Order Number 9510653

An integration of cognitive theory and psychometrics: Analogical reasoning

Diones, Ruth Ellen, Ph.D.

City University of New York, 1994

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

AN INTEGRATION OF COGNITIVE THEORY AND PSYCHOMETRICS: ANALOGICAL REASONING

by

RUTH ELLEN DIONES

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1994

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Sept 16, 1994
Date

Roger E. Millsap
Chair of Examining Committee

SEPT 27, 1994
Date

Executive Officer

Alan Gross

David Rindskopf

Carol Kehr Tittle
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

AN INTEGRATION OF COGNITIVE THEORY AND PSYCHOMETRICS:

ANALOGICAL REASONING

by

Ruth Ellen Diones

Advisor: Professor Roger Millsap

The purpose of this dissertation was to explain the difficulty of SAT Verbal analogy items with cognitively relevant, model-based variables by using a componential item response model, the Linear Logistic Test Model (LLTM; Embretson, 1985; Fischer, 1973; Sheehan & Mislevy, 1990; Spada, 1978) in two phases: exploratory and confirmatory. Six main variables were selected from the cognitive literature, based on a model of analogical reasoning, for a priori empirical operationalization. There were three semantic memory variables: written vocabulary knowledge, common semantic relations, and a dichotomy of relations (intensional or pragmatic), and three working memory variables: rationale difficulty, alternative choice influencing decision-making and stem contextualization, and structure-mapping. These variables reflected the following different aspects of a prototypical examinee solving analogy items: 1) the knower, or declarative and procedural memory, 2) the processor, or the utilization of information processes, strategies, or induction/inference within working memory, and 3) the experiencer, or episodic memory.

The data for this study originated from three forms (November 1988, 1989 and 1990) of the SAT used by the Educational Testing Service for equating purposes. Three consecutive forms were necessary in order to provide a large enough set of analogy items: 80 items. The sample of 30,907 was selected from among nondisabled college bound seniors for whom English was the first language with a representative proportion of males and females and ethnicities. The study proceeded in several steps: the cognitive variables were first operationalized and the LLTM model assumptions were checked. Next, an exploratory phase was run on a random subsample in order to test out the LLTM cognitive variables and additionally, to ascertain the importance of the Educational Testing Service's test development taxonomies. Lastly, a final LLTM model was fit on the remaining sub-sample of examinees in a confirmatory phase, thereby assessing the generalizability of the results.

Several variables were important in explaining item difficulty: two semantic memory variables, common Semantic Relations and the key vocabulary frequency; two working memory variables, Level of Relation for structure-mapping and Stem Rationale Difficulty; and one taxonomic variable, Level of Abstraction. Implications for the information processing model, construct validity and test design were discussed.

Acknowledgements

In order to have accomplished this daunting and at times seemingly insurmountable task, I had to have and did receive tremendous support and encouragement from those around me. My husband provided warm arms of comfort and undertook more than his fair share of the household duties. My daughter learned to accept and then enjoy some time in child care. All members of my dissertation committee were accessible, helpful and supportive. I wish especially to thank my chairperson, Roger Millsap. He was great in so many ways. My thoughts also turn to my understanding and flexible boss, Howard Everson. In addition, David Rindskopf helped sort out some software problems and challenged my thinking. Very importantly, Alan Gross patiently answered endless lists of questions throughout my entire graduate career. Through all of her roles, Carol Kehr Tittle offered enthusiasm, interest and support. Further thanks are due to Isaac Bejar -- not only was he very influential in my choice of dissertation topic, he additionally participated as an outside reader. Special thanks must also go to Gerhard Fischer for "saving my life" -- he recompiled his program, LLTM, to meet my specific data requirements. I further wish to acknowledge Roger Chaffin's help and interest as well as the ETS and graduate student raters.

Table of Contents

Abstract iv

Lists of Tables ix

Lists of Figures xi

Chapter 1: Introduction 1

Chapter 2: The Model 11
Item Response Theory Background 11
Rasch Models 19
Linear Logistic Test Model 24
Conclusions 52

Chapter 3: Characteristics of Analogical Reasoning . . 56
SAT Analogy Items 56
Defining Analogical Reasoning 59
Approaches to Analogical Reasoning 62
Information Processing Approach 63
The Problem Approach 86
Memories and Concepts 94
The Model 102
Conclusions 105

Chapter 4: Variable Selection and Research Hypotheses . 108
Model Components 109
Semantic Memory Variables 109
Vocabulary Knowledge 109
Semantic Relations 110
A Dichotomy of Relations 114
Working Memory Variables 119
Stem Rationale Difficulty 119
Alternative Choice Variables 121
Structure-Mapping Variables 125
Other Hypotheses 128
Summary 129

Chapter 5: Methods 131
Subjects and Materials 131
Design and Procedures 134
Task Design Phase 134
Vocabulary Knowledge 135
Common Semantic Relations 135
Dichotomy of Relations 136
Stem Rationale Difficulty 136
Alternative Choice Variables 137
Structure-Mapping Variables 138
Other Variables 139
Initial Analyses 140
Descriptive Statistics 140

Testing of Model Assumptions 140
Exploratory Phase 142
Hypothesized Linear Models 142
Confirmatory Phase 144

Chapter 6: Results 147
LLTM Variables 147
Model Assumptions 152
The Analogy Test 152
Dimensionality and Independence 152
Number of Item Parameters 153
Exploratory Phase 154
Confirmatory Phase 166

Chapter 7: Discussion and Implications 171
Results 171
Implications 178
Construct Validity 178
The Processing Model Revisited 180
Psychometric Concerns 183
Conclusions 184

Appendix A 186

References 194

Lists of Tables

Embretson's Analogy Components 79

Semantic Taxonomy of Relations 112

An Example of the Intensional/Pragmatic Distinction . . 116

The Intensional/Pragmatic and Semantic Class Hierarchy 118

Contextualization Effects of the Alternative Pairs . . 122

An Example of Propositional Maps 124

Equating the Three SAT Forms Over Four Groups 131

Sample Sizes for the Four Groups 133

Coding of the Cognitive Variables for F2 143

Coding of the Other Variables for F3 144

Frequency Counts of the Qualitative Variables 148

Means, Standard Deviations and Ranges for the Quantitative Variables 149

Correlation Coefficients of the F2 Matrix 151

Analogy Test Psychometric Statistics across the Four Groups 152

Fit of Item Response Theory Models using Likelihood Ratios 154

Exploratory and Confirmatory Sample Sizes 155

Model Fit for the Three F Matrices 158

Likelihood Ratios Comparing the F Models to the Rasch Model 158

Component Difficulty Results for all Three F Matrices . 159

Results of Backwards Stepwise Elimination from F2 and F3 161

Final Exploratory Component Difficulties 162

The Relative Importance of the F2 Variables 163

Rasch Model and LLTM Predicted Difficulty Values . . . 164

Final Exploratory and Confirmatory Fit Indices 168

Final Component Difficulties for F2 and F3 169

The Semantic Taxonomy Check Off Form 190

Lists of Figures

Empirical Item Characteristic Curve 12

ICC's for Three IRT Models 13

Item by Raw Score Group Data Matrix 22

Defining a Variable using Item Difficulty 24

Construct Validation Research 34

Levels of Analysis 37

Paragraph Comprehension Model 2 45

Document Literacy Model 51

Hierarchical Levels 51

Spearman's Model 66

Sternberg's Model I 67

Sternberg's Model III 71

Pellegrino and Glaser's Interactive Model 76

Bejar et al.'s Normative Model 81

IPA Model 84

IPA Model with Memory 103

Exploratory Plot 167

Confirmatory Plot 170

A Reconceptualized Processing Model 182

Chapter 1: Introduction

Analogical reasoning has long been considered a primary component of intelligence (Spearman, 1923; Sternberg, 1977; Dawis & Siojo, 1972; Bejar, Chaffin & Embretson, 1991). In fact, the beginning of intelligence research was tied to tests, of which an integral component was analogies, and the psychometric tool of factor analysis, in order to make a study of differential psychology or individual differences. There were two foci of this work: 1) ranking individuals in terms of their abilities and 2) interpreting the meaning of factors derived from tests, which was often equivalent to discovering what kinds of tests load most strongly on the general g factor for intelligence (analogies, of course!) (Dawis & Siojo, 1972).

With the importance of analogical reasoning acknowledged, research was initiated in several domains. Some researchers have investigated the information processing components of analogy problem-solving (Sternberg, 1977; Barnes, 1980; Embretson & Schneider, 1989); others have used computers to model solution processes within the artificial intelligence paradigm (Holyoak & Thagard, 1989; Gentner, 1989; Dierbach, 1990; Shinn, 1989). These researchers presented global models of an average analogy problem solver; the models were designed to explain the behaving or thinking aspects of a prototypical individual -- differences between people were considered error.

Another branch of relevant research has focused on the "structural" properties of the mind. This approach was concerned with semantic/lexical memory (Rumelhart, 1989; Johnson-Laird et al., 1984), concept representation (Smith, 1988; Barsalou, 1989; Medin, 1989), and relations between concepts (Chaffin & Herrmann, 1987, 1988). These theories involved the organization of internal memory structures and described the knowing individual engaged in all kinds of activities needing declarative knowledge -- not necessarily the particular case of analogical reasoning. Instead, this research helped complete the global analogy problem-solving information processing model formulated for this study. Other contributing fields of research were concerned with inductive, inferential and hypothesis testing processes (Holyoak & Nisbett, 1988; Goldman & Pellegrino, 1984). For example, inductive or inferential processes revealed an individual utilizing working memory to test viable analogy rationales. Again, the overall goal of these research programs was to explain or model person performance in general.

Yet, analogies have continued to be of importance in testing because analogical reasoning is still construed as being fundamentally representative of aptitude or intelligence. Therefore, analogy items have comprised a part of the SAT's, GRE's, Miller Analogies, and so on. These tests were designed to select or to help in decision-

making; the most important function of these tests was to rank people. Since the tests have played a critical role in people's lives politically, socially, academically and emotionally, much time and money have been invested in understanding the psychometric properties of such tests. Research topics have included construct and predictive validity (Embretson, 1983), item characteristics, measurement error, scoring, equating (Hambleton & Swaminathan, 1985), differential item functioning (Millsap & Everson, 1993) and so on.

However, no matter what the purpose, cognitive researchers and test developers both have traditionally used analogy items of the form A:B::C:? and A:B::?:? to measure analogical reasoning. Therefore, research on persons and psychometrics have always been inextricably intertwined. Further, there are always two sides to every response: the person and the item. The purpose of a study determines the focus, person or item, but the person side could not be measured without the item and vice versa. Experimental research has focused on general operating principles of persons while, in contrast, psychometric research has focused on both inter- and intra-individual differences and item/test properties. Hence, both cognitive and psychometric research domains have contributed to an overall understanding of a person interacting with these particular item formats.

As a result, for the last fifteen years there has been a call for an explicit integration of cognitive research with psychometric analyses (Carroll, 1976; Whitely, 1977; Embretson, 1985; Mislevy, 1993; Yamamoto, 1990). These fields should not be evolving independently as they can inform and promote each other's research endeavors. On one side, cognitive theory can suggest some factors that may cause an item to be difficult. On the other side, if cognitive variables can explain item difficulty, then cognitive theory is informed (Glaser & Pellegrino, 1978; Diones et al., in preparation). Or, in helping to further cognitive research, perhaps another item format should be used in experimental research, a format that may more successfully extract analogical reasoning from a problem solver.

In fact, several articles have described psychometric properties of analogy items from a cognitive perspective (Embretson, 1984; Chaffin & Pierce, draft; Bejar et al., 1991; Schmitt, Dorans & Lawrence; Scheuneman and Steinhaus, 1987). In other content domains, Fischer (1973), Embretson & Schneider (1981), Embretson & Wetzel (1987) Mislevy (1981) and Sheehan and Mislevy (1990) have explained item difficulty in terms of relevant structural features of the item using a Linear Logistic Test Model (LLTM), a specialized Item Response Theory model (IRT). Usually, the cognitive literature provided person variables generally

important in content-specific problem-solving while IRT psychometric models predicted item performance in terms of person and item characteristics, ability and item difficulty. An advantage of IRT models is that each model contains both the item and person side of problem-solving.

However, the tie between cognitive research and psychometric models has rested on item difficulty. It is extremely difficult to model directly information processing components in terms of ability, under typical test conditions; in contrast to laboratory settings, there is no real control of actual problem-solving mechanisms. Only Embretson (1984, 1985) has attempted this, and her procedure required multiple testings -- she had to collect data for each cognitive subtask as well as for each item as a whole. Thus, it has been tacitly assumed that difficult items create more effortful cognitive processing in an individual. It has additionally been assumed that items have certain features or qualities "causing" difficulty. Therefore, item difficulty has borne the burden of both the item and person sides of problem-solving. The tie must then rely on demonstrating that chosen item features are indeed related to general cognitive processing model(s) found in the relevant literature. Indeed, the above studies utilized variables that differed greatly in how cognitive they were. There seems to have been levels in describing cognitive variables: a level relying on superficial features of the

item such as, "this is a fraction problem requiring addition and multiplication components" or a level representing, for instance, recall of information from semantic memory (Glaser & Pellegrino, 1978).

The present study continues this trend by uniting psychometric and cognitive research in a LLTM to explain and predict analogy item difficulty with cognitively relevant item features, which reflect a cognitive theory/model of analogical reasoning. The psychometric part of the analysis implements a LLTM (Embretson, 1985; Fischer, 1973; Mislevy, 1981; Sheehan & Mislevy, 1990; Spada, 1978) on a traditional standardized test, the Scholastic Aptitude Test (SAT). Fortunately, within an IRT model, the way the person and item characteristics relate is mathematically explicated (θ-β or person ability minus item difficulty). For the cognitive part, capitalizing on item difficulty parameters embedded within the IRT model, variables selected from the literature on analogical reasoning for a priori empirical operationalization are used to help explain item difficulty. Therefore, the general research question is: do the cognitive variables, three of which are semantic memory variables (i.e. written vocabulary knowledge, common semantic relations and a dichotomy of relations (intensional or pragmatic)) and three of which are working memory variables (i.e. rationale difficulty, alternative choice variables influencing decision-making and stem

contextualization, and structure-mapping), explain analogy item difficulty? These variables reflect the following different aspects of a prototypical examinee solving analogy items: 1) the knower, or declarative and procedural memory, 2) the processor, or the utilization of information processes, strategies, or induction and inference within working memory, and 3) the experiencer, or episodic memory, as demanded by a particular item.

Despite the current background of controversy surrounding multiple choice versus constructed response versus performance assessment approaches to testing, this study nonetheless focuses on SAT items. Traditional multiple choice tests do offer advantages: ease, time and cost of administration, objective scoring, and pragmatically, continued widespread usage. Thus, these items will continue to be used and must be better understood. Further, a multiple choice item type, which remains constant over all items, lends itself to IRT and LLTM analyses.

Therefore, similar to past research with LLTM, the approach to this research problem is multifaceted. Initially, the cognitive literature is reviewed to extract an overall, representative cognitive model or theory of analogy problem-solving, e.g. a normative model (Bejar et al., 1991). Then, variables consistently deemed important across the entire literature base are selected to exemplify

aspects of the overall model. Since these variables have some known characteristics, their effect on item difficulty is predicted in the form of research hypotheses.

Next, the cognitive variables are operationalized and the LLTM model assumptions are checked. Subsequently, an exploratory phase is conducted on a random subsample of examinees in order to test out the LLTM cognitive variables and additionally, since they were available, to ascertain the importance of ETS test development taxonomies. Lastly, the final "best fitting" LLTM model is fit on the remaining sub-sample of examinees in a confirmatory phase, thereby assessing the generalizability of the results.

Such an approach has had many positive uses. A LLTM analysis comprises an important element in construct validation research, so issues of construct validity have been developed within the context allowed by this particular psychometric model (Embretson, 1983). That is, the analyses can clarify exactly what is being measured by SAT analogy items. Additionally, each component, since it has been selected from cognitive theory, can provide psychologically relevant information about analogy problem-solving (Fischer, 1973). Also, guidelines and procedures for test and item construction can be enlarged and refined (Scheuneman and Steinhaus, 1987). For instance, important item features can be manipulated (Whitely & Schneider, 1981; Embretson & Wetzel, 1987). Further, perhaps a response to current test

format concerns would be to modify multiple choice analogy tests in ways that can incorporate the advantages of the other approaches (i.e. having more realistic types of problems) and provide diagnostically relevant scores without giving up the past advantages of standardized multiple choice items. In addition, this approach may be useful when checking forms for constant construct validity during equating (Mislevy, 1981; Mislevy, Sheehan & Wingersky, 1993). Further, the results can inform the psychometric community of the efficacy of modelling item difficulty using a fairly simple model; it may be that more complex models would be required such as multidimensional IRT (Bock, Gibbons & Muraki, 1988; McKinley, 1989), the Hybrid (Yamamoto, 1989, 1990; Mislevy & Verhelst, 1990), or an IRT model with a covariate(s) (Muthén, Kao & Burstein, 1991). Such models may be required when the data fail to conform to the underlying assumptions of LLTM.

In sum, this research project is designed to integrate a specialized IRT model, LLTM, and a cognitive model of analogical problem-solving by examining features that predict item difficulty. In Chapters 2 and 3, the requisite framework is developed for the study, which then lead into the research questions and hypotheses in Chapter 4 and methods and procedures in Chapter 5. In Chapter 2, some basics of IRT are presented, along with a formal explication of the mathematics of the LLTM and a survey of the previous

uses of the LLTM. Important issues such as admissible inferences implied by an IRT model, the meaning of item difficulty, intricacies of construct validity, complications due to strict model assumptions, and the ramifications if assumptions are violated, are an interwoven part of the discussion. Chapter 3 continues with a comprehensive overview of the cognitive literature relevant to analogical reasoning and describes the model used in this study. At this point, Chapter 4 lays out the selected cognitive variables and research questions. Then Chapter 5 presents the research design by describing the student population sampled and the SAT test and forms, delineating the procedures and methods used to operationalize the variables, and outlining the exploratory and confirmatory phases. Chapter 6 summarizes the results, and finally conclusions are discussed in Chapter 7.

Chapter 2: The Model

Naturally, an understanding of Item Response Theory (IRT) models is necessary in knowing what kinds of conclusions can be drawn from outcomes. Since the linear logistic test model (LLTM) is derived from the Rasch model, which in turn is a constrained IRT model, this section first presents a general overview of IRT models. Then the specifics of the Rasch model are delineated in terms of the meaning of the item difficulty parameter, the meaning of a measured variable and the implications for construct validity. Finally, the LLTM is described and explained. Issues of parameter estimation and model fit are covered. Examples of previous research using the LLTM are surveyed, with a focus on the variables used, the fit with cognitive theory, purported purposes and expressed conclusions. Then, insights are summarized.

Item Response Theory Background

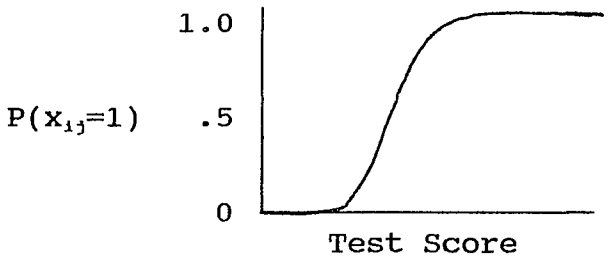
Like classical test theory, the purpose of all IRT models as a testing endeavor is the same: to measure or assign people a number representing their standing on an ability of interest in order to rank or compare (Andrich, 1988). Other valued uses of IRT include improved methods of equating test forms and checking for differential item functioning across population subgroups.

But IRT differs from classical test theory in a number of important ways. First, as its name suggests, item

response theory deals primarily with item and person level data. Second, an IRT function seeks to model mathematically the form of typical empirical item characteristic curves:

Figure 2.1

Empirical Item Characteristic Curve



where $P(x_{ij}=1)$ represents the probability of person j getting item i correct and test score is an observed measure of ability, $\Sigma_i x_{ij}$. Note that as the ability score increases, the probability of a correct response also increases, i.e. the curve is monotonic. Third, the IRT logistic functions[1] are defined by $\underline{\theta}$ and $\underline{\kappa}$, two vectors of one or more underlying *latent* or *unobservable*, but estimable, ability and item parameters, respectively. Assuming one ability, θ, Equation 1 depicts a general IRT function.

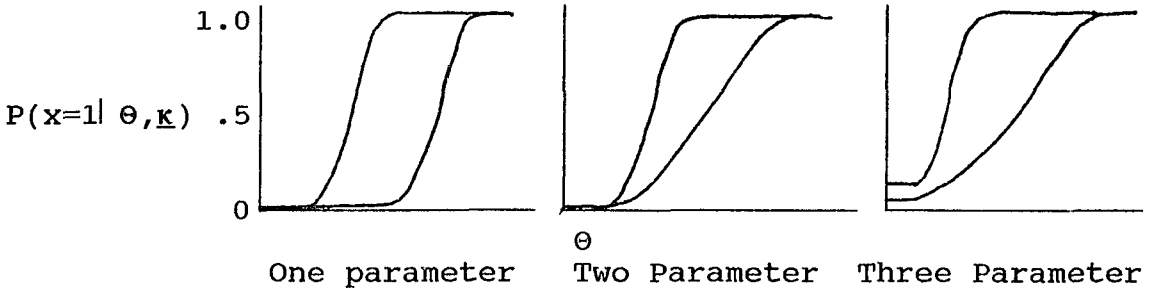
$$P(x_{ij}=1|\theta_j,\underline{\kappa_i}) = \varepsilon_i+(1-\varepsilon_i)\frac{e^{\alpha_i(\theta_j-\beta_i)}}{1+e^{\alpha_i(\theta_j-\beta_i)}} \qquad (1)$$

Item characteristic curves (ICC) can vary in distinct

ways depending on the number and kind of item parameters included in the model. Typically, three kinds of parameters are distinguished (Equation 1.): (1) item discrimination, α_i, which is proportional to the slope of the ICC, (2) item difficulty, β_i, or the ICC's point of inflection, and (3) a guessing indicator, ε_i, or the asymptote of $P(x_{ij}| \Theta,\underline{\kappa})$ as $\Theta \rightarrow -\infty$. A one parameter model, or Rasch model, is described only by item difficulty; its discrimination parameter is constrained to be 1 and the guessing parameter is set at 0; therefore, Rasch model item characteristic curves vary only in their location along the ability continuum. The two parameter model allows both item discrimination and difficulty to vary while the probability of guessing remains fixed at 0. The three parameter model additionally frees the guessing parameter; ε_i can be larger than 0. Figure 2.2 illustrates three possible IRT models.

Figure 2.2

ICC's for Three IRT Models



An added benefit particular to IRT is that precision at each

point along the ability continuum can be calculated in the form of test or item information functions (Hambleton & Swaminathan, 1985).

As is usual, IRT models are based on several assumptions, all of which are implied by the mathematical function describing the model. Dimensionality and local independence between items and examinees comprise two related assumptions. Hence, given a set of k items, local independence signifies

$$P(X_1, X_2, \ldots X_i \ldots X_k | \underline{\theta}_j, \underline{\kappa}_i) = \prod_{i=1}^{k} P(X_i | \underline{\theta}_j, \underline{\kappa}_i) \qquad (2)$$

or, "an examinee's responses to different items in a test are statistically independent (Hambleton & Swaminathan, 1985, p. 23)." Here, however, since most IRT models assume that the data are unidimensional[2], or one Θ ($\underline{\Theta}=\Theta$ -- it is scalar), only unidimensional IRT models will be discussed. With unidimensionality, all covariation between items is explained by one latent ability and after controlling for this Θ, there is local independence between all possible item pairs. For purposes of test design, this then requires a collection of items which draws on the same single ability.

To understand the meaning of a person score and thus also construct validity, however, unidimensionality must be further explored. The exact meaning of unidimensionality is

a bit controversial. It has been variously interpreted as: 1) a common ability, θ, underlying performance and 2) a property of the data (Bejar, 1992, personal communication). But what is meant by common ability? Whitely (1980) and Scheuneman and Steinhaus (1987) discuss it in terms of facets or multiple constructs. "..., unidimensional stimulus sets are not necessarily theoretically singular, since performance may depend on more than one construct (Whitely, 1980, p.99)" or "... a number of different abilities or attributes will affect the response to different items, but the resultant test score may still be 'unidimensional' in the sense that it meets the quantitative criteria usually applied... (Scheuneman & Steinhaus, 1987), p.4)." Such issues have important consequences for the kinds of LLTM's developed and implications for inferences and conclusions.

Following from the unidimensionality assumption, an additional theoretical property general to all IRT models is invariance. That is, across subpopulations, at any given level of ability, no matter how the subpopulations' underlying ability distribution varies, all groups have the same predicted probability of success. In effect, the ICC must be the same for all subgroups; item parameter estimates may differ only at a chance level across all subpopulations. This can be seen by looking at the expression $P(x=_{i,}1|\theta,\beta)$. The probability of success depends on a given level of

ability and item difficulty and nothing else. An alternative model could be $P(x_{ij}=1|\Theta,\beta,V)$, where V represents a demographic designation (Millsap & Meredith, 1993). In fact, IRT assumes that $P(x_{ij}=1|\Theta,\beta,V) = P(x_{ij}=1|\Theta,\beta)$.

Another commonality between IRT models is an inherent scale indeterminacy which can be arbitrarily fixed by setting either Θ or β to have a mean of 0 and standard deviation of 1. This becomes apparent in the following way: if Θ and β are transformed by adding a constant k to each to get $\Theta^*=(\Theta+k)$ and $\beta^*=(k+\beta)$, it doesn't matter -- $p(\Theta)=p(\Theta^*)$ because $\Theta^* - \beta^* = \Theta - \beta$. Therefore, it follows that the item characteristic curves are invariant with respect to additive transformations of the ability and difficulty parameters. An implied offshoot is that item difficulty and ability are on the same scale. However, this indeterminacy in scale must be resolved in order to compare people or items.

Another general property of IRT models is specific objectivity. With specific objectivity, estimates of item parameters do not depend on the sample used to calibrate the items, and estimates of person ability do not depend on which items are used for measuring. "...a comparison is specifically objective if it depends exclusively on properties residing in the objects (persons/items) and is invariant with respect to the instruments by means of which

a comparison is made (Scheiblechner, 1977)." For Rasch models particularly, this means that there is a separation between item and person parameters in the model.

Given the above, IRT models may be conceptualized in various ways. For instance, the model is probabilistic, $P(x_{ij}=1|\Theta)$, a Bernoulli with only x=1 or 0 as possible outcomes, and a nonlinear regression model, i.e. an exponential mathematical function which explains or predicts the probability of a correct response given the confrontation of a person of a certain latent ability level with an item with certain characteristics. Or, it is like a nonlinear one factor model -- an unidimensional latent trait model where Θ is the common factor. Further, given its form, the model is comparable to a loglinear model with two fixed main effects, that due to persons and the other due to items; note that there is no interaction term, only $(\Theta - \beta)$.

