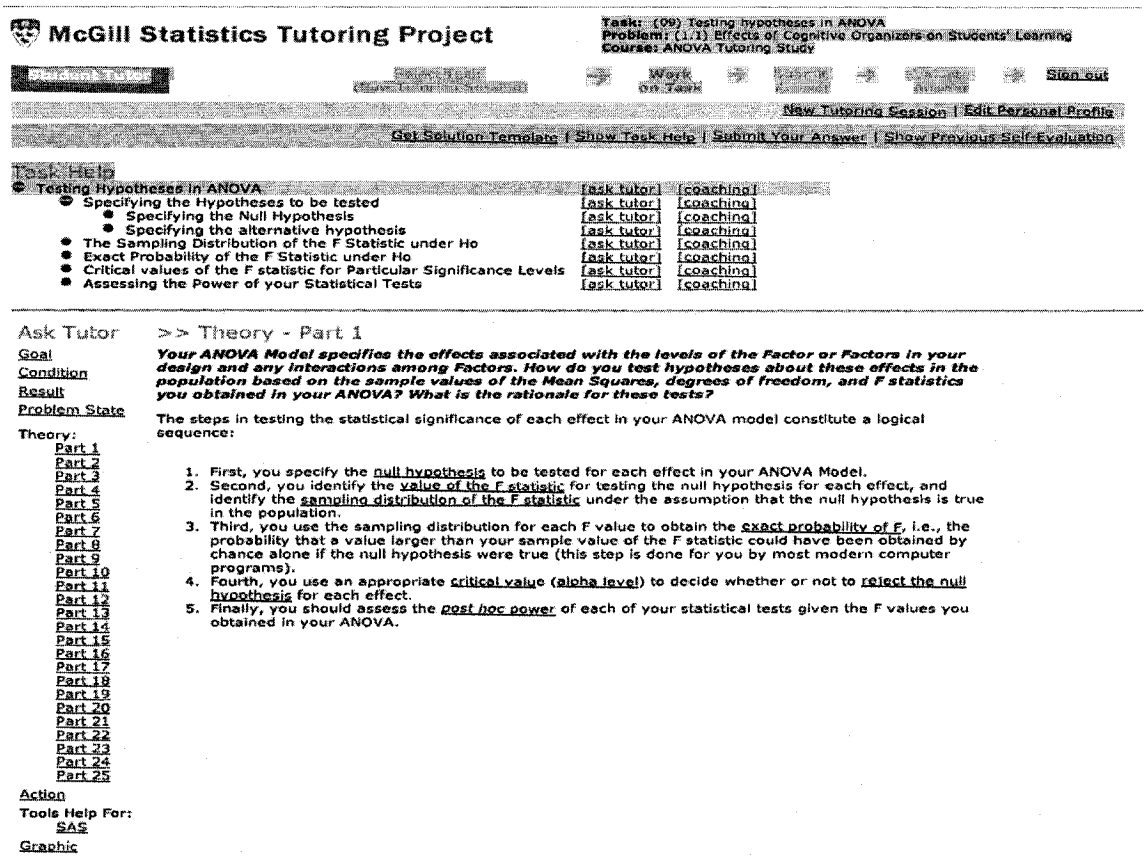

Figure 5.6. Theory part one in test of hypotheses task

In *Figure 5.6*, the tutor has posed a "Deep Question" about hypothesis testing in ANOVA. The tutor scaffolds the student by presenting five steps involved in testing hypotheses. Consecutively, a question about them has been posed and then a series of steps have been suggested.

In the coaching phase, knowledge is organized into questions, deep questions, clarifications and hints. The purpose of the tutor is not to directly explain and describe any concepts, definitions, theoretical mechanisms, models, and their relations. Rather the "Coaching" phase seeks to trigger a chain of deep thinking in learners. For example, in *Figure* 5.7, a deep question prompts students to consider the components of an ANOVA model, and then raises the questions of how to test their significance. Students receive three hints to prompt their thinking about solving this component (i.e., testing the hypothesis).

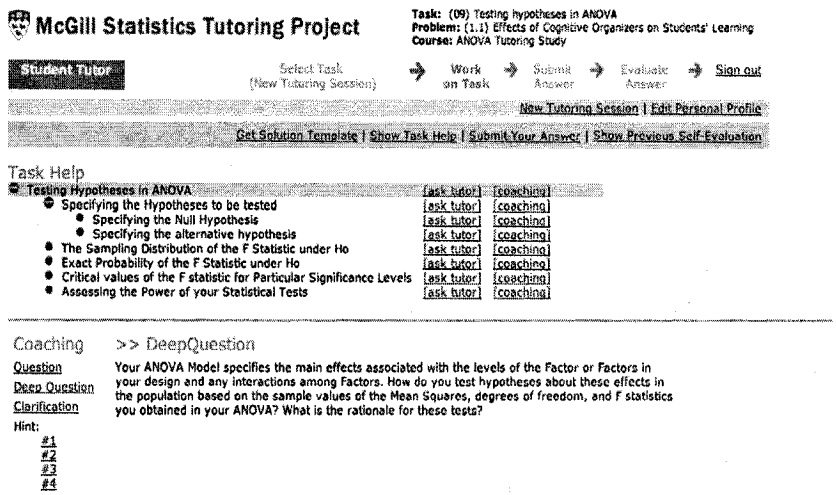

Figure 5.7. Deep question in test of hypotheses task (Task 9)

In brief, the tutorial system provides a research environment in which data can be collected as learners finish ANOVA problem solving tasks. During on-line tutorial-assisted learning sessions, students must select indexed components to obtain useful information. Their trajectories or combinations of sessions help complete the task. As students progress they will need to consult the tutor less.

5.2 A Stand-Alone Test as an Alternative Task Environment

A set of stand alone performance tasks corresponding to ANOVA tutor tasks were implemented as questions in a Performance Assessment of Statistical Learning Test (PASLT) (see Appendix A). PASLT simplifies the task of assessing cognitive processes, learner proficiency and ANOVA problem solving mastery without help from the Tutoring System. Since the tasks match those of the tutor, the cognitive model implemented in tutor explanations, coaching, and task structure can be used as a knowledge model in the assessment model.

5.2.1 The Structure of the Stand-alone Performance Assessment Test

The fundamental test structure simulates the structure of the statistics tutor system. PASLT consists of 13 tasks:

1. Task 1: research method and data collection
2.* Task 2: the sample data file
3. Task 3A: descriptive analysis of the data using sample statistics
4. Tasks 3B: interpretation of graphic representation of the means
5. Task 4: ANOVA design
6. Task 5: ANOVA score model
7. Task 6A: estimating effects
8. Task 6B: estimating residual scores
9. Task 7: analysis of variance table
10. Task 8: calculating and using ANOVA statistics

11. Task 9: testing hypothesis in ANOVA

12.* Task 10: testing contrasts among groups

13. Task 11: conclusion from the ANOVA

Tasks with an asterisk are optional and may be skipped if learners encounter difficulties when solving them.

These tasks directly reflect the complete ANOVA problem solving process and scaffold learners as they try to learn ANOVA problems. If learners can follow this procedure, they will be able to write a relatively complete report reflecting general task difficulty. In the current stand-alone test, Task 2, "the sample data file" and Task 10, "testing contrasts among groups" can be omitted.

The performance assessment test emphasizes two aspects of student cognitive competency: ability to apply knowledge to solve each subtask, and ability to use their knowledge to explain task components. Each task has two types of questions. Table 5.1 summarizes their distribution across 13 subtasks.

Table 5.1 indicates that there are 2 to 6 questions in each task. The numbers of performance and semantic explanation questions are unbalanced across tasks, reflecting different cognitive demands for each task.

Table 5.1. Two Types of Cognitive Tasks Distribution

Number	Tasks	Performance	Semantic explanation	Total
1	Task 1	3	3	6
2	Task 2	1	2	3
3	Task 3A	3	2	5
4	Task 3B	0	5	5
5	Task 4	0	3	3
6	Task 5	3	2	5
7	Task 6A	4	1	5
8	Task 6B	1	1	2
9	Task 7	4	1	5
10	Task 8	6	0	6
11	Task 9	3	2	5
12	Task 10	1	3	4
13	Task 11	1	1	2
Total		30	26	56

5.2.2 The Structural Features of Each Task Worksheet

Task work sheets consist of two parts. On page one there is a vignette followed by several questions. Vignettes can be ANOVA problem descriptions, SAS data steps, SAS program statements, SAS text outputs, or SAS graphic outputs. For example in task five, the ANOVA score model, the vignette is a segment of a SAS statement:

```
Proc   anova data=kirk;
Class group duration;
Model attitude=group duration group*duration;
Run;
```

This four-line statement gives implicit and/or explicit information to which learners can refer in completing the required subtasks.

5.3 Data Collection and Participants

Data was collected from twenty student responses to the stand-alone test.

Twenty participants were selected from students enrolled in a graduate intermediate statistics course covering ANOVA models and designs or who were not currently enrolled but had equivalent statistics background. Participants were informed about the experimental process and that the results would be used only to improve a web-based computer coaching systems and in research on building cognitive diagnostic assessment frameworks based on it (see Appendix I).

While the participants responded to items, they were not allowed to refer to the tutor system. The assumption is that they had some experience using the coaching system and that they had some knowledge and relevant problem solving skills in statistics (including SAS). When they focused on each task, they were allowed to refer back to previous task sections and to their previous response products and previous task vignettes. Student performance records on the stand-alone test were used to develop a scoring rubric that defined observable assessment variables. These variables were used to test cognitive and statistics models in the cognitive diagnostic assessment.

5.4 Development of Assessment Rubrics, Evidence Rules and Assessment Constructs

An assessment (score) rubric represents a step-by-step process for decomposing the ANOVA score model components. The assessment rubric reflects criteria and demands of the stand alone test questions, solution features and relevant expertise embedded in the tutorial system database. Task 5 stand alone test questions implicitly demand students to respond to some aspects of the ANOVA score model knowledge. Solution features are a set of critical points and rules of solutions to tasks representing possible problem spaces. Relevant expertise is widely distributed across tutor modules in help and coaching.

The assessment rubric is a "bridge structure" in building assessment and evidence models. It was developed based on cognitive and content analysis and provides a basis for assessment criterion for student ANOVA task performance. The assessment rubric was elaborated to acquire evaluation rules and then to develop evaluation variables. Rubric structure follows possible decompositions of ANOVA score models into components of which can become rubric categories which can be broken down into sub-components. For example, the error term is decomposed into 3 sub-terms: $e_{(\ )}$, i(jk) and $e_{i(jk)}$ which provides a basis for representing a part of a cognitive model. Then, a set of evidence variables can be defined.

On the basis of the assessment rubric, evidence rules are developed. They represent all fundamental features of this assessment rubric and develop the terms and sub-terms of the assessment rubric into "fine grain" components

which serve as evidence variables. Assessment rubric categories are explanatory variables which provide a basis for developing an elaborate assessment model. The feasibility of applying Bayesian networks was considered and a Bayesian network model, assessment construct, was used to assess an ANOVA score model.

5.5 Fundamental Features of Bayesian Networks

Once the Bayesian assessment network was built, fundamental features were examined using simulated data. Some clique patterns were examined in terms of Bayesian network structures. It probably assumed that prior probabilities were all binomial in consideration of acquiring simplicity and clearness of student responses to the items. Prior probabilities of parents were tested at different levels in mastery states (such as p at 0.50, 0.67 and 0.75) which indicated the probability of correctly answering a question. Prior probability is dispensable for examining student responses on the basis of Bayesian network models. Elaborating a Bayesian network requires a huge sample size of evidence to update it. A more reasonable measure of prior probability is needed based on experimental and practical experiences.

Iteration tests were conducted based on prior parent node levels in Bayesian networks. A probability value was pre-defined as 0.95. Combinations of simulated response patterns were shown. In the final run, a last updated probability was produced. If the value was less than 0.95, the parent prior probability would be updated for the next run. Iteration was ceased when the

posterior probability attained was equal to or larger than 0.95. During iterations, the updating processes were observed and some characteristics were examined.

Types of Bayesian net cliques such as simple and complex were summarized. Some of them were analyzed in consideration of assessment models. Clique patterns included one parent—one child Bayesian net clique, one parent-multi-child Bayesian net cliques, and mixed (multi level and multi component) Bayesian net cliques. The outcomes of these cliques were applied to assessment models.

5.6 Examination of Bayesian Assessment Models on the Basis of the Features of Bayesian Net Cliques

Updating trajectories of posterior probabilities in Bayesian assessment models were examined based on an examination of Bayesian net cliques. In one assessment model, the evidence variable space was estimated. In each class of the evidence model space, one case of the entire combination was sampled and posterior probabilities were carried out. All pattern classes were observed. Thus, for a given explanatory variable, a continuous change pattern was observed. The robustness and mastery level of the Bayesian assessment model were observed. Internal relations of selected explanatory variables were examined.

Finally, 20 student performances on the ANOVA score model were tested with the established assessment models. Diagnostic assessment information was reported in posterior probabilities and linear transformation of these probabilities to represent extent of student mastery.

5.7 Methodological Framework of Model-Based Assessment as a Summary Framework

Figure 5.8 summarizes the method of model development. It is divided into two parts by a dotted line. In the tutor development phase, Tutor Knowledge Data Base was a fundamental basis for theoretical and technical aspects of the tutorial system. It constitutes full content and structure information. Tutor tasks are introduced in Tutor Tasks and Questions so learners can go through the tutor help and learning environment. Questions designed in the tutorial system ask learners to respond to cognitive tasks. The Questions connect tutor designers and instructors, and learners have consistent expectation of what to complete in problem solving. Solution Features are a group of key points and principles to complete cognitive tasks. For example, items 1, 3, and 4 of Solution Features of Task 5 are: the score model is a linear equation; right of the equal sign is a sum of terms; and the first term is the symbol μ for the population grand mean respectively. They outline key points and clues for completing the cognitive tasks.

Tutor development phase as a submerged platform provides robust theory and database basis which can be transferred into parameters of expert knowledge models to guide assessment construct and assessment model.

The model-based assessment phase is below the dotted line in Figure 5.8. Assessment Tasks and Questions are the tasks and questions developed in the Stand-alone Performance Assessment Test. Based on assessment purposes,

these questions are related to those in the tutorial system, but they have been refined.

Scoring rubrics and evidence rules for performance assessment are two very close versions of documents. Scoring rubrics depict the categories of potentially observable components. Evidence rules are more the operational version used to decide what knowledge point has been applied or what is still deficient in student response processes and results. In the aid of evaluation and evidence rules, observable variable are dynamically instantiated based on data samples.

Assessment constructs were built based on expert knowledge models and tutor knowledge databases. Assessment models (probability networks) were used to test and validate assessment constructs.

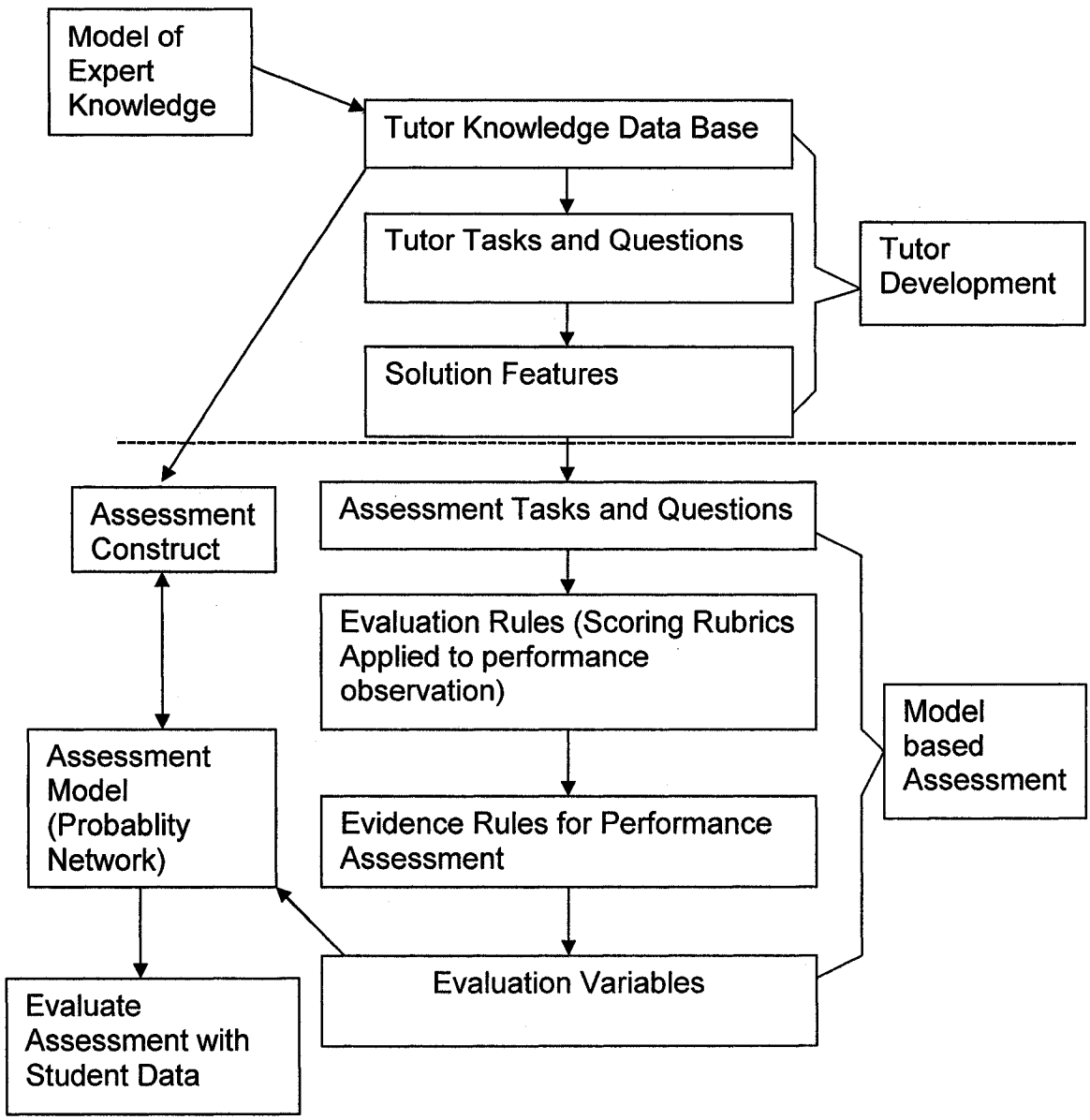

Figure 5.8. Methodological framework of model-based assessment

CHAPTER SIX: CONSTRUCTION AND EVALUATION OF THE ASSESSMENT MODEL

The structure of the model-based assessment in this study of statistics learning is task-based. The assessment system is designed to focus on learning tasks and assessment models are developed on the basis of these tasks. "PASL: ANOVA" is a stand-alone test (Appendix A) which is based on learning tasks used in the ANOVA tutor. "PASL: ANOVA" consists of 13 learning tasks and the structure simulates the ANOVA tutorial system. Therefore, the tasks in both the ANOVA tutoring system and "PASL: ANOVA" have the same structure and content. This study focuses on the development of an assessment model of one of these 13 learning tasks. The procedure for developing such an assessment model has been designed so as to be generalized to the other 12 ANOVA tasks (and other tasks as well). Task 5, the ANOVA score model task, was chosen as the content domain for developing an assessment model and procedure because it is a familiar topic to students in intermediate and advanced statistics course.

This chapter describes how the assessment model was developed. This will involve: (a) building an assessment structure and model, (b) developing evidence variables and a probability model connecting evidence variables to explanatory variables representing component knowledge and skills, and (c) establishing some basic characteristics of the network and applying the model to simulated data to examine its behaviour.

6.1 Establishment of an Assessment Structure and Model

An assessment structure is a framework for representing an arrangement of knowledge components in a hierarchical network. The network structure will define links among potential explanatory variables (constructs), which will be used to evaluate and interpret student mastery of these learning objects (constructs) and changes in student mastery over the course of learning. These knowledge and skill components cannot be observed directly and it may be necessary to decompose them into more fine-grained components for given assessment purposes. They must also be linked to evidence variables derived from observations of student performance.

Normally, an assessment structure can be established through a semantic analysis of the content of verbal problem-solving or tutoring protocols, combined with a cognitive task analysis of the problems solving. Assessment purposes and the desired "grain" of analysis influence the precision of the analysis carried out to build procedural and semantic models of required problem-solving knowledge. For the ANOVA score model, (a) writing the ANOVA score model components, and (b) explaining what these components refer to semantically, constitute two different aspects of knowledge and skill. Hence, the assessment structure consists of two submodels: (1) writing an ANOVA score model (a procedural model) and (2) explaining the ANOVA score model (a semantic model). Whereas 1 is about a procedure and 2 is about the meaning of the ANOVA score model. An ANOVA score model can be decomposed into several component procedures which become the basis for assessment model development:

$$Y_{i(jk)} = \mu + \alpha_j + \beta_k + \gamma_{jk} + e_{i(jk)}. \qquad \text{(Formula 6-1)}$$

$Y_{i(jk)}$ is the score of individual i in group jk. To the right of the equal sign, there are five components: Grand mean μ, the Main effect for group α_j, the Main effect for duration β_k, the Interactive effect γ_{jk}, and Residual score $e_{i(jk)}$. These terms are defined differently (see Table 6.1) to distinguish them in the procedural and semantic phases. The terms are listed for both phases in terms of rubrics in the columns.

Table 6.1. Examples of Scoring Categories for both Process and Semantic Aspects of the ANOVA Score Model

Category	Procedural phase	Semantic phase
Score	Score $Y_{ij(k)}$	Score of individual i in group jk
Grand mean	mu	Population GrandMean of scores
Main effect for group	alpha $_j$	MainEffect:Level j of factor A
Main effect for duration	beta $_k$	MainEffect:Level k of factor B
Interaction effect	$Gamma_{(jk)}$	Interaction of level j of A x level k of B
Residual score	$e_{i(jk)}$	Error in an individual score

Figure 6.1 presents a hierarchical frame representing the knowledge required to complete an ANOVA score model (for a two-way classification) and will be referred to as the "ANOVA Score Model (2way)" Frame.

Term components are arranged hierarchically in Figure 6.1. Procedural phase parameter ("ModelEquation" submodel) are used to represent knowledge of how to write the Grand mean μ, the Main effect for group α_j, the Main effect for duration β_k, the Interaction effect γ_{jk}, and the Residual score $e_{i(jk)}$. Semantic phase parameters ("Score Model" submodel) are used to represent knowledge of how to explain: "Score", "GrandMean", "MainEffectLevelofA", "MainEffectLevelofB", "InteractionAXB", and "Error". In the semantic explanation phase, the node "EffectOfFactors" refers to knowledge of effects. "GrandMean" refers to knowledge of the grand mean, and "Error" refers to the error in a person's score. "ScoreDecomposition" refers to knowledge of how an individual's score has been decomposed.

The 2-way "ANOVA score model" represents the fundamental constructs corresponding to components of student knowledge and skills required to complete and explain an ANOVA score model. These components can be diagnostically assessed. In a cognitive perspective, these network components are defined as different knowledge and skill components. The goal is to build an assessment network that reflects knowledge and skills revealed by a cognitive analysis of expert knowledge representations.

The assessment construct (Figure 6.1) provides a basic theoretical assessment network for explaining and tracking the development of knowledge and skills. Network nodes correspond to explanatory variables in Bayesian networks. Components such as "IndexValues" were added to ensure the completeness of the model to be tested in a Bayesian network. In addition, such

components as "Score Y_{ijk}" are further decomposed to include an explanatory variable "ijk" representing knowledge of how to write and apply the index to the score variable (see Figure 6.2).

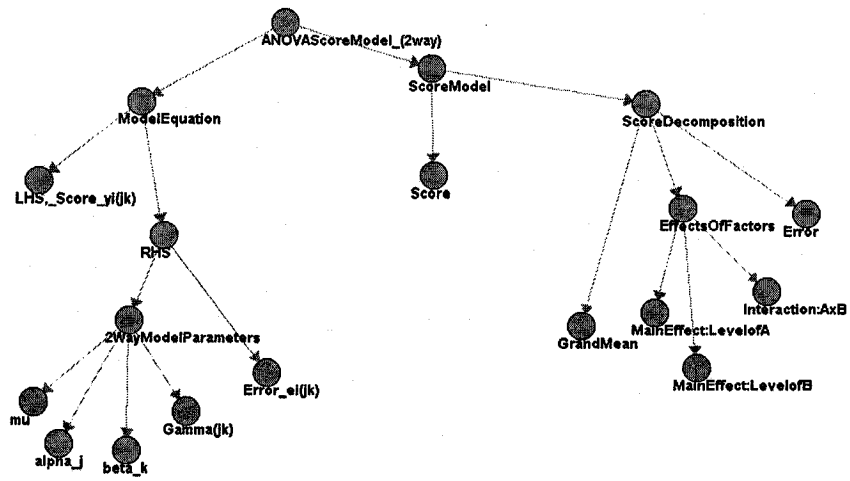

Figure 6.1. Basic assessment constructs of the ANOVA score model

Nodes in Figure 6.1 are potential explanatory variables in the Bayesian network used in assessment. These nodes reflect components of student problem solving knowledge and skills. However, in building the assessment construct, it is unclear when current network nodes are potential explanatory variables. A node found to be unlinked to any supporting evidence cannot be a potential explanatory variable. To function as an explanatory variable, nodes must be linked to evidence nodes. Evidence nodes are variables linked directly to performance by means of scoring rules.

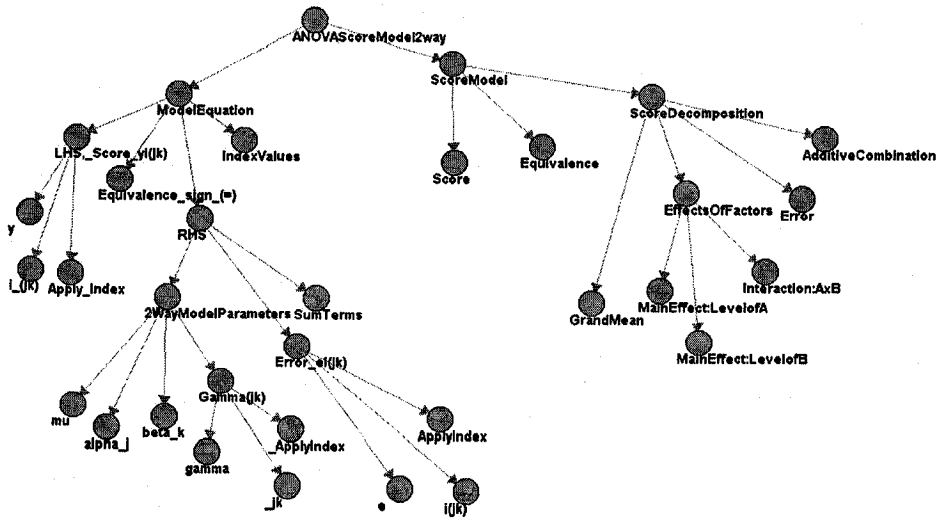

Figure 6.2. Assessment expansion construct of the ANOVA score model

6.2 Development of the Evidence Variables and Probability Model

Evidence variables require specification of scoring rules that extract values of observable variables from individual performance in relevant task environments. In this study, two sources were used to develop evaluation and evidence variables: (a) assessment tasks and questions, and (b) solution features. Both sources are needed to define evidence variables and to establish their values based on student performance.

6.2.1 Assessment Tasks and Questions

Assessment questions are developed in the assessment task framework, which roughly mimics tutorial learning system tasks. Assessment tasks and questions were developed into a test booklet, "Performance Assessment of

Statistics Learning" (PASL, see Appendix A). The structure of the tasks was described in section 5.2.1.

In Task 5, examinees have six sub-tasks:

(5.1) Write the ANOVA score model for your design.

(5.2) Explain the formula used in the ANOVA score model.

(5.3) What are the main effects? How are they interpreted?

(5.4) What are the interaction effects? How are they interpreted?

(5.5) What is the grand mean? How is it interpreted?

(5.6) Identify the residual or error score. How is it obtained from subject's observed score on the dependent variable?

Assessment questions and solution features of Task 5 have the same cognitive construct in the assessment model. As a set of leading statements, these sub-tasks appeal to student cognitive processes, and to products of knowledge and skill development. Sub-tasks 1-6 represent cognitive components of the Solution Feature and the Score Rubric. Sub-tasks 1-6 attempt to lead students to an understanding of the ANOVA score model.

Sub-task 5.1, "Write the ANOVA score model for your design," asks students to symbolize variables and parameters of the ANOVA score model. Students must use these terms and their relations to describe the ANOVA score model. Sub-task 5.1 asks for the ANOVA score model to be written in both a general and a specific way. For example, $X_{(\ )}$ can be written as X_{ijk} or as X_{223}, with different indexing.

Subtask 5.2 asks for a single explanation of $X_{(\ )}$ and asks students to decompose the score model terms expressed in semantic explanation competence. For $X_{(\ )}$, students must describe the relations expressed in X indexing of ijk or in 223.

Sub-task 5.3 and 5.4 involve an analysis of relations between terms. In two-way ANOVA, main effect: group, main effect: duration and their interactions are represented. Students must know how to express relations between α_j, β_k, and γ_{jk}, between α, β, γ and their indexing j, k and jk, and how to describe relations between α_j, β_k, and γ_{jk}.

Sub-task 5.5 asks students to define and interpret the grand mean which involves specifying the relation between the group mean and the grand mean.

Sub-task 5.6 asks students to describe $e_{i\,(jk)}$ which requires them to realize that the residual score has the same indexing as the score variable $Y_{i\,(jk)}$, and relations between $e_{i\,(jk)}$ and others variables.

Questions for ANOVA score model may vary. However, a main rule in designing sub-task 5.1-5.6 was for students to demonstrate their ability to acquire knowledge of the ANOVA score model.

6.2.2 The Relations of Solution Features and Score Rubrics

Most tasks respond to components of Solution Features list. Solution Features form sets of solution characteristics applied to individual task performances. They specify response steps and provide a comprehensive basis for responding to tasks. They facilitate students to answer questions correctly

and they help assessors work out score rubrics for judging student task performances.

When task score rubrics are represented, Solution Features help explain relations among model components based on them.

Section 6.2.2 emphasizes relations between task score rubrics and Solution Features. Developmental details will be introduced in 6.2.3. The complete task score rubric for Task 5 (Appendix C) is composed of seven items: (a) Score, (b) Grand Mean, (c) Main Effect (Group), (d) Main Effect (Duration), (e) Interactive effect, (f) Residual score, and (g) Complete model equation. Solution Features can represent the detail needed to elaborate more fine-grained rubric components. They can cover components across score rubric items. One Solution Feature may be involved in several score rubrics.

Table 6.2 shows relations between score rubric and solution features. There is no one-to-one relation between the score rubric and solution features. In fact they cross each other. One rubric rule may correspond to more than one Solution Feature description, and a Solution Feature may be linked to more than one rubric. Table 6.2 is a practical illustration of list items between rubric and Solution Features.

Table 6.2 presents Solution Features which are a relatively flexible way to provide assessment information. Score rubrics consist of rules which may be used to develop cognitive constructs and evidence variables.