A major task in IRT is providing "good" estimates of the unknown parameters -- there are N Θ's, or one for each person, and k α_i's, β_i's and ε_i's, an ordered set for each item. Estimation has not been an easy task and many methods have been proposed. Various methods of maximum likelihood estimation are usually used: (1) conditional maximum likelihood (CML; Fischer, 1973, 1983), (2) joint maximum likelihood called unconditional maximum likelihood by Wright and Panchapakesan (1969), (3) marginal maximum likelihood utilizing prior distributional assumptions (Baker, 1992) and

(4) a Bayesian estimation procedure building on marginal maximum likelihood (Baker, 1992). Since the LLTM uses CML, only the relevant method will be discussed below.

Estimation issues aside, an important concern is whether the model reasonably fits the data or if its underlying assumptions are upheld. One approach to model fit is to determine the congruence of the empirical item characteristic curve (using observed data) and an IRT model's predicted item characteristic curve (using estimated item and ability parameters to calculate the P values). Another overall goodness of fit procedure is a likelihood ratio test, but, given large samples, the test is nearly always significant. Factor analysis can be used to check for unidimensionality. Stout (1987) also provides a nonparametric procedure which has been quite useful in verifying dimensionality assumptions. In addition, Hattie (1984) surveyed item fit statistics most often used in conjunction with a Rasch approach -- infit and outfit statistics are useful at the item level. Also, Hambleton, Swaminathan & Rogers (1992) provide a summary of common fit methodologies; they further noted that IRT fit is an underdeveloped area which must be further researched. Details of the methods used to check model assumptions and fit will be discussed throughout the LLTM literature review and in Chapter 5.

Rasch Models

Rasch models comprise a family of constrained IRT models[4] (Hambleton & Swaminathan, 1985). As noted above, one ability parameter is estimated for each person and only item difficulty is allowed to vary -- all item discrimination parameters are set to 1 and it is further assumed that no one guesses. Therefore, the function takes a simpler form:

$$P(x_{ij}=1|\theta_j,\beta_i) = \frac{e^{(\theta_j-\beta_i)}}{1+e^{(\theta_j-\beta_i)}} \qquad (3)$$

Given the importance of item difficulty in Rasch models and LLTM, defining item difficulty and determining how to measure it becomes imperative. As the name suggests, item difficulty should indicate, in comparison to the other items, how difficult an item is. Item difficulty has been operationalized in several ways -- some definitions are intuitive and some are purely statistical. In classical test theory, **p**, the observed proportion of examinees answering an item correctly, is a measure of item difficulty. While intuitively this makes sense, the value of **p** depends on the ability of the group administered the item. For example, an easy item in a high ability subpopulation may be quite difficult for a low ability sub-

group. Delta is a statistic developed by ETS, that like **p**, is sample-dependent, but that has better measurement properties, mostly having to do with linearity issues. Another definition, and the one focused on here, is the Rasch IRT ß parameter. To reiterate, ß is statistically defined as the threshold point on the ability continuum where 50% of examinees will get the item correct or alternatively, the point on the plot of the item characteristic curve where concavity changes or the point of inflection. Further, if the model fits the data and if the model's assumptions are met, the ß estimate, $\hat{\beta}$, offers other benefits over **p** and delta: (1) due to specific objectivity, estimation of ß has been separated from the ability of the calibration group, (2) ß and θ are on the same scale, and (3) ß is on an interval measurement scale. There are, of course, other ways to model item difficulty or to represent item difficulty not mentioned here[5].

In addition, CML estimation of item difficulty and person parameters is possible because of a special property of Rasch models. Using an argument by mathematical induction, the assumption of independence of all people and items, and conditional probability, Rasch (1960) showed a separation of parameters and proved sufficiency of r_j, a person's raw score, as a statistic for ability, and of s_i, an item's raw score, as a statistic for item difficulty.

Therefore, when estimating a person's ability, the item parameter drops out and the sufficient statistic s_i may be used in its place. This is also true when estimating difficulty; the ability parameter drops out and r_j fills in instead. As a consequence, difficulty and ability parameters can be estimated separately just by using the appropriate conditional probability expressions and sufficient statistics. For instance, the conditional likelihood of a response matrix given all possible raw scores, all item parameters $\underline{\beta}$ and observed data is (Baker, 1992, p.139)

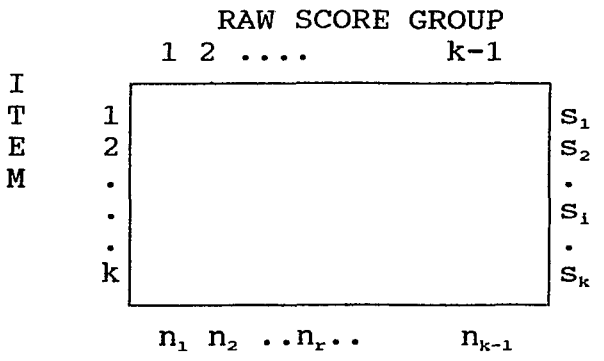
$$P([x_{ij}]\,|n_0, n_1, \ldots, n_k, \beta_1, \beta_2, \ldots, \beta_k) = \frac{\exp\left(-\sum_{i=1}^{k} s_i \beta_i\right)}{\prod_{r=0}^{k} [\gamma(r,\beta)]^{n_r}} \quad (4)$$

Note that $\gamma(r,\beta)$ is a symmetric function taking into account all possible ways, C^{n}_{r} ways, a particular r may occur, n_r is the number of people having score r, the product is over all possible r scores 0 through k, and the numerator shows in part the probability of a response matrix, $[x_{ij}]$. Further, CML estimates are consistent, fully efficient and asymptotically normal (Mislevy, 1981).

A further result of the above properties is that anyone with the same raw score ends up being assigned the same θ estimate. As a consequence, the 1,0 data can be presented in an item by raw score matrix (See Figure 2.3):

Figure 2.3

Item by Raw Score Group Data Matrix



where s_i represents item raw score, n_r symbolizes the number of individuals in each raw score group and matrix cells contain frequencies of passing the item within each group. This fact yields simpler estimation of ability parameters for CML estimates; in other IRT models, ability is an incidental parameter in that as $N \rightarrow +\infty$ the number of parameters also gets increasingly larger. By using raw score groups, the number of which are determined by the number of items, the number of Θ estimates is finite. Note further that the scores 0 and k are not included in the matrix as they do not lead to estimable 0 values. Further, items which every person answered either correctly or incorrectly need to be discarded for the same reason.

Although all IRT models allow for a fundamental level of specific objectivity, Rasch models permit additional conclusions. For example, this benefit appears when

comparing two people on item i using the logistic form odds ratio of a correct response[3]:

$$\ln O_{i1} - \ln O_{i2} = \Theta_1 - \beta_i - (\Theta_2 - \beta_i)$$

$$\ln(O_{i1}/O_{i2}) = \Theta_1 - \Theta_2 \ .$$

Note that β_i drops out. If it is desirable to compare items, the log-odds ratio of person j on two items is formed. Again, in this case, the Θ conveniently drops out. For Mislevy (1981), this occurrence is an indicator of an interval scale. "For example, the difference between β_i and β_j specifies in log-odds units how much more likely item j is to be answered correctly than item i. The result holds regardless of the absolute locations of β_i and β_j along the continuum. (p. 10)" Baker (1992) and Fischer (1973) also noted that the β's lie on a difference scale.

The development of the above details now allows for a meaningful definition of a measured variable (Wright & Stone, 1979; Mislevy, 1981). First, notice that the model predicts the outcome of a person with a given level of ability, confronting an item of a certain level of difficulty. For example, a person randomly sampled from a subgroup having an ability of Θ=1 confronting an item with a difficulty of β=-0.5 has a probability equal to .82 of correctly answering the item. In addition, recall that each test measures one variable; it is unidimensional and can be visualized as one line. Since ability and difficulty are both on the same scale, they can be defined in terms of each

other: [Please see Figure 2.4.]. Thus, the goal of a test is to locate each person, in terms of her measured level of ability, on this variable line. That is, the point on the line at which this person has a 50-50 chance of getting a correct answer, as in Figure 2.4:

Figure 2.4

Defining a Variable using Item Difficulty

"A person's ability is indicated as a point along the dimension defined by the item thresholds, ... (Mislevy, 1981, p. 10)". Hence, the difficulties of items that compose a test define the dimension or measured variable. Item difficulties are in fact "the operational definition of what the variable measures" (Wright & Stone, 1979). Therefore, any meaning attached to item difficulty in turn gives meaning to the measured variable and, further, to construct validity[6]! The next section describes a preferred method of studying item difficulty, the LLTM.

Linear Logistic Test Model

The LLTM was initially developed by Fischer (1973, 1983) and others (Scheiblechner, 1972) and expanded by Embretson (Embretson & Schneider, 1981; Embretson & Wetzel, 1987) and Mislevy (1981; Sheehan & Mislevy, 1990; Mislevy et

al., 1993). The focus in LLTM is on the Rasch item difficulty parameter and the goal is to find cognitively relevant item features or components that explain item difficulty. LLTM has been a useful model in diverse contexts and serves several goals.

In a review of the then existent literature, Fischer & Formann (1982) noted that in Europe, IRT was designed as a theory of "psychological measurement" which was used to model problem-solving processes and the LLTM was primarily applied in situations requiring an explanation of item difficulty in terms of "hypothetical cognitive structures and operations." Hence, difficulty is expressed in terms of "components" or features of the item called task processes. Fischer did not explicitly define these terms so their meaning must be inferred from research examples.

According to Fischer (1973), the LLTM approach is best conducted on content area tests that have a limited number of rules or cognitive operations. For example (Fischer, 1973), curriculum development could be furthered by parsing instructional units into discrete psychological pieces and estimating the difficulty of the whole in terms of the difficulty of smaller, meaningful pieces. With this diagnostic information, instructional theory and practice could be improved. One school subject examined was calculus differentiation rules, e.g. for polynomials, products, quotients, compound functions, trigonometric, exponential

and logarithmic functions. Each item was coded 1 or 0 if it required or did not require the operation of a rule and each item was coded on all rules; an item could require several rules for correct solution. So $x^3(x^2 + 1)^5$ would be coded 11010000 as it is the **product** of a **polynomial** with a **compound function.**

Cognitive theory enters the analysis with the choice of components or operations and the decision of how they should be coded. Here, Fischer (1973) originally had decided to count the number of times a rule was applied, a quantitative coding of items, but the results did not make sense. It seemed that once a rule was learned and could be successfully applied, it did not matter how many times it was needed. This provided some psychological feedback on problem-solving. Fischer finally concluded that "the difficulty of (differentiation) problems can be approximately explained through the assumption of seven psychologically meaningful cognitive operations. (p. 370)" Yet, Fischer's 'cognitive' coding was based on surface features of the item. Others (Fischer & Formann, 1982; Tatsuoka, 1990) have done similar kinds of analyses on different types of items, all easily coded on surface features. Formann (1973) examined Raven's Progressive Matrices, Scheiblechner (1972) studied logic items and Smith and his colleagues (Smith & Green, 1985; Kramer, Smith, & Kubiak, 1989; Kramer & Smith, 1990; Smith, Kramer & Kubiak,

in press) investigated spatial/perceptual items.

Mathematically, Fischer's LLTM is a direct extension of Rasch's latent trait item response model. The Rasch model is extended into the LLTM by using a mathematical function to express a transformed difficulty parameter:
$\beta^*_i = \Sigma_m f_{im}\tau_m + c$, where β^*_i is expressed as a weighted, linear, additive function of cognitive components (τ_m's are the unknown weights and f_{im}'s are the coded components) and should be, as nearly as possible, predictive of the Rasch β_i's. This linear combination of components replaces the original β_i's.

$$P(x_{ij}=1|\theta) = \frac{e^{(\theta-\beta_i^*)}}{1+e^{(\theta-\beta_i^*)}} \qquad (5)$$

$$P(x_{ij}=1|\theta) = \frac{e^{(\theta-(\Sigma_m f_{im}\tau_m+c))}}{(1+e^{(\theta-(\Sigma_m f_{im}\tau_m+c))})} \qquad (6)$$

This formulation may be further clarified:
$\beta^*_i = \Sigma_m f_{im}\tau_m + c$, or preferably, $\underline{\beta^*} = F\underline{\tau}$, where $\underline{\beta^*}$ is the vector of all k transformed β^*_i's, F is a k by m+1 design matrix of full rank (Baker, 1992) delineating the cognitive processes as well as a unit vector to take into account the normalization constant, c, where k is the number of items and m is the number of cognitive components, $m \leq k$; f_{im} is an element of the F matrix; $\underline{\tau}$ is a vector of estimated weights, or component difficulty parameters (Baker, 1992); and c is

similar to the intercept in a traditional general linear model, designed to solve the indeterminacy problem: $\Sigma\beta^* = \Sigma\beta = 0$ (Fischer, 1973).

$$\underline{\beta^*} = \begin{bmatrix} 1 & f_{11} & \cdots & f_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & f_{k1} & \cdots & f_{km} \end{bmatrix} \begin{bmatrix} c \\ \tau_1 \\ \vdots \\ \tau_m \end{bmatrix} \tag{7}$$

$$\underline{\beta^*} = \begin{bmatrix} c+f_{11}\tau_1+f_{12}\tau_2+\cdots+f_{1m}\tau_m \\ \vdots \\ c+f_{k1}\tau_1+f_{k2}\tau_2+\cdots+f_{km}\tau_m \end{bmatrix} \tag{8}$$

Again, in traditional general linear model terms, F is like an X design matrix and τ's are like Beta weights. Therefore, each β^* is expressed as a "best" linearly weighted combination of variables using a weighted least squares criterion. Each item is coded on all components and the coding can be qualitative (dummy variables) or quantitative. Note also that the LLTM is a restricted form of the Rasch model. Since $m \leq k$, a reduced number of components are being estimated, meaning that there are fewer item parameters to estimate. Thus, the Rasch model is, in comparison, a saturated model in that all β parameters are estimated. When F is a modified k x k identity matrix (to fix the scale indeterminacy), then LLTM becomes the Rasch model.

Procedures for estimating the τ's extend directly from

the usual CML methodology. Taking the conditional likelihood above, equation (4), but substituting in $\Sigma_m f_{im}\tau_m$ (c may be set to 0 because the conditional likelihood is independent of the normalization of item parameters), provides

$$P((s_m)\,|\,(n_r),((f_{im})),(\tau_m)) = \sum_{(s_i)} P((s_i)\,|\,(n_r),(\beta_i)) = [C]\frac{\exp(\Sigma_m s_m \tau_m)}{\Pi\gamma_r^{n_r}} \tag{9}$$

where $s_m=\Sigma_i s_i f_{im}$ is a minimal sufficient statistic for each component summed over all items, (n_r) is a marginal vector of raw score totals (see Figure 2.3), $[C]$ is a combinatoric representing the number of all possible matrices with the observed raw score marginals, and γ_r^{nr} is again a symmetric function. However, since this expression is proportional to equation (4) on page 21, the derivative of the log likelihood of (4) may be taken with respect to τ_α, set equal to 0, and solved for τ_α. This is not a simple computational task. To simplify the calculation of symmetric functions, Fischer and Formann (1972) developed a gradient method. Further details may also be found in Fischer's 1983 article.

CML estimation offers a major advantage over other estimating procedures in model fitting: "The CML method allows the computation of the conditional likelihood as a function of the τ_m's and the direct comparison of this with

the conditional likelihood function of the Rasch Model (Fischer & Formann, 1982, p. 400)." A likelihood ratio, given the model, is thus formed following a χ^2 distribution with df=k-1-m. Other alternative models can be tested using likelihood ratios (Fischer & Formann, 1982; Lane, 1991; Whitely & Schneider, 1981; Embretson & Wetzel, 1987). Hierarchical nesting is also possible. Although phrased in a likelihood ratio terminology, a model that fits well would minimize the sum of the squared differences between $\hat{\beta}$ and β^*, similar to a traditional general linear model. Programs like LLTM (Fischer, 1988) and LINLOG (Embretson, 1989) start with raw correct/incorrect data, and provide τ estimates as well as log likelihoods.

Spada (1978, 1982), working within Fischer's framework, modeled two alternate views of the human development of proportional reasoning by using balance problems. His models were based on "task structure hypotheses, that is, on hypotheses about cognitive operations..."; he proposed that any given item could require up to eight cognitive operations. For example, one cognitive operation was "attention to and deduction from different amounts of weights". Note that whereas Fischer's (1973) cognitive skills were superficial item features which represented operations, Spada's were far more cognitively oriented. The F matrix coded which operations were deemed necessary for

each item.

While the F matrix was the same under each theory, expectations for response patterns and ICC's did vary. The first theory, a common deterministic one originating with Piaget, was that a person has or does not have a skill and that learning is incremental or qualitative in all or nothing steps. Such a theory suggested a particular pattern of responses:

0 0 0 0
1 0 0 0
1 1 0 0
1 1 1 0
1 1 1 1

and items would have Guttman ICC's, which go in a step-like manner from $P(x=1)=0$ to $P(x=1)=1$. This view was not modelled with LLTM. The other addressed theory of development viewed learning as degrees of incremental change. So while the same cognitive operations were postulated, the probability of a correct response would be $0 \leq P(x=1) \leq 1$ and ICC's would be similar to Figure 2.2; LLTM was an appropriate model in this case. Results indicated that the deterministic model was unsuitable (rejected) and while the probabilistic model fit significantly better, it still needed to be improved.

Spada (1982) further pointed out a drawback of LLTM: the components are sequentially dependent, not independent. That is,

$$\frac{e^{(\theta-\beta^*)}}{1+e^{(\theta-\beta^*)}} \neq \prod_{j=1}^{m}\left[\frac{e^{(\theta-\tau_m)}}{1+e^{(\theta-\tau_m)}}\right]^{f_{im}} \qquad (10)$$

If this equivalency was in fact true, modelling and test construction would be greatly simplified.

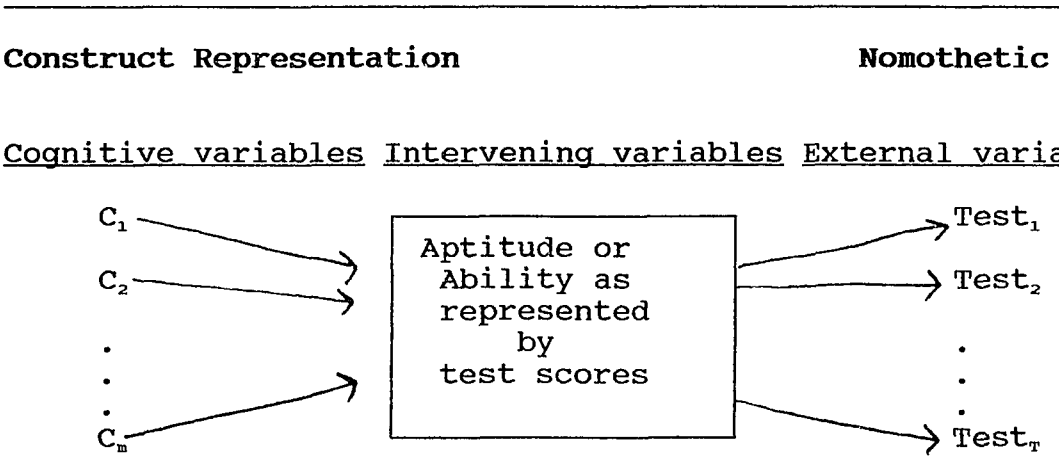
Embretson (Whitely & Schneider, 1981; Embretson & Wetzel, 1987) expanded upon Fischer's work in several ways. First, she explicitly connected her cognitive variables to information processing models and discussed different levels of cognitive variables. She also developed new models: a Multicomponent Latent Trait Model and General Multicomponent Latent Trait Model (Embretson, 1982, 1984, 1985), a dynamic testing model (Embretson, 1992a) and a model for learning and change (Embretson, 1991). The only model of additional interest for this project is the General Multicomponent Latent Trait Model as it is a multicomponent model with a LLTM model for each component. (This model will be discussed further while surveying the analogical reasoning literature.) Notwithstanding, the purpose behind her research has always been test validation and she viewed LLTM as an important tool in this process. In fact she considered any cognitive component analysis of aptitude, like LLTM, part of "a new type of validity research [--] cognitive component analysis of aptitude seeks to understand test validity by identifying information processing components that contribute to performance" (Whitely &

Schneider, 1981).

Embretson (1983, 1985) portrayed construct validation research as being comprised of two main parts, nomothetic span and construct representation, where "*construct* refers to a theoretical variable that may or may not be a source of individual differences (Embretson, 1983, p. 180)." Traditional validation research falls under a *nomothetic span* approach; it is in fact an individual differences analysis exploring correlations of an examined test with other external measures. Nomothetic span focuses on inter-test subject variability. In contrast, "*Construct representation* is concerned with identifying the theoretical mechanisms that underlie task performance. (Embretson, 1983)" The goal is task decomposition and task variability -- an investigation that cognitive investigators have been engaged in for a while. Here again is the contrast between individual differences research and experimental research. These two aspects of validation research are connected endeavors which can be pictorially depicted (C's designate components; Embretson, 1985, p. 197):

Figure 2.5

Construct Validation Research



Note. From Test Design: Developments in psychology and psychometrics (p. 197) by S. E. Embretson, 1985, Orlando: Academic Press. Copyright 1985 by Academic Press. Reprinted by permission.

Thus, the LLTM is ideal because it combines aspects of both subject and task variability "since it specifies how person and item parameters on a latent construct combine to determine response accuracy (Whitely, 1981) [$P(x_{ij}=1|\Theta)=\{1+\exp-(\Theta-\beta^{*})\}^{-1}$]." In addition, embedded in LLTM is a cognitive model, the F matrix, which specifies the task structure in terms of item difficulty, or said another way "item difficulty can be factored into contributions from separate processing operations (Whitely & Schneider, 1981, p. 386)", $\beta^{*}=\Sigma f_{im}\tau_{m}+c$ -- hence, it is part and parcel of construct representation research. Further, by decomposing the item into processing components, the processing demands of the test can be assessed or a new test can be constructed having the required processing demands (Whitely, 1981).

Hence, LLTM research develops the construct representation part of construct validation work. A fundamental assumption of construct representation work is that "the stimulus characteristics of the test items determine the [processing] components that are involved in its solution. (Embretson, 1983)" and "[i]tems vary in stimulus content, which in turn influences processing difficulty... (Embretson, 1992b, p. 129)." Therefore, "[t]he cognitive characteristics of items directly influence construct representation by controlling the item stimulus factors that influence cognitive processing (Embretson & Wetzel, 1987, p. 190)." Note that these statements

implicitly assume a common cognitive structure for all examinees.

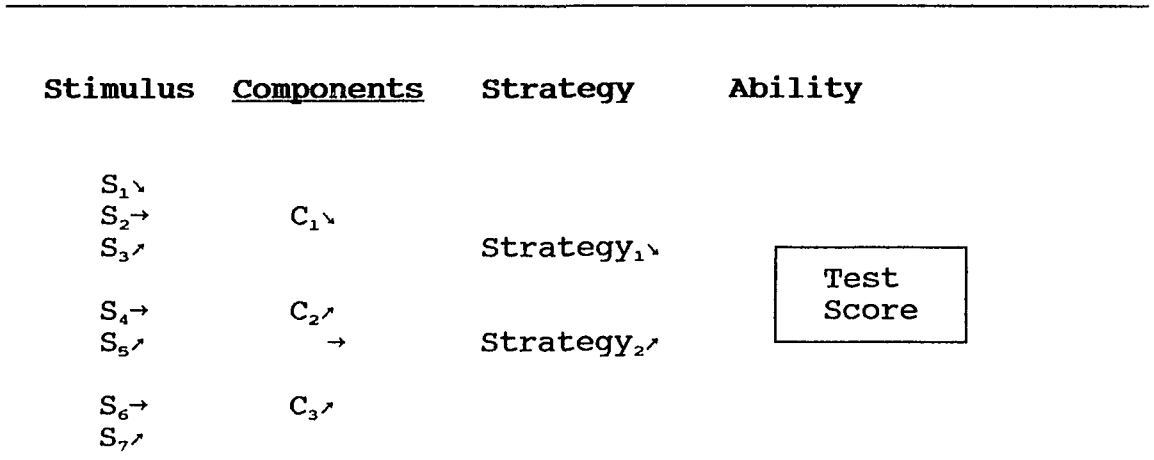
In order to advance construct representation research, four ingredients must be present: (1) a relation between person performance and item characteristics must be specified, i.e. $\ln O=(\Theta-\beta^*)$, (2) it must be possible to test alternative hypotheses, e.g. likelihood ratios, (3) a quantitative estimate for the theoretical construct must be available, for instance, in the form of τ's and (4) the methodology should also provide person scores on the construct, Θ. Pellegrino (1985, p. 49) took this line of reasoning even further:

> "Each [LLTM-like study should involve] delineating the details of a theory of the knowledge and processes involved in solving problems. Such a theory of cognitive components yields a theory of problem types and their mappings to components. These theories permit the design of sets of items and tasks which can be used to validate the theoretical constructs. Systematic, theory-based variations in problem characteristics constitute the basis for hypothesis testing,..."

Construct representation research can be further refined by examining different levels of cognitive variables in a hierarchical structure (see also Glaser & Pellegrino, 1978). Embretson (1985, p. 199) elaborated on the C's in Figure 2.5:

Figure 2.6

Levels of Analysis



Note. From Test Design: Developments in psychology and psychometrics (p. 199) by S. E. Embretson, 1985, Orlando: Academic Press. Copyright 1985 by Academic Press. Reprinted by permission.

For example, spatial ability is an aptitude having more than one possible strategy leading to a correct solution. Even so, at another level, all requisite processing components must be solved correctly in order to provide information needed for problem-solving under a particular strategy. At the "deepest" level, item stimulus features in turn influence the difficulty of processing components (Embretson, 1992b). Early LLTM's generally focused on the S level. Whether other levels can be inferred solely from item stimulus features depends on the quality of the background cognitive theorizing. Embretson's LLTM work (Whitely & Schneider, 1981; Embretson & Wetzel, 1987; Embretson, 1982, 1984, 1985) strived towards 'cognitive variables' by explicitly tying the items' representative stimulus features to information processing model components taken from the cognitive literature. To reiterate, construct representation research, or in this case LLTM, is seeking to explain person behaviors in terms of item characteristics.

In one research example, geometric analogies were explored in Whitely & Schneider's (1981) study. A typical geometric analogy has surface features which can be varied only in certain specified ways. After surveying and criticizing three information processing theorists' work on geometric analogy problem-solving, Whitely and Schneider settled on an extended version of two main sources of

processing complexity from Mulholland, Pellegrino and Glaser's (1980) study: the number of elements in A and the number of transformations required to convert A to B. Whitely and Schneider proposed three alternative hierarchical models. Model I, the most parsimonious, presented the number of elements in A as the only predictor of difficulty. Model II added two types of transformations, spatial displacements and distortions. Model III additionally incorporated other more specific types of transformations, like size and shape. The τ's were estimated for each model. Since the τ's were substantively different across models, the "best" fitting model was sought out by using log likelihood goodness of fit tests. Model III fit best and accounted for 73% of the variance in item difficulty, though only 7% more variance than Model II, but the Rasch model still fit significantly better so more complex information structures (F matrix) probably were needed. Results were discussed within the context of previous findings in cognitive theory and test development.

In summary, different kinds of transformations in going from A to B had differing effects on item difficulty. For instance, increased displacement increased difficulty while increased distortion lead to easier items. Whitely & Schneider suggested that processes implemented in solving displacement items may be qualitatively different than processes used in distortion problems and provided

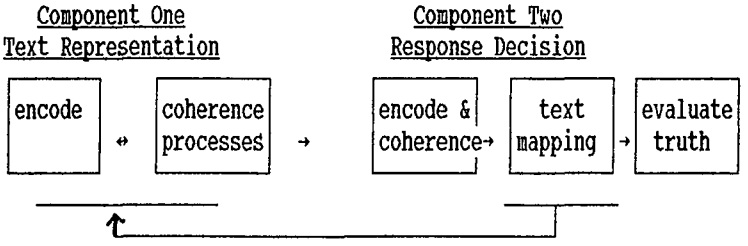
additional support from the cognitive literature. Also, test developers wishing to write an easier test can, given these results, put in more distortion items and fewer items requiring displacement transformations. Further, likelihood ratios were used to check for a possible gender difference in spatial ability, discounted in this study, but often found in research on spatial aptitude. Finally, a test of item homogeneity indicated only one underlying ability. Surprisingly, this disappointed Whitely and Schneider (p. 396), despite the model's unidimensionality assumption.

In another study, Embretson & Wetzel (1987) applied the LLTM to paragraph comprehension items so that estimates of the cognitive demands, in terms of item difficulty, could be made. Since sources of difficulty are often uncontrolled, this study sought to quantify sources of difficulty; items can thereby be screened for processing difficulty prior to piloting or test development. "Thus the impact of the cognitive processing variables on the psychometric properties of the test is quantified. (Embretson, 1992b, p.126)" -- stressing an additional goal of cognitively based test design (Pellegrino, 1985). Another purpose of the study was to compare two tests, ASVAB and CAT, to ascertain that they indeed measured the same construct.

A general processing model was formed from the literature (Embretson & Wetzel, 1987, p. 177):

Figure 2.7

Paragraph Comprehension Model 1



Note. From "Component Latent Trait Models for Paragraph Comprehension Tests" by S. E. Embretson and C. D. Wetzel, 1987, Applied Psychological Measurement, 11, p. 197.

In the text representation component, word meanings must first be encoded and coherent processes used to link words into meaningful propositions leading towards text comprehension, as they also must be in part of the response decision component. For example, Kintsch & van Dijk (1978) assumed that the basic unit of meaning was propositions. So Embretson and Wetzel used variables that took into account kinds of propositions in order to represent the encode and coherence processes sub-components (See Figure 2.6). These variables took advantage of the stimulus complexity features of the items. Alternately, in response decision, alternatives must be compared or mapped to the text and evaluated for match.

A major part of model selection required finding the most parsimonious model that best explained difficulty and that made the most sense cognitively. All models were first compared to the null model, where all item difficulties are constrained to equality, in order to calculate a χ^2, i.e. -2((log likelihood null)-(log likelihood model)). Then a fit index was calculated by the ratio {-2((log likelihood null)-(log likelihood model))}/{-2((log likelihood null) - (log likelihood Rasch))} = $\chi^2_{model}/\chi^2_{Rasch}$, or part to whole. This index is similar to Bentler & Bonnet's index of fit for structural equation models; it is also similar to R^2 in multiple regression, which is likewise a comparison of a null model, the mean of y as the best predictor of y, to a

linear model of 'best' weighted independent variables as the best predictor of y (Embretson, 1993, personal communication). Therefore, the larger the fit index, the closer the model comes to the baseline Rasch model and the better the fit. Finding a "best" model was an iterative process. First, several models for the text representation component were discarded from over eleven original sub-models. From these a "best" text coherence processes model was selected. Then a "best" decision processes model was chosen, again from several alternatives, by looking at the increment in χ^2 as compared to text processes. Finally, the "best" model, incorporating both text representation and decision processes, was selected.

During the course of the analyses, the stimulus complexity features for all variables were scored over all items and τ's estimated. Interpretations for the τ's were proposed and t-tests were calculated for each τ -- paying particular attention to the direction/sign and weight of the coefficients (similar to unstandardized regression coefficients). For example, modifier and connective propositional density and percent content words contributed significantly to predicting item difficulty, but in different ways. The higher the modifier density and percent content words the more difficult an item, suggesting that items with high semantic content were more difficult. However, items with high connective propositional density

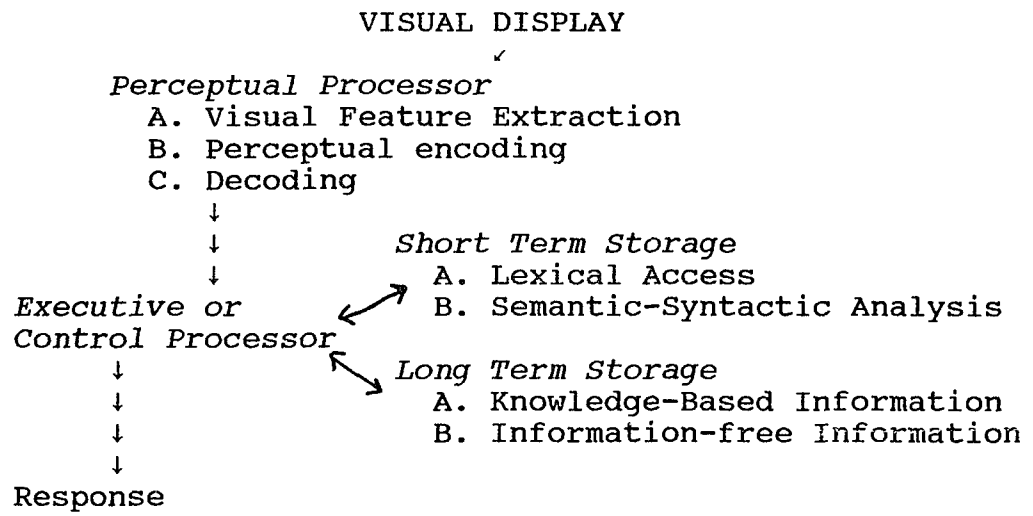
were easier, implying that connectives facilitated understanding and required fewer inferences.

Interestingly, the text representation and response decision components were found to be independent of each other, thereby representing two different abilities. That this violated the unidimensionality assumption was not addressed -- in fact this outcome was applauded as clarifying test development procedures. Instead, perhaps a more complex model than LLTM may be required. Further, in both of the above studies, the Rasch Model was considered a baseline comparison for the substantive hypotheses; however, not once was the fit of the Rasch model discussed.

In a similar though simpler study examining word knowledge and again paragraph comprehension, Mitchell (1983) very explicitly mapped her stimulus feature variables onto a cognitive processing model. Her model of verbal performance follows:

Figure 2.8

Paragraph Comprehension Model 2



Any of the italicized components could be operationalized. For example, total word count provided an index of the entire perceptual processing component. It was hypothesized that as the perceptual load increased, item difficulty would also increase.

Mislevy (1981, 1988; Sheehan & Mislevy 1990) advanced Fischer's work in one direction for his dissertation and further elaborated in the direction initiated by Embretson in his later work. In accordance with Wright and Stone (1979), Mislevy (1981) believed that an explanation of item difficulty in terms of item structures was an important part of defining a measured variable. "An index of an item's difficulty is assumed to express its location along a variable and the presence or absence of salient features is

used to explain its location....it is the effects of item features that determine where an item is located along the continuum. [Further,] The resistance an item offers against ability depends on its constituent parts (ps. 36 and 38)." However, he believed that a more general model of item difficulty was necessary; the assumption that the same variable is being measured across multiple groups must be verified. Group differences could originate through ethnicity, gender, grade level/developmental level/age or educational background knowledge.

Therefore, Mislevy (1981) created a linear model examining the effects of item structures and group structures on item difficulty: $B = K\Theta F + E$, where B is a group by item matrix of estimated Rasch difficulty parameters with each group's parameters separately estimated (Any Rasch software could provide these estimates.), K is a design matrix designating the contrasts between groups, F is analogous to Fischer's F matrix and lays out hypotheses concerning item features, and Θ represents the matrix of estimable parameters, with each parameter having a unique and relevant meaning. As with any general linear model, the meaning of the parameters depends on the coding of the design matrices and therefore also on the substantive questions.

In developing this model, Mislevy needed to resolve certain issues. First, the model was designed after Roy's

general linear multivariate model which is similar to a multivariate ANOVA. As a result, Mislevy demonstrated that the dependent variables, B's, were continuous with certain properties. Second, and as discussed above, CML or joint maximum likelihood (JML) estimation can be problematic. Accordingly, Mislevy demonstrated that using a procedure of generalized least squares to calculate Θ estimates, a noniterative approach not requiring symmetric functions, sufficiently replicated results from CML or JML, with a large enough set of items and sample of examinees. Additionally, with some simplifying assumptions, computations became even more straight forward. The following expression may be used to solve for the unknown parameters:

$$\hat{\theta} = (K'NK)^{-1}K'NBS^{*-1}F'(FS^{*-1}F')^{-1},$$

where N is a diagonal matrix designating the number of examinees in each group and S^* is a common error covariance matrix for item thresholds. It is also possible to determine the variance of $\hat{\theta}$. As with CML, though in a different manner, hypothesis testing activities may be carried out.

Mislevy furnished three applications of his model: (1) a check for item bias (see also Becker, 1990), (2) an evaluation of an educational program, and (3), relevant for this study, a verification of a LLTM componential analysis

across supposedly parallel forms. He noted that in doing a componential analysis, "The idea is that stimuli that share salient features can lead to response processes that are similar. ... [T]he extent to which various components are present in an item may [also] be assumed to determine its chances of being answered correctly compared to other items in the domain.... The ability variable may thus be interpreted in terms of the more basic features of the items,... (ps. 131, 134 & 136)."

The domain Mislevy focused on was fraction problems; his hypotheses concerned operations within the fraction domain. He assumed that one item would be easier than another if both require the same solution algorithm, but one had fewer steps, or, if the algorithm of the easier item was a subroutine of the more difficult problem. For example, adding fractions with a common denominator should be less difficult than adding fractions without a common denominator. Mislevy arrived at a model affording a plausible explanation of item difficulty by coding eight fraction operations. He then endeavored, not with full success, to confirm his model on a 'parallel' form of the test and provided potential explanations for discrepancies (e.g. problems with the choice alternatives).

In a later paper (1988), Mislevy developed a different, less restrictive estimation approach. This approach used LLTM to provide auxiliary information for empirical Bayes

estimation in order to get improved item difficulty estimates. These estimates in turn contributed diagnostic information on model fit and/or unexpected item outcomes (e.g. too easy or difficult due to poor item construction or alternative response strategies). Empirical Bayes procedures obtain more accurate estimates by shrinking estimates toward their 'mean' in inverse proportion to the amount of error. Typically, the 'mean' is a global one, calculated from the set of all item difficulty estimates. Such an approach is justified only when there is no prior information distinguishing between sources of item difficulty; however, the LLTM F matrix furnishes a priori information from cognitive theory. With LLTM, an empirical Bayes estimate would shrink towards the 'mean' of a item subset, which included the item and all items with the same feature pattern. Three steps are required in utilizing this approach: (1) get either CML, JML or Marginal Maximum Likelihood estimates of item difficulty using any traditional program, symbolized β^{MLE}'s, (2) compute item difficulty estimates with the F generated estimated $\underline{\tau}$'s, β^{LLTM}'s, which are the 'means' here and finally (3) calculate the empirical Bayes estimates, β^{EB}'s, by implementing a typical EM algorithmic approach. Shrinkage was defined as $(\beta^{MLE}-\beta^{EB})/(\beta^{MLE}-\beta^{LLTM})$. The poorer the β^{MLE}'s were as estimates, the more shrinkage and the closer the β^{EB}'s come to the β^{LLTM}'s; if β^{MLE}'s were precisely estimated, less shrinkage

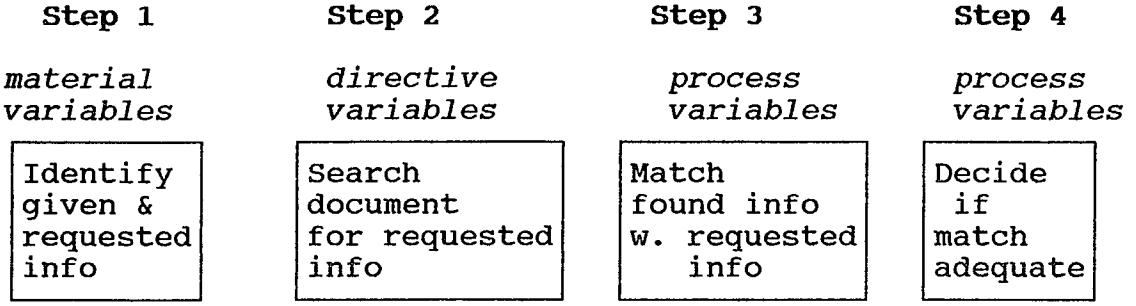
would occur and the closer β^{EB}'s would come to the β^{MLE}'s. Shrinkage was also directly related to error of estimation; shrinkage in conjunction with a large reduction in error implied greater precision.

Continuing his work on fractions, Mislevy (1988) tested out the same models examined in his dissertation, but with an empirical Bayes/LLTM approach. Three models were inspected: a global traditional empirical bayes approach and two LLTM fraction models. Both fraction models, 2 and 3, improved estimate precision over model 1, but model 2 indicated the need for model revision; some items had β^{EB}'s which differed unexpectedly, as indicted by a large standardized difference, from their predicted means, β^{LLTM}'s. An improved Model 2, Model 3, yielded an increase in precision equivalent to doubling the examinee sample size.

Since these results demonstrated the benefits of a combined LLTM within empirical Bayes technique of estimation for model support or revision, Mislevy with Sheehan (Sheehan & Mislevy, 1990) implemented this methodology on an Embretson-like LLTM analysis of document literacy. Sheehan and Mislevy began with Kirsch and Mosenthal's (1988) four step cognitive model:

Figure 2.9

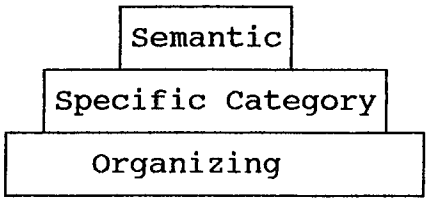
Document Literacy Model



Each step was represented by one of three types of variables, material, directive, or processing, each of which could be at one of three levels of hierarchical expository continuum:

Figure 2.10

Hierarchical Levels



Material variables characterized the length and organizational complexity of the document; directive variables depicted the length and organizational complexity of the document's directives; and process variables portrayed the difficulty of task solution processes. Two

examples of variable operationalization are (1) number of organizing categories in a document for both material and directive parts of text processing and (2) degree of correspondence or matching of text in document with text in directive for process variables. "The[se] elementary components are defined to reflect differences in the cognitive processing demands of the tasks (Sheehan & Mislevy, 1990, p. 262)." Again, item features are expected to call forth theorized kinds of processing from individuals which are more or less laborful/difficult. Sheehan and Mislevy, however, were careful to qualify by stating that it is unreasonable to expect a full explanation of reliable item variation in item difficulty parameters using a LLTM. Results indicated that each variable contributed significant and unique information in understanding item difficulty, though not as successfully with easier items. Like Embretson, Sheehan and Mislevy asserted that discovering manipulable item features that determine item difficulty, based on a cognitive model, supports and illuminates construct validity.

Conclusions

A survey of the LLTM literature has demonstrated its usefulness. The τ estimates can be quite informative in test construction (see also Hornke & Habon, 1986) -- items can be chosen and varied on the basis of component difficulties. Also, Fischer (1981) proved that the

difference between $\beta_i - \beta_l$ is unique and is reflective only of differences in cognitive structure of items i and l. Another advantage of using LLTM, rather than a traditional cognitive componential model (e.g. Sternberg, 1977), is that in using β as a measure of difficulty, right/wrong information is utilized, β and θ are on the same scale so that information exists regarding both individual differences and item difficulty (Embretson et al., 1986) and solution processes are not modelled for a group or globally as is usual in a general linear model, but instead for individuals, with estimation of the τ's or β's independent of the ability parameters (i.e. specific objectivity; Fischer, 1973). A typical R^2 "address[es] only average difficulties (p values) within populations ... and provide no link between [an] individual's overall performance on a set of tasks and their expected success [on tasks] (Sheehan & Mislevy, 1990)." Further, several fit indices exist to substantiate model fit: (1) likelihood ratios (Fischer & Formann, 1982; Whitely & Schneider, 1981) or (2) Embretson's fit index (Embretson & Wetzel, 1987).

Additionally, and importantly, LLTM furthers construct validation research. As an addendum, Carroll (1976) asserted that an item difficulty analysis provided evidence for construct identification. Taking this idea to the next logical step, Stenner, Smith and Burdick (1983, p. 305) stated

> "Just as a person scoring higher than another person on an instrument is assumed to possess more of the construct in question ..., an item (or task) that scores higher in difficulty than another item presumably demands more of the construct. The key question deals with the nature of the "something" that causes some persons to score higher and some items to score higher than other items."

They also listed several advantages to an LLTM approach: (1) falsifiable theories relying on constructs, (2) reliable and generalizable predictors and criterions and (3) ease of experimental manipulation of item characteristics.

However, in implementing a LLTM analysis, several possible problems could exist. It was questionable that underlying model assumptions were reasonable. Multidimensional data or data that is fit best by a two or three parameter model were likely outcomes and could be problematic. For instance, Diones, Bejar and Chaffin (in preparation) found that a two parameter IRT model fit SAT analogy items best. Also, it was reasonable to assume that instead of one ability, multiple abilities such as vocabulary and background knowledge could play a role in item variation. If any of these events were to occur, alternative analyses could be considered. Besides problems deriving from strict assumptions, a disadvantage of LLTM is the loss of a direct measurement of componential processing information, as is usual in reaction time experimental studies.

[1] There is also a multiplicative form, developed by Rasch (1960) and often used by European psychometricians. Most of the IRT literature deals with and assumes the logistic form.

[2] Multidimensional IRT models and software have recently become more common.

[3] Hence, since an odds ratio is defined as $O = P/(1-P)$, and P can also be expressed as $(1 + e^{-(\theta-\beta)})^{-1}$, and thus $1-P = (1 + e^{(\theta-\beta)})^{-1}$, then $O = e^{(\theta-\beta)}$ and $\ln O = (\theta-\beta)$.

[4] Rasch assumptions are severe and it is questionable if the assumptions are reasonable. It is unlikely that all test items discriminate between examinees equally or that guessing does not occur on multiple choice items.

[5] Measuring difficulty can present problems. Many other variables are intertwined with difficulty; therefore, it is important to control for confounding effects. For instance, is an item difficult because of cognitive item features or because of something about the way this kind of item is written? Or is it difficult due to ordering effects?

[6] However, the situation can be viewed in an exactly opposite manner, as Scheuneman and Steinhaus (1987) have pointed out: the item "difficulty parameter is still defined in reference to examinee ability levels." As mentioned in the Introduction, measuring persons cannot occur without items and vice versa. This paper will focus on Wright and Stone's perspective.

Chapter 3: Characteristics of Analogical Reasoning: A Literature Review

Analogical reasoning has long been researched (Dawis & Siojo, 1972). The purpose of the current review, however, was to define analogical reasoning, to ascertain which cognitive variables are significant in analogy problem-solving, to provide a global information processing model appropriate for analogy problem-solving and lastly to indicate the link between the information processing model and the chosen operational variables in order to justify a final LLTM. In addition, since a goal of this study was to integrate psychometrics and cognitive theory, and the final information processing model depended on the item format, the kind of items used had to guide and direct the cognitive literature review and model selection. Therefore, the kind of item used is described first. Next, the definition of analogical reasoning is elucidated. Then the literature is surveyed with the aims of (1) uncovering the important cognitive variables, (2) defining a final information processing model for the LLTM analysis and (3) linking these cognitive variables to the information processing model.

SAT Analogy Items

The Scholastic Aptitude Test[1] (SAT) often functions as a major component in college admissions decision-making procedures, offering some unique supplemental information over and above other typical measures of achievement, such

as high school grades. It is produced by the College Board and is designed to be a curriculum-free measure of developed verbal and mathematical abilities; abilities, however, are not considered innate characteristics. "Developed ability *is* an achievement, and aptitude test scores can and do change (Donlon, 1984, p. 38)." The purpose of the SAT's is to help predict first year college grades by taking a measure of an examinee's current aptitudes.

Each test is organized according to standardized specifications. There are six, thirty minute sections, two of which are verbal; the entire testing period is three hours. Verbal items, 85 in all, are comprised of analogies (20), antonyms (25), sentence completion (15) and reading comprehension (25). Within each sub-section, for example analogies, items are ordered by difficulty, easy to hard. In addition, efforts are made to maintain the same level and range of difficulty across sections.

For SAT analogy items, standard instructions are given along with an example in order to "suggest ... the mental processes involved ... and to stress that this process is fundamentally one of judgment -- of finding the *best* answer, rather than the only answer (Donlon, 1984, p. 41)":

> Each question below consists of a related pair of words or phrases, followed by five lettered pairs of words or phrases. Select the lettered pair that <u>best</u> expresses a relationship similar to that expressed in the original pair.

Example:

YAWN:BOREDOM:: (A) dream:sleep (B) anger:madness (C) smile:amusement (D) face:expression (E) impatience:rebellion

Over all items, the form remains constant: A:B::?:?; this format should be read, A is to B as what is to what. Here, the sought after relation is, an A is a physical expression of B. Further, there exists a typical terminology when discussing analogies. In this example, YAWN:BOREDOM is called the **stem** pair, (A) - (E) are named the **alternatives** with (C), the pair that expressed a relationship most similar to the stem, considered the **key** pair.

Although analogies all take the same form, item writers are given three taxonomies as guidelines for their item construction efforts: content (aesthetics/philosophy, world of practical affairs, science, and human relations), abstraction of terms (concrete, mixed, and abstract), and independence of stem and options (independent and overlapping); these "ideal" classification dimensions are <u>not</u> necessarily intrinsically meaningful (Donlon, 1984). Yet, there has been some empirical support for increased processing difficulty from concrete to abstract and independent versus overlapping analogies (Pellegrino & Glaser, 1979; Glaser & Pellegrino, 1979, e.g. semantic constraint); others have not found these taxonomic distinctions useful (Bejar et al., 1991). Examples of concrete, mixed and abstract analogies are house:roof,

sheriff:justice and electorate:democracy, respectively. In an independent analogy, the stem relation does not suggest the key relation (e.g. fire:ashes::event:memories) while an overlapping analogy's stem does (e.g. famine:food::drought:water). Not surprisingly, actual classification of items can be difficult.

Defining Analogical Reasoning

While the construct for the overall SAT is labelled "scholastic aptitude", analogies are considered verbal reasoning or more globally, inductive reasoning. But what is meant by 'reasoning', 'induction', and 'analogical reasoning'? Sternberg (1986) provided a general theory of reasoning. "A task is a reasoning task if and only if its solution involves the mediated and controlled application of inferential rules for the purposes of selective encoding, selective comparison, or selective combination (Sternberg, 1986, p. 293)." Inferential rules allow one or more of the three selective processes to come into operation when problem-solving. Reasoning activities, however, are further constrained in that the task must be effortful, or not automatized; it is therefore a considered task. Mediators, such as prior knowledge, may increase or decrease the examinee's problem-solving capacity. Whether a task is automatized or mediated is a person characteristic, built into ability, raising or decreasing the probability of a correct response. Sternberg (1986, p.293) then went on to

differentiate inductive from deductive reasoning. "[T]he difficulty of inductive problems mainly derives from selective encoding and selective comparison processes, both of which involve sorting of relevant from irrelevant information." An example applied to the case of analogy problem-solving will be presented below (please see * on page 72).

Coming from a different research paradigm, yet showing some convergence, Holyoak and Nisbett (1988) sought to elucidate induction. Induction, like reasoning, is a broad notion incorporating many facets. It involves categorization and concept formation, reasoning with rules in order to draw inferences in uncertain learning and problem-solving situations, utilizing and accessing existent, stored knowledge, and all with a goal towards further knowledge development.

Similar to Holyoak and Nisbett, several researchers approached analogical reasoning from the purview of creativity, learning and real life problem-solving. Fundamentally, "analogies involve reasoning about relations, in particular about relational similarity, so that a correspondence is established between one set of relations and another (Goswami, 1991, p. 1)." Or, "Reasoning by analogy has been thought to be a special kind of relational reasoning as it requires the ability to reason about *second-order* or higher-order relations. Analogies are used to map

information from a known to a new situation when the surface features of the two situations are dissimilar (Goswami, 1989, p. 251)." So a core definitional process of analogical reasoning is mapping or "the construction of orderly correspondences between the elements of a source analogy and those of a target (Holyoak & Thagard, 1989a, p. 295; Clement & Gentner, 1991)." As a result, analogies facilitate transfer of knowledge for learning and predicting (Holyoak, 1984; Clement & Gentner, 1991).

To summarize, testers use analogy items to measure individual differences on the construct of analogical reasoning. Thus, as noted earlier, verbal analogies are particularly attractive item types because of the complexity and importance of the skills drawn upon, their semantic richness, their uniform structure, and their historical connection to intelligence work and heavy loading on the **g** factor.

Therefore, as a first step in a LLTM construct validation research, the next section now summarizes and explicates the ideas and results emerging from appropriate cognitive research on analogy problem-solving, highlights important variables and draws up a final processing model linked to these variables. In Chapter 4, discovered relevant analogy item features, which relate to these variables and the final model, are presented, elaborated and justified, thus tying item format with prototypical

cognitive processes. The logically implied research hypotheses are also specified.

Approaches to Analogical Reasoning

Analogical reasoning has predominately been considered within two different research paradigms. The first empirical domain has been an information processing approach (IPA). As it is most closely tied to psychometric analogy items, this dissertation remains primarily within the IPA paradigm. Hence, investigators took this new methodology, which attempted to understand memory characteristics and processing components of problem-solving, and conjoined the methodology with previous intelligence paradigms, to produce a model(s) of analogical reasoning (Sternberg, 1977). The purpose of this line of research was to determine which information processing components are used by prototypical individuals solving analogy items.