Table 6.2. The Corresponding Relations between Solution Feature and Score Rubrics

Score rubric items	Solution features *
(a). Score:	
(1)Y a symbol for score variable "attitude toward minority".	1. The equation begins with a variable label representing the score on the dependent variable with appropriate subscripts to indicate the specific levels of factors used to classify a subject, and an index number for the subject.
(2) Index for score variable, ijk or i(jk)	Identical to solution feature 1.
(3) Complete expression $Y_{i\,(jk)}$ for score of individual i in group j and duration k	Identical to solution feature 1.
(b). Grand mean:	
(4) μ symbol for population mean. Parameter pooled in j and k	4. The first term is a symbol (Greek mu) for the population grand mean.
(c). Main effect (Group):	
(5) α symbol for main effect parameter	6. The next terms are symbols representing population main effect parameters for additional factors. Each effect parameter is indexed according to the level of this factor (as it was indexed on the score variable).

Table 6.2. (continued)

Score rubric items	Solution features *
(6) j index for group (j=1, 2, 3)	Identical to solution feature 6.
(7) α_j main effect for group j	Identical to solution feature 6.
(d). Main effect for duration:	
(8) β for main effect parameter	Identical to solution feature 6.
(9) k index for duration	Identical to solution feature 6.
(10) β_k main effect of duration k	Identical to solution feature 6.
(e). Interactive effect:	
(11) γ for interactive effect	7. If you have a crossed design with two or more factors, terms representing the population two-way interaction effects for pair wise combinations of factors are included in the model. Each effect parameter is indexed according to the levels of these factors (as they were indexed on the score variable).
(12) (jk) for index of cell in design	Identical to solution feature 7.
(13) $\gamma_{(jk)}$ for interactive effect of individual i in cell (j, k)	Identical to solution feature 7.
(f). Residual score:	
(14) e= residual score	8. The last term is a symbol (variable name) for the error (i. e., residual) score having the same indexing as the score variable.
(15) i(jk) same index as score variable	Identical to solution feature 8

Table 6.2. (continued)

Score rubric items	Solution features *
(g). Complete model equation:	
(16) $e_{i(jk)}$ error score in group j, duration k and subject i	Identical to solution feature 8.
(17) $Y_{i\,(jk)} = \mu + \alpha_j + \beta_k + \gamma_{jk} + e_{i\,(jk)}$	1. The score model is a linear equation. 3. Right of the equal sign is a sum of terms. 26. By decomposing the participants' score into a grand mean, effect components, and a residual score, we can systematically investigate the additive effect of each component as a contribution to the subjects' scores.

* The item numbers of the solution feature column indicate the one in Appendix B: Solution Feature of Task Five

6.2.3 Development of Evaluation Rules for the Performance Assessment

Evaluation Rules, also referred to as Scoring Rubrics (for Task 5), were developed based on Solution Features and assessment task questions. The Scoring Rubric provides basic criteria for connections between explanatory and task variables and for defining ANOVA score model symbols.

Task 5 Scoring Rubrics consist of the 7 components of its model equation: score component, grand mean, main effect for group, main effect for duration, interaction effect, residual score, and complete model equation (see Appendix C).

Three features will be demonstrated for the score component: (a) to define "Y is a symbol for score variable"; (b) to clarify the index of score variable, ijk; and (c) to write the complete expression $Y_{i\,(jk)}$ for the score of individual i in group j, corresponding to level j of factor A and duration k corresponding to level k of factor B.

Component 2, the grand mean, the symbol for the population mean μ must be written and defined.

Component 3, main effect for group, the symbol alpha must be written; the j index for group must be written; and the expression alpha j written to represent the main effect for each level j of the group variable.

Component 4, the main effect for duration, the beta symbol must be defined; the k index for duration must be written; and the expression beta k must be written to represent the main effect for duration k.

Component 5, the interaction effect, the gamma symbol must be written; the index jk must be written for the combination of level j of group and level k of duration; and the expression gamma jk must be written to represent the interaction effect for group j and duration k.

Component 6, the residual score, a symbol for the error must be written. The index ijk for the score variable must be written for individual i in group j and after duration k; and the expression $e_{i(jk)}$ for the error score of individual i in group j and duration k, must be written to represent the residual score.

Component 7 is writing the complete model equation. Even though learners may know knowledge component 1-6, these must be presented in an appropriate equation using the correct symbols.

Scoring rubrics include seventeen items. Process and semantic explanations are two aspects of the score rubrics. Thus, there are thirty-four items ("Solution Features") in all. Written productions of symbols and expressions demonstrate student performance component in completing Task 5—writing an ANOVA score model. Semantic Explanation scoring components reflect student understanding of why they write the model as they do, and how they assign meanings to score model components. Scoring Rubrics depict relatively fine-grained patterns for learner performances and semantic explanations. Diagnosing student performance and knowledge requires the application of an assessment model that can account for patterns of the thirty-four student score components. Thus, student score components will be linked to evidence variables, which are represented as observable evidence nodes in the assessment network. Evidence nodes are linked to bottom level explanatory constructs of the assessment model in Figure 6.1.

6.2.4 From Rubrics to Diagnostic Assessment: Evidence Rules

Evidence rules are sets of highly decomposed knowledge and skill components that apply to both performance and semantic phases for Task 5. Evidence rules are used to determine whether students have mastered the knowledge and skills necessary for solving ANOVA score model problems. Evidence rules for the ANOVA Score Model contain 45 items categorized into the performance process and semantic comprehension phase. The performance phase focuses on knowledge and skill components used in the performance process. The semantic phase focuses on components required for understanding concepts related to corresponding processes. In Y "(1), Y," is an indexed variable. Students can write Y_{ijk} on the left side of the equal sign in the ANOVA score model equation. They will be assessed on having completely mastered the performance process. In item "(27) 18 Dep. Var," the student must know the definition of Y which refers to the observed dependent variable "attitude towards minorities." Tables 6.3 and 6.4 provide details of these knowledge and skill components. As observable variables, these components are decomposed in a Bayesian network and then associated with rules for judging whether they have mastered the knowledge and skills needed to solve an ANOVA score model problem.

Table 6.3. Evidence Rules for Scoring Procedural Components

Evidence node (scoring item)		Evidence node (name)	Solution feature description	Content category
Expression in ANOVA model equation	Item number			
Y	(1)	Y	$Y_{(indices)}$ (indexed variable)	Variable
i(jk)	(2) (3) (4)	2a_i, 2b_j, 2c_k	i (part of index) j (part of index) k (part of index)	Index
$Y_{i(jk)}$	(5)	3_ApplyIndex	Apply indices to Y	Apply Index
=	(6)	4_=	equal sign	Equivalence
μ	(7)	5_μ	μ (parameter)	Parameter
α	(8)	6a_alpha	α_{index} (indexed parameter)	Parameter
j	(9)	6b_j	j (index)	Index
α_j	(10)	7_applyindex	apply j to α	Index
β	(11)	8a_beta	β_{index} (indexed parameter)	Parameter
k	(12)	8b_k	k (index)	Index
β_k	(13)	9_ApplyIndex	apply k to β	Index

Table 6.3. (Continued)

Evidence node (scoring item)		Evidence node (name)	Solution feature cescription	Content category
Expression in ANOVA model equation	Item number			
γ	(14)	10_gamma	$\gamma_{(index)}$ (indexed parameter)	Parameter
jk	(15) (16)	11a_j, 11b_k	j (part of index) k (part of index)	Index
γ_{jk}	(17)	12_applyIndex	Apply jk to gamma	Index
e	(18)	13_e	e_{index} (variable)	Variable
i(jk)	(19) (20) (21)	14a_i (error), 14b_j (error), 14c_k (error)	i (part of index) j (part of index) k (part of index)	Index
$e_{i(jk)}$	(22)	15_ApplyIndex	apply i(jk) to error term	Index
Sum of score com-ponents	(23)	16_sum terms	Apply appropriate sign +	Function (sum)
Index ranges	(24) (25) (26)	17a_i=1...,n 17b_j=1...,J 17c_k=1...,K	i index value i= 1, 2,3...,n; n=5 j index value j=1, 2, 3 k index value k=1, 2, 3	Index

Table 6.4. Evidence Rules for Scoring Semantic Components

Evidence node (scoring item)		Categorical phases	Items	Content category
Description	Item Number			
Dependent variable	(27)	18_Dep. Var.	Y refers to the observed dependent variable "attitude towards minorities" as measured by a scale.	Variables
Case ID	(28)	19_Case ID	Index i refers to an integer/numerical index where i refers to an individual examinee	Case Indices
Levels of group	(29)	20_Level (j) of A	j refers to level j of the independent variable.	Levels of effects
Levels of duration	(30)	21_Level (k) of B	Group k refers to level k of the independent variable duration.	Levels of effects
Interaction of group with duration	(31)	22_Grp. (jk) in 2Way table	jk refers to the cross-classification cell of the table of subjects (police officers) where each officer is classified into a category (cell), one category for each combination of group j (one of 3 areas patrolled) and duration of program k (one of 3 durations)	Levels of effects
Equivqlence of score and score components	(32)	23_Equiva-lence	Equivalence means that the expression (sum of term) reflecting the decomposition of the score on the right side of the equal sign is equivalent to the individual's score on the dependent variable (the left side of the equal sign)	Equivalence relation

Table 6.4. (Continued)

Evidence node (scoring item)		Categorical phases	Items	Content category
Description	Item number			
GrandMean-pooled	(33)	24_GMRef	Grand mean refers to the pooled mean of all scores (pooling over groups and durations) in the population	Grand mean
GrandMean-Avg of GrpMeans	(34)	25_Mean of Grp. Means	The grand mean is the average of the group means.	Grand mean
Main effect of group	(35)	26_Main Effect A (j)	α_j "refers to" the main effect of the independent variable group on the dependent variable, independent of group in the population.	Effects and their expression
Main effect of Grp (definition)	(36)	27_Grp. Mean(j)-GM	$\alpha_j = \mu_j - \mu$. This term means main effect A can be written as a difference between the mean of group μ_j (pooling over duration) and the grand mean μ in the population.	Effects and their expression
Main effect of duration	(37)	28_Main Effect B(k)	β_k refers to the main effect of the independent variable duration on the dependent variable.	Effects and their expression
Main effect of duration (definition)	(38)	29_GrpMean (k)-GM	$\beta_k = \mu_k - \mu$. Main effect B refers to the difference between the mean of group k (μ_k) (pooling over group) and the grand mean (μ) in the population	Effects and their expression

Table 6.4. (continued)

Evidence node (scoring item)		Categorical phases	Items	Content category
Description	Item n number			
Interaction effect	(39)	30_Interac- tion effect AB(jk)	$\gamma_{(jk)}$ refers to the interaction effect of the combination of group j and duration k on the subject's score—that is, a value of the dependent variable	Effects and their expression
Interaction effect (definition 1)	(40)	31_GrpM(jk) -M(j)- M(k)+GM	$\gamma_{(jk)} = \mu_{jk} - \mu_j - \mu_k + \mu$. The interaction effect $\gamma_{(jk)}$ is the mean of the combination of group j with duration k minus the pooled (marginal) mean of group j and the pooled (marginal) mean of duration k plus the grand mean (population values).	Effects and their expression
Interaction effect (definition 2)	(41)	32_GrpM (jk)-Eff(j)-Eff(k)-GM	$\gamma_{(jk)} = \mu_{jk} - \alpha_j - \beta_k + \mu$ The interaction effect may also be written as the mean of the combination of group j with duration k (cell mean) minus the main effect of group (α_j) minus the main effect of duration (β_k) minus the grand mean.	Effects and their expression

Table 6.4. (continued)

Evidence node (scoring item)		Categorical phases	Items	Content category
Description	Item number			
Residual variable	(42)	33_Residual var	$e_{i(jk)} = Y_{i(jk)} - \mu - \alpha_j - \beta_k - \gamma_{jk}$ The error term is a variable that refers to the residual portion (part) of a subject i's, score on Y after all of the effects and the grand mean have been subtracted out.	Variables
Residual score (def1)	(43)	34_Score (ijk)-GrpM(jk)	$e_{i(jk)} = Y_{i(jk)} - \mu_{jk}$ The error term is the difference between a subject's score on Y and the subject's cell mean.	Relation of errors with other terms
Residual score (def2)	(44)	35_Score (ijk)-GM-Effects	$e_{i(jk)} = (Y_{i(jk)} - \mu) - (\alpha_j + \beta_k + (\alpha\beta)_{jk})$. The error score can be interpreted as that portion of a subject's observed score on the dependent variable (expressed as a deviation from the general mean), which is not predictable from the effects of the individual's particular combination of group and duration.	Relation of errors with other terms
Additive components	(45)	36_Additive Combination	$(\mu + \alpha_j + \beta_k + \gamma_{(jk)} + e_{i(jk)})$ The score decomposition consists of a sum of five components: main effect of group, main effect of duration, interaction of group and duration, and error.	Additive components

Evidence rules are classified in terms of knowledge contents in procedural and semantic scoring components. Procedural scoring components have 6 knowledge types: variable, index, apply index, parameter equivalence, and function (sum) (see Table 6.5). The dependent and error variables are both random variables. Indices are associated with scores (variables and parameters), and define specific observations of variables or particular parameter levels. "Apply Index" refers to student ability to apply indices in appropriate positions when referring to variables or parameters. Parameters are constants whose values are estimated as specific population properties. These values determine characteristics of model equations. The summation function defines the sum of the model parameter and error variables. The equivalence relation equates the scores on dependent variables with scores on model terms. All six classifications are used in examining student performance. Although detailed explanation is unnecessary, student must apply all procedures completely.

There are 8 classifications in the semantic scoring components: variables, case index, levels of effects, effects and their expressions, equivalence, grand means, relations of errors with other terms, and additive components (see Table 6.6).

Content classification "variables" reflect student understanding of the dependent and error variables. Students are required to explain what they refer to. The "case" index shows that students know that an individual can be identified in relation to a combination of group and duration and how these are reflected in subscript positions of the dependent and error variables. Student explanations of

"levels of effects" demonstrate their understanding of duration levels of independent variables in cross-classifying observations by group and duration factors.

There are three levels of groups, three levels of durations, and nine cross-classifications of j by k. Student ability to explain "Effects and their expressions" reflects their understanding of effects and their relations to other parameters. For example, main effect A tells what α_j refers to; and "$\alpha_j = \mu_j - \mu$" indicates that "main effect A" is the difference between the group mean and grand mean. Correct explanation of "Equivalence" reflects student understanding of the right part of the model equation interpretation, and decomposition of the score given in the left part of the equation. Successful explanation of the grand mean reveals two ways of understanding the grand mean as a pooled (marginal) mean and as an average of the group means. Explaining the "Relation of Errors with other terms" requires students to know how the error score is related to other ANOVA model components.

Table 6.5. The Content Classification of Procedural Scoring Components

Term	The order number in table 6.3
Variable	(1), (18)
Index	(2), (3), (4), (9), (12), (15), (16), (19), (20), (21), (24), (25), (26)
Apply Index	(5), (10), (13), (17), (22)
Parameter and indexed parameter	(7), (8), (11), (14)
Equal sign	(6)
Function (sum)	(23)

The relationship between errors and other terms demonstrate student understanding of the error variable which can be written in alternative term combinations.

$$e_{i(jk)} = Y_{i(jk)} - \mu_{jk} \qquad \text{(Formula 6-2)}$$

and $$e_{i(jk)} = Y_{i(jk)} - \mu - \alpha_j - \beta_k - (\alpha\beta)_{jk} . \qquad \text{(Formula 6-3)}$$

Although, the two errors are mathematically equivalent, they provide different explanations of error. The first characterizes errors as deviations of observed scores from group means and the second characterizes errors as residuals after all ANOVA score model parameters have been subtracted from the score. Table 6.5 and 6.6 summarize classifications of procedural and semantic scoring components with respect to knowledge contents.

Table 6.6. The Content Classification of Semantic Scoring Components

Terms	The order number in table 6.4
Variables	(27), (42)
Case index	(28)
Levels of effects	(29), (30), (31)
Effects and their expressions	(35), (36), (37), (38), (39), (40), (41)
Equivalence relation	(32)
Grand mean	(33), (34)
Relations of error with other terms	(43), (44)
Addition components	(45)

6.2.5 Defining Evaluation Variables

Evaluation variables can be observed directly or they can be decomposed of observed variables. Observed data and evidence are transferred to potential explanatory variables and assessment constructs through evaluation variables. The decomposability of evaluation variables is relative. If a variable can be further decomposed, then it can become a potential explanatory variable; if the variable does not require further decomposition, it is an observable variable.

In assessment models evidence variables can transfer evidence to potential explanatory variables. If a potential explanatory variable has no child node and is also observable, the assessment construct variable can be redefined as an evaluation variable. For example, if the variable "mu" in an Assessment Construct (Figure 6.1) can be decomposed into child nodes, then it can become a potential explanatory variable in assessment construct variables. In the ANOVA Score Model, "mu" is not a potential explanatory variable and is viewed as an evaluation variable.

After assembling the 45 assessment model evaluation variables, it is easy to determine whether they are evidence variables for transferring data from observations to potential explanatory variables. Tables 6.3 and 6.4 identify evaluation variables.

6.2.6 Generation of an Assessment Model: the Probability Network

The assessment model is composed of an assessment construct (network of explanatory variables) and a set of evidence variables which together define

an assessment model. Cognitively, assessment construct evidence variables (see Figure 6.2) cover performance and explanation. In the performance phase, students demonstrate their knowledge and skills in developing an ANOVA score model for a set of data. In the explanation phase, students give each component a semantic explanation. The assessment construct maintains 22 of 33 nodes (Figure 6.2 and Table 6.7) as explanatory variables.

Table 6.7. The Intermediate Explanatory Variables in the Networks

Object	Nodes in procedural network	Object	Nodes in semantic network
1	alph_j	13	Score
2	beta_k	14	GrandMean
3	12_jk	15	MainEffect:LevelofA
4	15_i(jk)	16	MainEffect:LevelofA
5	Gamma(jk)	17	Interaction:AXB
6	Error_ei(jk)	18	Error
7	2_i(jk)	19	EffectsOfFactors
8	2WayModelParameters	20	ScoreDecomposition
9	LHS	21	ScoreModel
10	RHS	22	ANOVAScoreModel2way
11	IndexValues		
12	ModelEquation		

The remaining 11 nodes are classified as evidence variables in the evidence model. Once explanatory and evidence variables have been determined, the organization and structure of the entire assessment network must be determined in order to transfer information from observations to potential explanatory variables. This structure is discussed in section 6.2.6.1.

6.2.6.1 Definitions of the Assessment Network

The assessment model requires a network structure to describe relationships linking evidence and explanatory variables. Network models serve to estimate knowledge and skills underlying the ability of students to solve ANOVA score model problems by inferring components of performance or of semantic explanation. Explanatory variables that underlie student abilities must be inferred. The hypothesis is that all explanatory variables are mutually exclusive, though in practice, there may be some correlation among examinees' estimated values for these variables. To represent causal relations among variables, a hierarchical model has been adopted in which top-down and bottom-up relations can be examined. Student abilities to solve ANOVA score models can be inferred from observations, sets of limited evidence variables, rather than from observations of all network variables. In so doing, some information will be lost; however, certain confounding relationships will be discarded. Such an approach will result in efficient causal explanations based on observations of group evidence variables to estimate values of explanatory variables.

Hierarchical models organize data into tree-like structures to limit the number of network relationships. Hierarchical models define a kind of parent-child relationship where each parent has at least two children. Conversely, this model restricts each child node to exactly one parent.

Hierarchical models permit assessment variables to use parent-child relationships repeatedly. The network assessment model can be trained on the collected data to provide robust diagnostic learner assessments.

6.2.6.2 Probabilistic Spaces of Potential and Evidence Variables in a Bayesian Network

Variables in Bayesian networks applied in educational assessment stand for knowledge and skills. In Bayesian inference, variables can be categorized into potential explanatory variables and evidence variables. Variable spaces in assessment network models can be defined as binary or multi-category values based on assessment purposes and assumptions. A binary variable space was selected to simplify the problem. For all networks variables, the variable space has been defined as two states called "mastery" where students have mastered an ANOVA score model component and "non-mastery" where students have failed to master an ANOVA score model component. The state of a child node depends on the state of its parent. For the top-level "parent" node, there will be a prior probability that the student has mastered the ANOVA model skill, and a probability of 1.0 minus the prior probability of mastery that the student has not mastered the skill. For each child node, the probability of mastery of the node depends on the state of its parent node. This dependency is represented by four conditional probabilities which can be arranged in a 2x2 table.

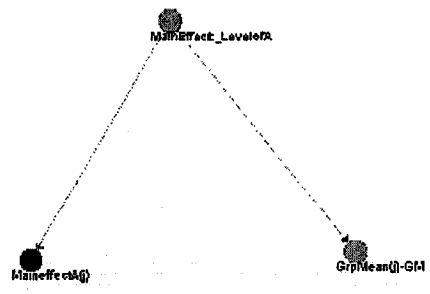

Figure 6.3. Main effect of A with two evidence variables

Figure 6.3 shows a parent node with two child nodes. The parent node, labelled "MainEffect:LevelA," and its two child nodes labelled "MainEffectA(j)" and "GrpMean(j)-GM" respectively. The prior probability for "MainEffect:LevelA" is defined as:

P(MainEffect: LevelA)= p_t if the variable instantiates as true (i.e. mastery)

P(MainEffect: LevelA)= p_f if the variable instantiates as false (i.e. non-mastery).

The prior probability of "MainEffectA(j)" is conditional on the probability of its parent node, "MainEffect:LevelA" which has two values, true (mastery) and false (non-mastery). "MainEffect A(j)" has two states for each condition listed above. Therefore, there are four conditional probabilities for each child node. The conditional probabilities for "MainEffectA(j)" are

P ("MainEffectA(j)"=true | "MainEffect: LevelA"=true)= p_{tt}

P ("MainEffectA(j)"=false | "MainEffect: LevelA"=true)= p_{tf}

P ("MainEffectA(j)"=true | "MainEffect: LevelA"=false)= p_{ft}

P ("MainEffectA(j)"=false | "MainEffect: LevelA"=false_= p_{ff}

The prior probabilities of "GrpMean(j)-GM" have the same structure. The conditional probability tables for the ANOVA score model Bayesian network will contain the same conditional probabilities.

6.2.6.3 Cliques and Levels in the Hierarchical Assessment Model

According to Xiang (2002), "a maximal set of nodes that is complete is called a clique; a clique is a maximal set of variables without graphically identifiable conditional independence" (p. 72). A clique in a hierarchical model is

an open walk (or path) with alternating sequence of vertices and edges whose first and last vertices are different. In a hierarchical network applied to an assessment model, a "sense-making clique" consists of at least one potential explanatory variable and its children. Therefore, the top level node of the clique should be a potentially explanatory variable. In addition, there should be at least one evidence variable in a clique at the bottom level of a hierarchical assessment Bayesian network. A "sense-making clique" is an open walk which begins at any explanatory variable going down to its evidence variables at the bottom without any disconnections.

In a Bayesian network, the length of a walk is the number of edges used; similarly, clique levels can be defined by the number of nodes used. If there are several consecutive parent-child relationships, the hierarchical model is multi-level. The maximum number of an adjacent node pairs in a route comprises the levels of the model. In the assessment construct for the ANOVA score model (Figure 6.2), there are six nodes connected one to the other from the top node "ANOVAScoreModel2way" to the bottom node "jk." Therefore, this network model has six levels. If "jk" had one more child, it would have 7 levels. Clique position in different levels of nodes hierarchically determines how a node functions when it occurs in different positions. The closer the node is to the top, the more weight the function has. Therefore, it is important to know the relative positions of nodes.

6.2.6.4 Evidence Spaces Considering Evidence Node Instantiations

ANOVA score model evidence nodes indicate the number and distribution of student responses on knowledge and skill components. Perhaps, one student responds to some evidence nodes and another may respond to all nodes. Different student responses constitute the evidence space. From the perspective of mathematical combination, it is necessary to know the evidence space in order to estimate the probabilities of different responses.

Problem solving assessment variables in the ANOVA score model can be in two states: true or false. Assuming there are n evidence nodes, and a student responds to r nodes. The response evidence space is based on the mathematical combination:

$$C_n^r = \frac{n!}{r!(n-r)!}. \qquad \text{(Formula 6-4)}$$

If one evidence node is instantiated, the evidence space is $2^1 C_n^1$.

If two evidence nodes are instantiated, the evidence space is $2^2 C_n^2$.

If three evidence nodes are instantiated, the evidence space is $2^3 C_n^3$.

If r evidence nodes are instantiated, the evidence space is $2^r C_n^r$.

If n evidence nodes are instantiated, the evidence space is $2^n C_n^n$.

The summation of these terms constitutes the evidence variable space:

$$\sum_{r=1}^{n} 2^r C_n^r = 2^1 C_n^1 + 2^2 C_n^2 + 2^3 C_n^3 + \ldots + 2^k C_n^k + \ldots + 2^n C_n^n \qquad \text{(Formula 6-5)}$$

The evidence variable space is huge. If students respond to all knowledge and skill components, the evidence variable space is $2^n C_n^n$. If there are 5 binary evidence variables, the spaces will be 2^5 X 1= 32. Thus, there will be 32 possible variations of 5 variable state combinations.

6.2.6.5 Hierarchical Structure for Assessment of ANOVA Score Model

Knowing the variable space, the entire assessment network model has been defined (Figure 6.4). The model is designed to reflect performance processes and semantic understanding. In the performance phase, the top-level variable is node "ModelEquation"; in the semantic phase, the top-level variable is node "ScoreModel". Both nodes share a parent node "ANOVAScoreModel2way" which is at the pinnacle, and can be viewed as a potential explanatory variable that describes the comprehensive ability of students to solve an ANOVA score model problem.

The assessment network model contains 67 nodes (see Figure 6.4). In performance processes, there are 38 nodes. In semantic explanation, there are 28 nodes. Adding the top parent node, there are 67 nodes in all.

There are 45 evidence variable nodes in performance and semantic phases. There are 26 performance process evidence nodes and 19 semantic comprehension evidence nodes.

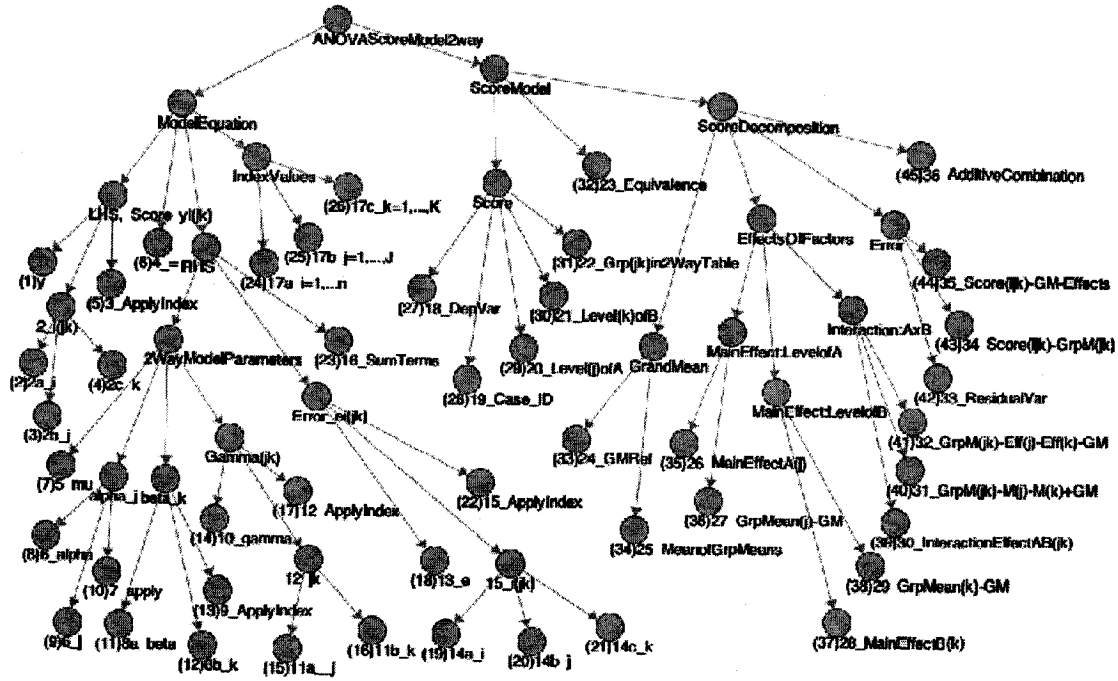

Figure 6.4. A Bayesian network for assessment of ANOVA score model

There are 22 "potential explanatory" variable nodes: 12 performance process phase nodes and 9 semantic comprehension phase nodes. Finally, there is one top level parent node.

6.3 Fundamental Structures and Characteristics of the Hierarchical Assessment Network: The ANOVA Score Model Assessment Network

Before applying the ANOVA score model assessment network to collected data, the fundamental features of the network will be examined. The entire network consists of cliques comprised of parent nodes and several child nodes. It is necessary to examine characteristics of the propagation of information through routes between a parent and children within cliques.

6.3.1 Definitions of Three Types of Network Cliques

Patterns can be found by observing cliques in the ANOVA score model Bayesian network. The concept of simple and complex net cliques can be defined. A simple net clique is a network structure segment that has only two levels: a simple parent node and one or more child nodes. If all child nodes are evidence variables, it is a "simple evidence net clique". A complex net clique is a network structure segment which has three levels. Third level nodes can have several explanatory nodes. Consequently, first level parent variables cannot be estimated simply from second level variables. Evidence must come from third level variables and must consist of instantiated evidence variables, so that their value pattern information can be propagated to the top parent variable. Once evidence variables are instantiated, the size of the problem space for the results will increase exponentially. Such net cliques are referred to as "complex net cliques." If the bottom level variables of complex net cliques consist of evidence nodes, the clique will be referred to as a "complex evidence net clique".

Cliques can also be complete or incomplete. All the children of a complete clique are evidence nodes and all the children of an incomplete clique are not evidence nodes. In practice, cliques are usually mixed or incomplete and atypical. For example, if a simple net clique consists of only two levels of nodes, and level two nodes are still potential (unobserved evidence nodes), it is incomplete. Therefore, simple net cliques, complex net cliques, incomplete cliques, and mixed cliques represent functionally different clique patterns. These

clique patterns involve different compound cliques. A complex clique is composed of at least two simple cliques. The ANOVA score model Bayesian network can have three different types of mixed cliques:

Type 1: one parent + multi children

Type 2: multi-parents + common children

Type 3: multi-parents + no common child

This classification will be useful in examining information propagation from the evidence variables to potential variables.

6.3.2 Nomenclature of Bayesian Net Cliques in the ANOVA Score Model Network

It would be convenient to define a naming system to describe a network clique as a written code to complement its representation as a graph segment. A graph segment is a clique within a Bayesian network graph (such as (JavaBayes) Cozman,1998 and (Netica) Norsys, 2006). For a potential node, if the parent node position in the clique is coded A, then the positions of its children can be coded B. The number of parents or children can be indicated by a natural number following A or B respectively. Any parent node can be designated as P, and any child node can be designated as C. Dashes designate relations between parent A and child B nodes. For example, clique "alpha j" has one parent and three children. That means in position A there is one parent, in position B there are three children. The dash is used to describe parent-child relations. This

relation is coded as: A1P-B3C. Following this notation, Table 6.8 shows the written code for each clique.