The second approach, a problem approach (PA), has viewed analogical reasoning as a vital piece of another larger concern, transfer in learning and creativity, and additionally has incorporated computer simulation. This domain sought to understand cognitive processes leading to transfer in learning; that is, transfer that occurs as a result of analogical reasoning. For the purposes of this study, only directly relevant aspects of the PA, related to analogy problem-solving and the evolving IPA model, were drawn upon. Lastly, several pieces of general cognitive

research, such as memory and concept formation work, were selectively reviewed to fill out the IPA model.

Information Processing Approach

IPA is now a well-known research paradigm, found in most cognitive psychology texts (e.g. Anderson, 1985), which can be applied to any content domain. A person's mind and thought processes are considered purely in terms of information processing capacities. Often, the path a task takes is serially considered. First, a problem task must be presented and perceived. Thus, a perceptual screen of huge capacity takes in or imprints for fractional seconds all stimuli from the outside world. However, a mechanism is in place that selectively focuses on information deemed important; this mechanism imposes constraints on the system -- confusion would reign if everything received equal attention weights; a consequence, though, is that information may be irretrievably lost if poor selection processes are in operation. Information finally concentrated on is said to be encoded and placed in short term memory (STM). STM has a limited capacity, and for information to be processed in STM, it must be attended to. Researchers now have expanded STM to include a working memory (WM)[2] component. There problem-solving activities can be carried out. STM and WM both have internal resources in the form of long term memory (LTM). All of a person's knowledge resides in LTM; knowledge may be procedural (e.g.

A list of the rules or steps involved in solving analogy items.), declarative facts (e.g. Analogy items appear on the SAT's.), episodic or experiential details (e.g. The day I took the SAT's it was pouring rain.), or abstract semantic/lexical information including concepts, definitions, categories and so on. LTM is generally considered to have a limitless capacity and a structure or organization -- though researchers discuss these issues at many levels (e.g. schemas, semantic networks and associative neural connections.) Difficulties may arise in accessing and retrieving information from LTM. Therefore, given an IPA paradigm, it is then each researcher's task to develop appropriate processing models for particular content domains. Note, however, that the item type utilized in analogy IPA research studies was primarily of the form A:B::C:?; a final IPA model will need to take into account the SAT format.

In a first phase of analogy research, Spearman (1923; cited in Bejar, Chaffin & Embretson, 1991; Sternberg, 1977; Barnes, 1980; Dawis & Siojo, 1972) initiated an IP type of analysis of analogy problem-solving. He listed three global solution processes, still considered fundamental: (1) encoding A:B or the apprehension of experience, (2) inferring a relation between A:B or the eduction of relationships, and (3) applying the A:B relation to C:D_i or the eduction of correlates. Spearman's notions can be

visually represented in an information processing flow chart taken from Bejar et al. (1991; See Figure 3.1).

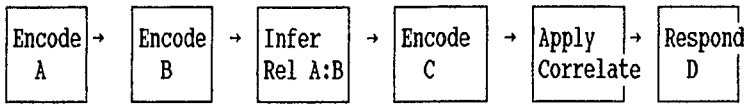
Researchers following Spearman and preceding Sternberg's (1977a, 1977b) seminal work generally supported the above processes (Barnes, 1980). Rumelhart and Abrahamson (1973) provided another unique view, based on a theory of memory and a definition of analogical reasoning. If memory is considered a multidimensional Euclidean space, then degree or magnitude of psychological similarity judgments of the A:B concepts and type of relation will depend on A:B's semantic distance and orientation in space, respectively. Semantic distance has been regarded as an important variable in research explaining item difficulty (Ingram & Pellegrino, 1977; Glaser & Pellegrino, 1979; Embretson & Curtright, 1982; Bejar et al., 1991).

In sum, Barnes (1980) portrayed models of this research period as holding three strong processing assumptions: (1) a person processes all components once, in their proper serial order, (2) an analogy is correctly and uncomplicatedly processed if word meanings are known, and (3) all examinees have procedural knowledge of analogy problem-solving. Therefore, this class of models is conceptually driven by top-down processing.

Sternberg's furtherance of Spearman's model and his integration of it with previous intelligence models, as well as his setting this work within an IPA framework, initiated

Figure 3.1

Spearman's Model



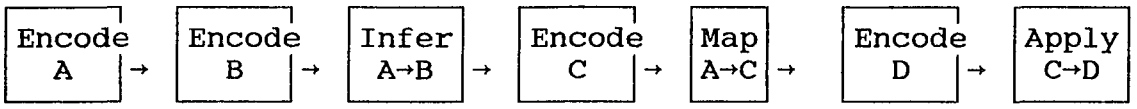
Note. From Cognitive and Psychometric Analysis of Analogical Problem Solving (p. 31) by I. I. Bejar, R. Chaffin, and S. E. Embretson, 1991, New York: Springer-Verlag. Copyright 1991 by Springer-Verlag. Reprinted by permission.

a next major phase of analogical reasoning research. As part of his development of a componential model, Sternberg[3] (1977a & 1977b) was interested in identifying component mental operations, clarifying the organization of the components and discovering how the components related to each other and to higher order abilities. In fact, he addressed analogy research primarily because it comprised one step along a path towards a new theory of intelligence. Interestingly, Sternberg considered this type of an approach helpful in construct validation efforts: componential analysis formed an internal validation while examining how components lead to individual differences constituted an external validation[4]. Further, a goal of an intensive analysis (i.e. componential) was to develop a psychological understanding of the task, while the goal of an extensive analysis was an integration of results into a meaningful, generalizable theory.

Using an A:B::C:?, multiple choice item format, Sternberg added three components to Spearman's model and ended up with a baseline model (Model I):

Figure 3.2

Sternberg's Model I



Here, when encoding analogy terms, each word was identified, and an as-complete-as possible list of attributes was retrieved from LTM, limited only by STM capacities; these attributes were concept features, deemed by the examinee as possibly useful for an analogy problem situation. Then, the examinee inferred relations between A and B exhaustively, that is, generated all possible relations between A's and B's attributes. Possible relations were stored. After C was encoded, possible relations or correspondences between A and C were formulated, or mapped, and also stored. Mapping was additionally exhaustive. Then for application, all retained attributes were applied to C in order to generate an ideal C:D', analogous to a stored A:B relation, while taking into account the mapping information and evaluating the possible D_j's for a match. Or as Barnes (1980, p. 19) phrased it "[a]pplication [was] a process that generate[d] an "image" of the correct answer and evaluate[d] the possible alternatives." Applying was implemented exhaustively. Added to this model was an orient/justify aspect. Orienting required attending to the problem, resolving to solve it and drawing up one's procedural rules for analogy problems. Justification occurred only when none of the alternatives allowed for a neat match with the stem rationale. When the justify component activated, the examinee checked for errors, attempted to recall information that could result in an unique solution, or chose the best,

though not perfect option. So in summary, Model I assumed sequential and exhaustive processing, with no room for backtracking. Note that the dependent variable here was item response latency, not correctness of response.

Sternberg's (1977a & 1977b) Models II, III and IV successively relaxed more and more of these strict assumptions. In Model II, the apply component was deemed self-terminating. That is, each attribute was individually considered for a unique solution. If one was found, processing ended; no further attributes were checked. For Model III, processing for the map component was additionally judged self-terminating. Model IV further regarded processing for the infer component self-terminating. Comparisons of R^2's derived from multiple regression analyses indicated that Models I and II provided poor explanation of response latency variance but that Model III accounted for 86% of latency variance while Model IV did not explain significantly more variance over and above Model III; it was difficult to discriminate between Models III and IV. Model III was selected as the most sensible model and is portrayed in Figure 3.3, where i is an attribute subscript and j an option subscript. Even though a self-terminating processing strategy was evidenced for two components, there was a trade off -- the probability of making an error increased.

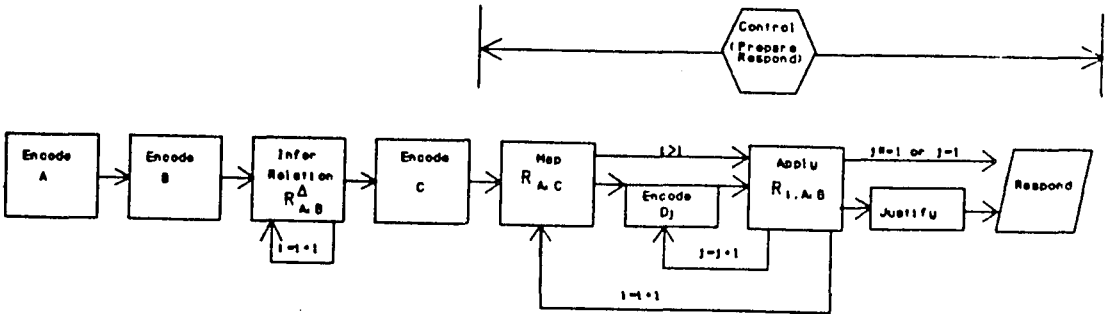
Sternberg asked other questions of his data. He

verified that Model III did not fit as well when the map component was omitted, but the effect of omitting the apply component was not clear. Figure 3.3 also indicates, as Sternberg discovered, that alternating option scanning occurred rather than sequential option scanning. That is, all options were processed while holding one attribute at a time constant. Further, he noted that over 50% of an examinee's time/effort went to encoding, while only 30% went to attribute comparisons. Barnes (1980) subsequently observed that Sternberg's Model III initiated with top-down processing, but as it permitted backtracking, bottom-up processing could predominate as characteristics of the particular problem, such as the alternatives, dictated.

Goldman and Pellegrino's (1984) conceptual organization of Sternberg's model is potentially useful for an IPA using A:B::?:? analogies. Since models for A:B::?:? item formats have not been as well developed or specified as A:B::C:? models, a system of global processing classes are useful for describing a larger, coarser componential model. Thus, they parsed the various processing activities into three global processing classes: attribute discovery, attribute comparison and evaluation. The first, attribute discovery, described encoding and internally representing semantic attributes in WM. Attribution comparison incorporated several of the processing components: inferring, mapping (the particular inference of attribute relations between A

Figure 3.3

Sternberg's Model III



i = attributive index
j = alternative index
j* = retained alternative index

Note. From Cognitive and Psychometric Analysis of Analogical Problem Solving (p. 32) by I. I. Bejar, R. Chaffin, and S. E. Embretson, 1991, New York: Springer-Verlag. Copyright 1991 by Springer-Verlag. Reprinted by permission.

and C) and one part of applying, the application of the A:B inference to C, to select an ideal D'. For simple analogies, evaluation gathered the second part of the application component, the evaluation of the D_j alternatives to confirm D', and response generation. For difficult, ambiguous analogies, evaluation also included the justification component.

*With the above as background, Sternberg's (1986) formulation of analogy problem-solving as inductive reasoning can now be shown. To properly solve an analogy problem primarily requires selective comparison processes: attributes are 'selected' from LTM by comparing and prioritizing them in terms of importance, thus requiring judgment and decision-making. Also, the solver is wielding selective encoding as she strives to infer relations; she must select from the attributes now stored in WM to infer a relation between words in an analogy term. In addition, both procedural and declarative inferential rules come into play. For example, a 'how to' map for analogy problem-solving is accessed and implemented which also incorporates strategy choices. Declarative rules take into account the nature of syntactic and semantic relations as well as word meaning. Sternberg stated that for inductive reasoning tasks, declarative rules were a large source of item difficulty. Mediators, however, can facilitate problem-solving.

Notwithstanding, Sternberg (1981) became concerned with possible theoretical deficiencies, which were reflected by a persistent high correlation of the regression constant with reasoning measures. "...[T]here may be one or more critical components of intelligent performance that are not now being extracted by componential procedures ... (Sternberg, 1981, p. 4)". He tackled this problem by using "nonentrenched" tasks, or unfamiliar reasoning tasks, that may better elicit reasoning behaviors, to assess the role of two possible metacognitive strategies, global and local planning. He did not suggest alternative processing components

After Sternberg's enormous effort, a number of researchers responded with replications, by asking specific questions designed to extend his results. Some studies asked if the mapping and inferring components were both necessary (Grundin, 1980; Sheard & Readence, 1988). For example, mapping seemed to come into play only when analogies were difficult or ill-structured; mapping was not an automatic processing component.

Other model modifications, however, originated with some of Pellegrino and Glaser's (1979; Glaser & Pellegrino, 1979) research. Like Sternberg, their work was part of a larger research goal: the analysis of inductive tasks using both "cognitive correlates" and "cognitive components" approaches, with an ultimate goal of instructional development. A cognitive correlates approach asked about

differences between high and low ability students' cognitive processing in problem-solving; Sternberg's model was an example of a cognitive components approach.

Pellegrino and Glaser not only emphasized a need for research concerning processing characteristics, but also research which considered content properties. Up to that point, content issues, like kind of semantic relation, memory organization, and so on, had been largely neglected (Exceptions were Ingram & Pellegrino, 1977 and Whitely, 1977.). This study will also pursue content issues through the avenues of a semantic taxonomy and word frequencies. In addition, Pellegrino & Glaser (1980) differentiated between models that explained latencies versus those that explained errors for A:B::C:?. For instance, models which sought to explain errors in terms of content issues (Pellegrino & Glaser, 1981) indicated that (1) while stem relational features constrained possible alternative solutions, there was tremendous variability in degree of constraint within any one semantic class and (2) alternatives varying in semantic appropriateness also affected processing difficulty in differing degrees.

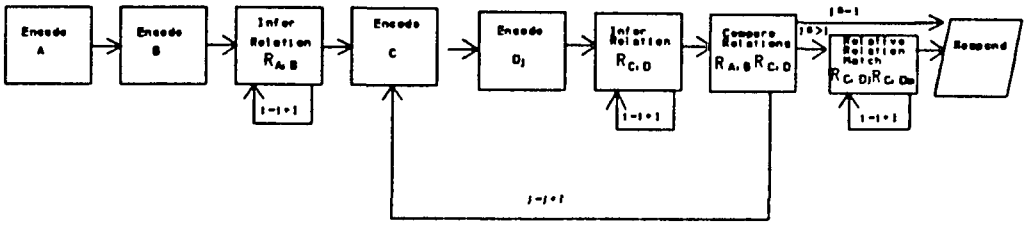
Within this context, Pellegrino and Glaser proposed a two stage processing model (1979, 1980; Glaser & Pellegrino, 1979). Only the first stage was needed when solving easy items; both stages were required for difficult items. They postulated that initially, an examinee gave each item a

global run through. Alternatives that were obviously incorrect could be discarded and if the solution was immediately salient, a response was given and processing terminated. If complexities emerged, the examinee slowed down and reconsidered the entire item, from a more complete consideration of the stem's rationale to a reevaluation and comparison of the alternatives. Stage one required only top-down serial processing; if stage two shifted in, data-driven processes like backtracking or working backwards from the alternatives became likely, requiring additional bottom-up processing -- that is, A:B, C:D_j's, and D_i:D_j's were reevaluated. Further, even though initial A:B inferring may have failed, recognition of A:B's relation was possible during the consideration of the C:D_j's (Goldman & Pellegrino, 1984; Alderton, Goldman & Pellegrino, 1985). In addition, if two D_j's were possible answers, they would be compared and one discarded before processing continued. Also, support was found for the importance of an application component. In sum, they proposed an encode-encode-infer A:B, encode-encode-infer C:D_j, compare A:B to C:D_j and then respond, model, as outlined for difficult items in Figure 3.4. This two phase model allowed for a substantially more complex portrait of analogy problem-solving and importantly, was more likely to generalize to an A:B::?? item format.

Pellegrino and Glaser (1979, 1980; Glaser & Pellegrino, 1979) also mentioned several factors that contributed to

Figure 3.4

Pellegrino and Glaser's Interactive Model



i = attribute index
j = alternative index
j* = retained alternative index

Note. From Cognitive and Psychometric Analysis of Analogical Problem Solving (p. 40) by I. I. Bejar, R. Chaffin, and S. E. Embretson, 1991, New York: Springer-Verlag. Copyright 1991 by Springer-Verlag. Reprinted by permission.

processing difficulty:

> "(a) processes of establishing a reasonably well-defined problem representation, (b) the subsequent utilization of that representation as a basis for selecting among alternatives, and (c) modifying the representation as necessary (Pellegrino & Glaser, 1979, p. 209)."

They additionally stressed the need for a "process theory of item difficulty" (Glaser & Pellegrino, 1979, p. 2) as part of an explanation of individual differences in task performance. Reported aspects that may contribute to item difficulty (Goldman & Pellegrino, 1984) were (1) encoding of uncommon words, (2) inferring, mapping and applying relations having multiple attributes or requiring multiple transformations of attribute features and (3) items eliciting justification. A natural implication was that more difficult items lead to a greater processing effort (Goldman & Pellegrino, 1984) -- again, the study of item difficulty is imperative.

Embretson and Curtright (1982) also found support for a two stage processing model, especially when the item's response format had five choice alternatives. Further, their results suggested exhaustive processing for the stem when the problem's requirements were perceived as difficult (e.g. five alternatives). In addition, many choice alternatives "contextualized" the stem and hence necessitated a reevaluation of the stem rationale; again, in this situation, bottom-up processing would be required. Further, put another way (Liu, 1981), alternatives can

constrain the number of attended to features in A:B by allowing the solver to discriminate between relevant and irrelevant features. These alternative choice context effects are important and will be examined as part of this study.

In a continuation of her latent trait IRT models (cf. Chapter 2), Embretson (1984, 1985) implemented a Multicomponent Latent Trait Model and a General Multicomponent Latent Trait Model on a simplified analogy IPA model. Without going into IRT model specification detail, her Multicomponent Latent Trait IPA model required simplification as she had to collect multiple subject responses: a response for each rule-oriented or reasoning component subtask, a (italicized) non-reasoning component response and a total item (bolded) response, rather than latencies (Table 3.1; Embretson, 1985, p. 204). Similar to Pellegrino and Glaser (1980), she sought probabilistically to model solution accuracy. Unlike others, she endeavored to acknowledge and include nonanalogical reasoning strategies in her processing model, that is, making associations and/or guessing. A third strategy, extrapolated from Pellegrino and Glaser's (1980) two stage model (i.e. recognition of A:B's rationale after processing option alternatives) was a partial rule strategy -- partial information increased the probability of a correct solution (Embretson, Schneider & Roth, 1986).

Table 3.1

Embretson's Analogy Components

Total Item	**Cat:Tiger::Dog:_____** **1) Lion 2) Wolf 3) Bark 4) Puppy 5) Horse**
Rule Construction Component	Cat:Tiger::Dog:_____ Rule:_____
Response Evaluation Component	Cat:Tiger::Dog:_____ 1) Lion 2) Wolf 3) Bark 4) Puppy 5) Horse Rule: A large or wild canine
Association Subtask	*Dog* *1) Lion 2) Wolf 3) Bark 4) Puppy 5) Horse*

Note. From Test Design: Developments in psychology and psychometrics (p. 199) by S. E. Embretson, 1985, Orlando: Academic Press. Copyright 1985 by Academic Press. Reprinted by permission.

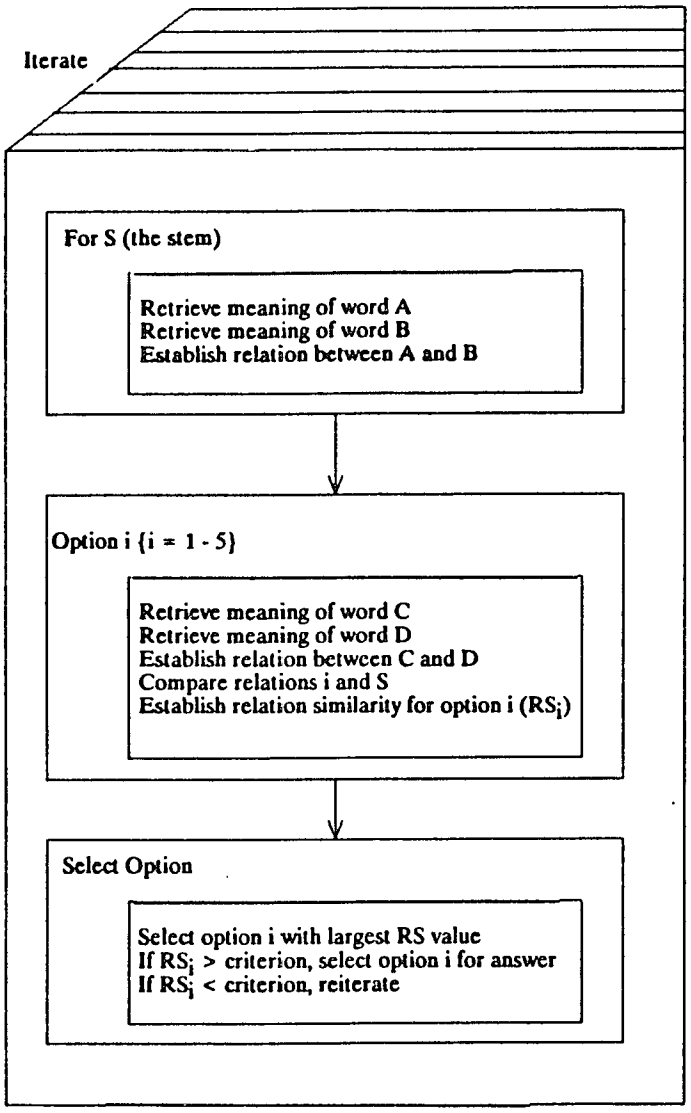
Moreover, metacognitive processes, it was postulated, determined strategy selection.

Furthermore, her General Multicomponent Latent Trait Model operationalization of analogy problem-solving integrated LLTM with the above Multicomponent Latent Trait Model. Utilizing the same components as in Table 3.1, item features within each component were proposed to explain component difficulty. For example, relational span (the rated semantic distance between analogy terms; Rumelhart & Abrahamson, 1973), inference saliency and syntactic complexity were features used to model the rule construction component (See also Embretson & Curtright, 1980).

As noted, all of these IPA models utilized analogies of the form A:B::C:?. However, in a study of GRE analogies, which take the same A:B::?:? form as SAT analogies, Bejar et al. (1991), provided a normative IPA model (Please see Figure 3.5.), transforming models from past research. Here, an examinee reads and encodes, that is, forms an internal representation of the meaning of A and B, and then infers a relation between the two concepts. Then she repeats this process for each pair alternative and compares the alternative relation to the stem relation. At this time, decision-processes are in effect; an alternative may be discarded as an improbable match, selected, or put on hold as other alternatives are deliberated. A C_i:D_i alternative is chosen once the mapping/comparing processes deem that it

Figure 3.5

Bejar et al.'s Normative Model



Note. From Cognitive and Psychometric Analysis of Analogical Problem Solving (p. 10) by I. I. Bejar, R. Chaffin, and S. E. Embretson, 1991, New York: Springer-Verlag. Copyright 1991 by Springer-Verlag. Reprinted by permission.

has surpassed a subjective threshold of relation similarity with A:B. At any time, the solver may reconsider the stem rationale or any of the alternatives or resort to nonanalogical strategies such as associational reasoning or guessing.

This model was an educated, make sense, best "guess" -- it took therefore a global, rather coarse form. Nonetheless, much of the literature has lent credence to its conceptualization. It makes sense to conceive of the A:B encoding and inferring as one whole and the encoding and inferring of each C_i:D_i iteration as another whole. Then the mapping of or applying or judging of similarity can be done using these two wholes. The model must necessarily remain global as details are not yet spelled out.

However, any processing model may break down at several points. The examinee may be unfamiliar with analogy problem-solving procedures. Or A, B, C_i and/or D_i may be comprised of unknown words. Or, finding a reasonable relation between A and B or C_i and D_i may be exasperating. In addition, several or none of the alternatives may come close to matching the stem relation. And so on.

In a content or semantic vein, historically analogies have regularly been categorized into dichotomies (Warren, 1921) -- for example by how internal/external, novel/common (Pellegrino & Glaser, 1979) or unfamiliar/familiar they were or degree of necessitated fluid/crystal intelligence. Bejar

et al. (1991) additionally noted that empirically GRE analogies fell into two clusters with a meaningful regularity. They therefore sought to formulate an historically integrative dichotomy, an intensional and pragmatic distinction, another important variable for this research endeavor.