Table 6.8. Description of Potential Cliques in Written Codes

Obj.	Cliques (by top node)	No. of parent	No. of children	Express code
1	alph_j	1	3	A1P-B3C
2	beta_k	1	3	A1P-B3C
3	12_jk	1	2	A1P-B2C
4	15_i(jk)	1	3	A1P-B3C
5	Gamma(jk)	2	2	A1P-B2C-B1P
6	Error_ei(jk)	2	2	A1P-B2C-B1P
7	2_i(jk)	1	3	A1P-B3C
8	2WayModelParameter	4	1	A1P-B1C-B3P
9	LHS	2	2	A1P-B2C-B1P
10	RHS	3	1	A1P-B1C-B2P
11	IndexValues	1	3	A1P-B3C
12	ModelEquation	4	1	A1P-B1C-B3P
13	Score	1	5	AIP-B5C
14	GrandMean	1	2	A1P-B2C
15	MainEffect:LevelofA	1	2	A1P-B2C
16	MainEffect:LevelofA	1	2	A1P-B2C
17	Interaction:AXB	1	3	A1P-B3C
18	Error	1	3	A1P-B3C
19	EffectsOfFactors	4	0	A1P-B3P
20	ScoreDecompositio	4	1	A1P-B1C-B3P
21	ScoreModel	3	1	A1P-B1C-B2P
22	ANOVAScoreModel2way	1	2	A1P-B2P

These "express codes" can be used to identify simple or complex net cliques. The parent A position is coded as A1P. If the child position B is coded as BXC, where X is a number, the clique is simple clique (such as clique 13-Score, and clique 18- Error; see Table 6.8). If the child position B is coded as BXP, (X=an integer), the clique is no longer simple since it includes children nodes that are parents in other cliques. Thus, complex and incomplete net cliques can be easily identified. The main advantage of this nomenclature system for Bayesian net cliques is that it recognizes the complexity of the net cliques.

6.3.3 Prior and Posterior Probabilities and Evidence Propagation in the ANOVA Score Model

Section 6.3.3 will introduce the calculations of Bayesian joint probabilities, posterior probabilities, and Bayesian updating processes with evidence. Basic Bayesian probabilities in Bayesian networks are prior and conditional probabilities which represent sets of beliefs that are determined by experts or other relevant researchers before any behavioural observations. These probabilities may be determined logically or on the basis of previous data. Prior probabilities are probability value vectors describing parent states. Probabilities of children are conditional on parent node states. Parent node posterior probabilities are probability value vectors that are calculated based on observations of child node states (evidence nodes). Updating is the computation process through which explanatory node posterior probabilities are calculated after entering the evidence values into child nodes (evidence variables).

6.3.3.1 Joint Probabilities as a Function of Prior and Conditional Probabilities

Once the prior probabilities of top parent nodes and the conditional probabilities of all other nodes have been defined, joint probabilities can be computed according to the chain rule which is a recursive representation of the probability distribution. Assuming there are N variables x_i, then the joint probability of the N variables can be expressed as (Pearl, 1988):

$$p(x_1,...,x_n) = p(x_n \mid x_{n-1},...,x_1)p(x_{n-1} \mid x_{n-2},...,x_1)...p(x_2 \mid x_1)p(x_1) =$$

$$\prod_{j=1}^{n} p(x_j \mid x_{j-1},...x_1) = \prod_{j=1}^{n} p(x_j \mid pa(x_j)) \qquad \text{(Formula 6-6)}$$

where $pa(x_j)$ is the product of the probabilities of the parents of x_j.

This indicates that the joint probability of N variables can be written as a product of N-1 conditional probabilities and one parent $p(x_1)$, i.e. the top level node. For example, if there are only two variables, one parent A and one child B, the formula can be shortened to:

$$p(AB) = p(B \mid A)p(A) \qquad \text{(Formula 6-7)}$$

If A and B are binary variables, Formula 6-7 can be specialized based on each state:

$$p([AB]^+) = p(B^+ \mid A = true)p(A = true) + p(B^+ \mid A = false)p(A = false) \quad \text{(Formula 6-8)}$$

$$p([AB]^-) = p(B^- \mid A = true)p(A = true) + p(B^- \mid A = false)p(A = false) \quad \text{(Formula 6-9)}$$

where $[AB]^+$ indicates event AB is true; $[AB]^-$ indicates event AB is false; where B^+ indicates event B is true; B^- indicates event B is false.

6.3.3.2 Posterior Probabilities Based on the Evidence Patterns

As conditional probabilities are produced and joint probabilities are calculated, posterior probabilities will be estimated using entered evidence. Suppose that we have N evidence elements comprising E, a vector of evidence values and potential vector variables of values of explanatory nodes designated as H (Pearl, 1988).

Evidences: $E = (e_1, e_2, ..., e_k, ..., e_n)$

Hypothesis: $H = (H_1, H_2, ..., H_k, ..., H_n)$

The posterior probability of a potential belief represented by hypothesis vector H_i can be expressed as:

$$p(H_i \mid e_1, e_2, ..., e_k, ..., e_n) = \frac{p(e_1, e_2, ..., e_k, ..., e_n \mid H_i) p(H_i)}{\sum p(e_1, e_2, ..., e_k, ..., e_n \mid H_i) p(H_i)} \qquad \text{(Formula 6-10)}$$

If we select an ANOVA score model net segment in which there is only one potential parent variable A and two evidence child variables B_1 and B_2, posterior probabilities can be determined based on Bayesian calculations. The number of combinations of B_1 and B_2 values is four since each has two states. Let the probability of A be designated as Φ_A, probabilities B_1 and B_2 be designated as Φ_{B1} and Φ_{B2} respectively, and the state of probability for true be indicated by "+" and for false be indicated by "-". The results and processes of evidence updating will be shown in section 6.3.3.3. The details will facilitate understanding of how evidence improves and changes beliefs (posterior probability) in the top parent level.

6.3.3.3 An Example of Updating: A Two-evidence and One Parent Clique

A simple net is selected from the ANOVA Score Model (Figure 6.4). The prior probabilities for parent P(A) and conditional probabilities for two children B_1 and B_2 are arbitrarily chosen here for illustrative purposes.

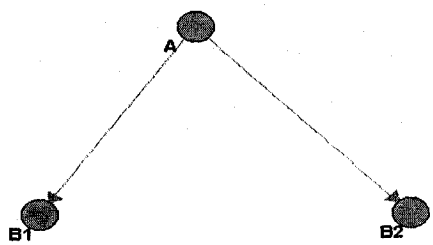

Figure 6.5. A Bayesian net with one parent and two children

Prior and conditional probabilities are:

$$\phi_A = \begin{pmatrix} true & false \\ 0.30 & 0.70 \end{pmatrix}$$

$$\phi_{B1} = \begin{pmatrix} & P(A^+) & P(A^-) \\ P(B_1^+ \mid A) & 0.80 & 0.10 \\ P(B_1^- \mid A) & 0.20 & 0.90 \end{pmatrix} \qquad \phi_{B2} = \begin{pmatrix} & P(A^+) & P(A^-) \\ P(B_2^+ \mid A) & 0.45 & 0.15 \\ P(B_2^- \mid A) & 0.55 & 0.85 \end{pmatrix}$$

Posterior and joint probabilities are obtained by running JavaBayes (Cozman, 1998) before entering any evidence:

For A: $\begin{pmatrix} Yes: & 0.3 \\ No: & 0.7 \end{pmatrix}$

For B_1: $\begin{pmatrix} Yes: & 0.31 \\ No: & 0.69 \end{pmatrix}$ For B_2: $\begin{pmatrix} Yes: & 0.24 \\ No: & 0.76 \end{pmatrix}$

Evidence combination results for P(A) after running JavaBayes are:

In B_1^- and B_2^- case: $\Rightarrow$ P(A): $\begin{pmatrix} Yes: & 0.05805 \\ No: & 0.94195 \end{pmatrix}$

In B_1^- and B_2^+ case: $\Rightarrow$ P(A): $\begin{pmatrix} Yes: & 0.2222 \\ No: & 0.7778 \end{pmatrix}$

In B_1^+ and B_2^- case: $\Rightarrow$ P(A): $\begin{pmatrix} Yes: & 0.6893 \\ No: & 0.3107 \end{pmatrix}$

In B_1^+ and B_2^+ case: $\Rightarrow$ P(A): $\begin{pmatrix} Yes: & 0.91139 \\ No: & 0.08861 \end{pmatrix}$

The computational processes have been decomposed into steps in order to examine these results. If evidence variables have been instantiated regardless of whether they are true or false, the four combinations will be realized by varying the parent variable A using the following rule.

$$P(A \mid B_1 B_2) = \frac{P(B_1 B_2 \mid A)P(A)}{\sum_A P(B_1 B_2 \mid A)P(A)} \qquad \text{(Formula 6-11)}$$

In the hierarchical model, nodes B_1 and B_2 are locally independent given A.

$$P(B_1, B_2 \mid A) = P(B_1 \mid A)\, P(B_2 \mid A). \qquad \text{(Formula 6-12)}$$

Consequently:

$$P(A \mid B_1 B_2) = \frac{P(B_1 \mid A)(B_2 \mid A)P(A)}{\sum_A P(B_1 \mid A)P(B_2 \mid A)P(A)} \qquad \text{(Formula 6-13)}$$

Here A has two values: A^- and A^+, B_1 has two values B_1^- and B_1^+ and B_2 has two values B_2^- and B_2^+, respectively. Theoretically, the combination space is 2^3=8:

$$P(A^- \mid B_1^- B_2^-) = \frac{P(B_1^- \mid A^-)P(B_2^- \mid A^-)P(A^-)}{[P(B_1^- \mid A^-)P(B_2^- \mid A^-)P(A^-)]+[P(B_1^- \mid A^+)P(B_2^- \mid A^+)P(A^+)]}$$

(Formula 6-14)

$$P(A^- \mid B_1^- B_2^-) = \frac{0.9 \times 0.85 \times 0.7}{0.9 \times 0.85 \times 0.7 + 0.2 \times 0.55 \times 0.3}$$

$$= \frac{0.5355}{0.5355 + 0.033} = \frac{0.5355}{0.5685} = 0.94195$$

$$P(A^+ \mid B_1^- B_2^-) = \frac{P(B_1^- \mid A^+)P(B_2^- \mid A^+)P(A^+)}{[P(B_1^- \mid A^-)P(B_2^- \mid A^-)P(A^-)] + [P(B_1^- \mid A^+)P(B_2^- \mid A^+)P(A^+)]}$$

(Formula 6-15)

$$P(A^+ \mid B_1^- B_2^-) = \frac{0.2 \times 0.55 \times 0.3}{0.9 \times 0.85 \times 0.7 + 0.2 \times 0.55 \times 0.3}$$

$$= \frac{0.033}{0.5355 + 0.033} = \frac{0.033}{0.5685} = 0.05805$$

$$P(A^- \mid B_1^+ B_2^-) = \frac{P(B_1^+ \mid A^-)P(B_2^- \mid A^-)P(A^-)}{[P(B_1^+ \mid A^-)P(B_2^- \mid A^-)P(A^-)] + [P(B_1^+ \mid A^+)P(B_2^- \mid A^+)P(A^+)]}.$$

(Formula 6-16)

$$P(A^- \mid B_1^+ B_2^-) = \frac{0.1 \times 0.85 \times 0.7}{0.1 \times 0.85 \times 0.7 + 0.8 \times 0.55 \times 0.3}.$$

$$= \frac{0.0595}{0.0595 + 0.132} = \frac{0.0595}{0.1915} = 0.3107$$

$$P(A^+ \mid B_1^+ B_2^-) = \frac{P(B_1^+ \mid A^+)P(B_2^- \mid A^+)P(A^+)}{[P(B_1^+ \mid A^-)P(B_2^- \mid A^-)P(A^-)] + [P(B_1^+ \mid A^+)P(B_2^- \mid A^+)P(A^+)]}$$

(Formula 6-17)

$$P(A^+ \mid B_1^+ B_2^-) = \frac{0.8 \times 0.55 \times 0.3}{0.1 \times 0.85 \times 0.7 + 0.8 \times 0.55 \times 0.3}$$

$$= \frac{0.132}{0.0595 + 0.132} = \frac{0.132}{0.1915} = 0.6893$$

$$P(A^- \mid B_1^- B_2^+) = \frac{P(B_1^- \mid A^-)P(B_2^+ \mid A^-)P(A^-)}{[P(B_1^- \mid A^-)P(B_2^+ \mid A^-)P(A^-)]+[P(B_1^- \mid A^+)P(B_2^+ \mid A^+)P(A^+)]}$$

(Formula 6-18)

$$P(A^- \mid B_1^- B_2^+) = \frac{0.9\times0.15\times0.7}{0.9\times0.15\times0.7+0.2\times0.45\times0.3}$$

$$= \frac{0.0945}{0.0945+0.027} = \frac{0.0945}{0.1215} = 0.7778$$

$$P(A^+ \mid B_1^- B_2^+) = \frac{P(B_1^- \mid A^+)P(B_2^+ \mid A^+)P(A^+)}{[P(B_1^- \mid A^-)P(B_2^+ \mid A^-)P(A^-)]+[P(B_1^- \mid A^+)P(B_2^+ \mid A^+)P(A^+)]}$$

(Formula 6-19)

$$P(A^+ \mid B_1^- B_2^+) = \frac{0.2\times0.45\times0.3}{0.9\times0.15\times0.7+0.2\times0.45\times0.3}$$

$$= \frac{0.027}{0.0945+0.027} = \frac{0.027}{0.1215} = 0.2222$$

$$P(A^- \mid B_1^+ B_2^+) = \frac{P(B_1^+ \mid A^-)P(B_2^+ \mid A^-)P(A^-)}{[P(B_1^+ \mid A^-)P(B_2^+ \mid A^-)P(A^-)]+[P(B_1^+ \mid A^+)P(B_2^+ \mid A^+)P(A^+)]}$$

(Formula 6-20)

$$P(A^- \mid B_1^+ B_2^+) = \frac{0.1\times0.15\times0.7}{0.1\times0.15\times0.7+0.8\times0.45\times0.3}$$

$$= \frac{0.0105}{0.0105+0.108} = \frac{0.0105}{0.1185} = 0.08861$$

$$P(A^+ \mid B_1^+ B_2^+) = \frac{P(B_1^+ \mid A^+)P(B_2^+ \mid A^+)P(A^+)}{[P(B_1^+ \mid A^-)P(B_2^+ \mid A^-)P(A^-)]+[P(B_1^+ \mid A^+)P(B_2^+ \mid A^+)P(A^+)]}$$

(Formula 6-21)

$$P(A^{+} \mid B_1^{+} B_2^{+}) = \frac{0.8 \times 0.45 \times 0.7}{0.1 \times 0.15 \times 0.7 + 0.8 \times 0.45 \times 0.3}$$

$$= \frac{0.108}{0.0105 + 0.108} = \frac{0.108}{0.1185} = 0.91139$$

Calculation results are identical to those from JavaBayes. Section 6.3.3.3 has illustrated the fundamental rules of how to calculate joint and posterior probabilities in terms of the evidence combinations. The above example can aid in understanding data propagation in the ANOVA score model Bayesian network. Data propagation includes inference from parent potential explanatory variables to the observable nodes, and updating from instantiated evidential variables to the parent potential explanatory variables.

This section focuses on demonstrating the rudimentary principles for building the ANOVA score model Bayesian network components and on specifying prior and conditional network node probabilities. Hence, it is important to examine the structure of the Bayesian network for the ANOVA score model domain.

6.3.4 Fundamental Structure in the Bayesian Cliques Contained in the ANOVA Score Model Network

ANOVA score model Bayesian networks contain many cliques. Several rudimentary clique patterns were chosen to examine network structures. There

are typically two to five children for the simple cliques in ANOVA score model networks. Although most simple network cliques are patterns with one parent and two or three children, we will focus first on several clique patterns. The one parent and one child models are very basic and were chosen though there are no such cliques in ANOVA score model nets (Figure 6.6). Two children (Figure 6.7), three children (Figure 6.8), and four children (Figure 6.9) cases will also be examined.

Compound clique patterns will also be explored. There were two levels of potential explanatory variables in compound cliques, with a single top node connected to two or three child nodes, each of which is the parent of a simple clique having evidence nodes as children. Lastly, a complicated net structure consisting of three levels of potential explanatory variable levels with four connected cliques is examined (Figure 6.13).

Different conditional probability values were tested in order to examine conditional probability effects on the calculation of posterior probability networks, prior probability and conditional probability networks. We hypothesized "No knowledge" of the state of the parent variable, which means that prior probabilities for true and false for top nodes were both set to 0.5. For the conditional probabilities, three levels were set specifying the conditional probabilities for child nodes states. Table 6.9 shows details.

All clique probabilities are arranged as in Table 6.9. The symmetric probability pattern and the expected symmetric probability changes can be noted.

Table 6.9. Conditional Probability Level Sets in Each Clique Type

Conditional probability level	Prior probabilities (at true=0.5/false=0.5) value of parent node			
	True		False	
	True	False	True	False
Level one	0.6	0.4	0.4	0.6
Level two	0.67	0.33	0.33	0.67
Level three	0.75	0.25	0.25	0.75

6.3.4.1 One Parent and One Child Bayesian Net

First clique was defined as a one parent and one child Bayesian net. Potential variable was set as true=0.5 level. This means that the student understands potential variables with 50% possibility to complete the task. Figure 6.6 illustrates this model graphically. Conditional probabilities have been set at three levels in Table 6.10.

Figure 6.6. One parent with one child Bayesian net model

The first test sets conditional probabilities at level one. Prior and conditional probabilities are as follows:

Prior probability of parent states: [true= 0.5 false=0.5]

		Parent State (Potential V)	
Conditional probability:	Evidence state	true	false
p(Evid | Potential V)	True	0.60	0.40
	False	0.40	0.60

Table 6.10 shows the resulting evidence states from updating the posterior probabilities with several evidence patterns.

Table 6.10. Updating Prior Probability with Consecutive Evidence Pattern for Clique having One Child Mode

#	Prior prob.		Estimated posterior prob. of the parent variable given			
	Posterior probability of parent		Zero evidence		Full evidence	
	True	False	True	False	True	False
1	0.5	0.5	0.4	0.6	0.6	0.4
2	0.6	0.4	0.5	0.5	0.6923	0.3077
3	0.6923	0.3077	0.6	0.4	0.7714	0.2286
4	0.7714	0.2286	0.6923	0.3077	0.8350	0.1650
5	0.8350	0.1650	0.7714	0.2286	0.8836	0.1164
6	0.8836	0.1164	0.8350	0.1650	0.9193	0.0807
7	0.9193	0.0807	0.8836	0.1164	0.9447	0.0553
8	0.9447	0.0553	0.9192	0.0807	0.9624	0.0376

Note: When the value larger than 0.95 is found, the iteration stops.

Test 2 sets conditional probabilities as level two. The prior and conditional probabilities are as follows:

Prior probability [true= 0.5 false=0.5]

Parent p (Potential V)

Conditional probability:		true	false
p(Evid | Potential V)	True	0.67	0.33
	False	0.33	0.67

The results of updating prior probabilities with consecutive evidence in the full evidence space are shown in Table 6.11.

Table 6.11. Updating Prior Probability with Consecutive Evidences for One Child Model

#	Prior prob.		Estimated posterior prob. of the parent variable given			
	Posterior probability of parent		Zero evidence		Full evidence	
	True	False	True	False	True	False
1	0.5	0.5	0.33	0.67	0.67	0.33
2	0.67	0.33	0.5	0.5	0.8048	0.1952
3	0.8048	0.1952	0.67	0.33	0.8933	0.1067
4	0.8943	0.1067	0.8048	0.1952	0.9444	0.0556
5	0.9444	0.0556	0.8932	0.1068	0.9718	0.0282

Note: When the value larger than 0.95 is found, the iteration stops.

The third test starts to set conditional probabilities as level three. The prior and conditional probabilities are as follows:

Prior probability [true= 0.5 false=0.5]

Parent p (Potential V)

Conditional probability:		true	false
p(Evid | Potential V)	True	0.75	0.25
	False	0.25	0.75

The results of updating prior probabilities with consecutive evidences in the full evidence space have been shown in Table 6.12.

Table 6.12. Updating Prior Probability with Consecutive Full Evidences

#	Prior prob.		Estimated posterior prob. of the parent variable given			
	Posterior probability of parent		Zero evidence		Full evidence	
	True	False	True	False	True	False
1	0.5	0.5	0.25	0.75	0.75	0.25
2	0.75	0.25	0.5	0.5	0.9	0.1
3	0.9	0.1	0.75	0.25	0.9643	0.0357

Note: When the value larger than 0.95 is found, the iteration stops.

"Excellent mastery level" has been set at 0.95. If the probability of Potential V is larger than 0.95, the learner has mastered this knowledge point with excellence. If evidence conditional probability equals 0.6 for true and false,

eight runs are required to top 0.95. If evidence conditional probability equals 0.67 for true and true, only 5 runs are required to top 0.95. If the evidence conditional probability equals to 0.75 for true and true, only 3 runs are required to top 0.95. This indicates that if conditional probabilities for mastery are set to a higher level, updating will occur more quickly than when it is set to a lower conditional probability level. Of course, the assumption is that for each run, positive evidence will be instantiated.

6.3.4.2 One Parent and Two Children Bayesian Net

This Bayesian net indicates that two evidence variables support one potential variable. If the potential variable receives full positive evidence propagation, the potential variable will have a higher posterior probability of mastery. Figure 6.7 is a clique taken from the Bayesian network for assessing the ANOVA score model. The α12_jk is a potential variable and α_j and b_k are two evidence variables. There are several possible evidence combinations. Zero evidence shows the learner fails to respond to either evidence variable. One true and one false for the evidence variables indicates that the student demonstrates partial knowledge and skill in responding to evidence variables resulting in mixed evidence. Finally, full positive evidence demonstrates that the learner performed successfully on both evidence variables.

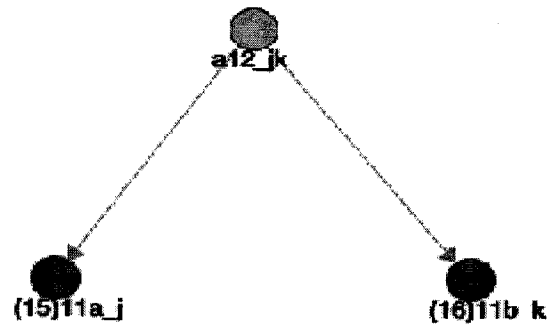

Figure 6.7. One parent with two children Bayesian net model

Test 1 starts by setting the conditional probabilities at level one (see table 6.13). The prior and conditional probabilities are as follows.

Prior probability [true= 0.5 false=0.5]

The two evidence variables (15)11a_j and (16)11b_k have the same conditional probability tables. Here, the conditional probability table of (15)11a_j is listed.

		Parent node state (α12_jk)	
Conditional probability:		true	false
p(Evid | α12_jk)	True	0.60	0.40
	False	0.40	0.60

The results of updating prior probabilities with consecutive evidences in the full evidence space are shown in Table 6.13.

Table 6.13. Updating Prior Probability (True = 0.6) with Consecutive Evidences for Two Children Model

#	Prior prob.		Estimated posterior prob. of the parent node given					
	Posterior probability of parent		zero evidence		One true& one false		All true evidence	
	True	False	True	False	True	False	True	False
1	0.5	0.5	0.3076	0.6923	0.5	0.5	0.6923	0.3077
2	0.6923	0.3077	0.5	0.5	0.6923	0.3077	0.8350	0.1650
3	0.8350	0.1650	0.6923	0.3077	0.8350	0.1650	0.9193	0.0807
4	0.9193	0.0807	0.8350	0.1650	0.9193	0.0807	0.9624	0.0376

Test 2 sets conditional probabilities as level two. The prior and conditional probabilities are as follows:

Prior probability [true= 0.5 false=0.5]

Two evidence variables (15)11a_j and (16)11b_k share the same conditional probabilities, here the conditional probability distribution of (15)11a_j is listed here as illustration.

		Parent p (α12_jk)	
Conditional probability:		true	false
p(Evid | α12_jk)	True	0.67	0.33
	False	0.33	0.67

The results of updating prior probabilities with several rounds of consecutive full evidence have been shown in Table 6.14.

Table 6.14. Updating Prior Probability (True = 0.67) with Consecutive Evidence for Two Children Model

#	Prior prob.		Estimated posterior prob. of the parent node given					
	Posterior probability of parent		zero evidence		One true& one false		All true evidence	
	True	False	True	False	True	False	True	False
1	0.5	0.5	0.1952	0.8048	0.5	0.5	0.8048	0.1952
2	0.8048	0.1952	0.5	0.5	0.8048	0.1952	0.9444	0.0556
3	0.9444	0.0556	0.8048	0.1952	0.9444	0.0556	0.9859	0.0141

Test 3 sets conditional probabilities as level three. The prior and conditional probabilities are as follows:

Prior probability [true= 0.5 false=0.5]

Two evidence variables (15)11a_j and (16)11b_k share the same conditional probabilities, here the conditional probability distribution of (15)11a_j is listed here as illustration.

		Parent p (α12_jk)	
Conditional probability:		true	false
p(Evid | α12_jk)	True	0.75	0.25
	False	0.25	0.75

The results of updating prior probabilities with several consecutive evidences in full evidence space are shown in Table 6.15.

Table 6.15. Updating Prior Probability (True = 0.75) with Consecutive Evidences for Two Children Model

#	Prior prob.		Estimated posterior prob. of the parent node given					
	Posterior probability of parent		zero evidence		One true& one false		All true evidence	
	True	False	True	False	True	False	True	False
1	0.5	0.5	0.1	0.9	0.5	0.5	0.9	0.1
2	0.9	0.1	0.5	0.5	0.9	0.1	0.9898	0.0122

When the conditional probability of true evidence variables is given and true mastery of parent node equals 0.6, the Bayesian net needs four runs to reach 0.95 probability. However, as conditional probability of true evidence and given mastery is set as 0.67, the net needs only three runs to reach 0.95 probability. When conditional probability is set to 0.75, only two runs are required to reach 0.95 probability.

The conclusion is almost the same as that for one parent and one child model. As true values of conditional probabilities increase, the updating process will occur faster with the lower conditional probabilities.

6.3.4.3 One Parent and Three Children Bayesian Net

The Bayesian net for one parent and three children was chosen from the Bayesian Network for Assessment of the ANOVA score model (see Figure 6.8). If a learner knows α and the index j very well and also knows how to apply j to α, the learner is believed to have acquired knowledge segment α_j very well. Figure 6.8 illustrates this situation.

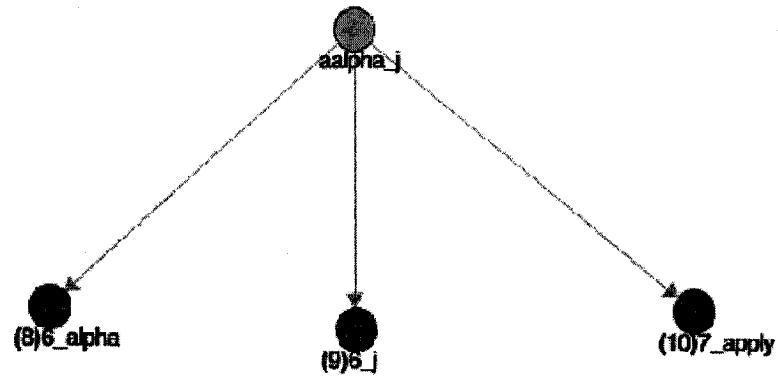

Figure 6.8. One parent with three children Bayesian net model

In this model, the conditional probabilities are identical to previous models. The "true and true" conditional probability has been set at the three values 0.6, 0.67, and 0.75. The first prior probability is still 0.5 to 0.5 for both true and false. In test 1, the "true and true" conditional probability=0.60, "false and false" conditional probability=0.6 (see Table 6.16).

Table 6.16. Updating Prior Probability (True = 0.6) with Consecutive Evidences for Three Children Model

#	Prior prob.		Posterior prob. of potential variables given							
	Post prob of evid.		Zero evid.		One true evid.		Two true evid.		All true evid.	
	True	False	True	False	True	False	True	False	True	False
1	0.5	0.5	0.23	0.77	0.4	0.6	0.6	0.4	0.77	0.23
2	0.77	0.23	0.5	0.5	0.69	0.31	0.83	0.17	0.92	0.08
3	0.92	0.08	0.77	0.23	0.88	0.12	0.94	0.06	0.97	0.03

Note. "Evid." is the abbreviation for evidence

Test 2 is about the prior probability, "true and true" conditional probability is 0.67, false and false is 0.67 (see Table 6.17).

Table 6.17. Updating Prior Probability (True = 0.67) with Consecutive Evidences for Three Children Model

#	Prior prob.		Posterior prob. of potential variables given							
	Post prob of evid.		Zero evid.		One true evid.		Two true evid.		All true evid.	
	True	False	True	False	True	False	True	False	True	False
1	0.5	0.5	0.11	0.89	0.33	0.67	0.67	0.33	0.89	0.11
2	0.89	0.11	0.5	0.5	0.80	0.20	0.94	0.056	0.99	0.01

Note. "Evid." is the abbreviation for evidence

Test 3 is about the conditional probability true and true is 0.75, false and false is 0.75 (see Table 6.18).

Table 6.18. Updating Prior Probability (True = 0.75) with Consecutive Evidences for Three Children Model

#	Prior prob.		Posterior prob. of potential variables given							
	Post prob of evid.		Zero evid.		One true evid.		Two true evid.		All true evid.	
	T	F	T	F	T	F	T	F	T	F
1	0.5	0.5	0.036	0.964	0.25	0.75	0.75	0.25	0.964	0.036

Note. "Evid." is the abbreviation of evidence. T is true and F is false

The three net tests show that posterior probability updating speeds up as the number of evidence variables increases and the conditional probability is set higher. For example, when conditional probability is set at 0.75 prior probability is expected to be over 0.95 after one test with full true evidence.

6.3.4.4 One Parent and Four Children Bayesian Net

This Bayesian net was created to mimic a clique of the ANOVA score model. One parent potential explanatory variable has four child evidence variables. The ideal state is that all four evidence variables receive positive responses. Figure 6.9 shows this situation.

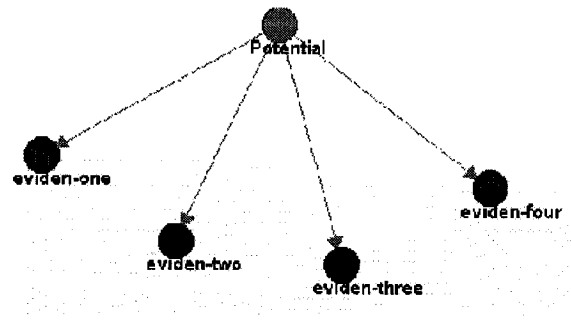

Figure 6.9. One parent with four children model

In this model the conditional probabilities are set at three levels. The "true and true" in the conditional probabilities has been set at values of 0.6, 0.67, and 0.75. The first prior probability is still 0.5 and 0.5 for both true and false.

The first test is about true and true of the conditional probability =0.60. Table 6.19 shows the results.

Table 6.19. Updating Prior-Probability (True = 0.6) with Consecutive Evidences for Four-Children Model

No.	Evidence	Posterior probability of potential	
	Prior probability:[true=0.5 false= 0.5]	True	False
1	zero true evidence, four false evidences	0.1649	0.8351
2	one true evidence, three false evidences	0.3077	0.6923
3	two true evidences, two false evidences	0.5	0.5
4	three true evidences, one false evidence	0.6923	0.3077
5	four true evidences, zero false evidence	0.8351	0.1649

Table 6.19 (continued)

No.	Evidence	Posterior probability of potential	
	Prior probability:[true=0.8351 false= 0.1649]	True	False
1	zero true evidence, four false evidences	0.5	0.5
2	one true evidence, three false evidences	0.6924	0.3076
3	two true evidences, two false evidences	0.8351	0.1649
4	three true evidences, one false evidence	0.9193	0.0807
5	four true evidences, zero false evidence	0.9625	0.0375

Test 2 is about the conditional probability when true and true is set at 0.67, and false and false is 0.67 (see Table 6.20).