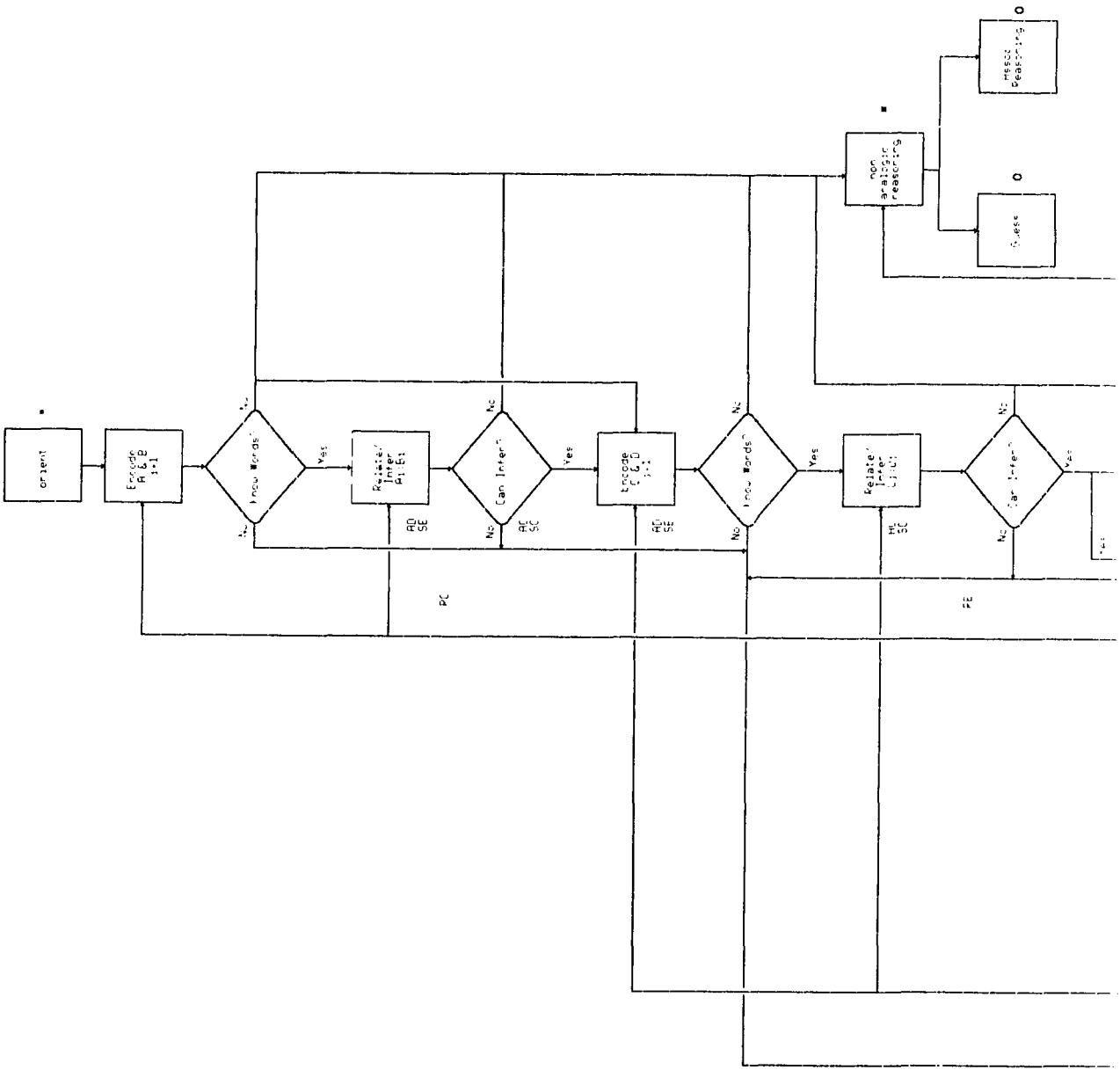
Summary

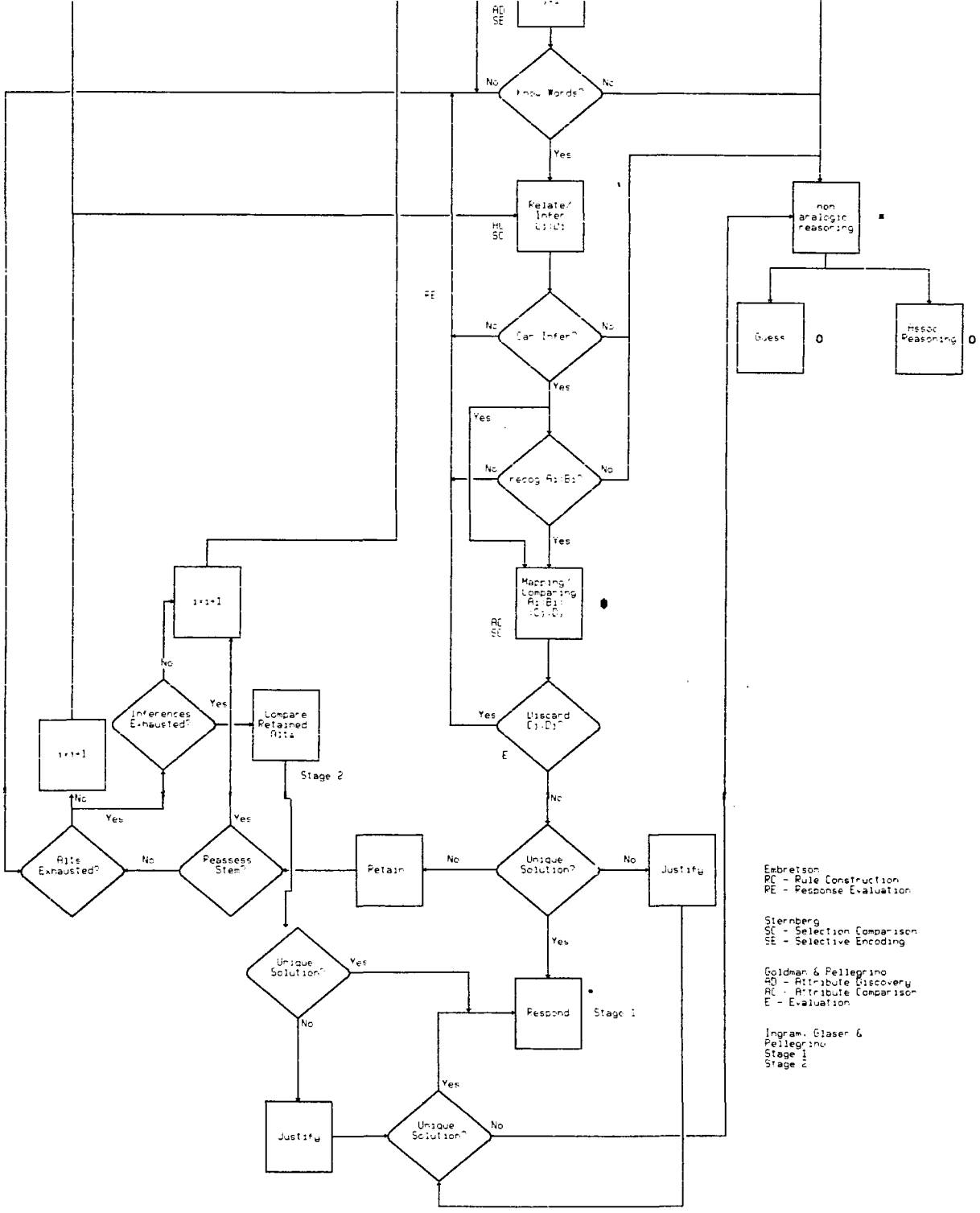
At this point, a specific model can be formulated integrating Bejar et al.'s (1991) portrayal with the other above-mentioned results (see Figure 3.6) -- a detailed description of the analogy problem-solving model will be presented later (p. 103). In reading Figure 3.6, note that squares depict actions or behaviors, triangles represent the memory systems and diamonds portray decision-making junctures. Problem-solving begins with the orient square, marked with an *, and ends either with the respond square or the non-analogical reasoning squares, all marked with an *.

Over time, analogy problem-solving models have changed in their details, complexity and completeness, the flexibility of processing (i.e. processing could be considered top-down (Stage 1), or bottom-up (Stage 2) or both), strategy usage and more (Barnes, 1980). Note that the model in Figure 3.6 **does** take into account points where problem-solving can break down (e.g. Gitomer et al., 1987), replaces the previous mapping and applying components with a new mapping/comparing/judging similarity component, ◆,

Figure 3.6

IPA Model





Embretson
RC - Rule Construction
RE - Response Evaluation

Sternberg
SC - Selection Comparison
SE - Selective Encoding

Goldman & Pellegrino
AO - Attribute Discovery
AC - Attribute Comparison
E - Evaluation

Ingram, Glaser &
Pellegrino
Stage 1
Stage 2

(Pellegrino & Glaser, 1982) and allows for reexamination of the stem rationale and alternatives (Stage 2; Goldman & Pellegrino, 1984; Gitomer et al., 1987) as well as a variation in responding strategies (O, Embretson et al., 1986). Also, pieces of Goldman and Pellegrino's conceptualization have been incorporated (attribute discovery, attribute comparison and evaluation). As this model evolves to its final form (see Figure 3.7 on p. 103), the cognitively relevant variables will be tied to the components, indicating how item features indeed relate to the IPA model and construct validity.

Initially, since analogical reasoning work implementing an IPA paradigm emerged from an intelligence factor analytic, individual difference outlook, throughout most of the early literature, analogical reasoning simply was what people did when they solved analogy problems. Sternberg's (1986) work and Goldman and Pellegrino (1984) supplied the first formal attempts to define analogical reasoning. In addition, coming from differential psychology, it became important to identify information processing components that explained variance in test scores (Barnes, 1980; Whitely & Schneider, 1981); Sternberg, Pellegrino et al., Embretson et al. and Gitomer et al. all focused on sources of individual differences in processing. However, Carpenter, Just and Shell (1990) offered a word of caution. There are important commonalities in intelligent behavior that do not emerge as

individual differences; these aspects of processing should not be neglected. So for instance, Carpenter et al. found that people have a common ability to decompose problems into manageable units and to iterate through these units for Progressive Raven Matrices problems. Differential abilities emerged in the ability to organize these units into subgoals and to manage the subgoals appropriately.

Further, in taking into account the entire IPA literature, several variables consistently emerged as being important (Ingram & Pellegrino, 1977; Embretson & Curtright, 1980; Bejar et al., 1991). Semantic or relational distance surfaced in some reports. Degree of semantic constraint imposed by the stem was another concern. Variables which will be further examined here are (1) two stem variables, rationale difficulty as well as syntactic order, (2) variables reflecting constraints or effects of option alternatives, like alternative similarity and contextualization of the stem and (3) three important content related variables, prototypical semantic relations, vocabulary and the intensional versus pragmatic distinction. All of these variables will be further discussed in Chapter 4.

The Problem Approach or Analogical Reasoning as Transfer in Learning and Problem-Solving Creativity

An entire other literature, also researching analogical reasoning, has evolved, the PA. Again, research in this

realm is closely tied to a problem format; there are two main prototypical problems. One (Holyoak and colleagues) is the fortress-ray problem: Subjects were first told a story about a general trying to overcome a fortress. A frontal attack, it was known, would doom the general to failure. Instead, successfully, the general split up her forces and attacked from several sides at once. Then, without cuing a connection, subjects were given a medical scenario. A patient had a tumor which must be destroyed with radiation therapy, but full strength radiation intensity would kill the patient. How could the tumor be destroyed with radiation? Proper solution necessitated noting the analogy between the two problem situations despite a lack of surface similarities. Another problem type (Gentner and colleagues) was pedagogical analogies. For example, a solar system is like an atom.

Some researchers believe that IPA and PA analogies tap different kinds of reasoning (Goswami, 1991). IPA analogies have four elements that need to be mapped and further, these mappings require within group, e.g. A is to B, and between group comparisons, i.e. A:B is to C:D, while PA analogies have only two, but substantially more complicated, elements, the base, e.g. the fortress situation, and the target, e.g. the medical situation. In addition (Holyoak, 1984), the IPA task set up explicitly requires analogical reasoning, while a primary challenge in a PA lies in noticing the need for

analogical reasoning -- problem types for the approaches function differently. Moreover, IPA analogies can be arbitrary and can reverse the natural syntactic order while PA analogies are usually causal and more real-life. PA analogies deal with larger, more complex constellations of ideas and require at times extensive transformation and restructuring of the base in order to map an analogy onto the target. Further the notion of mapping is extended to sets of ideas.

Yet IPA and PA analogy types also draw upon similar processes. That is, responses to both tasks would correlate positively (Holyoak, 1984). This literature is relevant for a SAT analogy item problem-solving model because it provides (1) an enlarged notion of the mapping/comparing component (Gentner, 1983; Clement & Gentner, 1991; Holyoak & Thagard, 1989a), (2) an archetype for propositionalizing the stem rationale (Holyoak, 1984) and determining levels of representation (Gentner, 1989), (3) a focus on "higher level" thought processes (Gentner, 1983, 1989), (4) an idea of abstract relation schemas (Holyoak, 1984; Holyoak & Thagard, 1989b) and system constraints (Gentner, 1983, 1989; Holyoak & Thagard, 1989a), (5) an incorporation of basic assumptions about the structural features of memory and the process mechanism of accessing and retrieving information from memory, and so on. Theories for PA analogies had to be more encompassing, due not only to problem complexity, but

because these investigators were cognitive scientists and wrote explicit problem-solving computer programs. Also note that the term 'mapping' is used differently here -- its meaning is defined below.

Gentner instituted a program of research into this domain. In an initial article (1983), Gentner laid out her assumptions and structure-mapping rules, based, she asserted, on psychological theory. First, simple comparisons of features or attributes do not work in an analogy situation; in the analogy, an electric battery is like a reservoir, comparisons made on the basis of size, color or shape are not useful. This was contrary to Sternberg's feature assumptions (1977a, 1977b). Second, knowledge, it was assumed, was represented as propositional networks. Further, structure-mapping makes 1 to 1 correspondences that are content free, relying only on the syntax of knowledge representations.

> "The central idea in structure-mapping is that an analogy is a mapping of knowledge from one domain (the base) into another (the target), which conveys that a system of relations that holds among the base objects also holds among the target objects. Thus an analogy is a way of focusing on relational commonalities independently of the objects in which those relations are embedded....Central to the mapping process is the principle of systematicity: People prefer to map connected *systems of relations* governed by higher-order relations with inferential import rather than isolated predicates (Gentner, 1989, p. 201)....Among the many predicate matches between a given base and target, it favors those that form coherent systems of mutually interconnecting relations (p.202)".

That is, higher-order predicates, or more meaningful

relations like causality, which are more deeply embedded in knowledge structures, are preferred. Thus, the principle of systematicity imposes constraints on the kinds of mappings permitted in the system (Clement & Gentner, 1991).

In addition, there are levels in the kinds of relations or similarities (Gentner, 1989, 1983): (1) literal similarity, (2) analogy and (3) abstraction. While Gentner focused on the analogy level, SAT analogies may fall at all levels, and perhaps most frequently at the lowest. Gentner suggested a representation procedure as a way to operationalize levels of relations, which will be useful for this study. Perhaps future SAT analogy items can all be at an "higher-order" level like planets:sun::electrons:nucleus. Moreover, Gentner, like Mitchell (in Chapter 2), explicitly portrayed memory in her flow diagram of analogy processing (cf. 1989, p. 216). Structure-mapping and level of relation are both critical variables and will be further considered in this study.

Now that structure-mapping has been defined, a last piece of IPA research can be described. Embretson & Schneider (1989) proposed a Model III revision (Sternberg, 1977a, 1977b), by adding a structure-mapping component and by emphasizing the image construction aspect of the application component. These additions reflected latency results, for total items and subtasks, in terms of relational ratings. They discovered that neither a solely

top-down, conceptually driven model, nor a mapping component, as defined by Sternberg, could fully explain latency results. Further, analogy solution could occur a small percentage of the time without a correct A:B inference. Additionally, decision context variables, reflecting the presence of choice alternatives, provided little additional explanation of latencies! Rarely did response alternatives prompt inference revisions. Instead, the A:B::C context was most important; inferences were substantially modified with the presentation of the C term. Thus, Embretson and Schneider first recommended the use of relational distance ratings only when the base has been contextualized by a target domain, and second, postulated a structure-mapping component, that is, assessing the base and ideal target relations for shared higher-order relations (cf. Embretson & Schneider, 1989, p.176 for new model). Further studies are needed to affirm processing order for this new model. Note, however, that this study's results need to be modified to take into account the A:B::?:? SAT format. Furthermore, on one hand, it seems unlikely that image construction, or using C to conjecture an ideal D, is a part of SAT analogy processing. On the other hand, it does seem likely that in A:B::?:? items, alternative options would indeed contribute to a substantial contextualizing effect.

Returning to a PA, for another prominent researcher,

Holyoak (1984, p. 200; Holyoak & Thagard, 1989b; Holyoak & Nisbett, 1988), analogical reasoning was a module in a much larger picture of research, induction. Induction research had to take into account concept and category formation and representation, mechanisms of memory search, reasoning, problem-solving, learning, and system constraints. This survey of Holyoak and colleagues' work necessarily focuses more narrowly.

After assuming a connectionist memory model (Holyoak & Thagard, 1989b; Holyoak, 1984), with parallel processing, Holyoak (1984) listed mapping as one of four necessary steps in analogy as transfer problem-solving. A mapping step can be directly applied to classical analogies at two levels, between A and B and C_i and D_i and between A:B and C:D. Thus, when mapping a set of correspondences, attributes for an $A_{base1}:B_{target1}$ comparison and object relations for a $(A_{base1}:B_{target1})^{BASE}::(C_{base2}:D_{target2})^{TARGET}$ comparison, must be found between base and target situations. "Mapping relationships hold between elements of mental representations of situations [BASE and TARGET], and depend on some form of propositional representation (Holyoak, 1984, p. 204)." This notion will be further developed in Chapter 5 for the operationalization of mapping. That is, the stem and key rationales and the key and alternative rationales will be propositionalized as a way to depict number of mapping elements between the stem and key and number of elements in

common with the key, across all remaining alternatives.

In another article, Holyoak & Thagard (1989a) extended Gentner's thoughts by making a further evaluation of mapping. They found that successful mapping required three constraints: (1) Gentner-like structural constraints, as well as (2) semantic and, (3) not applicable here, pragmatic constraints -- semantic correspondences, not just structural, facilitated the mapping process. Again, similar to Gentner, to compare and judge similarity, "[r]easoning by analogy implies a comparison of two concepts at the same level of abstraction ... In general, a comparison statement can be interpreted as an analogy whenever it is possible to divide each concept into causal antecedents and relevant consequences. Analogy [also] involves similarity, but it is [a] structured similarity with functional import (Holyoak, 1984, p. 201, 204)." In effect, analogical reasoning should involve non-trivial relations. Additionally, mapping often is not perfect, just as the key rationale may not perfectly match the stem rationale.

Holyoak (1984) continued by describing a necessary abstract induction of a schema for mapping and appropriate problem solution. In the fortress-ray problem, content specific schema are separately established for the military and medical situations. Then a general "convergence" schema is formed, permitting accurate mapping to ensue; the base and target schemas are transformed, allowing for

generalization and mapping. This may also occur for SAT analogies when the stem's surface features are vastly different than the key's. A distinction between items requiring extensive transformations and items that don't may be captured by the independent/overlapping ETS taxonomy mentioned earlier.

To summarize, a PA literature review has offered several insights for this research project. Other important variables emerged: structure-mapping, levels of relations and degree of transformation. In addition, ways to understand and conceptualize the map/compare component have been given which mesh nicely with the SAT analogy format. Further, the literature offered possible ideas for variable operationalization. Importantly, a more wholistic and realistic view of problem-solving has evolved as well, which explicitly included assumptions concerning memory and processing as well as concept structure and formation. The following sub-section elaborates on these themes by examining declarative memory related issues, i.e. theories of semantic memory, the structure of concepts, and relation element theory. This includes brief acknowledgements of procedural, working and episodic memories.

Memories and Concepts

Memory is composed of several **functionally** distinct subsystems, the semantic/lexical, procedural, episodic, and short term/working memories; these subsystems may or may not

be **structurally** distinct. For an analogy testing situation, all subsystems may be called upon. So for example, when confronted with an analogy test, an examinee first draws from her procedural memory a general, heuristic set of analogy problem-solving rules like those often provided in SAT preparatory books. One set was: (1) express the stem in as complete a sentence as possible, (2) think of an ideal answer before examining the choice alternatives, then (3) consider the alternatives for relations that could be expressed in the same 'ideal' way, at this time (4) discard inappropriate choices, and finally, (5) choose the best alternative.

When tackling an individual item, however, an examinee needs to perceive and encode each word, and retrieve them from LT semantic/lexical memory. Three major theories have attempted to explain semantic/lexical memory: (1) separate-trace models, or semantic networks, involving spreading activation, (2) episodic-trace models involving parallel activation and (3) composite/distributed or connectionist models (Raaijmakers & Shiffrin, 1992). An episodic trace model, (2), is very similar to a semantic network paradigm, (1), except that it assumes parallel processing over the separately stored, highly personalized, memory traces. Therefore in this discussion, these two models are appraised together.

To begin with Models 1 and 2, "Semantic networks

represent information as a collection of nodes connected by labelled arcs that express links or relationships between the nodes (Markowitz, Nutter & Evens, 1992, p. 377)." Nodes are concepts (e.g. each analogy term is represented by a concept); links express the semantic or episodic relations between the concepts (i.e. between analogy terms). Links are bidirectional and differentially weighted. Closely related concepts have multiple, heavily weighted links. Semantic nets are often visually diagramed, like maps; propositionalization is another usual form of semantic representation (Anderson, 1985). Both propositional and network representations assume mental structures. Networks, however, if frequently accessed, become more and more abstract or decontextualized. In this case, they are called schema.

As is usual, separate- and episodic-trace models have been criticized and modified over time (Johnson-Laird, Herrmann, & Chaffin, 1984; Collins & Loftus, 1975). For instance, Johnson-Laird et al. asserted that semantic network theory did not satisfactorily fulfill all of four fundamental requirements of a psychological theory: (1) to specify the processor's mental representation, (2 & 3) to explain intensional and extensional relations and (4) to interpret people's inferences. Intensions are composed of an expression's senses or meanings, while extensions go forth from intensions to the actually referred to, concrete

objects. In their critique, Johnson-Laird et al. asserted that semantic networks primarily clarify **intensional** relations, and only partially achieve that goal, but do not consider extensional relations. Note that a type of relation dichotomy has once more arisen.

Chaffin and Herrmann (1987, 1988) further criticized semantic network approaches. Network theory has treated relations, located on links between concepts, as unitary constructs. Instead, Chaffin and Herrmann argued, relations themselves are concepts, which need to be further decomposed into "semantic primitives", as some relations are more or less typical and more or less complex than others. One source of evidence was that subjects were able to distinguish between kinds of relations and classify them into 'most similar' categories. This, at the same time, provided empirical support for a semantic taxonomy of five categories: one was contrast relations, two categories, class inclusion and similarity relations, had logical or intensional characteristics -- that is, involving comparisons of properties, and lastly, two other categories, case and part whole relations, based on temporal associations, had pragmatic characteristics (Notice the dichotomy). Hence, relations within a category share similar properties; those in other categories differ in distinct ways. To make classification judgments, subjects had to compare these relational properties and, in doing so,

relations were decomposed. Chaffin & Herrmann systematized relation decomposition in their relation element theory (RET). RET was designed to account for people's abilities to identify, compare and discriminate between relations and to generate novel relations.

The third major thread of memory work, a connectionist approach, came to terms with some problems and conjoined many of the subsystems (Iran-Nerad, Wittrock & Hidi, 1992). One researcher in this domain, Rumelhart (1989, p.299; McClelland & Rumelhart, 1985) developed a memory theory called parallel distributed processing (PDP).

> "PDP models, like brains, consist of very large networks of very simple processing units, which communicate through the passing of excitatory and inhibitory messages to one another. All units work in parallel without a specific executive....Knowledge resides only in the connections, and all learning involves a modification of the connections."

Retrieval occurs when a previously activated memory trace is re-activated. New material is stored by strengthening appropriate units and weakening inappropriate units. An experience reflects a pattern of activation.

Although a micro-level theory, PDP can explain complex tasks like reasoning by similarity through pattern matching and mapping. Further,

> "semantic memory may be just the residue of the superposition of episodic traces....Over repeated experience with the same pattern in different contexts, the pattern will remain in the interconnections of the units relevant to the content subpattern, but the particular associations to particular contexts will wash out. [It becomes a composite (McClelland & Rumelhart, 1985).] However, material that is

> encountered only in one particular context will tend to be somewhat contextually bound (Rumelhart, 1985, p. 303)."

Thus, for example, intensional and pragmatic relations can be explained. Intensional relations refer to the oft used abstracted patterns; pragmatic relations remain context, reality bound. In addition, episodic memory was already incorporated into a PDP model. Hence, if an examinee fails to retrieve the meaning(s) of a word, she either had never learned it or it is at this time inaccessible. Therefore, a fall back could be retrieving episodic traces, or remembering the situations, times or contexts where this word had appeared. Also, addressing Chaffin and colleagues concerns, all concepts, relations or otherwise, are built up and represented from micro-units -- that is, relations originate from patterns of the most element connections. At this time, PDP research presents a strong theory of semantic memory and will be the assumed theory for this study. Notice that this requires an incorporation of a connectionist, parallel distributed view within a linearly presented IPA model -- a seeming contradiction. However, top-down and bottom-up processing have already been incorporated. Further, the structure of the memory components has been left as a "black box" for this study.

Notwithstanding, while a connectionist theory addresses and rectifies some problem areas in network theory, it does not consider concept representation in STM. Semantic

networks, either as maps or propositions, retain an heuristic value in this regard. To reiterate, propositionalizing declarative sentences (Johnson-Laird et al., 1984; Anderson, 1985) may be a useful technique in operationalizing alternative similarity and mapping.

Actual concept research has indicated evolving notions of concept development, formation and representation. This research begins with Bruner's classic hypothesis-testing formulation and moves on to Rosch's probabilistic prototype theory, illustrating extreme poles in concept theory, with Smith (1988) and Medin (1989) taking a middle-ground position. Concepts serve several important functions: they provide cognitive economy, allow inference beyond a given situation, link perceptual and non-perceptual information, and can be combined for complex thinking (Smith, 1988). Additionally, concept formation can be induced either through bottom-up or top-down processing (Holyoak & Nisbett, 1988). However, both the classic and probabilistic views have placed feature similarity as central in concept and category formation. Medin (1989) asserted that a theory based solely on similarity is too unconstrained; instead, concept formation is best conceived of as driven by a given individual's personal theories. Further, Barsalou's (1989) investigations have shown that retrieved concepts are not consistent. How a concept is retrieved or how it is represented depends on problem-solving requirements. In an

extension of Barsalou's work, Michalski (1989) focused in his studies on the dynamic, inferential nature of concepts. In addition, and related to Chaffin and Herrmann's findings, Henley (1989) extracted a subset of "primary relational concepts" (contains, in, has, is, makes, needs, on, part of, type of, uses). Notice that Henley views primary relations as concepts. For Henley, primary relations impose structure on concept formation, an internal constraint, and on kinds of transitions between concepts, or as he said, on "the flow of thought", an external constraint.

Once word concepts have been activated in LTM and represented in STM, they need to be 'worked on' in WM[4]. That is, the examinee attempts to infer a relation between A and B, or to map/compare higher-order predicates between A:B and C_i:D_i, or to make an answer decision. Baddeley (1992) recently has presented a model of WM composed of three subcomponents: a central executive, a visuospatial manipulator and a phonological processor. Functions of WM include retained activation of perceptually encoded information, coordination of resources, placement of information into LT storage and information manipulation. He further noted that a high correlation exists between working memory capacity and reasoning ability. Just and Carpenter's (1992) results supported that finding and they further concluded that WM capacity constitutes a source of individual differences. For example, those individuals with

a smaller WM capacity may not have resources available to contend with ambiguous analogy items, novel analogy relations or analogies with many similar-to-the-key alternatives.

In total, memory subsystem research has provided additional useful background for a fuller understanding of analogy problem-solving. A more detailed account is not necessary for this research effort. In fact, at this point, LLTM's can only model coarse portrayals of cognitive components. Nevertheless, a proposed IPA model must reflect current cognitive knowledge.