Table 6.20. Updating Prior Probability (True = 0.67) with Consecutive Full Evidences for Four-Children Model

No.	Evidence	Posterior probability of potential	
	Prior probability:[true=0.5 false= 0.5]	True	False
1	zero true evidence, four false evidences	0.0556	0.9444
2	one true evidence, three false evidences	0.1952	0.8048
3	two true evidences, two false evidences	0.5	0.5
4	three true evidences, one false evidence	0.8048	0.1952
5	four true evidences, zero false evidence	0.9444	0.0556

Table 6.20. (continued)

No.	Evidence	Posterior probability of potential	
	Prior probability:[true=0.9444 false= 0.0556]	True	False
1	zero true evidence, four false evidences	0.5	0.5
2	one true evidence, three false evidences	0.8047	0.1953
3	two true evidences, two false evidences	0.9444	0.0556
4	three true evidences, one false evidence	0.9859	0.0141
5	four true evidences, zero false evidence	0.9965	0.0035

Test 3 is about the conditional probability true and true is 0.75, and false and false is 0.75 (see Table 6.21).

Table 6.21. Updating Prior-Probability (True = 0.75) with Consecutive Evidences for Four-Children Model

No.	Evidence	Posterior probability of potential	
	Prior probability:[true=0.5 false= 0.5]	True	False
1	zero true evidence, four false evidences	0.0112	0.9878
2	one true evidence, three false evidences	0.1	0.9
3	two true evidences, two false evidences	0.5	0.5
4	three true evidences, one false evidence	0.9	0.1
5	four true evidences, zero false evidence	0.9878	0.0112

The problem space increases in the one-parent-four-child model. The evidence combination is five. As conditional probability levels increase, the range of true posterior also increases. For example, in Test 1, the conditional probability true and true is set at 0.6; the difference between posterior true between zero true-evidence and four true-evidences is 0.6702 (0.8351-0.1649). When the conditional probability true and true is set at 0.75, the difference between the posterior true between zero true evidence and four true evidence is 0.9756 (0.9878-0.0122).

6.3.4.5 Multi Level Multi Clique Bayesian Net Models

Multi-level and multi-clique complex models will be examined. These explorations will help in understanding more complicated models such as the ANOVA score model Bayesian network. Multi-level means that more layers are counted from the top parent to bottom evidence nodes. Multi-cliques show that more than one clique is connected to the up-level parent in a parallel fashion. Section 6.3.4.5 will include (a) three-level two-clique models, and (b) three-level three-clique models.

Three-level two-clique models have one top parent variable at level 1 and two potentials at level 2. Five observable variables at the bottom categorize two cliques, one with two evidential variables and another with three.

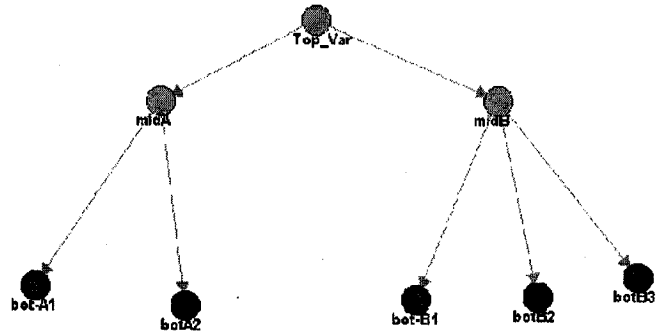

Figure 6.10. Three-level with two cliques Model

Top_Var is the parent of the net and midA and midB are two potentials. There are two cliques connected to top explanatory variable Top_Var: midA and midB. "midA" has two evidence nodes and "midB" has three. For each clique, evidence space is evidence node number plus one. The entire network evidence space is the product of evidence spaces of two cliques. Therefore, the space of this network is 3 X 4 = 12. The prior probability has been defined as true and false at both 0.5. The conditional probabilities are all set at p(true | true) = 0.67, p(true| false) = 0.33, p(false | true) = 0.33, and p(false | false) = 0.67. For future reference, any one of the four probability combinations may be specified at one time. Therefore, by default, prior probabilities for both true and false are set at 0.5; conditional probabilities for both true and true are set at 0.67. Table 6.22 displays the distribution of posterior probabilities of potential variables based on different evidence instantiated combinations for the model in Figure 10.

Table 6.22 shows that when all evidence is false, Top-Var value is 0.2751; when all evidence is true it is 0.7249. An interesting phenomenon is that values for each column have a complementary relationship, i.e. the values sum to 1,

about an imaginary central axis between Rows 6 and 7. For example, in the Top_Var true column, the values in Row 1 and Row 12 are complementary. Another interesting point is that values show gradually increasing trends from all false evidence to all true evidence for each true column.

Table 6.22. Updating Prior Probabilities with Different Combinations of Instantiated Evidences for a Three-level Two-Clique Model

#	Evidence of children of		Posterior Probabilities of					
	midA	midB	midA		midB		Top Var	
			true	false	true	false	true	false
1	FF	FFF	0.1682	0.8318	0.0940	0.9060	0.2751	0.7249
2	FF	TFF	0.1832	0.8168	0.2996	0.7004	0.3424	0.6576
3	FF	TTF	0.2079	0.7921	0.6381	0.3619	0.4531	0.5469
4	FF	TTT	0.2255	0.7745	0.8790	0.1210	0.5319	0.4681
5	TF	FFF	0.4545	0.5455	0.1067	0.8933	0.3663	0.6337
6	TF	TFF	0.4803	0.5196	0.33	0.67	0.4422	0.5578
7	TF	TTF	0.5196	0.4803	0.67	0.33	0.5578	0.4422
8	TF	TTT	0.5455	0.4545	0.8937	0.1067	0.6337	0.3663
9	TT	FFF	0.7745	0.2255	0.1210	0.8790	0.4681	0.5319
10	TT	TFF	0.7921	0.2079	0.3619	0.6381	0.5469	0.4531
11	TT	TTF	0.8168	0.1832	0.7004	0.2996	0.6576	0.3424
12	TT	TTT	0.8318	0.1682	0.9060	0.0940	0.7249	0.2751

Note: T indicates that a node is true; F indicates a node is false.

Figure 6.11 shows a three-level three-clique model. It has one top parent variable at level 1 and three level 2 potentials: Middle_A, Middle_B and Midlle_C. Nine observable variables at bottom, level 3, are categorized into three cliques with 2, 3 and 4 evidence nodes respectively. Evidence node variables are called Bottom_A_1, Bottom_B_2, etc.

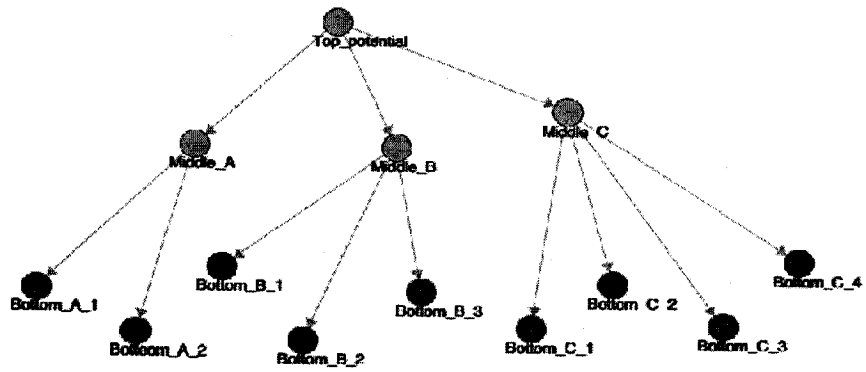

Figure 6.11. Three-level with three cliques model.

Figure 6.11 shows that the entire evidence space is a product of the evidence spaces of three cliques. The evidence space for each clique is the number of cliques plus one. Therefore, the space of this network is 3 X 4 X 5 = 60. In order to make the report briefer, several critical and typical values are shown in Table 6.23.

A complementary relationship about the central axis still exists in this model. It can be seen between numbers 1 and 60, numbers 2 and 59, numbers 29 and 32, and numbers 30 and 31. Another interesting point is that as the number of evidence variables increase, the column values between adjacent items decreased. For example, in Table 6.22, the difference between Row 1 and 2 in column Top-Var true is 0.0673 (0.3424-0.2751); while in Table 6.23 the

difference between Row 1 and 2 in Column Top_potential true is 0.0305 (0.1995-0.1690). This suggests that differences will decrease as the number of evidence variables increases. As the number of evidence variable approaches infinity, theoretically, potential variables will progress from discrete variables to become almost continuous.

Table 6.23. Updating Prior Probabilities with Different Combinations of Instantiated Evidences for a Three-level with Three-Clique Model

		Posterior Probabilities of							
	midA midB midC	midA		midB		midC		Top Potential	
		true	false	true	false	true	false	true	false
(1)	FF FFF FFFF	0.1445	0.8555	0.0792	0.9208	0.0414	0.9586	0.1690	0.8310
(2)	FF FFF TFFF	0.1513	0.8487	0.0834	0.9166	0.1513	0.8487	0.1995	0.8005
(3)	FF FFF TTFF	0.1682	0.8318	0.0940	0.9060	0.4235	0.5765	0.2751	0.7249
...									
(29)	TF TFF TTT F	0.5160	0.4840	0.3619	0.6381	0.7921	0.2079	0.5469	0.4531
(30)	TF TFF TTTT	0.5329	0.4671	0.3771	0.6229	0.9401	0.0599	0.5967	0.4033
(31)	TF TTF FFFF	0.4671	0.5329	0.6229	0.3771	0.0599	0.9401	0.4033	0.5967
(32)	TF TTF TFFF	0.4840	0.5160	0.6381	0.3619	0.2079	0.7921	0.4531	0.5469
...									
(58)	TT TTT TTFF	0.8318	0.1682	0.9060	0.0940	0.5765	0.4235	0.7249	0.2751
(59)	TT TTT TTTF	0.8487	0.1513	0.9166	0.0834	0.8487	0.1513	0.8005	0.1995
(60)	TT TTT TTTT	0.8556	0.1445	0.9208	0.0792	0.9586	0.0414	0.8310	0.1690

Note. The node is designated T when true and designated F when false.

6.3.4.6 Mixed Model Combining Potential Nodes and Evidence Nodes as Children

The model being used to assess the ANOVA score model problem has many mixed models including two or more potential variables connected in parent-child relations. These potential variables have their own evidence variables. From the entire assessment network model (Figure 6.4), one mixed model has been selected for analysis.

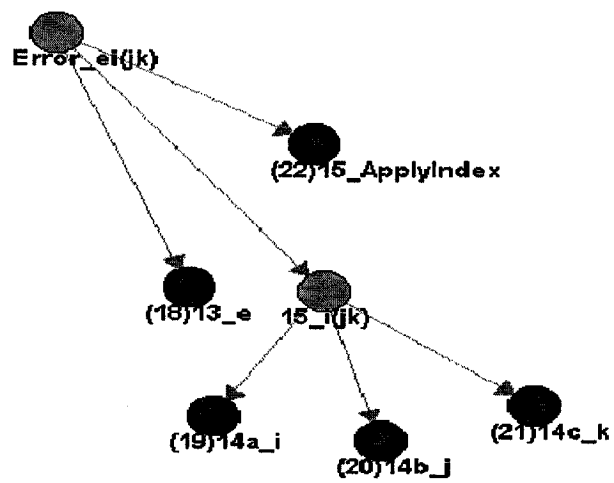

Figure 6.12. A mixed model with two potential variables and multi-evidences

The mixed model in Figure 6.12 consists of two connected potential variables in parent-child relations. They have two and three evidence variables respectively. The evidence space of this network is calculated in two steps. In Step 1, evidence nodes 18 and 22 are considered together with three combinations (TT, TF, FF). In Step 2, evidence nodes 19, 20, and 21 are

considered with four combinations (TTT, TTF, TFF, FFF). The product of Steps 1 and 2 is twelve different combinations. Updating results are shown in Tables 6.24 and 6.25.

Table 6.24. Updating Prior Probabilities with Consecutive Evidences for Mixed Model

Evidences of					Posterior Probability of			
(18)	(22)	(19)	(20)	(21)	Error-ei(jk)		15-i(jk)	
					True	False	True	False
F	F	F	F	F	0.1230	0.8770	0.0728	0.9272
F	F	F	F	T	0.1613	0.8387	0.2444	0.7556
F	F	F	T	T	0.2343	0.7657	0.5714	0.4286
F	F	T	T	T	0.2956	0.7044	0.8461	0.1539
F	T	F	F	F	0.3663	0.6337	0.1067	0.8933
F	T	F	F	T	0.4422	0.5578	0.3300	0.6700
F	T	F	T	T	0.5578	0.4422	0.6700	0.3300
F	T	T	T	T	0.6337	0.3663	0.8933	0.1067
T	T	F	F	F	0.7044	0.2956	0.1539	0.8461
T	T	F	F	T	0.7657	0.2343	0.4286	0.5714
T	T	F	T	T	0.8387	0.1613	0.7556	0.2444
T	T	T	T	T	0.8770	0.1230	0.9272	0.0728

Note: The node is designated T when true and designated F when false.

*Table 6.25. Updating Prior Probabilities with Consecutive Evidences for Mixed Model **

Evidences of					Posterior Probability of			
(18)	(22)	(19)	(20)	(21)	Error-ei(jk)		15-i(jk)	
					True	False	True	False
F	F	F	F	F	0.2533	0.7467	0.1598	0.8402
F	F	F	F	T	0.4874	0.5126	0.4395	0.5605
F	F	F	T	T	0.7588	0.2412	0.7637	0.2363
F	F	T	T	T	0.8981	0.1019	0.9302	0.0698
F	T	F	F	F	0.5831	0.4169	0.3558	0.6442
F	T	F	F	T	0.7968	0.2032	0.6948	0.3052
F	T	F	T	T	0.9284	0.0716	0.9037	0.0963
F	T	T	T	T	0.9732	0.0268	0.9748	0.0252
T	T	F	F	F	0.8522	0.1478	0.5156	0.4844
T	T	F	F	T	0.9417	0.0583	0.8144	0.1856
T	T	F	T	T	0.9816	0.0184	0.9476	0.0524
T	T	T	T	T	0.9934	0.0066	0.9868	0.0132

Note: T indicates the node is true, F indicates the node is false.

**potential nodes: Error-i(jk) with the posterior probabilities starting at true=0.8770 and false=0.1230; 15-i(jk) with the posterior probabilities starting at true=0.9272; false=0.0728)*

Evidence states start from all false states in which the posterior probability for Error-ei(jk) being true is 0.1230 and for 15-i(jk) is 0.0728. As the number of true states increases, the probabilities of true for these potentials increases to 0.8770 for Error-ei(jk) and to 0.9272 for 15-i(jk). Another interesting fact is an "inversely complementary phenomenon", where any summation of probabilities of the combination is 1. In the second run, this phenomenon does not exist

because the starting parent and potential probabilities are not symmetrical (true 0.5 to false 0.5).

Section 6.3.4.6 has discussed the fundamental structures of Bayesian net from one-parent one-child structures, to one-parent four-child structures. Following that, complex Bayesian net models were explored. In fact, two models are involved: three-level two-clique models and three-level three-clique models. These two models reveal that the top potential probability can receive sufficient updating as the number of evidence variables and the number of cliques increase. Finally mixed models have two connected potentials. This submodel indicates the complexity of transferring knowledge from one state to another. In other words, assessment of the ability to complete a knowledge task becomes more robust as the Bayesian net becomes more complicated and the number of cliques and evidence variables increases if they are efficiently validated.

6.3.5 Assessment Models Used to Examine the Knowledge and Skills Underlying Mastery of ANOVA Score Models

The basic purpose of the assessment model is to examine the ability of learners to manage ANOVA score model problems. The model assesses knowledge and skills with respect to procedural and semantic process explanation. Bayesian networks, which express these assessment purposes, have two phases which can be integrated into a unified model. Three practical models will be tested and named for top parent variable codes: ModelEquation,

ScoreModel, and ANOVAScoreModel2way. Before applying the models to survey data, their characteristics will be examined.

6.3.5.1 The ModelEquation Sub-Network Assessment of Performance Process

The ModelEquation phase is a part of the ANOVA score assessment model. It has 38 variables: 12 potential variables and 26 evidence variables. The maximum node layer is five. There are 5 connecting nodes from top parent node to bottom evidence nodes. The minimum number of node layers is 2, e.g. from ModelEquation to "=", the equal sign. There are two evidence nodes in the smallest clique "12_jk" and 3 evidence nodes in the largest clique "IndexValues." Figure 6.13 displays the detailed structure of this model.

In this model, the prior probability of the top parent variable being true is 0.5, and being false is 0.5. This assumes that in the absence of evidence, prior probability reflects complete uncertainty with respect to mastery of the ANOVA score model. Conditional probabilities are all set for "true given true" at 0.67; "false given false" at 0.67. This probability is satisfied by an assessment judgment that, give knowledge needed to complete the task represented by a child node, learners have a two-thirds chance of completing the specific task component.

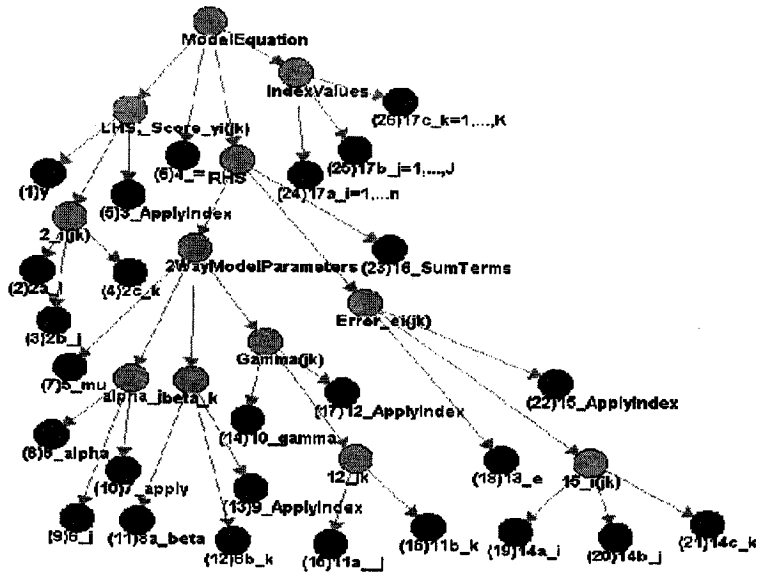

Figure 6.13 ModelEquation phase to assess performance process

The problem space is designated by:

$$\sum_{r=1}^{26} 2^r C_{26}^r = 2^1 C_{26}^1 + 2^2 C_{26}^2 + 2^3 C_{26}^3 + \ldots + 2^k C_{26}^k + \ldots + 2^{26} C_{26}^{26} \quad \text{(Formula 6-22)}$$

This is a huge problem space. In order to simulate evidence patterns instantiated in practice, a random sampling method was applied to test evidence states from one true evidence observation to 26 true evidence observations. When k true evidence observations were chosen, 26--k false evidences were automatically produced. After sampling, random evidence combinations were recorded (see Appendix D). For example, one true evidence instantiation with 25 false evidence instantiations was defined by randomly selecting the number 9 for the number of true evidence nodes. As soon as the evidence series was determined, the potential nodes for assessment were chosen.

Table 6.26. Updating Probabilities of Random Evidence combinations for ModelEquation Phase

No.	LHS	alpha_j	beta_k	Gamma(jk)	Error_ei(jk)	IndexValues	ModelEquation
1	0.08	0.22	0.07	0.09	0.08	0.07	0.09
2	0.11	0.07	0.22	0.09	0.08	0.07	0.10
3	0.11	0.22	0.07	0.13	0.08	0.07	0.10
4	0.12	0.06	0.06	0.08	0.17	0.31	0.37
5	0.30	0.07	0.07	0.12	0.27	0.58	0.25
6	0.11	0.24	0.24	0.14	0.15	0.08	0.14
7	0.34	0.54	0.07	0.14	0.11	0.25	0.18
8	0.27	0.36	0.36	0.72	0.37	0.08	0.16
9	0.19	0.25	0.25	0.39	0.11	0.57	0.22
10	0.61	0.68	0.14	0.74	0.49	0.09	0.25
11	0.55	0.29	0.59	0.43	0.64	0.12	0.51
12	0.57	0.69	0.69	0.24	0.15	0.37	0.56
13	0.18	0.69	0.69	0.25	0.80	0.26	0.21
14	0.48	0.16	0.90	0.77	0.83	0.38	0.60
15	0.50	0.74	0.74	0.84	0.21	0.61	0.35
16	0.82	0.92	0.75	0.85	0.21	0.10	0.29
17	0.90	0.65	0.65	0.71	0.54	0.74	0.78
18	0.44	0.78	0.93	0.87	0.88	0.65	0.48
19	0.83	0.77	0.93	0.72	0.66	0.76	0.84
20	0.89	0.44	0.92	0.70	0.66	0.93	0.89
21	0.92	0.93	0.93	0.53	0.67	0.93	0.90
22	0.88	0.93	0.93	0.64	0.91	0.69	0.63
23	0.59	0.94	0.94	0.92	0.91	0.89	0.60
24	0.92	0.73	0.91	0.88	0.92	0.93	0.90
25	0.89	0.94	0.94	0.92	0.92	0.93	0.90
26	0.92	0.94	0.94	0.91	0.92	0.93	0.91

Note: No. indicates the number of evidence combinations. For details on combinations, refer to Appendix D.

Given all the explanatory nodes in the network, seven were selected: LHS, alpha_j, beta_k, Gamma(jk), Error_ei(jk), IndexValues, and ModelEquation, These potential nodes reflect most of the knowledge components for assessing performance of the ANOVA score model problem task.

JavaBayes with incremental evidence combinations (each step increases positive evidence by one node) was used to update posterior probabilities of potential variables. The results are in Table 6.26.

Table 6.26 leads to several conclusions. First, as the number of instantiated evidence nodes increases, their probability values gradually increase. However, the probability values of potential are non-monotonically increasing.

Second, increases of instantiated evidences do not necessarily increase the entire performance level.

Third, the posterior network probabilities tend to become more stable as the number of instantiated evidences nodes increases.

Fourth, when the number of instantiated evidence nodes approaches 20, network posterior probabilities gradually tend towards a stable state. Probability values in each column almost maintain a monotonous increase except for random fluctuation due to step-wise random sampling of positive nodes.

Random incremental conditions in network node performances indicate that learners can, in principle, master the ANOVA score model problem incrementally. Students must master at least 76.92% (20/26) of knowledge and

skill nodes for their ability to solve the problem to become robust. Students demonstrate excellent mastery as they approach 92.30% (24/26).

6.3.5.2 The ScoreModel Sub-Network: Assessment of Semantic Explanation

The ScoreModel phase is the second part of the ANOVA score model assessment network. This submodel consists of 28 nodes: 9 potential variables and 19 evidence variables. The longest chain consists of four levels of nodes. The sub-network structure is in Figure 6.14.

In this model, the prior probability that the top parent node variable was true was set at 0.5 and the prior probability for false was also set at 0.5. This assumes that the prior probability reflects an assumption of complete uncertainty about mastery of the ANOVA score model in semantic explanation, in the absence of evidence from observed performance variables. The conditional probability that a child node is true given a true parent node was again set at 0.67; and the conditional probability for false was also set at 0.67. These probabilities satisfy the assessment judgment that learners have a two-thirds chance of completing this problem component, given the knowledge corresponding to its Bayesian parent node.

Before examining evidence effects, the problem space will be examined. Formula 6-23 gives the number of possible evidence space patterns.

$$\sum_{r=1}^{19} 2^r C_{19}^r = 2^1 C_{19}^1 + 2^2 C_{19}^2 + 2^3 C_{19}^3 + \ldots + 2^k C_{19}^k + \ldots + 2^{19} C_{19}^{19} \quad \text{(Formula 6-23)}$$

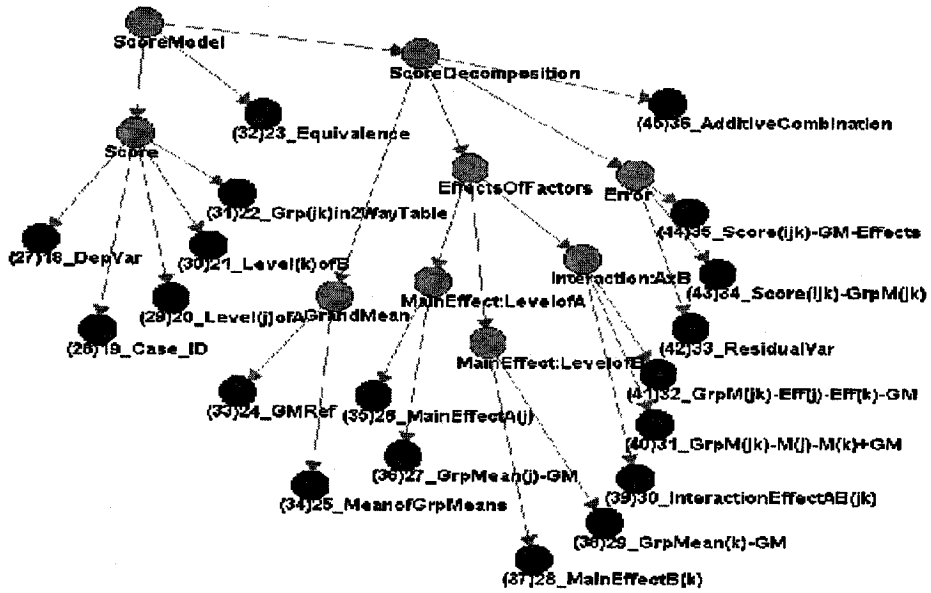

Figure 6.14. ScoreModel phase to assess semantic explanations

The instantiated evidence space is enormous. The same sampling method used for as the ModelEquation sub-network was used here, and a similar table of random evidence combinations was generated (see appendix E). For example, one true evidence instantiation with 25 false evidence instantiation was obtained by randomly selecting the number 31 for the number of true evidence nodes. The selected potential nodes to be assessed were Score, GrandMean, MaineffectA, MainEffectB, InteractAB, and Error. The updating probabilities of potential nodes were obtained based on the provided evidence combinations. Results are given in Table 6.27.

Table 6.27. Updating Probabilities of Random Evidence Combinations for ScoreModel

No.	Score	GrandMean	MainEffectA	MainEffctB	InteraAB	Error	ScoreModel
1	0.08	0.12	0.14	0.14	0.07	0.07	0.14
2	0.02	0.37	0.15	0.40	0.08	0.07	0.14
3	0.08	0.17	0.16	0.16	0.26	0.09	0.19
4	0.08	0.40	0.14	0.14	0.08	0.57	0.18
5	0.13	0.21	0.43	0.17	0.10	0.33	0.52
6	0.12	0.42	0.17	0.74	0.10	0.27	0.45
7	0.26	0.14	0.48	0.48	0.63	0.25	0.21
8	0.39	0.54	0.46	0.19	0.31	0.38	0.64
9	0.86	0.17	0.46	0.19	0.62	0.58	0.39
10	0.71	0.48	0.48	0.77	0.19	0.64	0.68
11	0.73	0.55	0.82	0.82	0.40	0.15	0.74
12	0.88	0.59	0.52	0.23	0.67	0.91	0.55
13	0.71	0.51	0.83	0.83	0.72	0.66	0.71
14	0.73	0.82	0.83	0.57	0.71	0.90	0.75
15	0.97	0.79	0.57	0.83	0.91	0.67	0.51
16	0.74	0.87	0.86	0.86	0.75	0.93	0.81
17	0.98	0.87	0.85	0.85	0.75	0.77	0.86
18	0.92	0.88	0.86	0.86	0.93	0.93	0.86
19	0.98	0.88	0.87	0.87	0.93	0.93	0.87

Note: No. indicates the number of evidence combination. For details on combinations, refer to Appendix E.

ScoreModel network results are similar to those obtained with ModelEquation network. As the updated posterior probability values increase, some instability and fluctuations in probabilities accompany the process. When

the number of instantiated evidence nodes approaches 17, the situation begins to change. When the network reaches 17, 18, or 19 true instantiated evidence nodes, the updated posterior probabilities indicate a robustly increasing trend. Quantitatively, if 89.47% (17/19) of knowledge nodes are mastered, a robust mastery level is established. In the case of the ModelEquation sub-network, only 76.92% of knowledge nodes have be mastered to establish robust mastery. The ScoreModel network requires mastery of 12.55% more nodes to attain a stable mastery level. This phenomenon indicates that a robust assessment result requires mastery of more components of the entire knowledge network when the total number of knowledge nodes is smaller. This implicitly suggests that we need to consider the size of the network when we require robust assessment results.

6.3.5.3 The Full Assessment Model to Examine both Performance and Semantic Explanation to ANOVA Score Model Learning

The full assessment model is composed of the ModelEquation phase and the ScoreModel phase. The purpose of the assessment is to examine the mastery of knowledge and skills for solving an ANOVA score model problem. This model has 68 evidential and potential variables in all: 1 top level node, 22 potential variable nodes, and 45 evidence variable nodes. The longest chain spans 5 nodes. The structure is illustrated in Figure 6.15.

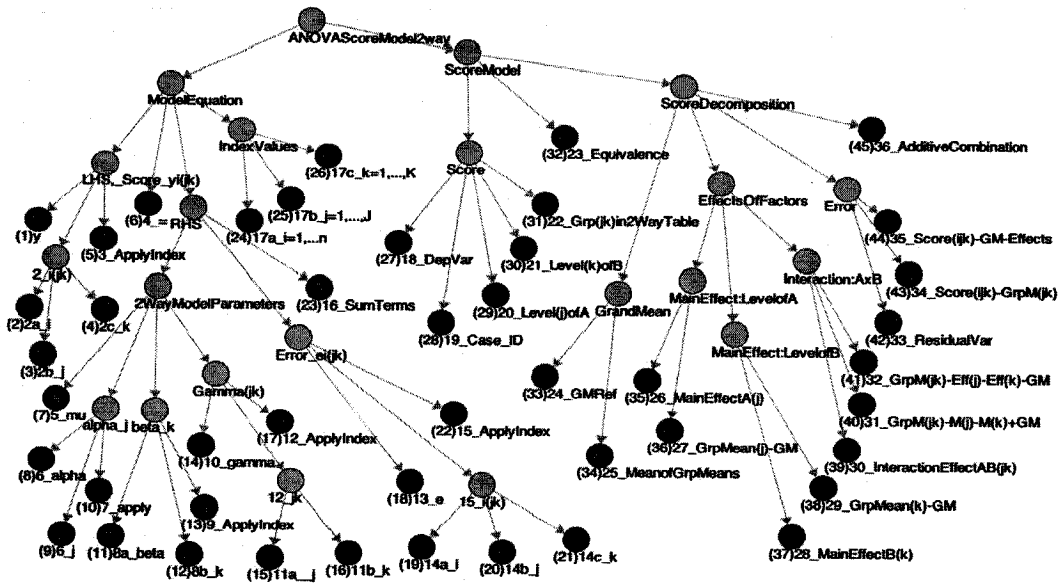

Figure 6.15. The full assessment model to examine the performance and semantic explanations.

Formula 6.24 represents the evidence space.

$$\sum_{r=1}^{45} 2^r C_{45}^r = 2^1 C_{45}^1 + 2^2 C_{45}^2 + 2^3 C_{45}^3 + \ldots + 2^k C_{45}^k + \ldots + 2^{45} C_{45}^{45} \quad \text{(Formula 6.24)}$$

This is a very large evidence space. In order to test the full assessment model, a sampling method comparable to that used for the score model was applied, and similarly a table of random evidence combinations was written (see appendix F). The potential nodes for estimation chosen were: ModelEquation (MEq), ScoreModel (SMo), LHS, RHS, IndexValues (IVa), Score, ScoreDecompostion (ScDe) and ANOVAScoreModel2way (ASMo).

The prior probabilities of ASMo for true and false were both 0.5. Conditional probabilities are set at 0.67 for true and true combination.

Table 6.28. Updating Probabilities of Random Evidence Combinations for Full Model

#	Top node ModelEquation (Performance Model)				Top Node Score Model (Semantic Model)			Top node of joined network
	Meq	LHS	RHS	IVa	Smo	Score	ScDe	ASMo
1	0.08	0.08	0.09	0.07	0.12	0.07	0.07	0.25
2	0.08	0.08	0.09	0.07	0.11	0.02	0.07	0.25
3	0.08	0.08	0.09	0.07	0.13	0.07	0.11	0.26
4	0.08	0.10	0.09	0.07	0.12	0.02	0.11	0.26
5	0.08	0.08	0.10	0.07	0.12	0.02	0.13	0.26
6	0.12	0.26	0.13	0.07	0.17	0.24	0.17	0.28
7	0.21	0.10	0.12	0.83	0.17	0.08	0.25	0.31
8	0.09	0.10	0.15	0.07	0.11	0.02	0.09	0.26
9	0.14	0.32	0.11	0.08	0.37	0.11	0.13	0.35
10	0.12	0.08	0.11	0.23	0.43	0.11	0.32	0.37
11	0.12	0.11	0.11	0.23	0.33	0.60	0.41	0.33
12	0.14	0.11	0.22	0.24	0.17	0.25	0.11	0.29
13	0.45	0.36	0.23	0.33	0.26	0.09	0.55	0.41
14	0.33	0.15	0.20	0.10	0.47	0.34	0.15	0.44
15	0.22	0.76	0.18	0.09	0.26	0.27	0.43	0.34
16	0.26	0.47	0.41	0.28	0.19	0.25	0.15	0.33
17	0.41	0.13	0.36	0.32	0.20	0.25	0.15	0.38
18	0.14	0.17	0.12	0.24	0.42	0.63	0.68	0.37
19	0.43	0.43	0.76	0.63	0.16	0.08	0.17	0.37
20	0.21	0.18	0.60	0.09	0.53	0.36	0.35	0.42
21	0.64	0.60	0.60	0.39	0.35	0.30	0.63	0.50
22	0.63	0.20	0.59	0.69	0.48	0.65	0.71	0.53
23	0.28	0.55	0.24	0.28	0.75	0.73	0.75	0.51
24	0.56	0.47	0.77	0.88	0.67	0.70	0.39	0.57
25	0.26	0.47	0.33	0.28	0.46	0.64	0.77	0.42

Table 6.28. (continuted)

#	Top node ModelEquation (Performance Model)				Top Node Score Model (Semantic Model)			Top node of joined network
	Meq	LHS	RHS	IVa	Smo	Score	ScDe	ASMo
26	0.78	0.64	0.46	0.91	0.76	0.91	0.42	0.66
27	0.68	0.68	0.34	0.71	0.40	0.62	0.45	0.52
28	0.80	0.65	0.57	0.92	0.76	0.91	0.42	0.67
29	0.79	0.65	0.82	0.74	0.46	0.87	0.40	0.58
30	0.77	0.32	0.71	0.91	0.79	0.92	0.58	0.67
31	0.79	0.82	0.88	0.44	0.59	0.89	0.78	0.62
32	0.28	0.25	0.50	0.28	0.84	0.98	0.84	0.54
33	0.68	0.89	0.78	0.90	0.59	0.89	0.82	0.58
34	0.77	0.39	0.75	0.91	0.47	0.64	0.64	0.57
35	0.89	0.89	0.87	0.93	0.42	0.32	0.79	0.60
36	0.87	0.89	0.83	0.77	0.88	0.93	0.92	0.73
37	0.80	0.71	0.66	0.74	0.88	0.98	0.90	0.71
38	0.91	0.92	0.89	0.93	0.60	0.89	0.78	0.65
39	0.91	0.92	0.83	0.93	0.87	0.93	0.89	0.74
40	0.92	0.92	0.91	0.93	0.82	0.92	0.66	0.73
41	0.86	0.73	0.90	0.76	0.88	0.98	0.89	0.72
42	0.88	0.68	0.90	0.93	0.89	0.98	0.92	0.73
43	0.92	0.92	0.91	0.93	0.88	0.93	0.93	0.74
44	0.92	0.92	0.91	0.93	0.89	0.98	0.93	0.75
45	0.92	0.92	0.91	0.93	0.89	0.98	0.93	0.75

Table 6.28 shows that knowledge and skill levels of the full assessment model network display in two stages. Stage 1 is basic mastery, which is above Row 36. The results will be found in column ASMo which stands for the general

problem solving level. In this column, except for Row 38 which is 0.65, all values increase in a stable fashion. Stage 2 approaches excellent mastery which is above Row 41. ASmo keeps increasing with stability. Quantitatively, Level 1 is greater than 80% of the nodes (36/45); the advanced level is above 91.11% of the nodes (41/45). These results indicate that the network can provide stable benchmarks for assessing mastery levels of problem solvers.

The fundamental structures of Bayesian nets contained in the ANOVA score model network have been explored. This covers simple cliques including one-parent one-child, one-parent two-child, one-parent three-child, and one-parent four-child cliques. Multi-level cliques and complex Bayesian networks were also examined. The features acquired from these net segments are very useful for understanding genuine Bayesian assessment systems developed for recognizing knowledge components and skills. Finally, a two-parent parent-child relation compound model was examined and an inverse symmetry was observed.

Following this, three assessment networks were tested by sampling evidence instantiated from possible evidence spaces. The first two networks used simulated data to examine assessments of performance and semantic explanations for solving ANOVA score model problems independently. The complete model, a combination of these two sub-networks was also tested using simulated data to assess performance and semantic explanations. In Chapter 7, these networks will be applied to data to examine the knowledge and skills of 20 students learning to solve an ANOVA score model problem.

CHAPTER SEVEN: APPLICATION OF THE BAYESIAN NETWORK TO COGNITIVE ASSESSMENT OF STUDENTS' PERFORMANCE

Chapter seven will present the frequency distributions of participant observed performance scores and will introduce estimated log-odds of mastery as measures of student proficiency, that is, of their progress toward achieving mastery of component skills and knowledge in the domain of performance. Sections 7.1 to 7.5 examine a natural way for representing estimated posterior probabilities as transformations of odds ratios to log odds ratios. Section 7.1 reports the frequency distribution and descriptive statistics of observed participants' performance scores. Section 7.2 analyzes model estimates of log odds of mastery for high-level explanatory variables in the Bayesian model. Section 7.3 focuses on log odds ratios for mastery of explanatory variables corresponding to complex net-cliques in the hierarchical model. Section 7.4 investigates relationships of log odds ratio estimates of proficiency to external variables. Section 7.5 examines correlations between raw scores and log odds ratios. Finally, section 7.6 provides an analysis of robustness and reports results of a global neighbourhood robustness analysis with ε-contaminated probabilities that produce upper and lower bound posterior probability estimates. The analysis uses the JavaBayes program and robustness algorithm (Cozman, 1999).

7.1 Frequency Distribution of Participants' Observed Performance Scores and Estimated Log-Odds of Mastery (ELOM)

Twenty intermediate statistics students voluntarily took part in this study of building an assessment model. All students registered in an intermediate statistics course had opportunities to solve ANOVA problems. They had time and opportunity to access an on-line tutoring system to assist their learning (Frederiksen and Donin, 2005). Participants with two types of background experience defined the two external variables used to predict log odds of mastery of ANOVA score model knowledge and problem solving skills (in Section 7.4). Section 7.1 reports the pooled frequency distributions of (a) observed participant performance scores, and (b) estimates of participant log-odds of mastery based on using their new performance scores as evidence variables to update Bayesian network estimates of posterior probabilities of explanatory variables (model constructs).

7.1.1 Frequency Distribution of Participant's Observed Performance Scores

The following is a summary of the observed scores of the 20 participants on the performance and semantic explanation components of the ANOVA score model task (see Table 7.1). The scores indicate the distribution of student performance and semantic explanation scores as in a traditional test.

Table 7.1. Participants Performance and Semantic Explanation Score

Student No.	Performance	Semantic Explanation	Total
1	6	6	12
2	23	7	30
3	15	5	20
4	17	5	22
5	12	4	16
6	23	12	35
7	23	3	26
8	23	5	28
9	22	7	29
10	23	6	29
11	19	5	24
12	23	5	28
13	23	4	27
14	23	6	29
15	6	2	8
16	23	6	29
17	23	7	30
18	23	5	28
19	23	6	29
20	8	4	12

Table 7.2 presents average student scores on performance and semantic explanation subtasks. Standard deviation, minimum and maximum values are also reported. Distribution of data was not expected to conform closely to a normal distribution because there were only 20 students in the sample.

Table 7.2 Descriptive Statistics of Performance and Semantic Explanations

Variable	N	Mean	Ste Dev	Minimum	Maximum
Performance	20	19.00	6.16	6.00	23.00
Explanation	20	5.45	1.85	2.00	12.00
Total proficiency	20	24.55	7.23	8.00	35.00

Tables 7.1 and 7.2 present score distributions of 20 participants regarding performance, semantic explanations, and general proficiency scores. In the performance phase, minimum, maximum and mean values are 6.00, 23.00, and 19.00 respectively. In the semantic explanation, minimum, maximum, and mean values are 2.00, 12.00, and 5.45 respectively. For general proficiency, minimum, maximum and mean values are 8.00, 35.00, and 24.55 respectively. The largest standard deviation is for general proficiency and the smallest for explanations.

The raw scores (shown in Tables 7.1) consist of the total number of Solution Feature components scored as complete in scoring students. Student scores were treated as if they had taken a conventional test and received raw total scores reported above. However, a model-based assessment is based on an organized knowledge and skill-based assessment framework. The assessment takes the form of estimates of posterior likelihoods that students have mastered the knowledge and skill components represented by explanatory variable nodes in the Bayesian assessment network. This estimate is based on evidence variable nodes that reflect scored Solution Feature components reflected in student performance. Evidence variables have different weights based on their network positions. A knowledge or skill component is not an

isolated piece of data; it is closely related to other knowledge and skill network components. Therefore, raw scores cannot properly discriminate student mastery of different knowledge and skill components which must be considered as model-based knowledge network components.

7.1.2 Log-Odds of Mastery as Measures of Proficiency

Raw scores on performance and semantic explanation subtasks identified in the model-based assessment framework do not adequately explain the student performance and behaviour as they solve the ANOVA score model problem. Raw scores are based on sums of values of evidence variables. In a Bayesian network, network variables can be used to infer posterior probabilities by updating the network using instantiated evidence variables.

For example, as defined in Section 6.2.6.2, for the top-level parent, the variable space was defined as having two mutually exclusive states: (a) mastery, where students have mastered the ANOVA score model task; and (b) non-mastery, where students have not mastered the ANOVA score model task. The prior probabilities of mastery and non-mastery sum to one. The odds ratio or the probability of mastery divided by the probability of non-mastery is a measure of the probability of mastery relative to the probability of non-mastery.

The statistical distribution of observed responses is usually assumed to be binomial (or multinomial) in terms of option spaces for each individual evidence variable (i.e. dichotomous or polytomous response options). Two posterior probabilities in odds ratios are obtained for each explanatory variable after

updating the network based on evidence variables. Posterior probabilities are usually assumed to have a Dirichlet prior distribution (Fredette & Angers, 2002). Under these assumptions, asymptotic distributions of log odds ratios have been studied and the unit normal density has been found to approximate this distribution (Fredette & Angers, 2002).

Tables 7.3, 7.4 and 7.5 present several measures of student proficiency based on (a) procedural performance (7.3), (b) semantic explanations (7.4), and (c) general performance on the ANOVA score model tasks (7.5). In Table 7.3, Student 1 has an estimated posterior probability of mastery on ModelEquation sub-task of 0.4399, an odds ratio of 0.7854, and a log odds ratio of -0.2415. Student 2 has an estimated posterior probability on ModelEquation of 0.7653, an odds ratio of 3.2608, and a log odds ratio of 1.1820. If the posterior probability of mastery is used as a measure of proficiency, a difference of 0.3254 is obtained for Student 1 and Student 2. However, these two probabilities do not follow a normal distribution, and the distribution of their difference is complex. When log odds ratios are used as measures of proficiency (-0.2415 and 1.1820 respectively), the difference is 1.4235. Log odds scores follow an approximately normal distribution in the population, and differences are meaningful since log odds ratios constitute an interval scale. Thus, differences in log odds of mastery can be interpreted as differences in proficiency.

Advantages of using the log odds measure can be seen in comparing student performance on ModelEquation model, ScoreModel model, and

ANOVAScoreModel-2way. The scores are presented in Tables 7.3, 7.4 and 7.5 respectively.

Table 7.3. Raw Scores, Model-estimated Odds Ratios, and Log Odds Ratios of Procedural Performance Construct (ModelEquation Sub-task Model)

ID	Raw Score	Estimated posterior probability of Mastery	Estimated posterior probability of Non-mastery	Odds Ratios	Log Odds Ratio
1	6	0.4399	0.5601	0.7854	-0.2415
2	23	0.7653	0.2347	3.2608	1.1820
3	15	0.6422	0.3578	1.7949	0.5849
4	17	0.6188	0.3812	1.6233	0.4845
5	12	0.5142	0.4858	1.0585	0.0569
6	23	0.7654	0.2346	3.2626	1.1825
7	23	0.7654	0.2346	3.2626	1.1825
8	23	0.7654	0.2346	3.2626	1.1825
9	22	0.7653	0.2347	3.2606	1.1819
10	23	0.7654	0.2346	3.2626	1.1825
11	19	0.7549	0.2451	3.0780	1.1243
12	23	0.7654	0.2346	3.2626	1.1825
13	23	0.7654	0.2336	3.2626	1.1825
14	23	0.7654	0.2346	3.2626	1.1825
15	6	0.1800	0.8200	0.2195	-1.5164
16	23	0.7654	0.2346	3.2626	1.1825
17	23	0.7654	0.2346	3.2626	1.1825
18	23	0.7654	0.2346	3.2626	1.1825
19	23	0.7654	0.2346	3.2626	1.1825
20	8	0.1734	0.8266	0.2098	-1.5616

Table 7.4. Raw Score and Odds Ratio of Semantic Explanation (ScoreModel Sub-Task Model)

ID	Raw Score	Estimated posterior probability of Mastery	Estimated posterior probability of Non-mastery	Odds Ratios	Log Odds Ratio
1	6	0.4398	0.5602	0.7851	-0.2419
2	7	0.1914	0.8086	0.2367	-1.4410
3	5	0.1600	0.8400	0.1905	-1.6581
4	5	0.1550	0.8450	0.1834	-1.6961
5	4	0.1550	0.8450	0.1834	-1.6961
6	12	0.4371	0.5629	0.7765	-0.2530
7	3	0.1434	0.8566	0.1674	-1.7874
8	5	0.1551	0.8449	0.1836	-1.6950
9	7	0.2936	0.7064	0.4156	-0.8780
10	6	0.1710	0.8290	0.2063	-1.5784
11	5	0.1550	0.8450	0.1834	-1.6961
12	5	0.1550	0.8450	0.1834	-1.6961
13	4	0.1449	0.8551	0.1695	-1.7749
14	6	0.1609	0.8391	0.1918	-1.6513
15	2	0.1443	0.8557	0.1686	-1.7802
16	6	0.1710	0.8290	0.2063	-1.5784
17	7	0.2205	0.7795	0.2829	-1.2627
18	5	0.1551	0.8449	0.1836	-1.6950
19	6	0.1710	0.8290	0.2063	-1.5784
20	4	0.1420	0.8580	0.1655	-1.7988

Table 7.5. Raw Score and Odds Ratio of Pooled Mastery (ANOVA ScoreModel2way)

ID	Raw Score	Estimated posterior probability of Mastery	Estimated posterior probability of Non-mastery	Odds Ratios	Log Odds Ratios
1	12	0.3547	0.6453	0.5497	-0.5984
2	30	0.4847	0.5153	0.9406	-0.0612
3	20	0.4312	0.5688	0.7581	-0.2769
4	22	0.4216	0.5784	0.7289	-0.3162
5	16	0.3838	0.6162	0.6228	-0.4735
6	35	0.5694	0.4306	1.3223	0.2794
7	26	0.4643	0.5357	0.8667	-0.1431
8	28	0.4717	0.5283	0.8929	-0.1133
9	29	0.5171	0.4829	1.0708	0.0684
10	29	0.4774	0.5226	0.9135	-0.0905
11	24	0.4680	0.5320	0.8797	-0.1282
12	28	0.4717	0.5283	0.8929	-0.1133
13	27	0.4681	0.5319	0.8801	-0.1277
14	29	0.4739	0.5261	0.9008	-0.1045
15	8	0.2817	0.7183	0.3928	-0.9345
16	29	0.4775	0.5225	0.9139	-0.0900
17	30	0.4950	0.5050	0.9802	-0.0200
18	28	0.4718	0.5282	0.8932	-0.1129
19	29	0.4775	0.5225	0.9139	-0.0900
20	12	0.2950	0.7050	0.4184	-0.8713

Thus, although posterior probabilities do discriminate student problem solving performance on the ANOVA score model, log odds measures are

preferable for assessing student proficiency since they can be analyzed as normally distributed random variables. Therefore, log odds ratios were calculated from the odds ratios of estimated posterior probabilities. Log odds ratios are known to have an approximately normal distribution with a mean of zero (representing uncertainty about the student state of mastery). Increasingly negative values represent decreasing likelihoods of mastery, and increasingly positive values indicate increasing likelihoods of mastery. Estimated log odds of mastery were used to measure student levels of proficiency, where proficiency is thought of as progress toward achieving a state of mastery on any component skill of the knowledge structure.

7.2 Analysis of Model Estimates of Log Odds of Mastery of Top Level (general) and Sub-task Explanatory Variables Reflecting Different Evidence Patterns

Instantiated values of log odds ratios for explanatory variables approximately follow a normal distribution, although more accurate approximations of posterior distribution have been studied and found to have subtle differences in terms of different approximation methods (Fredette & Angers, 2002). The density of posterior log-odds ratios approaches the standard normal distribution as sample sizes increase.

Normalization of odds ratios enables meaningful estimations of student proficiency levels on interval scales, i.e., their log odds likelihood of mastery with respect to their knowledge and skill components (see Table 7.6). Note that a zero log odds value corresponds to a posterior probability of mastery of 0.50 which

represents complete uncertainty with respect to student mastery of knowledge and skill components. Thus, log odd ratios scale values measure the likelihood that students have mastered or failed to master knowledge and skill components associated with potential explanatory variables in the Bayesian network.

Table 7.6. Relationships of Posterior Probabilities, Odds Ratios and Log-Odds Ratios

Probabilities	Odds ratios	Log odds ratios
0.0474	0.0498	-3.000
0.0500	0.0526	-2.945
0.0705	0.0758	-2.580
0.1235	0.1409	-1.960
0.2690	0.3679	-1.000
0.5000	1.0000	0.000
0.7311	2.7183	1.000
0.8765	7.0993	1.960
0.9296	13.1971	2.580
0.9500	19.0000	2.945
0.9526	20.0855	3.000
0.9900	99.0000	4.5951

In log odds ratios, a negative value reflects the likelihood that students have not mastered knowledge and skill components, with negatively increasing values indicating greater certainty that they have not mastered the components. A positive value indicates likelihood that students have mastered the components, with increasing values indicating greater certainty that they have

mastered the components. Log odds ratio scores can be converted to normal percentiles as another measure for assessing student progress toward mastery.

Table 7.3 in Section 7.1.2 reports raw scores, posterior probabilities of mastery, odds ratios and log odds ratios for ModelEquation. The data represent the performance of 20 students on the ANOVA score model problem.

Table 7.4 in Section 7.1.2 reports raw scores, posterior probabilities of mastery, odds ratios and log odds ratios for ScoreModel, where the data represent semantic explanations of 20 students on the ANOVA score model problem. Table 7.5 in Section 7.1.2 reports raw scores, posterior probabilities of mastery, odds ratios and log odds ratios for ANOVA ScoreModel2way.

The data show that with respect to student mastery of the ANOVA score model problem, conclusions that are drawn from log odds ratios based on posterior probabilities, estimated by updating the Bayesian network, differ from raw score based conclusions. The former provides measures of cognitive components of performance, and semantic explanation, that provide more plausible assessment information on which to base diagnostic decisions. Several cases in Tables 7.3 and 7.4 illustrate how they different from traditional raw scores.

In Table 7.3 Student 3 received a raw score of 15 and Student 4 received a raw score of 17. However, the mastery probability of Student 3 is higher than that of Student 4. The log odds ratio of Student 3 (0.5849) is larger than the log odds ratio of Student 4 (0.4845).

Table 7.4 shows a similar situation in which Students 2 and 9 receive the same raw score of 7, but their mastery probabilities are 0.1914 and 0.2936 respectively, and their log odds ratios are -1.4410 and -0.8780 respectively. The likelihood of mastery for Student 9 is higher than that of Student 2.

Table 7.4 also shows that Student 1 received a raw score of 6, and Student 2 received a raw score of 7. However, their mastery probabilities are 0.4398 and 0.1914 respectively, and their log odds ratios are -0.2419 and -1.4410 respectively. The likelihood of mastery of Student 1 is higher than that of Student 2, although Student 2's raw score is higher than that of Student 1. These two cases illustrate that log ratios based on posterior probabilities estimated by updating the Bayesian network provides model-based assessment information that differs from raw scores, rendering the data potentially much more valid and effective as it reflects the structure of knowledge represented by the Bayesian network.

7.3 Log-Odds of Mastery of Explanatory Variables (net cliques) that Correspond to Components of the Hierarchical Model

Students' log odds ratios for such high level components as: (a) performance, (b) semantic explanation, and (c) general mastery of ANOVA score models were presented in section 7.2. The results presented correspond to "macro" level nodes in the Bayesian assessment network. The estimate of log odds ratios at the level of cliques or net cliques are of importance since they provide cognitive diagnostic information that is a principal aim for using Bayesian

assessment networks. In this section, examples of net cliques will be chosen and estimated log odds ratios for parent nodes of these structures will be examined. Parent nodes reflect proficiencies associated with the clique as a whole. Two net cliques were selected: Error_ei(jk), a component of ModelEquation and EffectsOfFactors, a component of ScoreModel.

7.3.1 Estimated Log Odds Ratios for a Complex Net Clique: Erro_ei(jk)

Error_ei(jk) is a complex net clique in performance model, ModelEquation. It consists of seven nodes: two explanatory variables: Error_ei(jk) and 15_i(jk). 15_i(jk) has three evidence nodes: (19)14a_i, (20)14b_j, and (21)14c_k. Error-ei(jk) has one explanatory variable, 15_i(jk) and two evidence variables, (18)13_e and (22)15_ApplyIndex. If Error-ei(jk) is an up-explanatory variable, then there are five evidence variables sending information to it. The Bayesian network of Error-ei(jk) is presented in Figure 7.1.

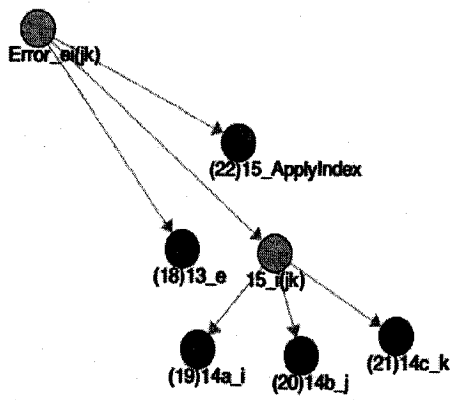

Figure 7.1. Complex clique error_ei(jk) with evidence notes

Raw performance scores from the 20 students are listed in Table 7.7 along with mastery probabilities, odds ratio and log odds ratios.

Table 7.7. Raw Score, Odds Ratio and Log Odds Ratio of Clique Error_ei(jk)

ID of the Students	Performance Score	Posteriior Probabilities	Odds Ratios	Log Odds Ratios
1	0	0.0904	0.0994	-2.3086
2	5	0.9172	11.0773	2.4049
3	3	0.6409	1.7847	0.5793
4	1	0.3510	0.5408	-0.6147
5	3	0.5127	1.0521	0.0508
6	5	0.9172	11.0773	2.4049
7	5	0.9172	11.0773	2.4049
8	5	0.9172	11.0773	2.4049
9	5	0.9172	11.0773	2.4049
10	5	0.9172	11.0773	2.4049
11	5	0.9118	10.3379	2.3358
12	5	0.9172	11.0773	2.4049
13	5	0.9172	11.0773	2.4049
14	5	0.9172	11.0773	2.4049
15	1	0.2823	0.3933	-0.9332
16	5	0.9172	11.0773	2.4049
17	5	0.9172	11.0773	2.4049
18	5	0.9172	11.0773	2.4049
19	5	0.9172	11.0773	2.4049
20	2	0.4631	0.8625	-0.1479

Log odds ratios of student performance on clique "Error_ei(jk)" represent results similar to results presented in section 7.2. The fact that two examinees have the same raw scores does not mean that the log odds ratios of their performance will be the same. For example, Students 3 and 5, both have

performance scores of 3, but their log odds ratios are 0.5793 and 0.0508, respectively. This indicates that raw scores only represent how many evidence variables have been successfully instantiated. They cannot reflect the details and the importance of individual evidence nodes. Log odds ratios can represent patterns of evidence information. Based on this information, diagnostic details can be traced back.

7.3.2 Estimated Log Odds Ratios for a Complex Net Clique: EffectsOfFactors

EffectsOfFactors is a complex net clique in the semantic explanation submodel: ScoreModel. It consists of eleven nodes: 4 explanatory variables: EffectOfFactors, MainEffect:LevelofA, MainEffect:LevelofB, and Interaction:AxB. MainEffect:LevelofA has two evidence nodes: Numbers 35 and 36; MainEffect:LevelofB has two evidence nodes: Numbers 37 and 38; and Interaction:AxB has three evidence nodes: Numbers 39, 40 and 41. EffectsOfFactors is a parent explanatory variable in a net clique in which there are 7 evidence variables sending information to EffectsOfFactors as shown in Figure 7.2. Raw performance scores from the 20 students as listed in Table 7.8 together with estimated posterior probabilities of mastery, odds ratios, and log odds ratios.

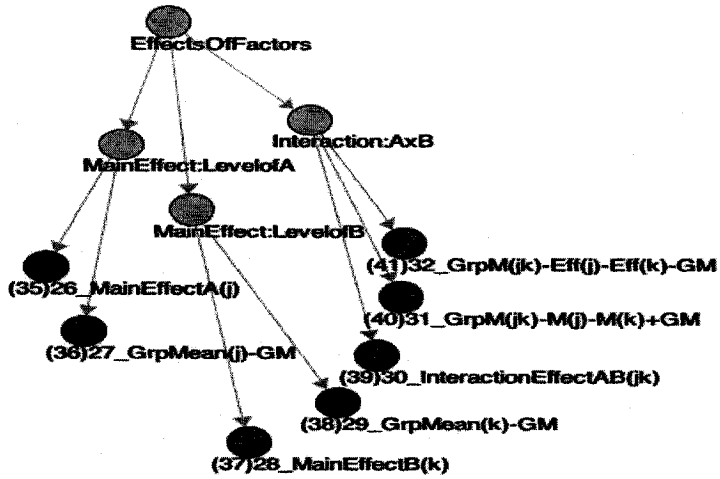

Figure 7.2. Complex clique EffectsOfFactors with its evidence nodes

These assessment variables for clique EffectsOfFactors provide different assessment information (see Table 7.8). Three examinees had raw scores of 5, 4, and 0. The other seventeen had raw scores of 3. However, the log odds ratios were different for almost all students ranging from –1.7958 to 0.4322 revealing again that raw scores do not represent differences in cognitive components underlying student performance.

In summary, using Bayesian networks to assess general mastery in writing ANOVA score models, including performance and semantic explanation submodels, the analyses of log odds ratios show that they discriminate details of student proficiency that raw scores are unable to detect. Thus, log odds ratios produced by updating the Bayesian network with student evidence patterns produces diagnostic rich information about the cognitive components of student proficiency.

Table 7.8. Raw Score, Odds Ratio and Log Odds Ratio of Clique EffectsOfFactors

ID of the Student	Semantic Explanation	Posterior Probability	Odds Ratios	Log Odds Ratio
1	3	0.3539	0.5477	-0.6020
2	3	0.3753	0.6008	-0.5095
3	3	0.3319	0.3319	-0.6996
4	3	0.3457	0.5284	-0.6379
5	3	0.3307	0.4941	-0.7050
6	5	0.6064	1.5407	0.4322
7	3	0.3153	0.4605	-0.7754
8	3	0.3457	0.5284	-0.6379
9	3	0.3440	0.5244	-0.6455
10	3	0.3488	0.5356	-0.6244
11	3	0.3457	0.5284	-0.6379
12	3	0.3457	0.5284	-0.6379
13	3	0.3307	0.4941	-0.7050
14	4	0.4567	0.4567	-0.1736
15	0	0.1424	0.1660	-1.7958
16	3	0.3472	0.5319	-0.6313
17	3	0.3517	0.5425	-0.6116
18	3	0.3457	0.5284	-0.6379
19	3	0.3472	0.5319	-0.6313
20	3	0.3264	0.4846	-0.7244

7.4 Relationships of Log-Odds Estimates of Proficiency to External Variables

In this section, scores on assessment variables (i.e., log odds ratio estimates of proficiency for particular explanatory constructs) were predicted

from two external variables characterizing participants. The focus is on relationships between student background variables and seven critical assessment variables. LHS, RHS and Index Values (IMD) are top-level components of ModelEquation (MEQ). Score, and ScoreDecompostion (ScDe) are top-level components of ScoreModel (Smo). ModelEquation and ScoreModel are the two submodels of ANOVAScoreModel2Way. These critical explanatory variables that were investigated correspond to the seven explanatory nodes listed above. Data analyzed consisted of estimated log odds ratios based on updating the Bayesian network for each student.

7.4.1 Effects of External Variables and Explanatory Variables on Estimated Log Odds Ratios: ModelEquation, LHS, RHS and IndexValues

In the performance model, ModelEquation (Meq), LHS, RHS and IndexValues (IND) were the four within-subject target variables. Dependent variables consisted of log odds ratios scores. Between-group independent variables were student background variables: using the tutor system (UT: 1=yes, 2=no) and taking a statistics course (SC: 1=yes, 2=no). Repeated measures MANOVA were carried out with explanatory variables (Meq, LHS, RHS and IND) as the within-subject (repeated measure) factor. The results indicated that UT and SC were no statistically significant effects of the dependent variables, nor were there any significant interactions with explanatory variables. However, there were significant differences within-subject among LHS, RHS and IND. Table 7.9 reports results of subject contrasts between the explanatory variables. The

results show that there were no significant differences between LHS and RHS, or between LHS and IND. However, there was a significant difference between RHS and IND. The differences between Meq and the other three variables were all significant. Table 7.10 provides mean and relevant descriptive details to help understand relations among these variables.