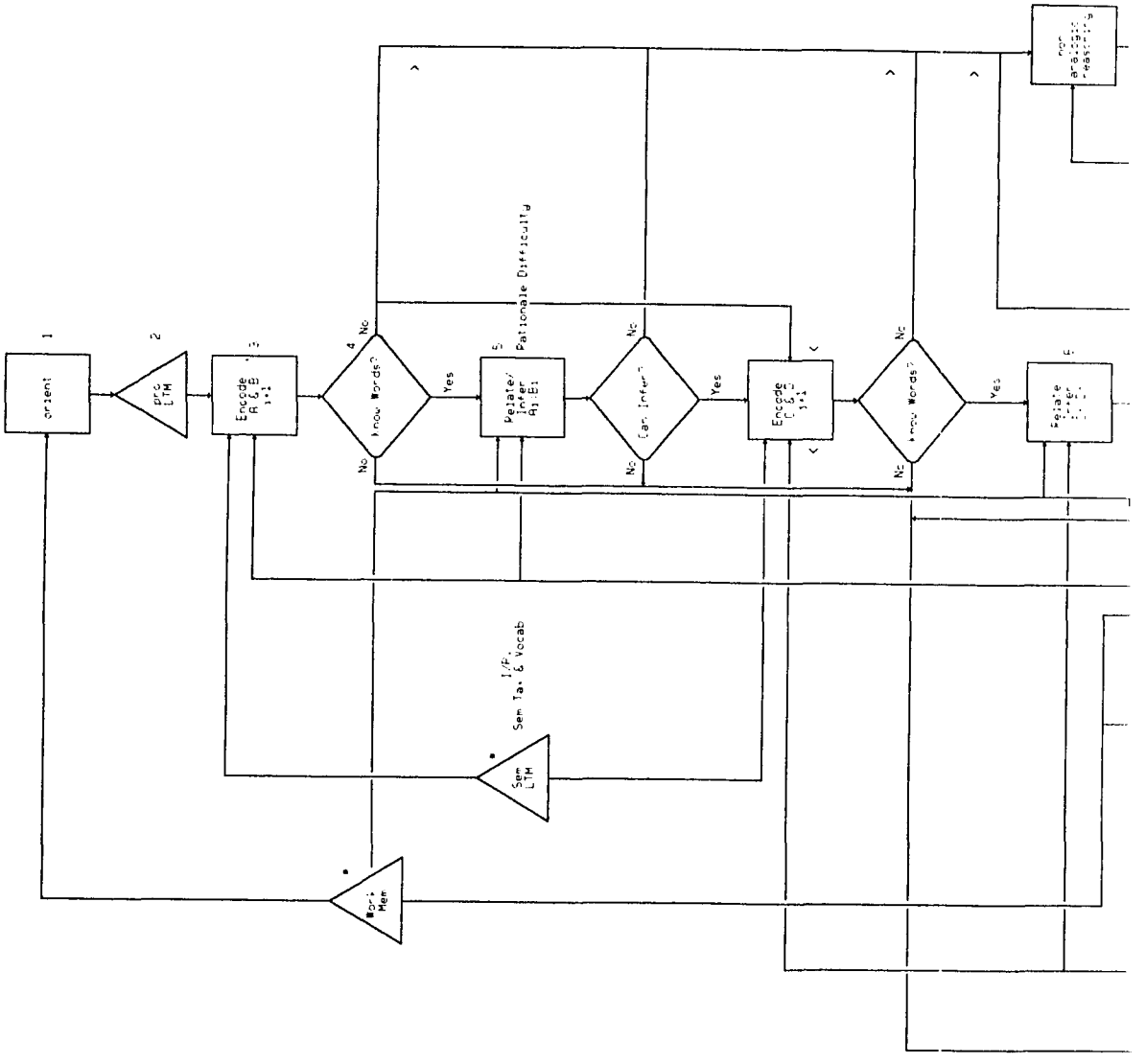
The Model

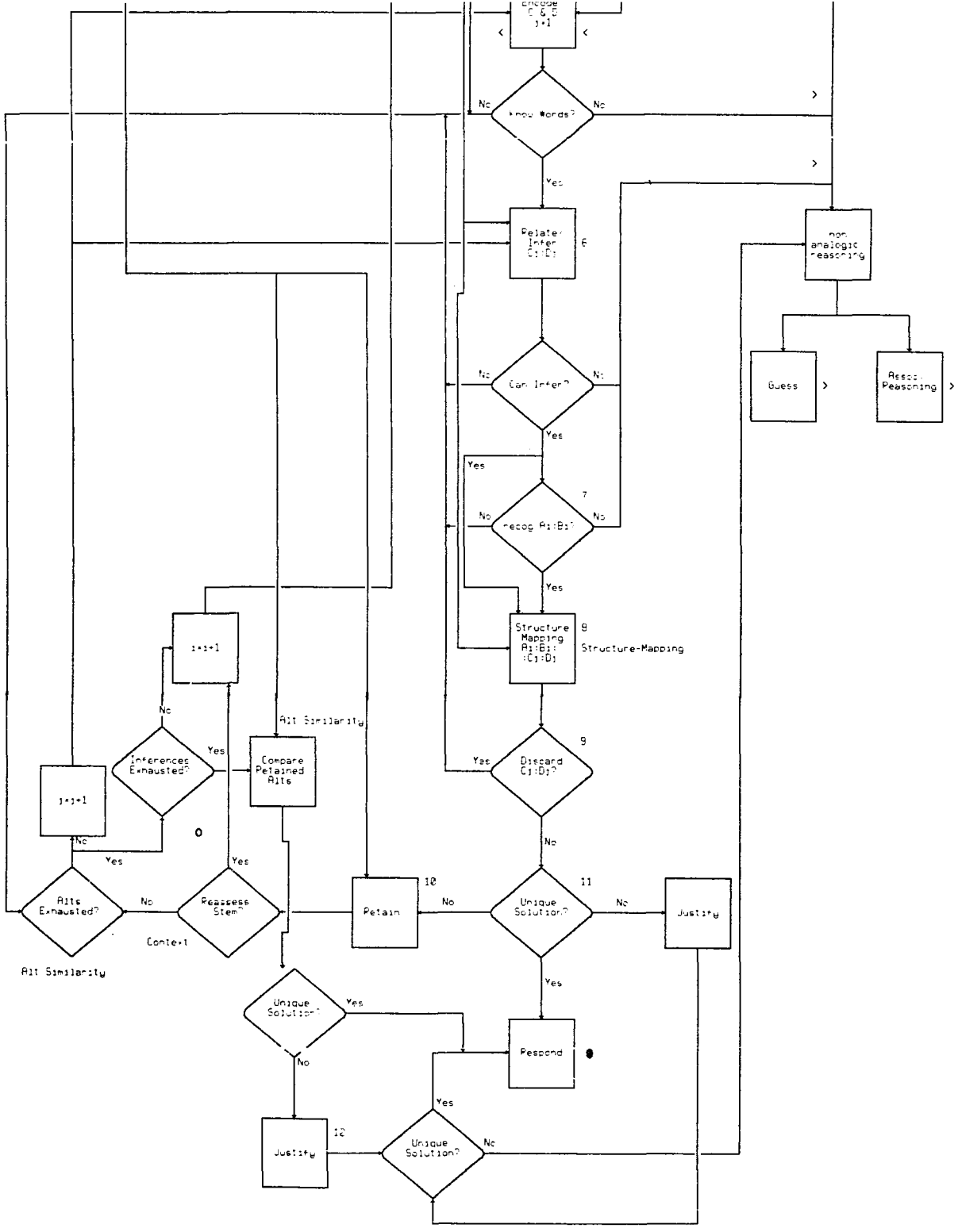
Therefore, a final integrated, global IPA model is now presented in Figure 3.7. Notice that memory subsystems are incorporated into the model, that the mapping/compare component is now called structure-mapping (See 8 on Figure 3.7) and that the appropriate components are now labelled with the discovered variables in a manner similar to Embretson & Wetzel's (1987) and Sheehan & Mislevy's (1991) models. A more detailed discussion of the variables follows in Chapter 4.

To run through the problem-solving model, a hypothetical individual is confronted with an analogy item. First she orients to the problem (See 1 on Figure 3.7) and draws up from her long term procedural memory (2) a list of analogy problem-solving rules. Then she encodes A and B (3)

Figure 3.7

IPA Model with Memory and Linked Variables




Encode C & D j+1
Know Words?
No
Yes
Relate: Infer C1:D1
6
non analogic reasoning
Guess
Assess Reasoning
Can Infer?
No
Yes
7
Recog A1:B1?
No
Yes
Structure Mapping A1:B1:: :C1:D1
8
Structure-Mapping
i+i+1
Alt Similarity
Inferences Exhausted?
No
Yes
Compare Retained Alts
9
Discard C1:D1?
Yes
No
j+j+1
0
Alts Exhausted?
No
Yes
Reassess Stem?
Yes
Context
Alt Similarity
Retain
10
Unique Solution?
No
11
Justify
Unique Solution?
Yes
No
Respond
Justify
12
Unique Solution?
Yes
No

and seeks to retrieve them from semantic long term memory *. In this study, three variables have been selected to represent this aspect of the processing model: vocabulary, a semantic taxonomy and the intensional/pragmatic distinction. If she doesn't know (4) the meaning of these words, she may either guess an answer or free associate to an answer, >, and move on to the next problem or jump ahead to the alternatives, < -- maybe A:B's relation will become evident, 7. If she successfully encodes A and B, then in WM, *, she infers the most complete, best guess A:B relation, 5; this is a function of the variable stem rationale difficulty. If she is unable to formulate a relation, again she has the option of exiting through guessing/free associating, >, or addressing the alternative options, <. Or, if the examinee has successfully inferred an A:B relation, the alternatives, C_i:D_i's, are perused, 6, and the same encoding and inferring activities are repeated. For each C_i:D_i, A:B and C_i:D_i are structurally mapped and compared in WM, * and 8. Note that there are three structure-mapping variables: an independent versus overlapping differentiation, number of common elements between the stem and key and level of relation. The examinee may now decide to discard an alternative, 9, to place it in WM for future consideration, * and 10, or, if a unique solution, 11, has been arrived at, to respond, •. If necessary, WM processing continues: the examinee may now reassess the stem rationale or other alternatives

iteratively or check and compare retained alternatives for a 'large enough' similarity to the stem rationale, ○ -- some subjective threshold must be surpassed. Selecting from alternatives creates new challenges and the context effect and alternative choice variables are intended to reflect them. If a subjective threshold is not surpassed, then the justification component is activated, 12. Depending on justification success/failure, again, the examinee may resort to guessing or free associating, >. If the threshold is exceeded, the examinee responds, •.

Conclusions

This chapter has reviewed many facets of the cognitive literature -- facets which all relate to analogy problem-solving. Several relevant variables have been uncovered. First, rationale difficulty determines the relative difficulty of inferring the stem relation. Next, alternative choice variables can contextualize the stem, necessitating a re-evaluation of the stem rationale, and/or lead to a choice decision-making process. Structure-mapping indicates the difficulty of mapping a relation between $A{:}B^{BASE}$ and $C_i{:}D_i^{TARGET}$. In addition, an intensional-pragmatic distinction between types of relations was formulated by integrating past relation dichotomies. In terms of a connectionist memory theory, items which have intensional relations are accessed differently than pragmatic items. Also, a semantic taxonomy may reflect repeatedly accessed

kinds of relations stored in semantic memory. As analogy relations may be novel, however, a semantic taxonomy can at most provide a partial explanation of item difficulty or, maybe semantic relations, as expressed in a taxonomy, represent an important variable with few individual differences (Carpenter et al., 1990). In addition, word frequency, either written or oral, can affect item difficulty -- students may not know or be unable to retrieve infrequently occurring words. Lastly, the non-theoretic variables, syntactic structure and/or ETS test development taxonomies, may be related to item difficulty. With this information, research hypotheses are presented in the next chapter and variable operationalization is described in Chapter 5.

Footnotes

[1] Note that a revised version of the SAT's, the Scholastic Assessment Test, came out in the spring of 1994. Analogy items, however, will be retained in their original form.

[2] Whether WM and STM are distinct memory subsystems, or not, varies throughout the literature. Without resolving this dilemma, it is useful for discussion purposes to keep them distinct.

[3] Sternberg's (1977a) book provided a complete and detailed analysis of his componential theory of analogy problem-solving. The Psychological Bulletin article offered an excellent summary of results, models and methodology.

[4] In his book (1977a), Sternberg provided an interesting depiction of a hierarchy leading from theory to model to component (Table 4.2, p. 69); the body of the table indicted, for all levels, sources of individual differences.

Chapter 4: Variable Selection and Research Hypotheses

Out of the works surveyed, six variables have been selected as being of key importance in analogy problem-solving. They were three semantic/episodic memory variables: written vocabulary knowledge, common semantic relations and a dichotomy of relations (Intensional or Pragmatic), and three working memory variables: stem rationale difficulty, alternative choice variables influencing decision-making and stem contextualization, and structure mapping. As mentioned earlier, these variables reflected different aspects of the individual solving analogy problems: 1) the knower, or declarative and procedural memory, 2) the processor, or the utilization of information processes, strategies, or induction and inference within working memory, and 3) the experiencer, or episodic memory. Other available and perhaps useful variables derived from Educational Testing Service (ETS) test development taxonomies (content and level of abstractness) and syntactic structure. Characteristics of analogy item representing these variables must be selected; that is, the variables must be operationalized. Hence, this chapter further describes and develops each variable, retaining the context of the LLTM processing model and states the formal research hypotheses; Chapter 5 explicates variable operationalization.

Model Components

Semantic Memory Variables

As noted in Chapter 3, a major part of successful analogy problem-solving relies on previously acquired knowledge -- knowledge stored in LTM. This section will cover the three semantic memory variables: vocabulary knowledge, prototypical semantic relations and the Intensional/Pragmatic (I/P) dichotomy.

Vocabulary Knowledge

Several studies (Enright, Duran & Pierce, 1986; Bejar et al., 1991; Carroll, 1979) have noted the importance of vocabulary as a source of difficulty in solving analogy items. For example, a student cannot hope to reason analogically unless a certain level of vocabulary proficiency has been attained (Enright et al., 1986). Additionally, even if an examinee "knows" a word, vocabulary knowledge must be accessed and retrieved from LT semantic memory -- sometimes a difficult task (Embretson, 1985; Embretson & Curtright, 1982). Therefore, if an examinee does not have sufficient vocabulary knowledge or cannot access it, she will never reach the inferring or mapping components noted in the analogy problem-solving model. This is true even though analogies are designed to measure reasoning. Therefore,

Hypothesis 1: It is expected that less frequently occurring words for the stem, key and alternative choice analogy terms

will result in more difficult items.

Semantic Relations

Researchers have developed taxonomies of semantic relations for two primary reasons: (1) to find "a limited number of relations ... which can function as explanatory primitives (the most 'primitive' or basic set of relations, to which all others can be reduced. [from Sowa, p. 13]) in theories of mental function (Chaffin & Pierce, draft)" and (2) to understand a memory structure (Whitely, 1977). Hence, taxonomies contend with cognitively relevant and frequently accessed **kinds** of relations. For example, "[g]lobal analyses of verbal analogy items reveal that the majority of verbal analogies can be classified as representing a limited set of basic types of semantic relations (Pellegrino & Glaser, 1979, p. 201)." Barnes (1980) also commented that taxonomic semantic relations are prototypical relations. As these are typical relations, it is assumed that they are stored in abstract, decontextualized composites in LTM (Rumelhart, 1989; cf Chapter 3).

Thus, many taxonomies of semantic relations have been developed (e.g. Chaffin & Pierce, draft; Freedle & Kostin, 1991, 1988, 1987; Whitely, 1977; Ingram & Pellegrino, 1977; Chaffin & Herrmann, 1987, 1988; Spearman, 1923; Sternberg, 1977; Markowitz et al., 1992). Henley's (1989) primitives can be regarded as another semantic taxonomy. In one case,

Pellegrino and Glaser (1980) listed class membership, function, location, conversion, part-whole, temporal order and property; similarly, Whitely (1977b) derived eight categories using latent class partition analysis: opposites, functional, word pattern, quantitative, similarities, class membership, class-naming and conversion. In fact, a great deal of overlap exists across taxonomies. Hence, if semantic taxonomies have in fact evolved out of researchers' needs to explain lexical or semantic memory, and if a semantic taxonomy indeed represents a set of explanatory primitives, there should be convergence to a limited number of kinds of relations. Thus, as some of the above-mentioned taxonomies (Markowitz et al., 1992; Freedle & Kostin, 1991) provide all-encompassing, minute subclasses, they may accordingly be rejected on the grounds of parsimony. Additionally, it must be noted that many of the taxonomies were developed empirically; therefore, the classes may also relate to the kinds of items that test writers have developed.

Of additional interest for this project, however, is a taxonomy that would be most relevant for SAT items. As one possibility, Chaffin and Pierce (draft) developed a taxonomy of relations to categorize, as parsimoniously as possible, analogies within a GRE context. Table 4.1 shows the result of extensive analyses of 179 GRE analogies.

Table 4.1

Semantic Taxonomy of Relations

Relation	Example	Rationale
Class Inclusion	robin:bird	A is a member of class B
Part-Whole	engine:car	A is a part of B
Similar	breeze:gale	B is a more intense A
Contrast	default:pay	A is the opposite of B
Attribute	beggar:poor	B is an attribute of A
Non-Attribute	harmony:discord	B is not an attribute of A
Case Relation	tailor:suit	A works on B
Cause/Purpose	hunger:eat	A is the cause of B
Space/Time	judge:court summer:harvest	A can be found in B B occurs during A
Representation	building:print	B is a representation of A

While the classes were developed to be cognitively general, GRE's are designed for a select population and must range toward the difficult end of kinds of relations; so some classes may be particular to GRE's (Chaffin, 1992, personal communication). Diones, Bejar and Chaffin (in preparation) did find that when drawing from a large set of disclosed SAT items, items could be classified into the ten classes and that, if items were clustered by taxonomic classes, the test was unidimensional. However, Schmitt and Bleistein (1987) needed to use a reduced form of the taxonomy. Therefore,

the above semantic taxonomy was used, but, if need be, the number of classes would be reduced.

Notwithstanding, there have been mixed results in using a semantic taxonomy to explain item difficulty (Diones et al., in preparation; Glaser & Pellegrino, 1979; K. Tatsuoka, personal communication, 1993), differential item functioning (Schmitt & Bleistein, 1987) or individual differences (Whitely, 1976). "One problem with ... a classification scheme is that it only captures the most salient relational feature .... Furthermore, it does not indicate differences in the ease or likelihood of identifying the relational features (Pellegrino & Glaser, 1980, p. 202)". Moreover, many analogy items require the formulation of novel or atypical relations. In addition, simpler relations may be concatenated or alternatives may dictate a reappraisal of the stem rationale (Bejar et al., 1991). Thus, a mutually exclusive approach to relation classification may be inappropriate and therefore may offer at most a partial explanation of item difficulty. In addition, while previous studies (Bejar et al., 1991; Diones et al., in preparation) have found rank order differences in item difficulty by class, taxonomic class still did not explain item difficulty (Diones et al., in preparation).

Therefore, if taxonomic relations are frequently accessed, prototypical relations, known by everyone, implying that there would be few individual differences,

and, given that many errors in performance arise because people do not always use what they know (Gitomer et al., 1987), then,

Hypothesis 2: it is expected that as a whole, the semantic taxonomy will not explain item difficulty. Accordingly, it is predicted that an hypothesis of no relationship would fail to be rejected[1].

If an examinee fails to access a relation, it may be due to problems accessing information in LTM or other properties of the item: infrequently used words, ambiguity, concatenated relations, a novel relation or the dichotomy of relations described next.

A Dichotomy of Relations

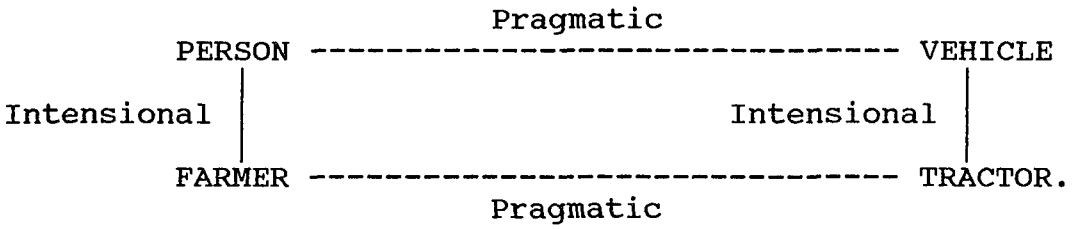
As discussed, dichotomies of relations have been noted previously, both theoretically and empirically. Since the late 1800's, researchers have parsed analogy relations into those capitalizing on the "internal" meaning of the word pair's rationale versus those depending on factors "external" to the pair's definitional meanings (Warren, 1967). Based on a theory of memory, Klix (1980) defined an intra-concept versus inter-concept distinction. Intra-concept relations are results of internal matching processes that are abstracted from perception (e.g. chair:furniture) while inter-concept relations call on observing and experiencing worldly events (e.g. sun:set). Note that intra-concept relations can be thought of as

decontextualized within semantic memory while inter-concept relations can be conceived of as fixed in memory together with their episodic traces, relating back to Rumelhart's connectionist memory theory. Looking only at single word meanings, rather than analogy relations, Sowa (1984) discussed intensions and extensions. As part of semantic memory, the intension of a word is its dictionary or universal principles meaning, including general facts about the concept. A word's extension, which is considered part of episodic memory, comprises the "set of all existing things to which the word applies." (p.11).

Therefore, an I/P distinction may be defined. An intensional relation is composed from a pair of concepts, say X:Y, by noticing overlapping attributes or properties, intrinsic to their semantic definitional meanings, which allow them to be conjoined, compared or matched. They also draw on a knowledge of language and are a temporal and universal (i.e. cross-cultural). In contrast, pragmatic relations are based on an external co-occurrence of the concepts through perception or experience. They also rely on knowledge of the world and are temporal. An illustration of this distinction (Bejar et al., 1991 p. 68) is portrayed in Table 4.2.

Table 4.2

An Example of the Intensional/Pragmatic Distinction



Note. From Cognitive and Psychometric Analysis of Analogical Problem Solving (p. 68) by I. I. Bejar, R. Chaffin, and S. E. Embretson, 1991, New York: Springer-Verlag. Copyright 1991 by Springer-Verlag. Reprinted by permission.

A farmer is by definition a person just as a tractor is a vehicle (These definitions can be found in a dictionary). These word pairs therefore describe intensional relations. Through personal experience, though, most individuals in technological societies know that farmers often use tractors when working the farm. Hence this relation is pragmatic. While this example simultaneously portrays intensional and pragmatic characteristics, typical analogy items express either an intensional or pragmatic relation.

Consequently, if intensional relations are stored in abstracted, decontextualized composites and pragmatic relations remain tied to their episodic traces (Chapter 3; Rumelhart, 1989), then,

Hypothesis 3: It is predicted that intensional items will be more difficult than pragmatic. For intensional items, an appropriate composite must be retrieved and then enriched and elaborated for each particular item; for pragmatic items, memory traces are retrieved in an already reality-bound form.

As additional background, an initial study (Bejar et al., 1991) suggested an hierarchical co-functioning of the I/P dichotomy and kind of prototypical relation, with the dichotomy presiding over the semantic classes. (Please see Table 4.3.) The kind of relations described by Class Inclusion, Similar, Contrast, Attribute, and Non-Attribute behaved similarly (i.e. data points clustered together) and

had intensional properties while the same was true of Case Relation, Cause-Purpose, Space-Time, Part-Whole, and Representation relations, except they had pragmatic characteristics. Given this theoretical scenario, relations were treated as belonging solely to one class and one type, but that type was determined once an item was assigned to a class.

Table 4.3

The Intensional/Pragmatic and Semantic Class Hierarchy

Relation Dichotomy	
Intensional	Pragmatic
Class Inclusion	Case Relation
Similar	Cause-Purpose
Contrast	Space-Time
Attribute	Part-Whole
Non-Attribute	Representation

Results from a recent study (Diones et al., in preparation) suggested that a reevaluation of this hierarchy is in order. A separate categorization scheme, independent of the taxonomy, needs to be developed or operationalized; that is, once an item is placed in a class, it is not automatically considered intensional or pragmatic. When Chaffin reclassified the items from the 1992 project as

Intensional or Pragmatic, independently of the taxonomy, the I/P distinction then significantly predicted item difficulty.

Working Memory Variables

A second major aspect of analogy problem-solving, and the part that draws most on an individual's reasoning abilities, occurs within the resources of WM. This section outlines three kinds of WM variables, stem rationale difficulty, alternative choice and structure-mapping. Both alternative choice and structure-mapping are respectively manifested through several variables: (1) context effects and alternative similarity and (2) independent/overlapping, number of common elements between the stem and key, and level of relation.

Stem Rationale Difficulty

Rationale difficulty, or the difficulty in inferring a stem's relation, has been approached in a number of ways -- as rationale complexity, relational distance and rationale difficulty. Often these variables have been considered representations of the same construct: stem processing difficulty. Up to now however, the variable stem rationale difficulty has taken various forms and has been only marginally successful. The fact that researchers persist in trying to manifest this variable highlights its importance. Additionally, operationalizations of rationale complexity have thus far not at all been successful in predicting item

difficulty (Bejar et al.; Schmitt, personal communication). For instance, Chaffin (Bejar et al., 1991 p. 22 and 130; see also Table 4.5) drew a propositional map representing the stem rationale. In his example *irrevocable:repeal*, three relations were implied: *repeal:law, law:revocable, revocable:irrevocable*. Nonetheless, an analysis based in part on a propositional approach and also number of syntactic elements (predicates and arguments) was not effective. In effect, stem rationale complexity may neither correlate with rationale difficulty nor relate to item difficulty. Lastly, relational distance was a variable emerging from Rumelhart and Abrahamson's (1973) Euclidean memory theory -- a view of memory rarely taken today. Hence, only stem rationale difficulty will be considered here.

Since theory has not suggested a way to operationalize stem rationale difficulty, ratings will be collected, as has been done previously (Embretson & Curtright, 1982; Bejar et al., 1991; Schmitt & Bleistein, 1987). Note that ratings will be collected only on the **stem** rationale; the ratings are not designed to reflect the entire item's difficulty. Accordingly,

Hypothesis 4: An item with a stem relation rated as difficult to infer will be more difficult than an item with a stem relation rated as easy to infer. Note that it will be the **final** stem rationale, as contextualized by choice

alternatives, that will be rated (cf Chapter 5).

Alternative Choice Variables

Two alternative choice variables emerged from the literature as being particularly important. They involved the possible context effects of the alternatives when inferring a stem rationale and the difficulties of selecting the key from several similar alternatives.

Context effects may affect item difficulty in systematic ways. While each examinee brings with her an internal context over which the tester has little control, choice alternatives provide an external context over which the tester can exert control. Some researchers (Bejar et al., 1991; Embretson & Curtright, 1982; Embretson, Schneider & Roth, 1986; Liu, 1981) have observed that multiple choice alternatives can affect the way a stem rationale is formulated or, in other words, the context provided by the alternatives constrains the possible stem rationales. Hence, the C_i:D_i's context may require a re-evaluation of the stem rationale, providing a **"contextualized"** final rationale (* in Table 4.4 is an example of an elaborated final key rationale (Bejar et al., 1991, p. 23)); that is, the stem rationale had to be modified after considering the offered alternatives. For other items, the stem pair itself is so constraining that there is only one possible, immediately apparent, stem rationale. Thus, in determining several of the variables' classification, it will be the

Table 4.4

Contextualization Effects of the Alternative Pairs

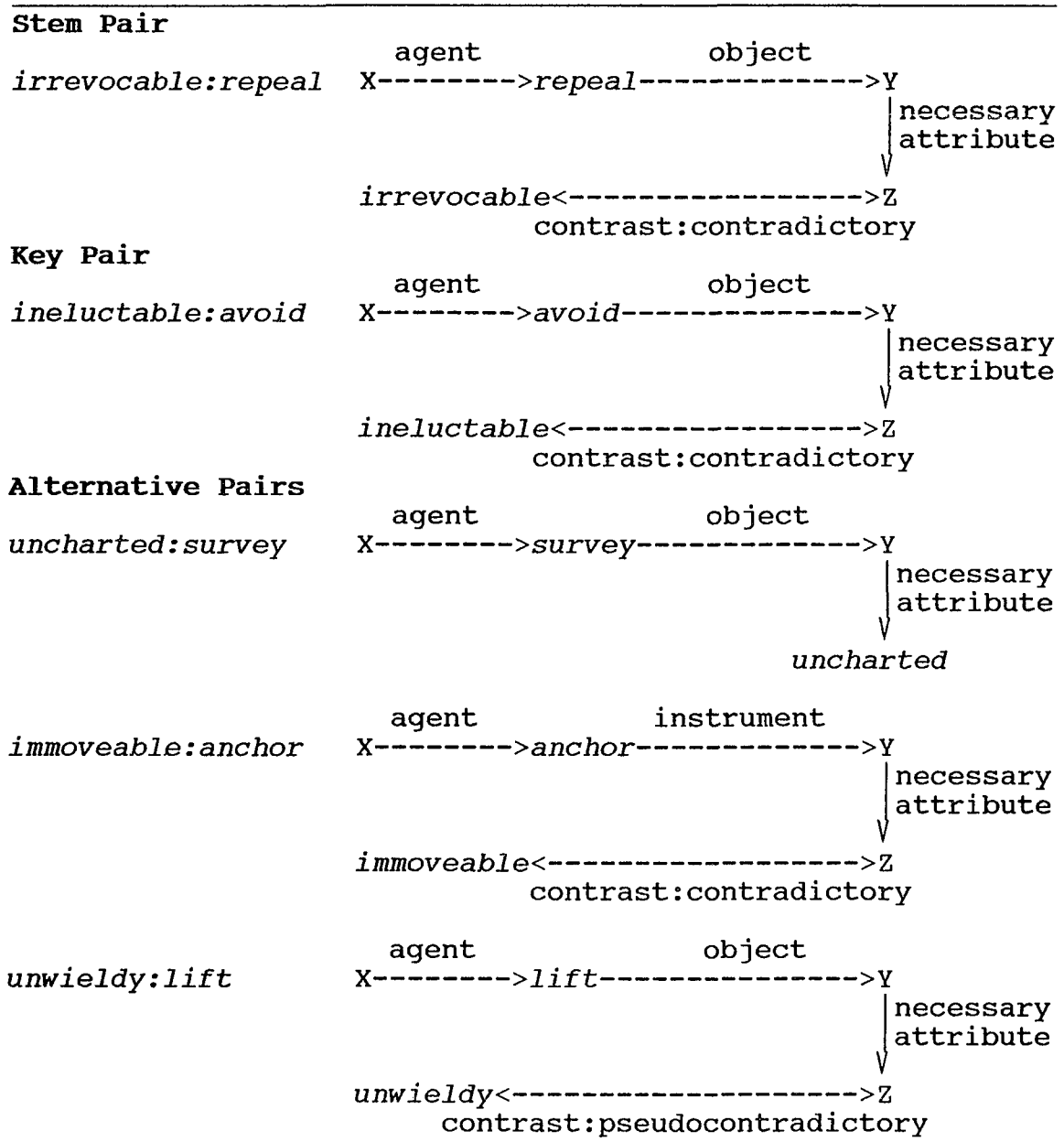
Stem Pair	Alternative Pair	Relation Expression
clapper:bell	tongue:mouth	A is in B
clapper:bell	horn:automobile	A is part of B
clapper:bell	speaker:radio	A is a component of B that produces sound
clapper:bell	needle:phonograph	A is a component of B that produces sound by contact with B
clapper:bell	hammer:piano	A is a component of B that produces sound by striking B*

Note. From Cognitive and Psychometric Analysis of Analogical Problem Solving (p. 23) by I. I. Bejar, R. Chaffin, and S. E. Embretson, 1991, New York: Springer-Verlag. Copyright 1991 by Springer-Verlag. Reprinted by permission.

final rationale, contextualized or not, that will be judged (See Chapter 5 for more details.).

Second, the task in multiple choice analogy problem-solving, and particularly as given by SAT directions, necessitates selecting the best alternative relation; that is, the examinee must select the alternative rationale most similar to the stem's. In effect, several alternatives may reflect varying degrees of concordance with the stem's relation, yet not be the best or exactly the same relation as the stem's relation (Test developers must take care that there are not two viable keys.). Further, a set of alternatives very similar to the key would also increase an item's decision-making difficulty. A propositional map approach illustrates one conceptualization of how the alternatives can differ from stem and key (Bejar et al., 1991, p. 22). Note that alternative rationales varied from the stem/key's in terms of number of simpler relations (3 or 4) and the kinds of simpler relations (e.g. contrast:contradictory versus contrast:pseudocontradictory). Consequently, a problem solver must ideally not only infer several relations per item, she must also contend with a decision-making task, choice selection from a set of similar alternatives.

Table 4.5
An Example of Propositional Maps



Note. From Cognitive and Psychometric Analysis of Analogical Problem Solving (p. 22) by I. I. Bejar, R. Chaffin, and S. E. Embretson, 1991, New York: Springer-Verlag. Copyright 1991 by Springer-Verlag. Reprinted by permission.

Therefore,

Hypothesis 5a: An item with a context effect will be more difficult than an item without a context effect.

Hypothesis 5b: Items with several similar alternatives will be more difficult than items with few or no similar alternatives.

Structure-Mapping Variables

As summarized earlier, structure-mapping is a major component of analogy problem-solving necessitating mapping or comparing elements of the base problem, or for SAT items, the stem, to the target problem, or the alternatives. In terms of structure-mapping success, items can vary on three variables: an independent/overlapping dichotomy, mapping of common elements between stem and key, and level of relation.

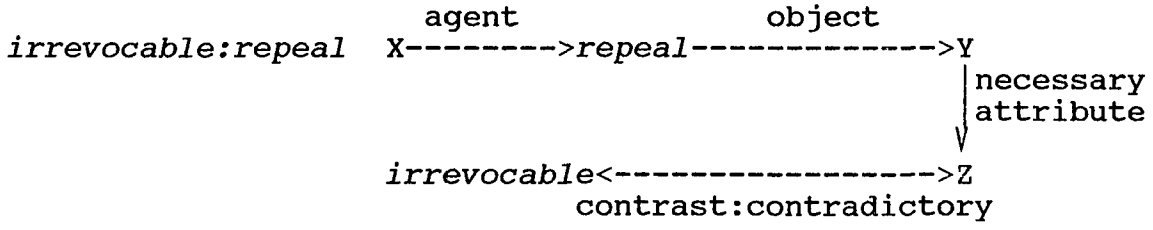
First, noticing the analogy between the base and target, A:B and C_i:D_i, can be difficult -- there may be few or many surface similarities between them. If few surface similarities exist, then the base and/or target must be transformed or abstracted (Holyoak, 1984) to allow for proper mapping. An independent/overlapping dichotomy, based on the ETS taxonomy, can manifest this distinction. To define these terms, "An analogy is said to be independent if the relationship in the stem would not normally directly suggest the relationship in the answer, but it is said to overlap if the stem suggests the key. Thus

FIRE:ASHES::EVENT:MEMORY demonstrates independent while FAMINE:FOOD and DROUGHT:WATER overlap (Donlan, 1984, p. 48)."

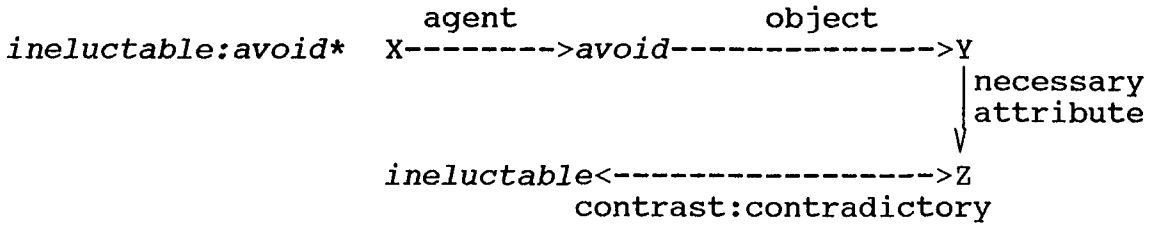
Second, the student must map the common elements between stem and key. For example, below are two propositional maps taken from the example in Table 4.5 (Bejar et al., 1991, p. 22), one for the key pair and the other for the stem pair. Here there are 4 common elements between the two maps: agent, object, necessary attribute and contrast:contradictory.

Example

Stem Pair



Key Pair



The more common elements there are between stem and key, the easier the mapping should be.

Third, recall that Gentner (1983, 1989) asserted that relations could be at many levels -- for example color attributes versus causality, but her principle of systematicity constrained mapping to higher order levels of

relations. Gentner (1989) proposed a set of representation conventions which can be illustrated. For example, when confronting a Water Flow problem (i.e. water flowing from a large beaker to a small vial through a pipe), three different levels can be portrayed: zero-order, first-order and second-order levels. LIQUID(water) is zero-order; an attribute of water is that it is a liquid. Another level of representation indicates that the diameter of the beaker is greater in size than the diameter of the vial or GREATER THAN(x,y); this is a first-order level. The highest level portrays a situation where the difference in pressure between the two vessels causes water to flow from the beaker to the vial or CAUSE{GREATER [PRESSURE(beaker), PRESSURE(vial)], [FLOW(water,pipe,beaker,vial)]}; this is a second-order level. As level increases from zero-order to second-order, the examinee's representation becomes more meaningful and the processing grows more challenging. As SAT items often have artificial, fabricated relations, however, stem rationales may not obey Gentner's systematicity constraint.

Thus,

Hypothesis 6a: An overlapping analogy will be less difficult than an independent analogy, as defined by the ETS taxonomy.

Hypothesis 6b: As the number of common elements between the stem and key increase, the item's difficulty level will decrease.

Hypothesis 6c: Items with higher-order mappings will be relatively more difficult than items with lower-order mappings.

At this point, all model driven theoretic variables and their respective hypotheses have been presented. Although diverse variables have emerged as important, the entire process of analogy problem-solving functions as a unified set of operations, hopefully resulting in a unidimensional data set.

Other Hypotheses

Other variables, while not explicitly connected to the processing model, will be evaluated. For example, relations which require inversion of the pair's word order may elicit more WM processing than relations not requiring an inversion (Embretson & Curtright, 1980; Embretson, 1984, 1985). For example, the pair *engine:car*, with a rationale "An engine is part of a car.", would be categorized as "retain the order of the stem's word order" while the pair *paint:artist*, with the rationale "An artist paints.", would be categorized as "reverse the order of the stem's word order". Further, since ETS has test construction taxonomies, content (e.g. philosophy, history, science) and level of abstractness (e.g. concrete, mixed or abstract), it would be interesting to see if they are useful in predicting item difficulty.

Hypothesis 7a: Relations requiring inversion of the stem word pair in order to formulate a rationale will be more

difficult than an item with a stem word pair that does not require an inversion to infer a rationale.

Hypothesis 7b: Level of Abstraction and content distinctions will relate to item difficulty.

Summary

In this chapter, the selected variables were further explicated and the research hypotheses were formulated. This LLTM study has extended previous work on psychometric analogy items by developing an as-complete-as-possible IPA model, explicitly showing the connection between the IPA model and the cognitively relevant variables, selecting and developing some new variables, extending some old ones and delineating very specific hypotheses. Next, Chapter 5 will describe this study's design, procedures and the methodology for operationalizing the variables.

[1] Though such a prediction, fail to reject the null hypothesis, is not considered typical of traditional inferential statistics, it has been used before (Hasher & Zacks, 1979) and it is regularly used for models like structural equation models, confirmatory factor analysis, and so on.

Chapter 5: Methods

Subjects and Materials

The data used for this study, responses to SAT[1] analogy items, were derived from three consecutive November[2] SAT administrations, 1988, 1989 and 1990; three forms were needed to furnish a large number of analogy items. Since these three forms needed to be equated or linked, Educational Testing Service (ETS) equating samples were used. In general, an ETS equating sample is comprised of approximately 10,000 SAT test takers. Equating samples are random and fairly representative across gender and ethnicity subgroups (Wendler, 1993, personal communication). Forms are linked as follows:

Table 5.1

Equating the Three SAT Forms Over Four Groups

1988		**1989**		**1990**
group 1		group 2	group 3	group 4
20 items	10 common items	20 common items	10 common items	20 items

To explain, each form contains 20 operational analogies, those items actually used in scoring the examinees, and 10 equating analogy items. However, the equating items are not the same each year, so that the items linking 1988 and 1989 November administrations (groups 1 and 2) are different than

the items linking 1989 and 1990 (groups 3 and 4). Thus, the middle year, 1989, has two groups (2 and 3), one with linking items for 1988 and the other with linking items for 1990; the 20 operational items are in common for all 1989 test-takers.

Therefore, given three forms, there were 60 disclosed, operational analogy items plus 20 non-disclosed equating items, yielding a total of 80 items. Certain item information was additionally collected: (1) the items' classifications on ETS test development taxonomies, (2) the items' text, (3) the answer key, (4) student responses on the Student Descriptive Questionnaire and (5) students' raw (formula scored) and scaled SAT-Verbal scores. Further, each item had six possible outcome scores, the five options and '0', which could signify multiple responses, omits, or not reached items. As noted earlier, students were explicitly instructed not to guess on the SAT's and to reinforce this the SAT's were formula scored (i.e. students were penalized for guessing and items scored as '0' were not counted in the student's final score). Reflecting this policy, within each analogy section, as items progressed from easy to hard, the percent 0's increased dramatically. Nonetheless, for the purposes of this study, all 0's were scored as being incorrect. Despite this, the correlation between item difficulty estimates with 0's omitted from the analysis versus 0's scored as incorrect was extremely high

(r=.98); thus the method of scoring was probably not significant for this LLTM analysis.

Summing across the four groups, there was a total sample of 43,929 students. However, as the population of interest was a college bound, English as first language, high school juniors and seniors non-disabled group, students were accordingly selected on grade level, USA citizenship, English as first language and non-disabled. Further, LLTM assumptions required that individuals with perfectly correct or incorrect scores be removed -- here a negligibly small number. Thus, the final group numbered 30,907. See Table 5.2 for details.

Table 5.2

Sample Sizes for the Four Groups

	Number of Examinees		
Group	Original	Selected	Post-LLTM
1	10,711	7,891	7,886
2	11,884	8,121	8,120
3	10,183	6,926	6,926
4	11,151	7,983	7,975
Total	43,929	30,924	30,907

The characteristics of the students comprising each group were quite similar: (1) Over 90% were 12th graders,

(2) the majority, 55%, were female, (3) around 85% of each group was Caucasian, 8% African American, and 2% Asian American; the percents of Native Americans and Hispanics varied somewhat more, (4) most mothers had a high school education, 28%, while fewer had some college, 16-17%, or a B.A. degree, 16%, while fathers were equally spread across three levels of educational accomplishment, high school or B.A. degree or graduate education, with around 20% in each category, (5) for every group, the modal income category was greater than $70,000 a year, (6) each group had exactly the same mean high school average, 8.4 or B, (7) the mean SAT Verbal score for the four groups was 438.3, (8) reflecting national trends, on average males did better than females on SAT Verbal; however the difference on average between males and females on the vocabulary subscore, which includes analogy items, was negligible across all groups (1 or 2 points), and (9) score distributions for SAT Verbal visually appeared to be normally distributed, with a very slight positive skew.

Design and Procedures

Task Design Phase

All 80 items were first rated and coded in terms of the six cognitively relevant variables. Thus, the variables were operationalized so that a reliable and structured procedure could be mechanically followed. There were six raters: three high verbal (GRE-V greater than 650 and

English as a first language) graduate students, two ETS verbal test development specialists and the researcher. The rating forms appear in Appendix A.

Vocabulary Knowledge: Vocabulary knowledge has often been operationalized in terms of word counts; that is, how often did each word appear over 14,360,884 words of text (Breland, Jones & Jenkins, 1994). Frequencies were extracted from Breland et al.'s (1994) word frequency data base -- the most recent and representative corpus. The frequency values actually used in the analyses were based on a statistic called the Standard Frequency Index (SFI) -- a transformed metric, preferred because of its interpretability (Breland et al., 1994). Thus, the SFI was noted for the least frequent word in each pair: the stem, key and alternatives. In mathematical terms, stemfreq=min(SFIstem), keyfreq=min(SFIkey) and altfreq= $((\min(SFIalt_1)+\min(SFIalt_2)+\min(SFIalt_3)+\min(SFIalt_4))/4)$. All unlisted words, which were assumed to have a frequency of 0, were arbitrarily given a value smaller than the smallest recorded value, 5, as ln(0) is undefined (The SFI statistic was in $\log_{10}$ units). Possible SFI values ranged mostly from 10 to 90.

Common Semantic Relations: When evaluating the analogy item's final stem rationale, the graduate student raters independently sorted all items into their respective taxonomic classes. Though a demanding task and perhaps

inappropriate[4], the items were classified into mutually exclusive classes. Then the raters resolved all classification disagreements through consensus. Please see pp. 190-191 in Appendix A[5]. An index of inter-rater agreement, Kappa (Fleiss, 1973), averaged .61 over all pairs of the three raters.

Dichotomy of Relations: As theorized, an intensional relation was composed from a pair of concepts, A:B, by noticing overlapping attributes or properties, intrinsic to their semantic definitional meanings, which allowed them to be conjoined, compared or matched. They also drew on a knowledge of language and were universal (i.e. cross-cultural). In contrast, pragmatic relations were based on an external co-occurrence of the concepts through perception or experience -- relying on worldly knowledge. Therefore, the researcher used WORDNET (Miller, 1990) as the basis for type of relation classification. WORDNET is a psychologically based lexical compendium, now distributed as software. Therefore, if in WORDNET the definition of A or B (the stem analogy terms) overlapped, in either direction (i.e. the definition of A referenced B or vice versa), then the A:B relation was deemed intensional; if there was no overlap, A:B's relation was considered pragmatic. If a word did not appear in WORDNET, the Oxford English Dictionary (Simpson & Weiner, 1989) was used as a substitute.

Stem Rationale Difficulty: The ETS specialists judged each

final stem rationale on how difficult/easy they thought it would be to formulate the given stem rationale. For each item, the ratings were averaged over both specialists to provide a concluding score. Before averaging, the ratings correlated r=.48. Note that the scale ranged from 1, very difficult, to 7, very easy. See p. 189 in Appendix A.

Alternative Choice Variables:

Context: The two ETS specialists made a yes/no judgment, determining which items were deemed "contextualized" or not. This was achieved through a number of steps. First, after viewing an item stem alone, a rationale was formulated. Then, after evaluating the entire item, the stem rationale was reassessed and if necessary revised. The ETS specialists provided a context judgment, based on their expertise, and having experienced the aforementioned process. Also, they decided on rationales for all alternative pairs, including the key. Further, all rationales were phrased in the form of declarative sentences. Any differences in scoring the context effect were resolved through discussion. Prior to consensus, the inter-rater agreement index, Kappa, was .07. See p. 187 and p. 188 in Appendix A.

Choice Similarity: Operationalizing this variable also required several steps. First, an expert graduate student propositionalized all final pair rationales, as provided by the ETS specialists, using Turner and Green's (1978)

propositional analysis technique. Next, the researcher counted the differences between the number of elements in the key and the number of elements common to the key in each alternative, across all alternatives. To standardize these values, each sum of differences was divided by the number of elements in the key. That is,

((No. of elements key - No. of common elements alt_1) + (No. of elements key - No. of common elements alt_2) + (No. of elements key - No. of common elements alt_3) + (No. of elements key - No. of common elements alt_4))/(No. of elements key) See p. 192 in Appendix A for an example. Therefore, the lower the score, the more similarity existed between the alternatives' and the key's rationales.

Structure-Mapping Variables:

Independent/Overlapping: Although these classifications were provided by ETS Test Development -- they were part of the ETS data set -- coding for this particular dichotomy has been notoriously unreliable. Therefore the researcher and a graduate student rater recoded all items on this dichotomy, using ETS's definition for independent/overlapping (Donlan, 1984; Bejar et al, 1991). Inter-rater agreement was Kappa=.45.

Number of Common Mapping Elements: The same technique used in operationalizing choice similarity was again implemented. In this case, however, the researcher counted the number of elements in the stem's propositional analysis, counted the

number of elements the key had in common with the stem, computed the difference (No. elements stem - No. of common elements), and to standardize across items, divided by the number of elements in the stem. Again, lower scores indicated greater similarity. See p. 192 in Appendix A.

Level of Mapping: The researcher used Gentner's representation conventions and the propositional maps for each stem rationale to code the level of relation. The representation conventions were described in Chapter 4. The item was categorized as being at either a zero-, first- or second-order level using the following rules:

zero-order was (isa, object, object)
pred(object, object)
pred(object, object(O:)

first-order was loc(pred, obj) S:(obj, O:)
pred(obj(obj), obj(obj))

second-order was pred(pred, pred)
pred(pred, obj), pred(obj, pred)

For example,

a zero-order level was: Candy has a sweet taste.
(Quality of: candy, (Quality of taste, sweet))

a first-order was: A student lives in a dormitory.
(Location: in(live, S:student), dormitory)

a second-order was: A briefcase is designed to carry papers.
(Purpose: design(isa, briefcase, receptacle),
(carry, O: papers))

Other Variables:

Syntactic Order: Given an item's stem pair and its final rationale, the graduate students judged if this particular final stem rationale required a reversal of the pair's word

order or not. See p. 193 in Appendix A.

ETS Taxonomies: Content and Abstraction classifications were provided by ETS Test Development -- they were part of the ETS data set. However, the abstraction classifications were recoded by the researcher and a graduate student rater (Kappa=.85) due to poor ETS coding reliability.

Initial Analyses

Descriptive Statistics:

Basic descriptive statistics were run: (1) frequencies were computed for background variables, as well as for the qualitative LLTM variables, I/P, semantic taxonomy, order level, independent/overlapping, context and syntactic order, and ETS taxonomic variables, (2) frequency distributions of items and people were plotted, (3) means and standard deviations of the quantitative scores, counts and ratings were calculated, (4) plots of empirical item characteristic curves (ICC's; The proportion correct was graphed for each ability level; Chapter 2.) and model predicted ICC's were considered, and so on. These preliminary analyses provided an initial "feel" of the data and helped to detect any coding or data entry errors, now corrected.

Testing of Model Assumptions

As noted in Chapter 2, a linear logistic test model (LLTM) approach demanded that the data satisfy several strict assumptions. This study addressed the concerns of dimensionality, local independence, and "best" fitting Item

Response Theory (IRT) model.

Dimensionality: A major LLTM assumption was that the data be unidimensional. That is, only one ability should explain test performance or, in other words, the test should measure only one construct. Therefore, a traditional one factor confirmatory analysis was run using TESTFACT (Wilson et al., 1987); all items were by default constrained to load on one factor, called analogical reasoning. Note that, TESTFACT is a program designed to handle item level data with dichotomous, right-wrong scoring[6]. This analysis was repeated for each group. Steiger's root mean square error of approximation (rmsea; Browne & Cudeck, 1993) was used to assess the fit of a one factor model.

Independence: Recall that this assumption was strongly tied to unidimensionality. In fact, if the data were to be unidimensional, the local independence of items assumption would be also implied; that is, after controlling for this unidimensional Θ, no relation would exist between any possible combination of item pairs. Thus, since tests of local independence were not easily available, if the data were unidimensional, it was then assumed that the independence assumption held.

Number of Item Parameters: Assuming unidimensionality, the data were run on BILOG to check for the "best" fitting model: the one, two or three parameter logistic IRT model. "Best" fitting meant using -2(ln likelihood) to find the

most parsimonious model; that is, the model that did not significantly worsen fit from the baseline three parameter model.

Exploratory Phase

The benefit of a large sample was the possibility of running first an exploratory phase and then following up with a confirmatory phase. The exploratory phase of LLTM was carried out on a random sub-sample, half the number of examinees, drawn from the total sample of examinees, across all 80 items. Once the final exploratory LLTM had been fit, this model was tested for replication on the remainder of the group during the confirmatory phase.

Hypothesized Linear Models: The F Matrices:

Cognitive Variables: After all variables were operationalized -- the categorical variables coded and quantitative variables left as they were -- they were all placed in F matrices, as mentioned in Chapter 2. LLTM software (Fischer, Formann & Wild, 1988) was then used to run the models in order to explain item difficulty.

Since the total number of variables, after coding, was quite large, all variables could not be entered into the model simultaneously. Therefore, three sub-LLTMs were run. Hypothesis 2, kind of relation, was first tested, requiring an F matrix, F1, with nine dummy variables using 1,0 coding and a default unit vector for an intercept constant. Further, it made sense to keep the semantic classes

together; the taxonomy as a whole was hypothesized to have little explanatory effect.

For the second run, the remaining cognitive variables, three vocabulary frequencies, Intensional/Pragmatic (I/P), stem rationale difficulty rating, context, alternative choice similarity and structure-mapping (overlapping/independent, number of common mapping elements and levels of relation), were entered into an F matrix, F2, and tested (Hypotheses 1, 3-6c). Recall that some measures were quantitative and others were qualitative:

Table 5.3

Coding of the Cognitive Variables for F2

Qualitative Variables:

a. I/P with 1,0 coding; 1=intensional, 0=pragmatic
b. context with 1,0 coding; 1=context, 0=no context
c. two dummy variables for level of relation with 1,0 coding, the zero order level will be omitted
d. independent/overlapping with 1,0 coding; 1=independent, 0=overlapping

Quantitative Variables:

a. three text frequency counts
b. an averaged stem rationale difficulty rating
c. alternative choice similarity and structure-mapping counts of common elements

---

Other Variables: Since they were not part of the cognitive IPA model, the third LLTM run, F3, independently studied the effects of syntactic order and the other ETS test development taxonomies on item difficulty (Hypotheses 7a and

7b).

Table 5.4

Coding of the Other Variables for F3

---

a. Syntactic Order with 1,0 coding; 1=reverse, 0=original
b. Three dummy variables for content; world of practical affairs omitted; science, human relations and aesthetic/philosophical 1,0 coding
c. Level of abstraction with two dummy variables; concrete omitted; mixed and abstract 1,0 coding

---

At this point, Hypotheses 1 through 7b had been considered. Since the three sub-LLTM's were not nested, Embretson and Wetzel's (1987) fit index was used to judge model fit. Further, likelihood ratios were constructed to ask if each sub-LLTM substantially worsens fit from the Rasch model. Both of these methods were discussed in Chapter 2. In addition, the meaning and significance of the component difficulties were discussed.