Table 7.9. Within-subject Contrasts among ModelEquation, LHS, RHS and IND Log Odds Ratio Scores from Repeated Measures Analysis MANOVA

Source	D F	Multivariate F value	Pr>F
Main effect explanatory variable	3, 14	44.31	<0.0001
Meq vs LHS	1, 16	25.92	<0.0001
LHS vs RHS	1, 16	0.92	0.3530
RHS vs IND	1, 16	69.86	<0.0001
Meq vs RHS	1, 16	5.93	0.0270
Meq vs IND	1, 16	137.69	<0.0001
LHS vs IND	1, 16	0.18	0.6798

Table 7.10. Descriptive Statistics for Estimated Log Odds Ratios for Four Explanatory Variables in the ModelEquation Clique

Variable	N	Mean	Std Dev	Minimum	Maximum
MEQ	20	0.7124540	0.8837510	-1.5715200	1.1825200
LHS	20	1.5348815	1.2306079	-2.2918500	2.1828600
RHS	20	1.2782640	1.2654482	-1.6258500	2.0298900
IND	20	-1.7679090	0.2531364	-2.4450700	-1.6397400

7.4.2. Effects of External Variables and Explanatory Variable on Log Odds Ratios: Score Model, Score and ScoreDecomposition

In the semantic explanation model, ScoreModel, Score, and ScoreDecomposition were the within-subjects variables. Dependent variables consisted of log odds ratios scores. The between-group independent variables were UT and SC. A repeated measure MANOVA was carried out with explanatory variables as the within-subject factor. The results indicated that effects UT, SC and the UT by SC interaction were not statistically significant, nor did they interact with the within-subjects variables. Table 7.11 indicates there was a significant main effect of Explanatory Variables (the within-subject factor), and there were significant contrasts between ScoreModel and Score, and between Score and ScoreDecomposition, but no significant difference between Scoremodel and ScoreDecompostion.

Table 7.11. Contrasts among Smo, Score and ScDe of Log Odds RatioScores from Repeated Measures MANOVA

Source	DF	F value	Pr>F
Main effect of explanatory variables	2, 15	3.24	0.0679
Smo vs Score	1, 16	6.84	0.0188
Score vs Scde	1, 16	6.77	0.0192
Smo vs ScDe	1, 16	1.08	0.3152

Table 7.12 provides mean and relevant descriptive details which help understand relations among variables.

Table 7.12. Descriptive Statistics for Estimated Log Odds Ratios for Three Explanatory Variables in the ScoreModel Clique

Variable	N	Mean	Std Dev	Minimum	Maximum
Smo	20	-1.4812515	0.4720172	-1.8626000	-0.2419700
Score	20	-2.7851400	1.6036146	-3.9072400	1.8643200
ScDe	20	-1.4491825	0.3718885	-2.1664900	-0.3285200

7.4.3. *Effects of External Variables and Explanatory Variables on Estimated Log Odds Ratios:* ModelEquation, ScoreModel, and ANOVAScoreModel2way

The full proficiency model includes ANOVAScoreModel2way (ASM), ModelEquation (Meq), and ScoreModel (Smo). These were the levels of the within group factor, and UT and SC were the between-group independent variables. A repeated measure MANOVA was carried out with Explanatory Variables as the within-subjects factor. The results indicated that effects of UT, SC and the UT by SC interaction were not statistically significant, nor did they interact with the explanatory variables. However, the main effect of Explanatory Variables (the within-subjects factor) was statistically significant. Table 7.13 reports pairwise contrasts among these within-subject variables. The results indicated that significantly different levels of mastery of the two submodels (Meq vs. Smo), and mastery levels of the submodels that differed from the mastery level for the entire model (ASM) contribute to the general mastery model in different weights.

Table 7.13. Contrasts among ModelEquation, ScoreModel and ANOVAScoreModel2way of Log Odds Ratio in MANOVA

Source	DF	F value	Pr>F
Main effect	2, 15	79.96	<0.0001
Meq vs Smo	1, 16	76.84	<0.0001
Smo vs ASM	1, 16	169.90	<0.0001
Meq vs ASM	1, 16	16.87	0.0008

Table 7.14 provides mean and relevant descriptive details which help in understanding the relations among those variables.

Table 7.14. Descriptive Statistics for Three Explanatory Variables in the ScoreModel Clique

Variable	N	Mean	Std Dev	Minimum	Maximum
Meq	20	0.5809360	0.8402908	-1.6814800	1.1532300
Smo	20	-1.4779565	0.4202720	-1.9450000	-0.1297800
ASM	20	-0.2159060	0.2969554	-0.9360400	0.2794000

On the basis of three repeated measure MANOVAs, several points may be summarized. First, the external variables UT, SC and their interaction UT*SC had no statistically significant effects on the log odds ratios, nor did they interact with the Explanatory Variables. The twenty student samples constituted a small data set and background variables on these subjects. This sample did not contain sufficient variation in background information to predict the log odds scores on the Explanatory Variables. Second, the submodels ModelEquation and ScoreModel are two significant contributors to assessing student mastery of both

performance and semantic knowledge in the ANOVA score model domain. Third, the two submodels are significant components in a unified assessment model, ANOVAScoreModel2way. They are indispensable for assessing student performance and semantic explanations.

7.5. Correlations between Raw Scores and Log Odds Ratios

Correlations between raw scores and log odds ratios were computed to examine their relationships as assessments of student performance and semantic explanations of the ANOVA Score Model. Three explanatory variables were selected: performance, semantic explanation, and pooled mastery. Raw scores and Log odds ratio for each variable constituted the variables in the correlation matrix (Table 7.15):

Rasg indicates raw score of General Model;

Rameq indicates raw score of ModelEquation;

RaSMo indicates raw score of ScoreModel;

Lodg indicates log odds ratio of General Model;

Lomeq indicates log odds ratio of ModelEquation; and,

LOSMo indicates log odds ratio of ScoreModel.

Table 7.15 presents correlations among (a) raw scores and (b) log odds ratios, for the general model performance, and semantic explanation.

The top-left cell of Table 7.15 reports inter-correlations of variables when raw scores are used. It can be seen that the "general model" is correlated at 0.9675 with ModelEquation, and 0.6286 with ScoreModel. This difference reflects

the fact that score levels on ModelEquation were much higher than ScoreModel. Consequently they contributed more to the total. In addition, there was a correlation of 0.4154 between ModelEquation and ScoreModel based on raw scores.

The lower-right cell of Table 7.15 reports the inter-correlations of these variables using scores consisting of log odds ratio estimates of likelihood of mastery. It shows that "general model" proficiency is correlated at 0.9403 with ModelEquation and at 0.2443 with ScoreModel. This result also shows that the "general model" proficiency estimate reflected high influence of ModelEquation on general proficiency estimates, and small influence of ScoreModel on general proficiency. Correlation of ModelEquation with ScoreModel was only 0.0654 which reflects the near independence of these submodels when model-based estimates of proficiency are used.

The upper-right cell of Table 7.15 reports cross-correlations between raw scores and log odds proficiency estimates for the three assessment variables: the correlation between general model raw scores and general model log odds ratio is 0.9531; the correlation for "Model-Equation" between raw score and log odds ratio is 0.9301; and the correlation for ScoreModel between raw score and log odds ratio is 0.7427. These correlations show that Bayesian estimates of general model proficiency are very high indicating high agreement, especially for the "general model" variable. Thus, the top-level proficiency estimate from the Bayesian model may serve as a measure of mastery that will agree well with other global assessments (total score, as an IRT mode estimate). However, the

component proficiencies estimated using the Bayesian network approach provides evidence of components of mastery that raw total scores do not provide.

Table 7.15. Correlations between Raw Scores and Log Odds Ratios of Three Model Explanatory Variables

Assessment	Raw score			Log-Odds ratio		
Variable	Rasg	Rameq	RaSMo	Lodq	Lomeq	LOSMo
Rasg	1.0	0.96752**	0.62856**	0.95306**	0.91964	0.12636
Rameq		1.0	0.4154	0.90212**	0.93010**	-0.9331
RaSMo			1.0	0.66055**	0.45396**	0.74266**
Lodq				1.0	0.94027**	0.24432
Lomeq					1.0	0.06539
LOSMo						1.0

** indicates p value < 0.01

7.6. Robustness Analyses of the Bayesian Network Assessment Models

After building a Bayesian assessment network, robustness is a concern that reflects both the quality and application efficiency of assessment networks. Due to possible imprecision in prior and conditional probability parameters that are often based on assumptions and input from experts or other sources, the results will depend on these parameters. Therefore, to evaluate the effects of changes in model parameters on assessment results, it is desirable to supplement point estimates of posterior probability values with probability intervals for these parameters. The size of estimates for conditional probabilities will indicate the extent to which estimates depend on conditional probabilities that

were used in the model. The idea of robustness analysis is to employ sets of distributions to represent perturbations and variations in probabilistic models. The goal is to generate bounds on expected values of posterior probabilities. Intervals between upper and lower bounds induced by sets of distributions reflect the quality of the model and the data; small intervals indicate robustness to effects of perturbations.

The Bayesian network assessment model represents a joint distribution through a collection of locally defined probability distributions. However, to understand the robustness of the entire model, it is necessary to check global neighbourhoods of Bayesian networks because they can describe the effects of global perturbations in model probabilities on estimates of posterior probabilities. JaveBayes software (Cozman, 1997) offers several different methods for evaluating robustness in relation to perturbations in global neighbourhoods of probability distributions. For any Bayesian network, there is a cover set of probability distributions that can be used, called a "credal set". One way to specify a credal set is to specify ε-contaminated and lower density bounded classes, which is a common way to describe posterior probability intervals in a global sense. An ε-contaminated class is generally characterized by a distribution p, q and a real number $\varepsilon \in (0,1)$. The p refers to the distribution of probabilities for the current model, a weight ε is specified (0<ε<1) which controls the amount of perturbation in the probabilities. Then, q, a probability of any weighted arbitrary distributions can be added.

$$r(x) = (1-\varepsilon)p(x) + \varepsilon q(x) \qquad \text{(Formula 7-1)}$$

The ε controls the amount of contamination by the arbitrary distribution. Cozman's (1997) method can be used to estimate upper and lower bounds for posterior probability estimates obtained when using the network to infer posterior probabilities of explanatory variables given evidence values. According to Cozman (1997), an "ε-contamination of 0.1" means that expectation being correct 90% of the time, but in 10% of the cases it is expected that other joint distributions are possible.

The robustness of global neighbourhoods with ε-contaminated class has been included in JavaBayes software (Cozman, 1997). Expected interval values can be calculated as the evidence is instantiated for the evidential variables in the entire Bayesian network. Parameters are output in the format of a lower-envelope and an upper-envelope. The lower envelope is usually designated by $\underline{p}$, and the upper envelope by $\overline{p}$. As an ε value is defined and variables are instantiated, bounds on all posterior probabilities can be generated. The ε's were set at 0.10, 0.05, and 0.01. Once ε is set, pairs of upper and lower bounds are produced for probabilities of mastery and non-mastery.

Table 7.16. Effects of ε-contamination on Explanatory Variables of Posterior Probabilities for Full Model

Subject	Posterior probability	Prob (mastery)		Prob (non-mastery)		Robust mastery decision overlap
		Lower bound estimate for p	Upper bound estimate for p	Lower bound estimate for q	Upper bound estimate for q	
			ε =0.10			
1	0.4592	0.4132	0.5133	0.4867	0.5867	*
2	0.4847	0.4363	0.5362	0.4637	0.5637	*
3	0.4312	0.3881	0.4881	0.5119	0.6119	
4	0.4216	0.3795	0.4795	0.5205	0.6205	
5	0.3838	0.3454	0.4454	0.5546	0.6546	
			ε =0.05			
1	0.4592	0.4362	0.4862	0.5138	0.5638	
2	0.4847	0.4605	0.5105	0.4895	0.5395	*
3	0.4312	0.4097	0.4597	0.5403	0.5903	
4	0.4216	0.4005	0.4505	0.5494	0.5994	
5	0.3838	0.3646	0.4146	0.5854	0.6354	
			ε =0.01			
1	0.4592	0.4546	0.4646	0.5354	0.5454	
2	0.4847	0.4799	0.4899	0.5101	0.5201	
3	0.4312	0.4269	0.4369	0.5631	0.5731	
4	0.4216	0.4174	0.4274	0.5726	0.5826	
5	0.3838	0.3740	0.3840	0.6100	0.6200	

* indicates it is ε -contaminated

Table 7.16 represents estimates for 5 subjects as examples of robustness analyses with ε set at 0.10, 0.05 and 0.01. A judgemental rule intuitively is that an overlap of an upper bound value in one boundary group with a lower bound value in another boundary group is judged as ε-contaminated. When ε is set at 0.10, there are two cases, Students 1 and 2 are ε-contaminated. The overlap value for

Student 1 is 0.0266 (from 0.5133-0.4867); the overlap value for Student 2 is 0.0725 (from 0.5362-0.4637). When ε is set at 0.05, only Student 2 is still ε-contaminated. However, the overlap value decreased from 0.0725 to 0.021. When ε is set at 0.01, there are no ε-contaminated cases. The change reveals that the number of ε-contaminated estimates decreases as the value of ε decreases. In other words, the network is becoming more robust as ε is set at smaller and smaller values. The ε-contaminated details for all 20 students for general mastery, performance and semantic explanations are reported in Appendix H.

CHAPTER EIGHT:DISCUSSION AND CONCLUSION

This study explored diagnostic cognitive assessment (DCA) using Bayesian networks and evidence-centred design in a statistics learning content domain. The initial motivation was to design an effective assessment methodology for statistics in a context of learning within a web-based tutorial environment. Therefore, the assessment environment simulates problem solving activities that occurred in a web-based statistics learning environment known as the McGill Statistics Tutoring Project (MSTP). A stand-alone test entitled the Performance Assessment of Statistical Learning Test (PASLT) was developed. PASLT mimics the task constructs and the structures of MSTP.

The approach to cognitive assessment may be summarized as follows. On the basis of cognitive and content analyses of expert tutoring and performance data, cognitive models of components of expert knowledge and performance skills were developed (Frederiksen & Donin, 2005). Assessment models composed of (a) assessment constructs and (b) evidence models were developed based on these cognitive models. In an assessment model, assessment constructs correspond to components of knowledge and procedural skill identified in the cognitive models. These constructs are represented as explanatory variables in the assessment model, and explanatory variables are linked to evidence variables. The evidence variables represent specific aspects of student task performance or semantic explanations which may be identified in a student's performance of assessment problems. The assessment model constructs and evidence variables represent different aspects of both task

performance and semantic explanations that are produced over the course of a student's performance of ANOVA tasks. Bayesian networks are used to connect the explanatory variables to evidence variables and these links enable the network to propagate evidential information to explanatory model variables in the assessment model.

Chapter 8 will be organized into three parts: (a) a summary of objectives, methods and results of the study; (b) conclusions concerning the appropriateness of the approach to Bayesian cognitive assessment and its relationship to IRT and to other Bayesian cognitive assessments; and (c) an evaluation of the significance and limitations of the findings, and recommendations for future research.

8.1. The Objectives, Methods and Results of the Study

8.1.1 Objectives, Purposes and Assumptions of Diagnostic Cognitive Assessment

This section summarizes the assumptions of the current study, and the research objectives and methods used to develop the Bayesian assessment model.

The purpose of diagnostic cognitive assessment (DCA) is to evaluate a student's mastery of the cognitive components (of knowledge and skills) that characterize expert/competent performance in a well-defined domain, and their successful application to problem-solving tasks in the domain. In the statistics learning domain, the objective of assessment was to design an effective

assessment procedure by which learning progress, proficiency, and mastery could be evaluated and recorded. The objective of the study was to develop a diagnostic model and methodology that implements the assessment in a specific domain of statistics, and evaluate it in relation to its potential to achieve the objectives of DCA.

The DCA model and methods investigated in the study were based on several assumptions. First, DCA should be based on student performance of authentic and complex tasks from a learning domain (in this case, a domain of statistics).

Second, task performance in solving or explaining a problem can be represented by measures of an exhaustive set of "fine-grained solution features" that can be observed and scored (e.g., as present or absent) by an observer competent in the domain.

Third, the results of the assessment will be restricted to inferences based on response features of tasks that have been identified and evaluation rules that have been applied to "score" records of performance.

Fourth, the purpose of DCA is to infer which specific cognitive components of knowledge and skill have been mastered by a student. These inferences will be realized probabilistically by a measure which indicates, more specifically, the likelihood that a student has mastered a specific component of knowledge or skill based on the observations of the student's response features during performance of an assessment task ("evidence").

Fifth, to make such diagnostic inferences will require that an assessment framework, assessment constructs, and an evidence model be developed. This can be accomplished using a Bayesian probability network (a) to represent cognitive components (of knowledge and skill) as assessment constructs (variables), (b) to represent scored components of performance as evidence variables, and (c) to make inferences about mastery of assessment constructs based on evidence from a student's performance.

Sixth, a Bayesian assessment is assumed to begin with complete uncertainty about a student's mastery of components, i.e., the prior probability of mastery of an assessment component is 0.5, and the prior probability of non-mastery is 0.5. The assumption is that without evidence the user of the assessment is completely uncertain about whether a student has mastered the components of the domain knowledge and problem solving skills at a very general level, i.e., at the top level of the Bayesian belief network. This assumption of uncertainty represents a lack of bias in the user's belief before the assessment has taken place. After the student completes the assessment problem-solving tasks, the evidence variables can be instantiated and then the assessment network can be updated, resulting in new posterior probabilities of mastery that replace the initial (prior) probabilities. The child nodes are fixed at 0.67 for mastery status and at 0.33 for non-mastery status, conditional to the parent node. This provides a well-defined basis for estimates of the posterior probabilities after updating using patterns of the instantiated evidence nodes. Even though these values are slightly arbitrary, sufficient and appropriate

reasons were provided for the selected values. Their effects can be further explored by setting the conditional probabilities to other values. The selection of these conditional probability values was found to affect the rate of change in posterior probabilities produced by introducing new evidence.

Seventh, mastery of cognitive components in a domain is a dynamic process in which mastery of components changes as the student develops the knowledge and skills underlying expertise in the domain of assessment. If a student's problem-solving progresses following a particular trajectory of development of proficiency, this trajectory can be assessed to evaluate a student's progress toward a state of mastery. Mastery is viewed as a "milestone" of the learning.

Therefore, from a developmental perspective, the purpose of cognitive assessment is to evaluate the student's trajectory in the development of components of knowledge and skill in a domain by updating the likelihoods of mastery (prob(mastery)) of cognitive components over repeated assessments. At any point in the assessment, the estimation of the student's posterior probabilities of mastery of specific components can be used to provide diagnostic information which can be used to provide feedback and guidance to the student, coaches, tutors or others involved in supporting the students' development of expertise.

8.1.2 Outcomes of the Study

This study applied the model development method to the ANOVA score model domain in order to attain the objectives of the study. The results document: (a) the process of model development in a specific domain; (b) the properties of a Bayesian assessment model; (c) the performance of the BN in tracing progress towards mastery by using the model to successfully update posterior probabilities; (d) the use of estimates of log odds ratios of mastery as measures of "progress toward mastery of cognitive assessment constructs;" (e) the robustness of diagnostic inferences based on the BN; and (f) the use of the Bayesian assessment model for diagnostic assessment with a sample of students who completed the assessment tasks. In general, the study demonstrated that an effective diagnostic cognitive assessment methodology for statistics learning could be established.

The study carefully documented the method for developing assessment models that meet the above objectives. The application of the method was documented for a specific domain of statistics learning, using cognitive models and performance tasks that had been previously developed in the MSTP.

The hierarchical Bayesian assessment network developed in this domain was examined to investigate model updating for network components ranging from simple cliques, to net-cliques, to assessment submodels, to the general model.

Simulated data were applied to study assessments resulting from the model. Trajectories of progress towards mastery obtained using the model were examined.

The use of log odds ratios as estimates of likelihoods of node mastery, as a "measure of progress toward mastery," were proposed and evaluated. The instantiated values of log odds ratios for explanatory variables follow a normal distribution. Log odds ratios rescale the posterior probabilities of mastery initially obtained by updating Bayesian assessment networks based on performance evidence. Rescaled measures provided improved measures of learning proficiency and mastery. The results demonstrated that log odds ratios are an appropriate measure of cognitive construct mastery, and indicated that mastery can be tracked by providing increasing evidence to the network and using it to update the log odds ratios. The results demonstrated the diagnostic value of log odds ratios for assessing model construct mastery.

Analysis of the robustness of Bayesian assessment networks was carried out. The results indicated that BNs were robust on the basis of both student and simulated data.

The diagnostic use of log odds estimates of mastery of explanatory constructs were investigated using data from a sample of students studying intermediate statistics. First, by examining log odds ratios for different network nodes, it was found that the assessments revealed differences in patterns of mastery of assessment constructs (components of knowledge and skill) that were not reflected in pooled raw scores for components.

Second, students with equivalent total scores were found to differ on the model-based estimates of general competency and mastery of specific model components. The log odds ratios estimated using the model provided diagnostic information that was sensitive to student patterns of solution features (evidence variables). Estimates of student log-odds mastery of specific model components were used for diagnostic assessment in situations in which pooled raw performance scores were useless.

Third, analysis of group data revealed that model-based assessments successfully detected differences in mastery of different model components.

Fourth, the external validity of the assessments were determined by using survey data from a group of students and examining variables as predictors of assessments resulting from the application of the Bayesian assessment network to the student data. The results showed that there were no significant differences related to the effects of background variables, i.e., using the tutor system (UT), or taking a statistics course (SC), on the log odds mastery of selected explanatory variables. This finding was attributed to the small sample size and the distribution of scores on performance tasks. However, there were significant differences among these explanatory variables. A further multiple comparison analysis using a repeated measure MANOVA procedure revealed differences between parent and child explanatory variables, and among child explanatory variables.

Fifth, correlations between raw scores and log odds ratios for three high-level assessment variables were examined: "ScoreModel," "ModelEquation" and "ANOVAScoreModel2way." There were two principal results: (a) the correlation

between estimated log odds mastery of the two submodels (i.e., the procedural and semantic knowledge submodels) was very small and non-significant (0.0654, see table 7.15), indicating the local independence of the submodel estimates; (b) the correlation between the log odds ratios of general mastery with the raw total score was very high (0.9531, see Table 7.15) , indicating that the two assessments of general proficiency are very closely related. Thus, the Bayesian assessment network provides independent diagnostic information about submodels, while providing a general assessment of mastery comparable to that obtained using a raw total score (or presumably, an IRT estimate).

Finally, the robustness of the Bayesian assessment network model was examined using student data. Results indicated that the BN estimates of posterior probabilities of node mastery were very robust.

8.2 Conclusions

Section 8.2 seeks to draw conclusions about the appropriateness of the evidence-centred design approach. The approach used to develop Bayesian assessment models in the domain studied, and potential applications of both to DCA in other complex domains of performance and learning. To put these conclusions into some perspective, similarities and differences between the approach developed in this thesis, and previous work in DCA using both BN models and IRT models, will be considered. Comparison of these different models as alternative methodological approaches to assessment can provide alternative possibilities that can help us design for (a) different kinds of

assessment situations: e. g., tests composed of multiple tasks (or "items") vs. performance assessments using complex and extended tasks; and (b) different assessment objectives: e.g., for use in conjunction with computer-based learning environments, in tracking learning, and in establishing evidence of mastery in a domain of expertise. It will be argued that construct validity and consequential validity should be central considerations in such situations of cognitive assessment.

Finally, it is suggested that there can be a complementary relationship between IRT approaches and Bayesian approaches to assessment. Bayesian approaches are particularly appropriate for diagnostic cognitive assessment, and IRT approaches are particularly appropriate for assessing generalizable proficiency in a domain in which many tasks can be designed to sample performance in a domain. Bayesian approaches are optimal when the domain is complex, performance of domain tasks requires extended problem solving and extensive conceptual knowledge, and diagnostic information is important given the purposes of the assessment.

8.2.1 Appropriateness of the Evidence-Centered Design Framework

Evidence-centered design (ECD) provides an appropriate general methodological framework for diagnostic cognitive assessment (DCA). It allows the use of assessments to infer individual progress towards mastery of components of knowledge acquisition and skill development from observations of student performance using assessment and learning tasks. These assessments

can track an individual's development of general proficiency and its cognitive components. The complexity of assessment design depends on how complex the tasks are and how learning environments provide affordances to the learning. Development of evidence variables from learning tasks to define measurable objects is a tactical process. Furthermore, the connection of evidence models to assessment constructs requires an efficient and powerful statistical engine.

In the evaluation model of ECD, Bayesian networks are often used to propagate information from evidential variables to explanatory variables. Both the evidential and the explanatory variables can be dichotomous or polytomous. This feature allows a great many possibilities for assessment designs to set flexible probabilistic state spaces. Such assessment designs can appropriately fit many different assessment purposes.

In addition, ECD provides a possibility of applying IRT in estimating explanatory variables. In other words, any local or general node of a Bayesian network can be explored with an IRT model. This implicitly represents a connection between IRT and Bayesian networks. On the basis of flexibility and a combination of IRT and Bayesian networks, ECD is appropriate for many different assessment situations. In the current study, the ECD framework was essential in designing the model blocks and assembling them effectively into DCA model.

8.2.2 Appropriateness of the Specific Bayesian Assessment Model (Developed in the Project)

This study developed a Bayesian assessment network model that was designed to accomplish both diagnostic and cognitive objectives. The assessment models incorporated performance (procedural) and semantic (declarative) knowledge, and applied these models to both real and simulated data. In order to simplify the evidence spaces and to implement a goal of assessing likelihood of mastery of specific cognitive components, binary state spaces for assessment constructs were chosen.

Emphasis was placed on assessments as tools for developing beliefs about the state of mastery of model components rather than for measuring individuals' levels on proficiency scales. Thus, very detailed assessments of cognitive components (procedural and declarative knowledge) that provide procedural and semantic explanations of student performance were obtained on the basis of the evidence variables updating to the explanatory nodes in the assessment model. Assessments trajectories based on this transfer of information can be used for reporting on such important process information as missing knowledge and skill components, and for tracking progress.

In other words, the approach taken was to estimate the likelihood of mastery based on evidence consisting of instantiated evidence variables. Linear scales of belief were obtained by transforming estimated probabilities of mastery into log odds ratios. These scales play critical roles in diagnostic judgments and in tracking learning process. Transformation of the likelihoods of mastery to log

odds ratios enables a linear scale of "proficiency" that is sensitive to subtle changes in patterns of evidence derived from a student's task performance.

Use of fixed conditional probabilities (vs. data-based estimates) allowed unambiguous estimates. Since weights were fixed, assessment estimates only reflect the structure of the BN. The choice of specific conditional probabilities effected the rate of change of estimated posterior probabilities of mastery of assessment constructs. These can be adjusted to "tune" the sensitivity of BN patterns of evidence obtained from simple assessment tasks.

Model development involved (a) the development of a cognitive model, (b) mapping it into an assessment model (a Bayesian network), and (c) the development of evidence rules. Thus, the development of a Bayesian network (BN) or a Bayesian belief network (BBN) assessment model using this methodology does not require large data sets to estimate parameters in an assessment model. This also provided benefits for the validity of assessments and their interpretation.

The BN assessments developed here allow direct linking of the assessments to authentic, challenging, and natural problem solving environments. Assessment networks can be flexibly adjusted on the basis of cognitive constructs and assessment purposes. They can be repeatedly updated to enable the direct tracing of developmental trajectories during learning, and the diagnostic reporting of dynamic progress toward mastery. Furthermore, the assessment network provides valid diagnostic information about specific components, and tracks development towards mastery of learning goals. By

using the assessment model to track progress across multiple domain tasks that are authentic and increasingly challenging, convincing evidence of developing competencies in domains can be acquired.

Therefore, the assessment network is appropriate with respect to content validity, construct validity, and consequential validity for many assessment purposes.

When an assessment network can be simplified into two layers, its structure is very close to an IRT structure. Thus, IRT models may be thought of as special cases of a BN (Junker, 1999; Yan, Almond, & Mislevy, 2003). Therefore, general assessments of developing proficiency using IRT models can provide complementary general assessments which can be supplemental with diagnostic information provided by BN models. There is no doubt that this diagnostic information will enrich the psychometric information.

8.2.3 Mastery and Proficiency from an IRT vs. a BBN Perspective

Mastery and non-mastery have been proposed as two states in the Bayesian student model construct (Desmarais & Pu, 2005) and in Bayesian performance assessments (VanLehn, 2001). Proficiency is often used to assess an ability or trait in a task domain when a problem solver successfully solves multiple tasks (test items). There are different theoretical bases and assumptions underlying mastery versus proficiency. BN assessment models can provide estimates of posterior probabilities of mastering assessment constructs by which learning progress can be reported by inference from evidence variables. We

investigated the utility of a log odds ratio transformation to produce linear and normally distributed measures of likelihood of mastering assessment constructs. IRT is a set of psychometric models that usually include such item parameters as item difficulty, discrimination power and latent assessment variables measured on continuous scales. A comparison may be useful in rethinking relationships between the two classes of assessment models.

First, the advantage of BN assessment models over IRT is that the assessment of the probability of mastery based on measurable objects in a Bayesian model does not rely on single traits representing an ability continuum. Bayesian assessment network models can contain both single and multiple dimensions of ability in the form of probabilities of mastery of model components at different levels.

Second, BNs can represent student model variables. Andes is one example of using a BN approach to assessment (Martin & VanLehn, 1995b; VanLehn & Niu, 2001). BNs provide a basis for further developing an evidence-centred assessment design. IRTs do not necessarily require complex and advanced cognitive model. However, deficits in the cognitive model would devastate inferences about explanatory constructs (variables) designed to model domain knowledge and skills.

Third, modeling is obviously different between the two classes of models. IRT models are derived from test data collected from student performance on samples of tasks from a domain. Consequently, model parameters reflect student knowledge levels and item parameters that are estimated for the particular sets

of test items used. IRTs do not rely on the construction of knowledge models based on cognitive task analysis in some domain. They are not good at representing complex cognitive assessments, and details such as detecting misconceptions and learning errors which occurred during the learning process are not easily addressed.

Most Bayesian assessments categorize student mastery or level of mastery in terms of a category. In the model presented here, the focus of the assessment of mastery was on the estimated log odds of mastery as a measure of the likelihood that a student has mastered a component of knowledge or skill. This approach yields: (a) approximately normally distributed "scores" on a continuous scale, and (b) a measure that is appropriate if the purpose of the assessment is viewed as a probabilistic judgement about a student's state of mastery. Such an assessment of general mastery (i.e. of a construct at the top of a BN) is probably nearly equivalent to the level of mastery of a student learning assessed by an IRT model.

In short, the BN model studied in this thesis can provide assessment with an approach that combines aspects of both psychometric and cognitive assessment, since the knowledge and skill components represented in the cognitive models can often be hierarchically structured. These models can be widely applied in complex learning environments (including computer-based learning environments), and they can provide both diagnostic information for use during learning, and assessments of progress in establishing mastery in domains of learning.

8.3. Contribution, Limitations and Future Research Directions

Section 8.3 presents a discussion of the contribution of diagnostic cognitive assessment design, limitation of the analyses, and future research directions related to the current study.

8.3.1 Contributions

This study explored a methodology for establishing diagnostic cognitive assessments using an evidence-centered assessment design (ECD) approach. As a theoretical framework, ECD has been applied in many different domains. However, the current study applying ECD as an assessment design is unique in how it combines cognitive model theory and BN in a complex statistics learning domain. The contributions of the current study are as follows:

First, the study examined the application of principles of ECD to diagnostic cognitive assessment in a statistics learning domain, and how it enabled assessment of student problem-solving competencies and deficiencies diagnostically and dynamically.

Second, the study explored and carefully documented an assessment development methodology appropriate to a specific situation of DCA. The assessment will allow students to learn statistics while receiving diagnostic self-assessments on the basis of their performance and semantic explanations for tasks and learning problems they are completing. This capability can be provided in computer-coached learning environments such as the MSTP application.

Although the study does not realize this engineering objective, it has laid the basis for achieving this objective of web-based assessment in a dynamic coached-learning environment.

Third, the study demonstrated how an inferential framework from evidence to assessment constructs can be implemented as a Bayesian assessment network based on a detailed cognitive model. The Bayesian model can be used to instantiate the evidence variables and to propagate assessment information between the assessment constructs and the evidence variables.

Fourth, after completing the Bayesian assessment networks, the study examined the basic structure of the assessment network from simple to complex net cliques. This allowed a better understanding of the characteristics of the Bayesian assessment network.

Finally, on the basis of student and simulated test data the Bayesian assessment network was successfully applied to assess the probability of the model. Mastery vs. non-mastery of cognitive model constructs have been proposed in the Bayesian student model construct. The result demonstrated that the assessment of mastery based on the estimated log odds of mastery provided a good diagnostic measure of the likelihood that a student had mastered specific components of knowledge and cognitive skills. Bayesian assessment networks of the kind studies here can be constructed for any well-structured domains of problem-solving competency, once a cognitive task and knowledge have been developed and validated for the domain.

8.3.2 Limitations

One limitation of the study was that student data was limited and not evenly distributed in terms of levels of performance on the assessment task. The attempt to used student background variables to explore the external validity of the assessments was not successful.

Second, the use of a dichotomous state space for BN nodes may have provided coarser information about mastery status unlike perhaps multiple mastery states. However the use of dichotomy as state spaces simplified the assessment design and the interpretation of assessment results. For example, appropriate assessment categories to consider for other assessment situations might include mastery, non-mastery, and partial mastery as an intermediate state.

Third, the testing of the assessment network using simulated data could be expanded to simulations of large samples of student data including a wide range of response patterns. Exploration of "inconsistent" response patterns needs to be undertaken to see how these patterns influence network assessments. Study of these "buggy patterns" might lead to the addition of bug detection features to the assessment model.

Fourth, a larger sample of student data would have enabled the fitting of an IRT model to the data, enabling stronger conclusions about the relationship of the Bayesian estimate of "general proficiency" to IRT-based estimates.

8.3.3 Future Research Directions

The current study suggests many possible future research directions.

1. Apply the DCA methodology to other complex domains of performance, learning and skill development.

2. Test the DCA model with larger student data sets.

3. Investigate the use of assessment to track learning and to provide feedback in instructional situations.

4. Investigate possible variations on the model, such as, the introduction of slip and guessing probabilities into evidence models.

5. Investigate the relationships of DCA estimates of measures of general proficiency (log odds likelihood of mastery) with IRT-based assessment of general proficiency using appropriate data sets.

6. Examine future effects of details of particular BN used in DCA. In particular, to investigate the use of multi-category state spaces to evaluate their relative advantages.

7. Implement long-term studies to investigate assessment strategies that use Bayesian DCA in combination with IRT-based psychometric assessment to provide both (a) psychometric estimates of general proficiency, and (b) DCAs to obtain diagnostic information useful in developing student knowledge and expertise in various domains.

REFERENCES

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, 32(8), 10-14.

Alexander, P. A., Jetton, T. L., & Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology*, 87(4), 559-575.

Alexander, P. A., Murphy, P. K., Woods, B. S., Duhon, K. E., & Parker, D. (1997). College instruction and concomitant changes in student s' knowledge, interest, and strategy use: A study of domain learning. *Contemporary Educational Psychology*, 22, 125-146.

Alexander, P. A., Sperl, C. T., Buehl, M. M., Five, H., & Chiu, S. (2004). Modeling domain learning: Profiles from the field of special education. *Journal of Educational Psychology*, 96(3), 545–557.

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *The Journal of Technology, Learning, and Assessment,* 1(5), 1-63.

Andersen, E. B. (1995). Polytomous Rasch model and their estimation. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch model: Foundations, recent developments, and applications* (pp. 271-292). New York: Springer-Verlag.

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

Andreassen, S., Jensen, F. V., & Olesen, K. G. (1990). Medical expert systems

based on causal probabilistic networks. Aalborg, Denmark: Institute of Electronic Systems, Aalborg University.

Andrich, D. (1978). A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, 31, 84-98.

Annett, J. (2000). Theoretical and pragmatic influences on task analysis methods. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis* (pp. 25-40). Mahwah, NJ: Lawrence Erlbaum.

Annett, J., & Cunningham, D. (2000). Analysing command team skills. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive task analysis* (pp. 401-415). Mahwah, NJ: Erlbaum.

Arcos, J. L. L., Muller, W., Fuente, O., Orue, L., Arroyo, E., Leaznibarrutia, I., & Santander, J. (2000). Lahystotrain: Integration of virtual environment and ITS for surgery training. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent tutoring systems* (pp. 43-52). Pittsburgh, PA: Springer.

Arocha, J. F. (1990). *Clinical case similarity and diagnostic reasoning in medicine*. Unpublished doctoral dissertation, McGill University, Montreal, Quebec, Canada.

Arocha, J. F., & Patel, V. L. (1995). Novice diagnostic reasoning in medicine: Accounting for evidence. *The Journal of the Learning Science*, 4(4), 355-384.

Ayala, C. C. (2003). On the cognitive validity of science performance

assessments. *Dissertation Abstracts International Section A: Humanities & Social Science*, 64 (10-A), Univ. Microfilms International, 3469.

Ayala, C. C., Ayala, M. A., & Shavelson, R. (2000, April). *New dimensions for performance assessments*. Paper presented at AERA, New Orleans, LA.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York, NJ: Marcel Dekker.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement*, 29(1), 1-17.

Bechtel, W., Abrahamsen, A., & Graham, G. (1998). The life of cognitive science. In W. Bechtel, & G. Graham (Eds.), *A companion to cognitive science* (pp. 1-104). Malden, Mass: Blackwell.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing*, 4(4), 295-301.

Bennett, R. E., Morley, M., & Quardt, D. (1998). *Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests* (RR-98-45). Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp.3-29). Norwell, Mass: Kluwer Academic Publishers.

Bitan, T., Karni, A., & Bitan, T. (2004). Procedural and declarative knowledge of word recognition and letter decoding in reading an artificial script. *Cognitive Brain Research*, 19(3), 229-243.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29-51.

Brown, A. J. (1994). The advancement of learning. *Educational Researcher*, 23(8), 4-12.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32-42.

Byrnes, J., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology*, 27(5), 777-786.

Caroll, J. B. (1993). *Human cognitive abilities*. Cambridge, NY: Cambridge University Press.

Carroll, J. B. (1998). Human cognitive abilities: A critique. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 5-23). NJ: Lawrence Erlbaum Associates.

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

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6, 271-315.

Clancey, W. (1997). *Situated cognition: On human knowledge and computer representations*. New York, NY: Cambridge University Press.

Collins, A. (1990). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp.75-87). Hillsdale, NJ: Lawrence Erlbaum.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnkck (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-491). Hillsdale, NJ: Lawrence Erlbaum.

Conati, C., & VanLehn, K. (1996). POLA: A student modeling framework for probabilistic on-line assessment of problem solving performance. *Proceedings of the fifth international conference on use modeling* (pp. 75-82). Kailua-Kona, Hawaii.

Confrey, J. (1995). How compatible are radical constructivism, social-cultural approaches and social constructivism? In L. Steffe & J. Gale (Eds.), *Constructivism in Education* (pp.185-226). Hillsdale, NJ: Lawrence Erlbaum.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction,* 4, 253-278.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 19-41). Hillsdale, NJ: Lawrence Erlbaum.

Corter, J. E. (1995). Using clustering methods to explore the structure of diagnosis tests. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 305-325). Hillsdale, NJ: Lawrence Erlbaum.

Cozman, F. G. (1997, August). *Robustness analysis of Bayesian networks with local convex sets of distributions.* Paper presented at the thirteenth annual conference on uncertainty in artificial intelligence, Providence, RI.

Cozman, F. G. (1998). *JavaBayes user Manual.* Retrieved June 20, 2004, http: //www.cs.cmu.edu/~javabayes/Home/

Derry, S. D. (1990). Learning strategies for acquiring useful knowledge. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 347-375). Hillsdale, NJ: Lawrence Erlbaum.

Derry, S. J., DuRussel, L. A., & O'Donnell, A. M. (1998). Individual and distributed cognitions in interdisciplinary teamwork: A developing case study and emerging theory. *Educational Psychology Review*, 10(1), 25-56.

Derry, S. J., & Lajoie, S. P. (1993). A middle camp for (un)intelligent instructional computing: An introduction. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 1-11). *Hillsdale*, NJ: Lawrence Erlbaum.

Desmarais, M. C., & Pu, X. (2005). A Bayesian student model without hidden

nodes and its comparison with item response theory. *International Journal of Artificial Intelligence in Education,* 15(4), 291-323.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/ psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), Cognitively diagnostic assessment (pp. 361-389). Hillsdale, NJ: Lawrence Erlbaum.

du Boulay, B. D. (2000). Can we learn from ITSs? In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Intelligent Tutoring Systems, 5th International Conference, ITS 2000* (pp.9-17). New York: Springer.

Duit, R. (1995). The constructivist view: A fashionable and paradigm for science education research and practice. In L. Steffe & J. Gale (Eds.), *Constructivism in education* (pp. 271-286). Hiilldale, NJ: Lawerence Erlbaum.

Eisner, E. (1993). Forms of understanding and the future of educational research. *Educational Researcher*, 22(7), 5-11.

Embretson, S. E. (1984). A general latent trait model for response process. *Psychometrika*, 49(2), 175-186.

Embretson, S. E. (1990). Diagnostic testing by measuring learning processes: Psychometric considerations from dynamic testing. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 407-432). Hillsdale, NJ: Lawrence Erlbaum.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for new generation of tests* (pp. 125-153). Hillsdale, NJ: Lawrence Erlbaum.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application and mathematical reasoning. *Journal of Educational Measurement* 32(3), 277-294.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305-321). NY: Springer-Verlag.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for Psychologist.* Mahwah, NJ: Lawrence Erlbaum.

Ernest, P. (1993). Constructivism, the psychology of learning and the nature of mathematics: Some critical issues. *Science & Education*, 1, 87-93.

Essens, P. J. M. D., Post. W. M., & Rasker, P. C. (2000). Modeling a command center. In J. M. Schraagen, S. F. Chipman & V. L. Shalin (Eds.), *Cognitive task analysis* (pp. 401-415). Mahwah, NJ: Lawrence Erlbaum.

Farhana, S., Evens, M., Michael, J., & Rovick, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes*, 33(1), 23-52.

Feltovich, P. J., Spiro, R. J., & Coulson R. L. (1993). Learning, teaching, and

testing for complex conceptual understanding, In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for new generation of tests* (pp. 181-217). NJ: Lawrence Erlbaum.

Fischer, G. H., & Seliger, E. (1997). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 323-345). NY: Springer-Verlag.

Fosnot, C. T. (1993). Rethinking science education: A defence of Piagetian constructivism. *Journal of Research in Science Teaching*, 30(9), 1189-1201.

Frederiksen, C. H. (1975). Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology*, 7(3), 371-458.

Frederiksen, C. H. (1986). Cognitive models and discourse analysis, In C. R. Cooper & S. Greenbaum (Eds.), *Studying writing: Linguistic approaches* (pp. 227-267). Beverly Hills, CA: Sage.

Frederiksen, C. H. (1999). Learning to reason through discourse in a problem-based learning group. *Discourse Processes*, 27(2), 135-160.

Frederiksen, C. H., & Breuleux, A. (1990). Monitoring cognitive processing in semantically complex domains. In N. Frederiksen, R. Glaser, A Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 351-391). Hillsdale, NJ: Lawrence Erlbaum.

Frederiksen, C. H., & Donin, J. (1999). Cognitive assessment in coached learning environments. *Alberta Journal of Educational Research*, 45(4), 392-408.

Frederiksen, C. H., & Donin, J. (2005). *Coaching and the development of expertise: Designing computer coachers to emulate human tutoring in complex domains*. Manuscript submitted for publication.

Frederiksen, C., Donin, J., Bracewell, R., Mercier, J., & Zhang, Z. (2002, April). *Human tutoring and the design of computer coaching systems*. Paper presented at the meeting of AERA, New Orleans, LA.

Frederiksen, C. H., Stemmer, B. (1993). Conceptual processing of discourse by a right hemisphere brain-damaged patient. In H. B. Hiram & Y. Joanette (Eds.), *Narrative discourse in neurologically impaired and normal aging adults* (pp.239-278). San Diego: Singular Publishing Group, INC.

Frederiksen, N. (1990). Introduction. In N. Frederiksnen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. ix-xvii). Hillsdale, NJ: Lawrence Erlbaum.

Fredette, M., & Angers, J. F. (2002). A new approximation of the posterior distribution of the log-odds ratio. *Statistica Neerlandica*, 56(3), 314-329.

Gadd, C. S., & Pople, Jr. H. E. (1990). Evidence from internal medicine teaching rounds of the multiple roles of diagnosis in the transmission and testing of medical expertise. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 89-111). Hillsdale, NJ: Lawrence Erlbaum.

Ganeshan, R., Johnson, W. L., Shaw, E., & Wood, B. P. (2000). Tutoring

diagnostic problem solving. In G. Goos, J. Hartmanis & J. V. Leeuwen (Series Eds.) & G. Gauthier, C. Frasson & K. Vanlehn (Vol. Eds.), *Intelligent tutoring system* (pp. 33-42). New York: Springer.

Gierl, M. J., Leighton, J. P., & Hunka, S. M. (2000). An NCME instructional model on exploring the logic of Tatsuoka's rule-space model for test development and Analysis. *Education Measurement: Issues and Practices*, 19(3), 34-44.

Gitomer, D. H., Steinberg, L. S., & Mislevy, R. J. (1995). Diagnostic assessment of troubleshooting skill in an intelligent tutoring system. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp.73-101). Hillsdale, NJ: Lawrence Erlbaum.

Greeno, J. (1998). The situativity of knowing, learning, and research. *American Psychologist*, 53(1), 5-26.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York, NY: Macmillan Library Reference USA.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement item. *Journal of Educational Measurement*, 26(4), 301-321.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-389). Hillsdale, NJ: Lawrence Erlbaum.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.

Hambleton, R. K., Swaminathan, H., & Roger, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications.

Hardy, M. D., & Taylor, P. C. (1997). *Von Glasersfeld's radical constructivism: A critical review. Science and Education*, 6(1-2), 135-150.

Hay, M. L., & McTaggart, J. A. (2003, March). *Defining computer proficiency: How to measure the immeasurable—part 2.* Paper presented at the Society for the Integration of Technology in Education Conference, Albuquerque, NM.

Heffernan, N. T. (2001). *Intelligent tutoring systems have forgotten the tutor: Adding a cognitive model of human tutors.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh.

Hmelo, C. E. (1998). Problem-based learning: Effects on the early acquisition of cognitive skill in medicine. *The Journal of the Learning Science*, 7(2), 173-208.

Hollnagel, E, (2003). *Handbook of cognitive task design.* Hillsdale, NJ: Lawrence Erlbaum.

Horn, J. (1998). A basis for research on age differences in cognitive capabilities. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp.57-91). NJ: Lawrence Erlbaum.

Hunt, E. (1995). Where and when to represent students this way and that way: An evaluation of approaches to diagnostic assessment, In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitive diagnostic assessment* (pp. 411-429). Hillsdale, NJ: Lawrence Erlbaum.

Ignizio, J. P. (1991). *Introduction to expert systems: The development and implementation of rule-based expert systems.* New York, NY: McGraw-Hill, Inc.

Jensen, F. V. (2001). *Bayesian networks and decision graphs.* New York, NY: Springer-Verlag.

Joseph, G.M., & Patel, V. L. (1990). Domain knowledge and hypothesis generation in diagnostic reasoning. *Medical Decision Making*, 10(1), 31-46.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Paper prepared for the National Research Council Committee on the Foundations of Assessment.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment model with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement,* 25(3), 258-272.

Kassim, A. A., Ahmed, K. S., & Ranganath, S. (2001, August). *Web-based intelligent approach to tutoring.* Paper presented at the international conference on engineering education, Oslo, Norway.

Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoka, K. K. (1993). *Extending*

the rule space model to a semantically-rich domain: Diagnostic assessment in architecture (RR-93-42-ONR). Princeton, NJ: Educational Testing Service.

Katz, I. R., Martinez, M. E., Sheehan, K. M., & Tatsuoka, K. K. (1998). Extending the rule space methodology to a semantically-rich domain: Diagnostic assessment in Architecture. *Journal of Educational and Behavioral Statistics*, 24(3), 254-278.

Kelderman, H., & Rijkes, C. P. M. (1994). Loglinear multidimensional IRT models for polytomously score items. *Psychometrika*, 59(2), 149-176.

Kyllonen, P. C. (1996). Is working Memory Capacity Spearman's g? In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 49-75). NJ: Lawrence Erlbaum.

Lajoie, S. P. (1993). Computer environments as cognitive tools for enhancing learning. In S. P. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp.261-288). Hillasale, NJ: Lawrence Erlbaum.

Lajoie, S. P. (2003). Transitions and trajectories for studies of expertise. *Educational Researcher*, 32(8), 21-25.

Lajoie, S. P., Azevedo, R., & Fleiszer, D. M. (1998). Cognitive tools for assessment and learning in a high information flow environment. *Journal of Educational Computing Research*, 18(3), 203-233.

Lajoie, S. P., & Lesgold, A. (1989). Apprenticeship training in the workplace: Computer-coached practice environment as a new form of apprenticeship. *Machine-Mediated Learning*, 3, 7-28.

Lajoie S. P., & Lesgold, A. M. (1992). Dynamic assessment of proficiency for solving procedural knowledge tasks. *Educational Psychologist*, 27(3), 365-384.

Lesgold, A., Lajoie, S., Logan, D., & Eggan, G. (1990). Applying cognitive task analysis and research methods to assessment. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 325-350). Hillsdale, NJ: Erlbaum.

Lesgold, A., Lajoie, S. P., Logan, D., & Eggan, G. M. (1990). Cognitive task analysis approaches to testing. In A. Frederiksen, R. Glaser, A. Lesgold & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 325-350). Hillsdale, NJ: Lawrence Erlbaum.

Lidz, C., Jepsen, R. H., & Miller. B. (1997). Relationships between cognitive processes and academic achievement: Application of a group dynamic assessment procedure with multiply handicapped adolescents. *Educational & Child Psychology*, 14(4), 56-67.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), 517-548.

Lord, F. M., & Norvick , M. (1968). *Statistical theories of mental tests*. New York: Addison-Wesley.

Maris, E. M. (1995). Psychometric latent response models. *Psychometrika*, 60(4), 523-547.

Marshall, S. P. (1990). Generating good items for diagnostic tests. In N.

Frederiksen, R. Glaser, A Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 433-452). NJ: Lawrence Erlbaum.

Marshall, S. P. (1995). Some suggestions for alternative assessments. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitive diagnostic assessment* (pp. 431-453). Hillsdale, NJ: Lawrence Erlbaum.

Martin, J., & VanLehn, K. (1995a). A Bayesian approach to cognitive assessment. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cogntively diagnostic assessment* (pp.141-165). Hillsdale, NJ: Lawrence Erlbaum.

Martin J., & VanLehn, K. (1995b). Student assessment using Bayesian nets. *Human-Computer Studies*, 42, 575-591.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometricka*, 47(2), 149-174.

Master, G. N. & Mislevy, R. J. (1993). New views of student learning: Implications for educational measurement. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for new generation of tests* (pp. 219-242). Hillsdale, NJ: Lawrence Erlbaum.

Mayer, R. E. (1996). Learner as information processor: Legacies and limitations of educational psychologies second metaphor. *Educational Psychologist*, 31(3), 151-161.

Mayo, M. & Mitrovic, A. (2000). Using a probabilistic student model to control problem difficulty. In G. Gauthier, C. Frasson & K. VanLehn (Eds.),

Intelligent Tutoring Systems, 5th International Conference, ITS 2000 (pp.524-533). New York: Springer.

McCalla, G., Vassileva, J., Greer, J., & Bull, S.(2000). Active learning modeling. In G. Goos, J. Hartmanis, J. van Leeuwen (Series Eds.) & G. Gauthier, C. Frasson, & K. VanLehn (Vol. Eds.), *Intelligent tutoring systems* (pp.53-62). New York: Springer.

Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher*, 23(2), 13-23.

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment*, 10(1), 1-9.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.

Mislevy, R. J. (1993). Foundation of a new test theory. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for new generation of tests* (pp. 19-39). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cogntively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Rep. 632). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Gitomer, D. H. (1996, June). *The role of probability-based inference in an intelligent tutoring system* (CSE Tech. Rep. 413). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999, Jan.). *On the roles of task model variables in assessment design* (CSE Tech Rep. 500). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2000). *Evidence-centered assessment design (A submission for the NCME award for technical or scientific contributions to the field of educational measurement).* Unpublished manuscript, Educational Testing Service in Princeton.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Haertel, G. D., & Penuel, W. R. (2001, February). *Leverage points for improving educational assessment.* (CSE Tech Rep. 534). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., Breyer, F. J., & Johnson, L. (2001, March). *Making sense of data from complex assessments* (CSE Tech. Rep. 538). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002, October). *Design and analysis in task-based language assessment* (CSE Tech Rep. 579). Princeton, NJ: Educational Testing Service.

Mumford, M. D., Baughman, W. A., Supinski, E. P., & Anderson, L. E. (1998). In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 75-112). Mahwah, NJ: Lawrence Erlbaum.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14(1), 59-71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159-176.

Murphy, P. K., & Alexander, P. A. (2002). What counts? The predictive power of subject-matter knowledge, strategic processing, and interest in domain-specific performance. *Journal of Experimental Education*, 70(3), 197-214.

Nandakumar, R. (1994). Assessing dimensionality of a set of items—Comparison of different approaches. *Journal of Educational Measurement*, 31(1), 17-35.

Newell, A., Rosenbloom, P. S., & Laird, J. E. (1989). Symbolic architectures for cognition. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp.93-131). Cambridge, Mass: The MIT Press.

Norsays Software Corp. (2006). Netica: Application for belief networks and influence diagrams [user's guide]. Retrieved June, 2006, from http://www.norsys.com

Ohlsson, S. (1990). Trace analysis and spatial reasoning: An example of intensive cognitive diagnosis and its implications for testing. In A. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 251-295). Hillsdale, NJ: Lawrence Erlbaum.

Patel, V. L., & Arocha, J. F. (1995). Cognitive models of clinical reasoning and

conceptual representation. *Methods of Information in Medicine*, 34(1-2), 47-56.

Patel, V. L., Evans, D. A., & Kaufman, D. R. (1990). Reasoning strategies and the use of biomedical knowledge by medical students. *Medical Education*, 24(2), 129-136.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science*, 10(1), 91-116.

Patel, V. L. & Groen, G. J. (1993). Reasoning and instruction in medical curricula. *Cognition and Instruction*. 10(4), 335-378.

Patel, V. L., Groen, G. J., & Arocha, J. F. (1990). Medical expertise as a function of task difficulty. *Memory & Cognition*, 18(4), 394-406.

Pearl, J. (1988). *Probabilistic reasoning in intelligent system: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. NY: Cambridge.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24). (pp. 307-353). Washington, DC: American Educational Research Association.

Pellegrino W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Pellegrino, J. W., Mumaw, R., & Shute, V. (1985). Analyses of spatial aptitude

and expertise. In S. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 45-76). New York, NY: Academic Press.

Pylyshyn, Z. W. (1989). Computer in cognitive science. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 49-91). Cambridge, Mass: The MIT Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test.* Chicago: University of Chicago Press.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York: Springer.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement, 14(4), 361-373.

Reusser, K. (1993). Tutoring systems and pedagogical theory: representational tools, and reflection in problem solving, In S. Lajoie & S. J. Derry (Eds.), *Computers as cognitive tools* (pp. 143-177). Hillsdale, NJ: Lawrence Erlbaum.

Richardson, M. W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika*, 1(2), 33.49.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural

knowledge of Mathematics: Does one lead to the other? *Journal of Educational Psychology*, 91(1), 175-189.

Rost, J. (1990). Rasch model in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.

Roth, K. J. (1990). Developing meaningful conceptual understanding in science. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction*. Hillsdale, NJ: Lawrence Erlbaum.

Sackett, P. R. (1998). Performance assessment in education and professional certification: Lessons for personnel selection? In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection* (pp. 113-129). Mahwah, NJ: Lawrence Erlbaum.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 34, (4, Pt. 2).

Samejima, F. (1996). The graded response model. In W. J. van der Linden & Hambleton, R. K. (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Scarr, S. (1998). How do families affect intelligence? Social environmental and behavior genetic predictions. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp.113-136). NJ: Lawrence Erlbaum.

Schacter, D. L. (1989). Memory. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 683-725). Cambridge, Mass: The MIT Press.

Schiffrin, D. (1994). *Approaches to discourse*. Cambridge, Mass: Blackwell.

Schwartz, D. L., Biswas, G., Bransford, J. D., Bhuva, B., Tamara, B., & Brophy, S. (2000). Computer tools that link assessment and instruction: Investigating what makes electricity hard to learn. In S. P. Lajoie (Ed.), *Computers as cognitive tools, volume two: No more walls* (pp. 273-307). Mahwah, NJ: Lawrence Erlbaum.

Segers, M. S. R. (1996). Assessment in a problem-based economics curriculum. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 201-224). Norwell, Mass: Kluwer Academic Publishers.

Shavelson, R. J, (2000). *On the cognitive interpretation of performance assessment scores.* Unpublished manuscript.

Shavelson, R. J, & Ruiz-Primo, M. A. (1998). *On the assessment of science achievement conceptual underpinnings for the design of performance assessments: Report of year 2 activities.* (CSE Tech. Rep. No. 491). CA: University of California, Center for the Study of Evaluation.

Shavelson, R. J, & Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2003) *Evaluating new approach to assessing learning.* (CSE Tech. Rep. No. 604). CA: University of California, Center for the Study of Evaluation.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4-14.

Shepard, R. N., & Feng, C. (1972). A chronometric study of mental paper folding. *Cognitive Psychology*, 3(2), 228-243.

Shepard, R. N. & Metzler. J. (1971). Mental rotation of three dimensional objects.

Science, 171(3972), 701-703.

Simon, H. A. & Kaplan, C. A. (1989). Foundation of cognitive science. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 1-47). Cambridge, Mass: The MIT Press.

Sireci, S. G., Wainer, H., & Braun, H. (1998). Psychometrics. In P. Armitage, & T. Colton (Eds.), *Encyclopedia of Biostatistics*. London: John Wiley & Sons.

Snow, R. E. (1998). Abilities as aptitudes and achievements in learning situations. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp.93-112). NJ: Lawrence Erlbaum.

Snow, R. E., Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: an introduction. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for new generation of tests* (pp. 1-18). Hillsdale, NJ: Lawrence Erlbaum.

Sowa, J. F. (1987). Semantic networks. In S. C. Shapiro, D. Eckroth & G. A. Vallasi (Eds.), *Encyclopedia of artificial intelligence* (pp.1011-1023). New York, NY:John Wiley & Sons.

Sowa, J. F. (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*. San Mateo: Morgan Kaufmann.

Sowa, J. F. (2000) *Knowledge representation: Logical, philosophical, and computational foundation*. Pacific Grove, CA: Brooks/Cole Publishing Co.

Sowa, J. F. (1992). Semantic networks. In S. C. Shapiro (Eds.), *Encyclopedia of Artificial Intelligence* (pp. 1493-1511). Wiley, New York.

Sowa, J. F. (2005). Semantic Network. Retrieved from http://www.jfsowa.com/pubs/semnet.htm

Stout, W. (1987). A Nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55(2), 293-325.

Sugrue, B. (2000). Cognitive approaches to web-based instruction. In S. P. Lajoie (Ed.), *Computers as cognitive tools, volume two: No more walls* (pp. 133-162). Mahwah, NJ: Lawrence Erlbaum.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345-354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*,12(1), 55-73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 452-484). Hillsdale, NJ: Lawrence Erlbaum.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive

diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327-359). Hillsdale, NJ: Lawrence Erlbaum.

ten Berge, T., & van Hezewijk, R (1999). Procedural and declarative knowledge: An evolutionary perspective. *Theory & Psychology*, 19(5), 605-624.

Thissen, D.(1993). Repealing rules that no longer apply to psychological measurement. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for new generation of tests* (pp. 79-98). Hillsdale, NJ: Lawrence Erlbaum.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, 11(1), 1-13.

Tweney, R. D., & Walker, B. J. (1990). Science education and the cognitive psychology of science. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction.* (pp. 291-306). Hillsdale: Lawrence Erlbaum.

VanLehn, K. (1989). Problem solving and cognitive skill acquisition. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 527-580). Cambridge, Mass: The MIT Press.

VanLehn, K. (2001, April). OLAE: *A Bayesian performance assessment for complex problem solving.* Paper presented at the National Conference on Measurement in Education, Seattle, WA.

VanLehn, K., & Martin, J. (1998). Evaluation of an assessment system based on

Bayesian student modeling. *International Journal of Artificial Intelligence and Education*, 8(2), 179-221.

VanLehn, K., & Niu, Z. (2001). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education*, 12(2), 154-184.

Vanlehn, K. Niu, Z. Siler, S., & Gertner, A. (1998). Student modeling from conventional test data: A Bayesian approach without priors. *Proceedings of the 4th ITS'98 conference* (pp. 434-443). Heidelberg, Berlin: Springer-Verlag.

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement*, 24(2), 151-162.

Virvou, M., & Moundridou, M. (2000). Modeling the instructor in a web-based authoring tool for algebra-related ITSs. In G. Goos, J. Hartmanis, J. van Leeuwen (Series Eds.) & G. Gauthier, C. Frasson & K. VanLehn (Vol. Eds.), *Intelligent tutoring systems* (pp.635-644). New York: Springer.

Voss, J. F., & Post T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, M. J. Farr, & Glaser, R. (Eds.), *The Nature of expertise* (pp. 261-286). NJ: Lawrence Erlbaum Associates.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wenger, E. (1987). *Artificial intelligence and tutoring systems.* Los Altos, CA: Morgan Kaufmann.

Wertsch, J. V. (1985). *Culture and cognition: Vygotsky perspectives.* Cambridge, MA: Cambridge University Press.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, 45(4), 479-494.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. *International Journal of Testing*, 4(4), 303-332.

Williamson, D. M., Steinberg, L. S., Mislevy, R. T., & Behrens, J. T. (2003). *Creating a Complex Measurement Model Using Evidence Centered Design.* Unpublished manuscript.

Xiang, Y. (2002). *Probabilistic reasoning in muitiagent systems: A graphical models approach,* NY: Cambridge University Press.

Yamamoto, K., & Gitomer, D. H. (1993). Application of a HYBRID model to a test of cognitive skill representation. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 275-295). Hillsdale, NJ: Lawrence Erlbaum.

Yan, D., Almond, R., & Mislevy, R. (2003). *Empirical comparisons of cognitive diagnostic models.* Unpublished manuscript.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing

local item dependence. *Journal of Educational Measurement*, 30(3), 187-213.

Yin, Y., Ayala, C., & Shavelson, R. (2001). *Students' problem solving strategies in performance assessment: Hands on and minds on*. Unpublished Manuscript.

Zhou, Y. (2000). *Building a new student model to support adaptive tutoring in a natural language dialogue system*. Unpublished doctoral dissertation, Illinois Institute of Technology.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement*, 24(4), 293-308.

APPENDICES

Appendix A: Performance Assessment of Statistics Learning

Performance Assessment of Statistical Learning

<ANOVA>

McGill Statistics Tutoring Project
The Faculty of Education

McGill University

February, 2005

First name:_____ Last name:_____

Statistics courses you have taken (or equivalent course):

- EDPE 575 (Introduction course in statistics): __yes __no
- EDPE 676 (Intermediate Statistics: ANOVA): __yes __no
- EDPE 682 (Regression analysis & the General Linear Model) __yes __no

Have you used the McGill Statistics Tutor __yes __no

If you answer is yes, please describe about how long or how many times you used it and give comments about your experience.

Instruction:

1). In this assessment you will solve a data analysis problem using analysis of variance. In this booklet, you will be given a description of the research problem, method, and data sets, as well as the purpose of the study.

2). Read the problem description (you can refer back to it if necessary).

3). You will "carry out " a sequence of tasks to analyze the data using SAS. However you will be provided with the SAS commands and results at each step.

4). For each task, please answer all the questions using the results that are provided as needed.

Task One – Research Method and Data Collection

Description of the Study

A statistics consultant has been called to assist the police department of a large metropolitan city to evaluate its human relations training program for 45 recently hired police officers. The officers come from different assigned areas that they patrol. The areas are classified as: the upper-income area, the middle-income area, and the lower-income area. The training program was designed to improve police attitude towards minorities. The researchers wanted to investigate the effect of the duration of the training program on the officers' attitudes. The program durations that were compared were five, ten, or fifteen hours of human relations training. The officers were randomly assigned to one of these three course durations. The attitude of the officer toward minority groups was assessed following the program using a test developed and validated previously by the consultant (the data will be given in next task).

1). What are the dependent variables?

2). What are the independent variables?

3). What research questions motivated the study?

4). How was the sample of subjects obtained?

5). How were the subjects assigned to the conditions?

6). What attempts were made to control possible effects of extraneous variables?

Task Two – The Sample Data File [this task has been deleted, and you can skipped]

SAS Data Step:

```
data kirk;
input group $ duration $ attitude;
cards;
upper          five          24
upper          five          33
upper          five          37
upper          five          29
upper          five          42
upper          ten           44
upper          ten           36
upper          ten           25
upper          ten           27
upper          ten           43
upper          fifteen       38
upper          fifteen       29
upper          fifteen       28
upper          fifteen       47
upper          fifteen       48
middle         five          30
middle         five          21
middle         five          39
middle         five          26
middle         five          34
middle         ten           35
middle         ten           40
middle         ten           27
middle         ten           31
middle         ten           32
middle         fifteen       26
middle         fifteen       27
middle         fifteen       36
middle         fifteen       46
middle         fifteen       45
lower          five          21
lower          five          18
lower          five          10
lower          five          31
lower          five          20
lower          ten           41
lower          ten           39
lower          ten           50
lower          ten           36
lower          ten           34
lower          fifteen       42
lower          fifteen       52
lower          fifteen       53
lower          fifteen       49
lower          fifteen       64
;
proc print;
run;
```

SAS Output:

Obs	group	duration	attitude
1	upper	five	24
2	upper	five	33
3	upper	five	37
4	upper	five	29
5	upper	five	42
6	upper	ten	44
7	upper	ten	36
8	upper	ten	25
9	upper	ten	27
10	upper	ten	43
11	upper	fifteen	38
12	upper	fifteen	29
13	upper	fifteen	28
14	upper	fifteen	47
15	upper	fifteen	48
16	middle	five	30
17	middle	five	21
18	middle	five	39
19	middle	five	26
20	middle	five	34
21	middle	ten	35
22	middle	ten	40
23	middle	ten	27
24	middle	ten	31
25	middle	ten	32
26	middle	fifteen	26
27	middle	fifteen	27
28	middle	fifteen	36
29	middle	fifteen	46
30	middle	fifteen	45
31	lower	five	21
32	lower	five	18
33	lower	five	10
34	lower	five	31
35	lower	five	20
36	lower	ten	41
37	lower	ten	39
38	lower	ten	50
39	lower	ten	36
40	lower	ten	34
41	lower	fifteen	42
42	lower	fifteen	52
43	lower	fifteen	53
44	lower	fifteen	49
45	lower	fifteen	64

1). How does the SAS data step organize the data for processing using SAS?

2). How are the data organized in the output?

3). Construct a 3x3 table displaying the attitude scores obtained for the subjects in the group for each combination of conditions (group and duration of training).

Subject Attitude Values for Each Combination of Conditions

	Duration of Training		
Group	Five Hours	Ten Hours	Fifteen Hours
Upper income			
Middle income			
Lower income			

Task Three (A) – Descriptive Analysis of the Data Using Sample Statistics

SAS Program Statements:

```
proc means data=kirk mean;
var attitude;
class group;
proc means data=kirk mean ;
var attitude;
class duration;
proc means data=kirk mean var std stderr;
var attitude;
class group duration;
proc means data=kirk mean;
var attitude;
run;
```

SAS Output:

The MEANS Procedure Analysis Variable : attitude

group	N Obs	Mean
lower	15	37.3333333
middle	15	33.0000000
upper	15	35.3333333

The MEANS Procedure Analysis Variable : attitude

duration	N Obs	Mean
fifteen	15	42.0000000
five	15	27.6666667
ten	15	36.0000000

The MEANS Procedure Analysis Variable : attitude

group	duration	N Obs	Mean	Variance	Std Dev	Std Error
lower	fifteen	5	52.0000000	63.5000000	7.9686887	3.5637059
	five	5	20.0000000	56.5000000	7.5166482	3.3615473
	ten	5	40.0000000	38.5000000	6.2048368	2.7748874
middle	fifteen	5	36.0000000	90.5000000	9.5131488	4.2544095
	five	5	30.0000000	48.5000000	6.9641941	3.1144823
	ten	5	33.0000000	23.5000000	4.8476799	2.1679483
upper	fifteen	5	38.0000000	90.5000000	9.5131488	4.2544095
	five	5	33.0000000	48.5000000	6.9641941	3.1144823
	ten	5	35.0000000	77.5000000	8.8034084	3.9370039

```
The MEANS Procedure

Analysis Variable : attitude

        Mean
--------------
35.2222222
--------------
```

1). Construct a 3x3 table displaying the mean attitude score for each combination of the conditions (group and duration) include the marginal (row and column) means and the grand mean in your table.

Mean Attitude Score for Each Combination of Conditions

	Duration of Training		
Group	Five Hours	Ten Hours	Fifteen Hours
Upper income			
Middle income			
Lower income			

2). Describe the effect of each of the independent variables (group and duration) on the mean attitude score.

3). Make a table (similar to your table of means) giving the variance and standard error of the mean for each cell in the table.

For the variance

Duration of Training

Group	Five Hours	Ten Hours	Fifteen Hours
Upper income			
Middle income			
Lower income			

For the Standard Error of the Mean

Duration of Training

Group	Five Hours	Ten Hours	Fifteen Hours
Upper income			
Middle income			
Lower income			

4). Are there any differences in the sample variances? Describe any differences. What is their importance for an Analysis of Variance of the data?

5). Give an example of how you can use the standard errors to judge the size of the differences between cell means or marginal means?

Task Three (B) - Interpretation of a Graphic Representation of the Means

SAS Graphical Output for Group:

[see the following page]

1). Interpret the plot of the means of the Groups (pooled over Duration) in terms of the effect of the Group variable on mean attitude score.

2). Use the error bars to evaluate the size of the difference between groups. State your conclusions.

3). What is the statistical basis for using the error bars in this way?

SAS Graphical Output for Duration:

[see the following page]

1). Interpret the plot of the means of the Duration conditions (pooled over Groups) in terms of the effects of the Duration variable on mean attitude score.

2). Use the error bars to evaluate the size of the differences between levels of Duration. State your conclusion.

Task Four – The ANOVA Design

SAS program statements:

Also refer to the Description of the Study (Task One) and the SAS data step (in Task Two).

1). State the ANOVA design for these data.

2). What are the factors in the design what are their levels?

3). Is this a balanced design or an unbalanced design and why?

Task Five – The ANOVA Score Model

SAS Statement:

```
proc anova data=kirk;
class group duration;
model attitude=group duration group*duration;
run;
```

1). Write the ANOVA score model for your design.

2). Explain the formula used in the ANOVA score model.

3). What are the main effects? How are they interpreted?

4). What are the interaction effects? How are they interpreted?

5). What is the grand mean? How is it interpreted?

6). Identify the residual or error score and how it is obtained from subject's observed score on the dependent variable.

Task Six (A) – Estimating Effects

SAS ANOVA program statements:

```
proc anova data=kirk;
class group duration;
model attitude=group duration group*duration;
means group duration group*duration;
run;
```

SAS output:

The ANOVA Procedure

Dependent Variable: attitude

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2907.777778	363.472222	6.09	<.0001
Error	36	2150.000000	59.722222		
Corrected Total	44	5057.777778			

R-Square	Coeff Var	Root MSE	attitude Mean
0.574912	21.94074	7.728015	35.22222

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	2	141.111111	70.555556	1.18	0.3185
duration	2	1554.444444	777.222222	13.01	<.0001
group*duration	4	1212.222222	303.055556	5.07	0.0024

Observed and Residual Scores

The ANOVA Procedure

Level of group	N	attitude Mean	attitude Std Dev
lower	15	37.3333333	15.2299830
middle	15	33.0000000	7.2702918
upper	15	35.3333333	8.1474507

Level of duration	N	attitude Mean	attitude Std Dev
fifteen	15	42.0000000	11.1419414
five	15	27.6666667	8.7722506
ten	15	36.0000000	7.0101967

Level of group	Level of duration	N	attitude Mean	attitude Std Dev
lower	fifteen	5	52.0000000	7.96868873
lower	five	5	20.0000000	7.51664819
lower	ten	5	40.0000000	6.20483682
middle	fifteen	5	36.0000000	9.51314880
middle	five	5	30.0000000	6.96419414
middle	ten	5	33.0000000	4.84767986
upper	fifteen	5	38.0000000	9.51314880
upper	five	5	33.0000000	6.96419414
upper	ten	5	35.0000000	8.80340843

1). Compute the estimate of the main effect for each level of group.

2). Compute the estimate of the main effect for each level of Duration.

3). Show how the estimate of the interaction effect is calculated for any single combination of levels of Group and duration (pick a particular combination of levels of Group and Duration).

4). Write the score model in terms of the estimates of the effects for a particular combination of levels of group and duration (pick one).

5). Give your interpretation of the main effect for these data.

Task Six (B) - Estimating Residual Scores

SAS ANOVA program statements:

```
proc glm data=kirk;
class group duration;
model attitude=group duration group*duration;
output out=tests r=residual;
run;
```

SAS Output:

Obs	group	duration	attitude	residual
1	upper	five	24	-9
2	upper	five	33	0
3	upper	five	37	4
4	upper	five	29	-4
5	upper	five	42	9
6	upper	ten	44	9
7	upper	ten	36	1
8	upper	ten	25	-10
9	upper	ten	27	-8
10	upper	ten	43	8
11	upper	fifteen	38	-0
12	upper	fifteen	29	-9
13	upper	fifteen	28	-10
14	upper	fifteen	47	9
15	upper	fifteen	48	10
16	middle	five	30	-0
17	middle	five	21	-9
18	middle	five	39	9
19	middle	five	26	-4
20	middle	five	34	4
21	middle	ten	35	2
22	middle	ten	40	7
23	middle	ten	27	-6
24	middle	ten	31	-2
25	middle	ten	32	-1
26	middle	fifteen	26	-10
27	middle	fifteen	27	-9
28	middle	fifteen	36	0
29	middle	fifteen	46	10
30	middle	fifteen	45	9
31	lower	five	21	1
32	lower	five	18	-2
33	lower	five	10	-10
34	lower	five	31	11
35	lower	five	20	-0
36	lower	ten	41	1
37	lower	ten	39	-1
38	lower	ten	50	10
39	lower	ten	36	-4
40	lower	ten	34	-6
41	lower	fifteen	42	-10
42	lower	fifteen	52	0
43	lower	fifteen	53	1
44	lower	fifteen	49	-3
45	lower	fifteen	64	12

1). This SAS output gives the residual score for each subject. How were the residual scores are computed?

2). Interpret the estimated residual scores in the SAS output. What is the importance of large residual scores (positive and negtive)?

Task Seven – Analysis of Variance Table

ANOVA Table:

Dependent Variable: attitude

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2907.777778	363.472222	6.09	<.0001
Error	36	2150.000000	59.722222		
Corrected Total	44	5057.777778			

R-Square	Coeff Var	Root MSE	attitude Mean
0.574912	21.94074	7.728015	35.22222

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	2	141.111111	70.555556	1.18	0.3185
duration	2	1554.444444	777.222222	13.01	<.0001
group*duration	4	1212.222222	303.055556	5.07	0.0024

1). Explain what the columns are in the above ANOVA table

2). Explain what the rows are in the ANOVA table

3). What is relationship between the total SS, the Model SS, and the Error SS?

4). What is the relationship of the Model SS to the Group, Duration, and Group*Duration Sums of Squares?

5). What is the relationship of a mean square to the Sum of Square and Design of Freedom in any row?

6). What is the relationship of the F statistic to the mean squares?

Task Eight – Calculating and Using ANOVA Statistics

ANOVA Table:

Dependent Variable: attitude

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2907.777778	363.472222	6.09	<.0001
Error	36	2150.000000	59.722222		
Corrected Total	44	5057.777778			

R-Square	Coeff Var	Root MSE	attitude Mean
0.574912	21.94074	7.728015	35.22222

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	2	141.111111	70.555556	1.18	0.3185
duration	2	1554.444444	777.222222	13.01	<.0001
group*duration	4	1212.222222	303.055556	5.07	0.0024

1). Write an equation showing the decomposition of the total SS. Based on ANOVA table above, show how the total SS is decomposed into a sum of other sums of squares in the table.

2). What are the degrees of freedom of these sums of squares and how do you calculate them?

3). What are Mean Squares and how do you calculate them?

4). How do you get R^2 based on the Sums of Squares in ANOVA table. How is it interpreted?

5) How do you obtain the Root Mean Square Error? How does it relate to the "SEM" (Standard Errors of the Means) for the cells in the ANOVA design table (Task 3A)?

6). What is the Coefficient of Variation and how do you interpret it?

Task Nine – Testing Hypothesis in ANOVA

ANOVA Table:

Dependent Variable: attitude

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2907.777778	363.472222	6.09	<.0001
Error	36	2150.000000	59.722222		
Corrected Total	44	5057.777778			

R-Square	Coeff Var	Root MSE	attitude Mean
0.574912	21.94074	7.728015	35.22222

Source	DF	Anova SS	Mean Square	F Value	Pr > F
group	2	141.111111	70.555556	1.18	0.3185
duration	2	1554.444444	777.222222	13.01	<.0001
group*duration	4	1212.222222	303.055556	5.07	0.0024

1). State all of the hypotheses to be tested.

2). What are the F values and how are they obtained?

3). How are the F values used to test hypotheses?

4). What are the expected values of the F statistics under the null hypothesis for each hypothesis to be tested?

Why are the expected F values important in the decision to use a particular F statistic to test each hypothesis?

5). What is the labelled as "Pr > F" in the ANOVA table? How is it determined from the F value and the F distribution with the appropriate degrees of freedom?

Task ten – Testing Contrasts among Groups (This task can be skipped)

SAS Program Statement:

```
proc glm data=kirk;
class group duration;
model attitude=group duration group*duration;
contrast 'dur 1 vs 3' Duration 1  0  -1;
contrast 'dur 2 vs 3' duration 0  1  -1;
means duration/ tukey alpha=.1 cldiff;
run;
```

SAS Output of Pre-planned Contrasts:

The GLM Procedure
Class Level Information

Class	Levels	Values
group	3	lower middle upper
duration	3	fifteen five ten

Number of observations
The GLM Procedure

Dependent Variable: attitude

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	2907.777778	363.472222	6.09	<.0001
Error	36	2150.000000	59.722222		
Corrected Total	44	5057.777778			

R-Square	Coeff Var	Root MSE	attitude Mean
0.574912	21.94074	7.728015	35.22222

Source	DF	Type I SS	Mean Square	F Value	Pr > F
group	2	141.111111	70.555556	1.18	0.3185
duration	2	1554.444444	777.222222	13.01	<.0001
group*duration	4	1212.222222	303.055556	5.07	0.0024

Source	DF	Type III SS	Mean Square	F Value	Pr > F
group	2	141.111111	70.555556	1.18	0.3185
duration	2	1554.444444	777.222222	13.01	<.0001
group*duration	4	1212.222222	303.055556	5.07	0.0024

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
dur 1 vs 3	1	270.0000000	270.0000000	4.52	0.0404
dur 2 vs 3	1	520.8333333	520.8333333	8.72	0.0055

SAS Output of Tukey Test:

The GLM Procedure
Tukey's Studentized Range (HSD) Test for attitude

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.1
Error Degrees of Freedom	36
Error Mean Square	59.72222
Critical Value of Studentized Range	2.99758
Minimum Significant Difference	5.9813

Comparisons significant at the 0.1 level are indicated by ***.

duration Comparison	Difference Between Means	Simultaneous 90% Confidence Limits		
fifteen - ten	6.000	0.019	11.981	***
fifteen - five	14.333	8.352	20.315	***
ten - fifteen	-6.000	-11.981	-0.019	***
ten - five	8.333	2.352	14.315	***
five - fifteen	-14.333	-20.315	-8.352	***
five - ten	-8.333	-14.315	-2.352	***

1). What pre-planned contrast were tested for these data and why?

2). State the hypothesis being tested for each pre-planned contrast. Interpret the results.

3). What is the Tukey post hoc contrast? Interpret the results.

4). Compare the results of the pre-planned contrasts to those for the Tukey post hoc contrasts. What analysis (pre-planned vs post-hoc) would you choose for these data and why?

Task Twelve –Conclusions from the ANOVA

SAS Results: See previous displays in the booklet.

1). State you conclusions based on the results of this study.

2). If there was any interaction between the effects of group and duration of training, explain your interpretation of this interaction.

3). What are the limitations of the study and its results?

Appendix B: Solution Features of Task Five

Question one:

1. The score model is a linear equation.
2. The equation begins with a variable label representing the score on the dependent variable with appropriate subscripts to indicate the specific levels of factors used to classify a subject, and an index number for the subject.
3. Right of the equal sign is a sum of terms.
4. The first term is a symbol (Greek mu) for the population grand mean.
5. The next term is a symbol for a population main effect parameter of the first factor indexed according to the level of this factor (as it was indexed on the score variable). [FOR ON FACTOR MODELS]

   (question one) the last term is a symbol (variable name) for the error (i. e. residual) score having the same indexing as the score variable. (MISSING

   [FOR ONE FACTOR MODEL]

[FOR TWO FACTOR MODELS:

6. The next terms are symbols representing population main effect parameters for additional factors. Each effect parameter is indexed according to the level of this factor (as it was indexed on the score variable).
7. If you have a crossed design with two or more factors, terms representing the population two-way interaction effects for pairwise combinations of factors are included in the model. Each effect parameter is indexed according to the levels of these factors (as they were indexed on the score variable).

8. the last term is a symbol (variable name) for the error (i. e., residual) score having the same indexing as the score variable.]

Question 2:

9. An individual's score Xij in an ANOVA design can always be written in terms of the sum of two terms: the cell mean μj (i. e., the population mean of the individual's group in a one factor model) plus an error or residual score eij. (the deviation of the individual's score on the dependent variable from the group mean). (This is the Means Version of the Score Model).
10. The residual score eij is the deviation of the individual's score on the dependent variable Xij from the group mean μj.
11. In ANOVA models, each individual's score Xij on the dependent variable is expressed as a deviation score by subtracting the population grand mean μ.
12. [FOR ONE WAY DESIGNS] In a one-factor models, the population grand mean μ is also subtracted from the group mean μj on the right hand side of the Means Model equation. This expresses the population group mean as a deviation from the population grand mean μj- μ.
13. [FOR TWO WAY DESIGNS] This deviation of the group mean from the grand mean is called a Main Effect and written as a single main effect parameter for level j of the Factor alpha j.
14. [FOR TWO WAY DESIGNS] Because they are deviation scores of group means from the grand means, the effect parameters for all the levels of a factor add up to zero. A large effect of level j means that the mean of group j deviates substantially from the grand mean.

15. Population effects must be estimated from the data. Since they are differences between population group mean μj and the grand mean μ, they can be estimated from the sample estimates of these means.
16. In the one way ANOVA score model, the effect parameter is substituted for the difference between the group mean and the grand mean in the Means Model. Then the grand mean is added back to both sides. This expresses a subject's score as a sum of the grand mean μ, a main effect parameter alpha j, and a residual (error) score eij.
17. The identification of comparison or control groups is important for planning and contrasts that will be tested in ANOVA.
18. By decomposing the participants' scores into a grand mean, effect components, and a residual score, we can systematically investigate the additive effect of each component as a contribution to the subjects' scores.
19. Individual with large residual scores can be identified as atypical "outlier" subjects. This can be important in many analyses.

[FOR TWO WAY DESIGNS]

question 2

20. In two-Way ANOVA models, an individual's score Xijk in an ANOVA design can always be written in terms of the sum of two terms: the cell mean μjk (i. e., the population mean of the individual's group in a one factor model) plus an error or residual score eijk. (the deviation of the individual's score on the dependent variable from the group mean). (This is the Mean Version of the Score Model).

21. the residual score eijk is the deviation of the individual's score on the dependent variable Xij from the cell mean μjk.
22. In Two-Way ANOVA models, each individual's score Xijk on the dependent variable is expressed as a deviation score by subtracting the population grand mean μ .
23. In ANOVA models with two (or more) factors, there is a population main effect parameter for the effect of a level of each factor on the subject's score. Each of these is a deviation of the marginal mean for particular level of a factor from the grand mean.
24. In a two way design with two factors, interactions, say of level j of one factor and level k of a second factor, are equal to: μjk- μj- μk+ μ (i. e., the cell mean minus the marginal means for the levels of the two factors plus the grand mean).
25. Interaction effects are large when there is a large difference between the cell mean μjk and the marginal means μj. and μ.k for levels j and k (of Factor A and B).
26. By decomposing the participants' score into a grand mean, effect components, and a residual score, we can systematically investigate the additive effect of each component as a contribution to the subjects' scores.
27. Individuals with large residual scores can be identified as atypical "outlier" subjects. This can be important in many analyses.

Appendix C: A Score Rubric of Task Five

	Write (Process)	Explain (Semantic)

(a). Score:

(1) Y is a symbol for score variable "attitude toward minority".

(2) Index for score variable, ijk or i(jk) .

(3) Complete expression $Y_{i\,(jk)}$ for score of individual I in group j and duration k

(b). Grand Mean:

(4) μ symbol for population mean. Parameter pooled in j and k.

(c). Main Effect (Group):

(5) α symbol for main effect parameter

(6) j index for group (j=1, 2, 3)

(7) α_j main effect for group j

(d). Main effect for duration:

(8) β for main effect parameter

(9) k index for duration

(10) β_k main effect of duration k

(e). Interactive effect

(11) γ for interactive effect

(12) (jk) for index of cell in design

(13) $\gamma_{(jk)}$ for interactive effect of individual i in cell (j, k)

(f). Residual score:

(14) e= residual score

(15) i(jk) same index as score variable

(16) $e_{i(jk)}$ error score in group j, duration k and subject i

(g). Complete model equation:

(17) $Y_{i\,(jk)} = \mu + \alpha_j + \beta_k + \gamma_{jk} + e_{i\,(jk)}$

*17 X 2=34 items

Appendix D: Random Sampling Evidence Nodes in ModelEquation

Runs	True Nodes
01	09
02	02 11
03	03 10 16
04	06 19 21 24
05	05 16 18 25 26
06	04 08 11 15 21 23
07	02 05 09 10 15 21 25
08	05 07 08 13 14 17 18 21
09	02 04 09 11 14 16 20 24 25
10	01 05 07 08 09 14 17 18 19 21
11	02 03 05 06 08 11 12 16 17 18 22
12	01 02 03 06 07 09 10 11 13 16 21 24
13	03 04 07 08 10 11 12 15 18 19 21 22 25
14	03 05 06 07 11 12 13 14 17 18 20 21 22 26
15	01 02 03 07 08 10 11 12 14 15 17 19 21 25 26
16	01 02 03 04 05 07 08 09 10 11 12 14 15 17 19 21
17	01 02 03 04 05 06 09 10 11 13 14 17 18 19 20 24 25
18	02 05 07 09 10 11 12 13 14 15 17 18 20 21 22 23 25 26
19	01 03 05 06 07 08 09 11 12 13 14 15 16 18 19 20 23 24 26
20	01 02 04 05 06 07 08 11 12 13 14 15 16 19 21 22 23 24 25 26
21	01 02 03 04 05 06 07 08 09 10 11 12 13 14 19 20 22 23 24 25 26
22	01 02 03 04 05 07 08 09 10 11 12 13 16 17 18 19 20 21 22 23 25 26
23	01 03 04 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
24	01 02 03 04 05 06 08 09 10 11 12 13 14 15 16 17 19 20 21 22 23 24 25 26
25	01 02 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
26	01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

Appendix E: Random Sampling Evidence Nodes in Score Model

Runs	True Nodes
01	31
02	33 38
03	27 40 45
04	28 34 42 44
05	29 32 35 43 45
06	27 32 34 37 38 42
07	28 29 36 38 40 41 43
08	28 30 32 34 36 40 44 45
09	27 28 30 31 36 40 41 42 44
10	27 28 30 32 34 35 37 38 42 43
11	27 28 29 32 34 35 36 37 38 40 45
12	28 29 30 31 33 36 39 41 42 43 44 45
13	27 29 30 32 34 35 36 37 38 39 40 42 43
14	28 29 30 32 33 34 35 36 37 39 41 42 43 44
15	27 28 29 30 31 33 34 36 37 38 39 40 41 42 44
16	28 29 31 32 33 34 35 36 37 38 39 41 42 43 44 45
17	27 28 29 30 31 32 33 34 35 36 37 38 39 40 43 44 45
18	28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
19	27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45

Appendix F: Random Sampling Evidence Nodes in Whole Model

Runs	True Nodes
01	27
02	15 19
03	09 28 33
04	03 21 35 42
05	08 10 34 37 41
06	05 12 22 29 30 44
07	13 17 24 25 26 29 45
08	02 12 13 15 16 18 20 42
09	01 03 12 14 21 30 32 37 39
10	08 13 26 29 32 35 36 42 43 44
11	03 12 19 24 27 28 30 35 39 42 45
12	02 07 17 18 19 21 25 29 31 35 36 38
13	05 06 08 10 11 22 25 29 33 34 36 42 45
14	04 06 09 12 13 14 16 19 20 30 31 32 37 38
15	01 02 04 05 07 14 19 21 27 28 33 35 37 39 45
16	02 04 05 08 13 14 17 20 23 24 27 29 35 36 40 43
17	06 12 13 14 15 16 17 18 19 22 25 27 31 35 36 37 42
18	02 03 09 10 16 21 25 28 29 30 33 34 35 36 37 40 43 45
19	04 05 07 08 09 13 16 17 18 19 20 22 23 25 26 29 41 42 44
20	02 03 07 08 09 18 19 20 21 23 27 29 32 34 35 36 37 40 41 43
21	02 03 05 06 07 10 13 20 21 23 24 29 31 33 34 35 36 37 39 40 45
22	04 06 10 11 12 13 16 17 19 21 23 24 26 28 30 31 33 34 38 42 44 45
23	02 03 04 05 08 09 10 13 17 18 19 26 27 28 31 32 33 34 37 39 40 43 45
24	01 03 07 11 12 13 17 18 21 22 23 24 25 26 27 28 31 32 33 34 36 37 39 40
25	01 03 04 08 11 13 14 15 17 18 21 22 26 28 29 31 33 34 35 36 38 39 42 44 45
26	02 03 05 06 08 09 12 13 18 19 22 24 25 26 27 28 29 30 32 33 35 37 38 40 41 43
27	01 02 03 04 06 08 11 13 15 18 20 21 24 26 28 29 31 33 35 36 37 38 39 40 41 42 44
28	01 02 04 06 07 08 09 12 13 15 16 18 22 24 25 26 27 28 29 31 32 33 35 36 38 39 41 42
29	02 03 05 06 07 08 09 12 13 14 15 16 19 20 22 23 24 25 28 29 30 31 33 36 37 39 40 43 44
30	02 04 06 08 09 12 14 15 16 17 18 20 23 24 25 26 27 28 29 30 32 34 35 38 40 41 42 43 44 45
31	01 04 05 06 07 08 09 10 11 12 14 16 17 18 20 21 22 23 25 27 28 29 31 33 34 37 38 41 42 43 45
32	02 03 04 07 08 09 10 11 12 13 14 15 18 19 20 21 22 24 27 28 29 30 31 32 33 34 37 39 40 42 44 45
33	01 02 03 04 05 08 10 11 12 14 15 17 18 21 22 23 24 25 26 27 28 29 31 33 34 35 37 38 39 40 42 44 45
34	02 03 04 06 07 08 09 10 11 12 13 14 15 17 19 23 24 25 26 27 28 29 33 34 35 36 37 38 39 40 41 42 43 44
35	01 02 04 05 06 08 09 10 11 13 14 15 16 17 18 19 20 21 22 23 24 25 26 30 31 33 34 35 37 39 40 41 43 44 45
36	01 02 03 05 06 07 08 09 10 12 13 14 15 17 19 22 23 24 26 28 29 30 31 32 33 34 35 36 37 39 40 41 42 43 44 45
37	01 02 03 04 06 07 10 11 12 13 14 15 16 17 18 19 20 21 22 24 26 27 28 29 30 31 32 33 34 36 39 40 41 42 43 44 45
38	01 02 03 04 05 06 07 08 09 11 12 13 14 15 16 17 18 20 22 23 24 25 26 27 28 29 30 35 36 37 38 39 40 41 42 43 44 45
39	01 02 03 04 05 06 08 09 10 11 12 14 15 16 17 18 19 20 21 23 24 25 26 28 29 30 31 32 33 34 35 36 37 39 40 41 42 43 45
40	01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 26 27 28 30 31 32 33 34 35 36 37 39 40 41 42 43
41	02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 26 27 28 29 30 31 32 33 34 35 37 38 39 40 41 43 44 45
42	03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 38 39 40 41 42 43 44 45
43	01 02 03 04 05 06 07 08 09 10 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
44	01 02 03 04 05 06 07 08 09 10 11 12 13 14 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
45	01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45

Appendix G. Diagnostic Scoring System Student Number:_____

Procedural Scoring Components		Semantic Scoring Components	
Item Number	Value	Item Number	Value
01		27	
02		28	
03		29	
04		30	
05		31	
06		32	
07		33	
08		34	
09		35	
10		36	
11		37	
12		38	
13		39	
14		40	
15		41	
16		42	
17		43	
18		44	
19		45	
20			
21			
22			
23			
24			
25			
26			
Total			

Appendix H. Examination of the Robustness of the Models for the Student Data

Student #	ModelEquation	ScoreModel	FullModel	ε
1	0*	0*	0	0.01
2	0*	0*	0	0.01
3	0	0	0	0.05
4	0	0	0	0.05
5	0	0*	2	0.01
6	0	0	0	0.05
7	0	0	0	0.05
8	0	0	0	0.05
9	0*	0*	0	0.01
10	0	0	0	0.05
11	0	0	0	0.05
12	0	0	0	0.05
13	0	0	0	0.05
14	0*	0	0	0.01
15	0	0	0	0.05
16	0	0	0	0.05
17	0*	0*	0	0.01
18	0	0	0	0.05
19	0	0	0	0.05
20	0	0	0	0.05

- indicates that the ε is at 0.05