Confirmatory Phase

The Exploratory Phase culminated with an acceptable LLTM model and the importance of several variables, cognitive or otherwise, were verified or refuted. The Confirmatory Phase sought to replicate this final model on the remaining subsample of examinees (those left after random selection for the exploratory phase). The exploratory and confirmatory values of Embretson and Wetzel's fit index, which were analogous to R^2's in regression analyses, were compared for percent of variance

explained in item difficulty by the cognitive variables. The exploratory and confirmatory component difficulties were also compared.

[1] A detailed description of an SAT test was given in Chapter 3. That is, the test's structure and administration procedures and the analogies' format and instructions were presented.

[2] The November test-taking group is known to be quite stable over years.

[3] There is typically a 95% response rate.

[4] For example, on one hand, the analogy pair *engine:car* has a rationale "An engine is part of a car." -- the Part-Whole class completely describes this rationale while on the other hand, *clapper:bell*, with a final rationale "A clapper is a part of a bell which creates a sound when it strikes the sides of the bell.", incorporates more than one relation, Part-Whole and Cause-Purpose. However, estimation problems will be incurred if a design matrix of non mutually exclusive classes is provided. Therefore, the graduate students must chose the class that indicates a rationale's **predominant** relation.

[5] The graduate students criticized the given definitions of the semantic classes. Their feedback could be used to strengthen future work with the taxonomy.

[6] Recall that omitted responses were considered wrong and were scored accordingly.

Chapter 6: Results

LLTM Variables

As previously described, all items received a score for each variable. Frequency counts were then calculated for all qualitative variables asking: How many items fell into which categories? Most variables had a fairly well balanced number of items in each category. For example, 39 items were scored as having no context effect while 41 were categorized as having been contextualized by the alternatives. A notable exception was syntactic order -- 78.8% of the items retained the order of the stem word pair. This was not surprising given the well-defined format used by the ETS test development raters to formulate the rationales. In addition, ETS Taxonomic variables followed strict test development specifications. Thus, for level of abstraction, there were fewer concrete items than mixed and abstract; concrete items were in general too easy. Otherwise, only the Part-Whole class within the semantic taxonomy had relatively few items. Table 6.1 shows the frequency distributions for the qualitative variables.

Table 6.1

Frequency Counts of the Qualitative Variables

Variable	Count
I/P	
Intensional	44
Pragmatic	36
Contextualization	
No Context Effect	39
Context Effect	41
Level of Relation	
Level 0	20
Level 1	31
Level 2	29
Independent/Overlapping	
Independent	40
Overlapping	40
Semantic Taxonomy	
Attribute	6
Contrast	10
Representation	9
Similar	8
Class Inclusion	7
Cause - Purpose	11
Case Relations	9
Nonattribute	8
Part - Whole	3
Space - Time	9
Level of Abstraction	
Concrete	20
Mixed	32
Abstract	28
ETS Content Taxonomy	
Aesthetic/Philosophic	21
Human Relations	19
Science	21
Practical Affairs	19

As several of these variables required inter-rater consensus, indices of agreement were also computed. For the Context, Independent/Overlapping, Level of Abstraction, and Semantic Taxonomy variables the Kappa statistics (Fleiss, 1973) were .07, .45, .85, and, averaged over all possible pairs of the three raters, .61 respectively. Clearly, there was a problem with the context variable; with a Kappa of .07, there was essentially little agreement between raters before a forced consensus. The values for Level of Relation and the Semantic Taxonomy were acceptable.

Descriptive statistics were run on the quantitative variables. The means, standard deviations and ranges are listed in Table 6.2.

Table 6.2

Means, Standard Deviations and Ranges for the Quantitative Variables

Variable	Mean	Std.Dev.	Range
Stem Frequency	38.35	12.37	51.1
Key Frequency	39.27	12.23	54.2
Alt. Frequency	43.29	6.09	37.1
Stem Difficulty	4.52	1.29	6
Alt. Similarity	1.75	.73	3.18
No. Common Elements	.07	.14	.58

Large values were to be expected for the frequency variables -- some words were quite common and thus appeared frequently in text. A mean stem difficulty rating of 4.52 indicated that the ETS Raters felt the inference of the stem rationales were on average towards the easy end of the rating scale. In addition, the correlation between the two raters for this variable, another measure of inter-rater agreement, was $\underline{r}$=.48. The troublesome variable here was Number of Common Elements. Unfortunately, the form of the raters' stem and key rationales rarely differed, indicating an almost complete overlap in elements; thus, this variable had a mean close to 0 and a small standard deviation. Again, these raters were test development specialists and were accustomed to formulating the rationales in a common format.

A correlation matrix, depicting the relationship between the F2 variables, is shown in Table 6.3 -- neither the semantic classes nor the "other" variables strongly inter-correlated. As to be expected, the frequency variables did covary as did the level variables. Further, few variables were significantly correlated, though stem difficulty did have a moderate correlation with I/P, context and independent/overlapping as well as Level1 with stem/key. Nonetheless, a condition index was calculated to uncover possible multicollinearity problems (Dillon & Goldstein, 1984); there were none ($\delta_{max}/\delta_{min}$=17.99).

Table 6.3

Correlation Coefficients of the F2 Matrix

	I/P	CONTEXT	IND/OVER	STEMDIFF	STEMFREQ	KEYFREQ
I/P	1.0000					
CONTEXT	-.0729	1.0000				
IND/OVER	.1005	.0750	1.0000			
STEMDIFF	.3784*	-.3171*	-.2289*	1.0000		
STEMFREQ	.0121	.0467	-.0305	.1314	1.0000	
KEYFREQ	.0188	.2121	.0553	.0700	.2406*	1.0000
ALTFREQ	-.1454	-.0343	.0447	.0321	.3411*	.2645
STEM/KEY	.1643	.2178	.1404	.0067	-.0992	.0639
KEY/ALT	.0671	.0132	-.0134	.0667	-.0727	.0221
LEVEL1	.0542	-.0969	.0770	-.0116	-.0958	.0021
LEVEL2	.0497	.0072	-.0780	-.0009	.0061	-.0654
	ALTFREQ	STEM/KEY	KEY/ALT	LEVEL1	LEVEL2	
ALTFREQ	1.0000					
STEM/KEY	-.0925	1.0000				
KEY/ALT	.1066	-.0687	1.0000			
LEVEL1	.0403	.2535*	.0754	1.0000		
LEVEL2	-.1624	-.0537	-.0619	-.5998*	1.0000	

Note. I/P=Intensional/Pragmatic, CONTEXT=Contextualization, IND/OVER=Independent/Overlapping, STEMDIFF=Stem Rationale Difficulty, STEMFREQ=Stem Frequency, KEYFREQ=Key Frequency, ALTFREQ=Alternative Frequency, STEM/KEY=No. of Common Elements, KEY/ALT=Alternative Similarity, LEVEL1=First-Order Relation, LEVEL2=Second-Order Relation.

Model Assumptions

The Analogy Test

Although analogies comprised only a portion of SAT Verbal items, each group was treated as having taken a discrete analogy test, where the groups were composed of the entire sample, as described in Table 5.2; groups were linked through common items. TESTFACT provided some background classical item statistics for this test (Please see Table 6.4). Although groups did perform somewhat differently, that should not matter given the invariance of IRT parameter estimates. The reliability of each test also differed.

Table 6.4

Analogy Test Psychometric Statistics across the Four Groups

Group	Mean	Std.Dev.	Alpha	rmsea[a]	% Variance[b]
1	16.4	4.8	.78	.03	25.4%
2	15.3	4.3	.74	.03	21.6%
3	13.7	4.3	.74	.04	20.0%
4	14.2	5.2	.82	.04	26.9%

Note. [a]Root Mean Square Error of Approximation [b]Percent of Variance Explained

Dimensionality and Independence

TESTFACT was additionally used to confirm the LLTM unidimensionality assumption. As hoped, an item level, one

factor confirmatory analysis was supported for each group; Steiger's rmsea was .04 or less across the groups (Please see Table 6.10; rmsea values should be less than .08 for model fit (Browne & Cudeck, 1993).). Rmsea was the preferred fit statistic as it was not as affected by large n's, a problem with this data set. The total percent of variance explained across each group's respective items (Table 5.1) using a one factor model is additionally displayed in Table 6.10. Further, given these results, the local independence of items was considered upheld.

Number of Item Parameters

As a first approach, the four groups were equated and the various IRT models were fit using BILOG. In order to ascertain the "best" fitting model, likelihood quotients were evaluated comparing the more restrictive models to the least constrained three parameter with a variable $\underline{c}$ model, the baseline. Thus, the resulting χ^2 statistics were examined for significance after the differences between -2loglikelihoods were calculated. A significant χ^2 indicated that a more restrictive model, that is, a model requiring fewer item parameters, worsened the fit over the baseline model. The results are listed in Table 6.11. Using this criteria, the three parameter model with a variable $\underline{c}$ did fit "best"; the Rasch or one parameter model significantly worsened fit. This outcome, however, was expected as these χ^2's are influenced by sample size.

Table 6.5

Fit of Item Response Theory Models using Likelihood Ratios

Model	-2loglikelihood	Δ df	χ^2
1 Parameter	226,787.66	80	1578.81*
2 Parameter	225,208.85	1	341.83*
3 Parameter with constant c	224,867.02	79	206.51*
3 Parameter with variable c	224,660.50		

Note. * A significant χ^2

Despite this, there was still a good case for continuing with the Rasch model. A global χ^2 statistic provided by BIGSTEPS (Wright & Linacre, 1993), indicating fit of data to the Rasch measurement model, supported the fit of the Rasch model. Further, the correlation between the Rasch and the three parameter item difficulty estimates was r=.97. In addition, regressing the three parameter estimates onto the Rasch estimates explained 94% of the variance and there were no extreme residuals -- all standardized values were less than 3 and only six items had values larger than | 2 | .

Exploratory Phase

At this point the actual LLTM analyses were initiated. Despite the a priori hypotheses implicit in the F matrices, the purpose of the exploratory phase was to ascertain the

"best" fitting model. That is, several sub-models could be examined, modified and then accepted or rejected. The final exploratory model would then be verified or reconfirmed using the confirmatory sample. Therefore, subsets, halving each of the four groups, were randomly drawn in order to provide the exploratory and confirmatory samples. The N's are listed in Table 6.6.

Table 6.6

Exploratory and Confirmatory Sample Sizes

	Exploratory	Confirmatory
Group		
1	3942	3944
2	4059	4061
3	3462	3464
4	3987	3988

Then, as described in Chapter 5, three F matrices were written: F1 was composed of the semantic taxomonic classes (i.e. Hypothesis 2), F2 contained the remaining theoretic variables (i.e. Hypotheses 1, 3-6c), and F3 held the "other" variables (Hypotheses 7a and 7b). LLTM (Fischer, Formann & Wild, 1988) was the software used to complete the actual data runs; it has the capability of estimating the τ's for all 80 items across the four groups.

Initially, several analyses were undertaken. Embretson's fit index was computed for each F model (See Chapter 2 for more details; {-2((log likelihood null)-(log likelihood model))}/{-2((log likelihood null) - (log likelihood Rasch))} = $\chi^2_{model}/\chi^2_{Rasch}$). Additionally, likelihood ratios compared each F model to the baseline Rasch model asking if the LLTM constraints significantly worsened the fit from the Rasch model? Further, for all models, a $\underline{t}$ statistic was calculated for each component difficulty. The outcomes follow in Tables 6.7-6.9.

According to Embretson's Fit index (Please see Table 6.7), models F1 to F3 all fit quite acceptably with values of .28, .47 and .31 respectively; these values were comparable to those achieved and reported in past research (Embretson & Wetzel, 1987). It was thus immediately clear that the semantic taxonomy did to some extent explain item difficulty. Therefore, Hypothesis 2 was rejected, contrary to the a priori hypothesis. Although the component difficulties are listed in Table 6.9, they did not add substantive information in terms of Hypothesis 2.

In addition, the χ^2's for all likelihood ratios (Table 6.8) and all $\underline{t}$ statistics for model F2 (Table 6.9) were significant. This meant that the constraints imposed by each model significantly worsened fit from the Rasch model and that the component difficulties were significantly different than 0. However, these outcomes were considered

products of the large sample; there was statistical significance rather than practical significance. For example, the I/P component had a weight of .08, very close to 0, yet was significantly different than 0, according to the t statistic.

Therefore, variables were removed from F2 one at a time and the change in the fit index noted (Table 6.7; Recall that multicollinearity had not been a problem.). If the index dropped by .05 or more, the particular variable was considered an important contributor to the model. It quickly became apparent that several variables were unnecessary; that is, the fit index remained unchanged. For instance, when I/P was deleted from the F2 matrix, the fit index remained .47.

Table 6.7

Model Fit for the Three F Matrices

Model	Log Likelihood	df	χ^2	Fit
Null	-267,868.26			
Rasch	-173,355.41	79	189,025.7	
F1	-241,019.93	9	53,696.66	.28
F2	-223,139.62	11	89,457.28	.47
F2d	Model F2 with item frequencies removed			.35
F2e	Model F2 with I/P removed			.47
F2f	Model F2 with stem/key removed			.47
F2g	Model F2 with context and stemdiff removed			.36
F2i	Model F2 with stem freq removed			.47
F2j	Model F2 with level variables removed			.42
F2k	Model F2 with ind/over removed			.44
F2l	Model F2 with context removed			.47
F2m	Model F2 with stemdiff removed			.36
F2n	Model F2 with keyfreq removed			.41
F2o	Model F2 with altfreq removed			.45
F3	-238,135.65	6	59,465.22	.31

Table 6.8

Likelihood Ratios Comparing the F Models to the Rasch Model

Likelihood Ratios	df	χ^2
$LR_{F1\text{-}Rasch}$	70	135,329
$LR_{F2\text{-}Rasch}$	68	99,568
$LR_{F3\text{-}Rasch}$	73	129,560

Table 6.9

Component Difficulty Results for all Three F Matrices

F1 -- Hypothesis 2

Variable	Component Diff.	Std Err	t
Space-Time	-.25	.019	-13.66
Case Relation	-.54	.02	-27.66
Cause Purpose	.57	.018	32.22
Non Attribute	1.91	.019	99.76
Class Inclusion	.72	.019	38.58
Similar	1.61	.019	85.47
Representation	.32	.019	16.72
Contrast	1.22	.018	66.43
Attribute	1.04	.019	55.00

F2

Variable	Hypothesis	Component Diff.	Std Err	t
I/P	3	.079	.008	9.89
Context	5a	-.18	.008	-22.7
Lev 1	6c	.61	.009	65.21
Lev 2	6c	.90	.009	99.17
Ind/Over	6a	.55	.007	75.52
Stem Freq	1	-.004	.0003	10.67
Key Freq	1	-.03	.0003	104.1
Alt Freq	1	-.04	.0006	58.58
Stem Diff	4	-.48	.003	140.64
Stem/Key	6b	-.47	.03	17.65
Key/Alt	5b	-.12	.005	25.4

F3

Variable	Hypothesis	Component Diff.	Std Err	t
Syntactic Order	7a	.31	.008	39.97
Mixed	7b	1.00	.008	119.09
Abstract	7b	1.90	.01	198.21
Science	7b	-.38	.009	-41.21
Aest/Phil	7b	.05	.01	4.84_{ns}
Hum Rel	7b	-.002	.01	$-.2_{ns}$

Thus an approach analogous to backwards stepwise regression was initiated. One after another, variables were incrementally removed from the F2 matrix until the index had dropped by .1. The process is shown in Table 6.10. In total, seven variables needed to be dropped from the F2 model. The remaining four were Levels of Relation, Key Frequency, and Stem Difficulty. The same strategy was used for F3; ultimately, only the Level of Abstraction variable was needed. For another perspective, when only the four "important" variables were removed as a block from F2, the fit of F2 diminished to .18.

Finally, both the F2 and F3 models were rerun with only the "important" variables included. The resulting fit values for F2 and F3 were .41 and .29 respectively; the final component difficulties are listed in Table 6.11. The direction and thus the meaning of the retained component weights were as hypothesized: (1) as the level of relation incremented, the degree of item difficulty increased (Hypothesis 6c), (2) when stem rationale was rated as being easier to infer, item difficulty went down (The negative weight reflected the direction of the rating scale; Hypothesis 4), (3) as words in the key became more common, that is, as they appeared more frequently in text, the item's difficulty decreased (Hypothesis 1) and (4) as the level of abstraction increased, concrete to abstract, the measure of difficulty also increased (Hypothesis 7b).

Table 6.10

Results of Backwards Stepwise Elimination from F2 and F3

Step	Model	Log Likelihood	χ^2	Fit	Variables Removed
	F2	-223,139.62	89,457.28	.47	
(1)	F2f	-223,295.75	89,145.03	.47	stem/key
	F2f1	-223,710.31	88,315.91	.467	stem/key, context
	F2f2	-223,333.65	89,069.23	.47	stem/key, stemfreq
(2)	F2f3	-233,331.67	89,073.19	.47	stem/key, I/P
(3)	F2f4	-233,366.46	89,003.60	.47	stem/key, I/P, stemfreq
(4)	F2f5	-233,791.81	88,152.91	.466	stem/key, I/P, stemfreq, context
(5)	F2f6	-224,128.318	87,479.89	.463	stem/key, I/P, stemfreq, context, key/alt
	F2f7	-226,900.56	81,935.41	.43	stem/key, I/P, stemfreq, context, ind/over
(6)	F2f8	-227,241.40	81,253.73	.429	stem/key, I/P, stemfreq, context, key/alt, ind/over
	F2f9	-232,963.76	69,809.01	.369	stem/key, I/P, stemfreq, context, key/alt, ind/over, level
	F2f0	-233,818.36	68,099.80	.36	stem/key, I/P, stemfreq, context, key/alt, ind/over, key freq
(7)	F2fa	-229,473.07	76,790.38	.41	stem/key, I/P, stemfreq, context, key/alt, ind/over, alt freq
	F3	-238,135.65	59,465.22	.31	
	F31	-240,573.74	54,589.04	.289	content, syntactic order

Table 6.11

Final Exploratory Component Difficulties

Variable	Component Diff.	Std. Error	t
Model			
F2			
Level 1	.6622	.0086	76.74
Level 2	.9818	.0086	113.95
Key Frequency	-.0395	.0003	-133.33
Stem Difficulty	-.4922	.0028	-175.00
F3			
Mixed	1.0094	.0086	117.75
Abstract	1.9408	.0089	218.36

An examination of the component weights, similar to unstandardized regression coefficients, needed to be put in perspective. For example, note that the weights of the frequency variables were small relative to the other weights, probably because the range and size of values for the frequencies were relatively so much larger. The key issue was, which variables were relatively more important? To examine this, each remaining variable was removed one at a time from the final F2, and the difference in the fit index computed, analogous to a semi-partial correlation. If the difference was large, the variable was relatively more important. These results are reported in Table 6.12.

Table 6.12

The Relative Importance of the F2 Variables

Model	Log Likelihood	χ^2	Fit	Removed
Null	-267,868.26			
Rasch	-173,355.41	189,025.7		
2F8	-229,473.07	76,790.38	.4062	
2F81	-236,247.97	63,240.58	.3346	Level
2F82	-239,196.32	57,343.88	.303	Key Freq
2F83	-246,964.8	41,806.92	.2212	Stem Diff

The differences in fit were .072, .103 and .185 for Level of Relation, Vocabulary and Stem Rationale Difficulty respectively. Therefore, the rank order of "importance" was Stem Difficulty, Vocabulary and Level of Relation.

As a last exploratory step, the differences between the LLTM predicted difficulty values and the Rasch item difficulties were examined -- the actual values are printed in Table 6.13. Comparing the difficulties across models can give another, perhaps more intuitive sense of how discrepant the LLTM's specifications were from the baseline Rasch model. Further, the Rasch and LLTM predicted difficulty values were correlated for each F matrix; they were $\underline{r}_{Rasch\text{-}F1}=.54$, $\underline{r}_{Rasch\text{-}F2}=.63$ and $\underline{r}_{Rasch\text{-}F3}=.60$.

Table 6.13

Rasch Model and LLTM Predicted Difficulty Values

Item	ExpRASCH	ExploreF1	ExploreF2	ExploreF3	ConRASCH	ConfirmF1	ConfirmF2	ConfirmF3
1	-2.52	-.92	-1.36	-1.08	-2.50	-.91	-1.34	-1.08
2	-2.13	-.92	-.94	-1.08	-2.06	-.91	-.92	-1.08
3	-2.13	-1.21	-.08	-1.08	-2.19	-1.21	-.08	-1.08
4	-1.48	-.35	-.57	-1.08	-1.47	-.35	-.56	-1.08
5	-.95	-.35	-.55	-.07	-1.01	-.35	-.54	-.08
6	.56	.94	-.27	.86	.48	.94	-.26	.86
7	.79	-.10	.60	-.07	.74	-.11	.60	-.08
8	1.34	1.24	-.56	.86	1.33	1.24	-.55	.86
9	1.42	.55	.33	.86	1.42	.56	.33	.86
10	2.19	.05	.31	.86	2.07	.40	.32	.86
11	-1.75	-.92	-.73	-1.08	-1.81	-.91	-.72	-1.08
12	-1.80	-.10	-1.12	-.07	-1.93	-.11	-1.10	-.08
13	-1.44	-.92	-.17	-.07	-1.55	-.91	-.16	-.08
14	-1.28	-1.21	.59	-1.08	-1.34	-1.21	.59	-1.08
15	-.28	.94	-.05	.86	-.29	.94	-.04	.86
16	.34	.37	-.74	-.07	.29	.35	-.73	-.08
17	1.52	.55	-.33	.86	1.55	.56	-.32	.86
18	1.25	.05	-.17	-.07	1.21	.04	-.16	-.08
19	1.14	.94	-.01	.86	1.25	.94	-.05	.86
20	1.46	1.24	1.73	.86	1.46	1.24	1.72	.86
21	-2.86	.37	-1.53	-.07	-2.80	.35	-1.53	-.08
22	-1.62	-.35	-.65	.86	-1.57	-.35	-.66	.86
23	-2.26	.55	-.39	-.07	-2.21	.56	-.38	-.08
24	-1.47	-1.21	-.08	-.07	-1.44	-1.21	-.10	-.08
25	-1.76	-.92	-.11	-1.08	-1.67	-.91	-.10	-1.08
26	.92	-.67	.80	-.07	.86	-.70	.78	-.08
27	1.13	.94	.51	-.07	1.09	.94	.50	-.08
28	2.09	1.24	1.41	.86	2.17	1.24	1.40	.86
29	1.57	.55	1.41	.86	1.58	.56	1.40	.86
30	2.30	-.10	1.74	-.07	2.23	-.11	1.72	-.08
31	-2.55	-1.21	-1.83	-1.08	-2.53	-1.21	-1.82	-1.08
32	-1.81	-.10	-1.66	-.07	-1.82	-.11	-1.65	-.08
33	-.79	.37	-.41	.86	-.80	.35	-.40	.86
34	-.90	-.35	-.35	-.07	-.82	-.35	-.34	-.08
35	.28	.05	-1.43	-1.08	.26	.04	-1.43	-1.08
36	.59	.05	-.57	.86	.59	.04	-.58	.86
37	1.20	-.92	1.49	-1.08	1.23	-.91	1.47	-1.08
38	1.50	1.24	-.07	.86	1.58	1.24	-.07	.86
39	2.25	.37	.61	.86	2.31	.35	.61	.86
40	1.77	.55	.50	.86	1.79	.56	.50	.86
41	-3.13	-.67	.03	-1.08	-3.20	-.70	.02	-1.08

Table 6.13 Continued

Item	ExpRASCH	ExploreF1	ExploreF2	ExploreF3	ConRASCH	ConfirmF1	ConfirmF2	ConfirmF3
42	-1.76	-.92	-1.47	-1.08	-1.72	-.91	-1.44	-1.08
43	-2.06	-.92	-.19	-1.08	-2.01	-.91	-.18	-1.08
44	-1.83	.05	-.46	.86	-1.80	.04	-.46	.86
45	.50	1.24	1.22	-.07	.48	1.24	1.21	-.08
46	.96	-.10	.29	-.07	.97	-.11	.29	-.08
47	.63	.05	-1.69	-1.08	.70	.04	-1.69	-1.08
48	1.41	.94	.82	.86	1.49	.94	.82	.86
49	2.33	1.24	.86	.86	2.33	1.24	.85	.86
50	2.98	-.10	.69	.86	2.97	-.11	.68	.86
51	-2.84	-.10	-1.37	-1.08	-2.79	-.11	-1.37	-1.08
52	-1.97	-.10	-.23	-1.08	-2.02	-.11	-.22	-1.08
53	.11	-.92	-1.20	-1.08	.13	-.91	-1.18	-1.08
54	.98	1.24	.37	.86	.99	1.24	.38	.86
55	.62	.94	-.05	-1.08	.63	.94	-.04	-1.08
56	1.72	.37	1.66	.86	1.70	.35	1.64	.86
57	1.55	.94	.31	.86	1.55	.94	.31	.86
58	2.10	.55	-.08	-.07	2.06	.56	-.09	-.08
59	1.82	.94	1.91	-.07	1.79	.94	1.88	-.08
60	2.44	.37	.34	-.07	2.30	.35	.32	-.08
61	-2.52	.05	-1.07	-1.08	-2.51	.04	-1.07	-1.08
62	-1.51	-1.21	-.70	-.07	-1.48	-1.21	-.71	-.08
63	-1.37	-.67	-1.55	-.07	-1.32	-.70	-1.55	-.08
64	-.68	.55	.51	-.07	-.53	.56	.51	-.08
65	.01	-.35	.87	-.07	-.01	-.35	.86	-.08
66	.55	-1.21	-.07	-.07	.59	-1.21	-.07	-.08
67	.80	-.10	-.01	-.07	.86	-.11	.00	-.08
68	1.27	-.10	2.47	-.07	1.30	-.11	2.45	-.08
69	1.38	.55	1.10	.86	1.35	.56	1.09	.86
70	2.12	-.35	.64	.86	2.04	-.35	.64	.86
71	-2.24	-1.21	-.66	-.07	-2.30	-1.21	-.64	-.08
72	-2.06	-.10	.66	-.07	-2.12	-.11	.66	-.08
73	-1.51	-.35	-1.17	-.07	-1.38	-.35	-1.16	-.08
74	-1.47	-1.21	-.50	-1.08	-1.44	-1.21	-.51	-1.08
75	-1.10	.55	.60	-.07	-1.07	.56	.60	-.08
76	.96	.55	-.07	-.07	.94	.56	-.09	-.08
77	.50	-1.21	.60	.86	.56	-1.21	.60	.86
78	1.12	1.24	.50	.86	1.10	1.24	.50	.86
79	1.63	-.35	.43	.86	1.57	-.35	.44	.86
80	1.68	-.35	.35	-.07	1.66	-.35	.33	-.08

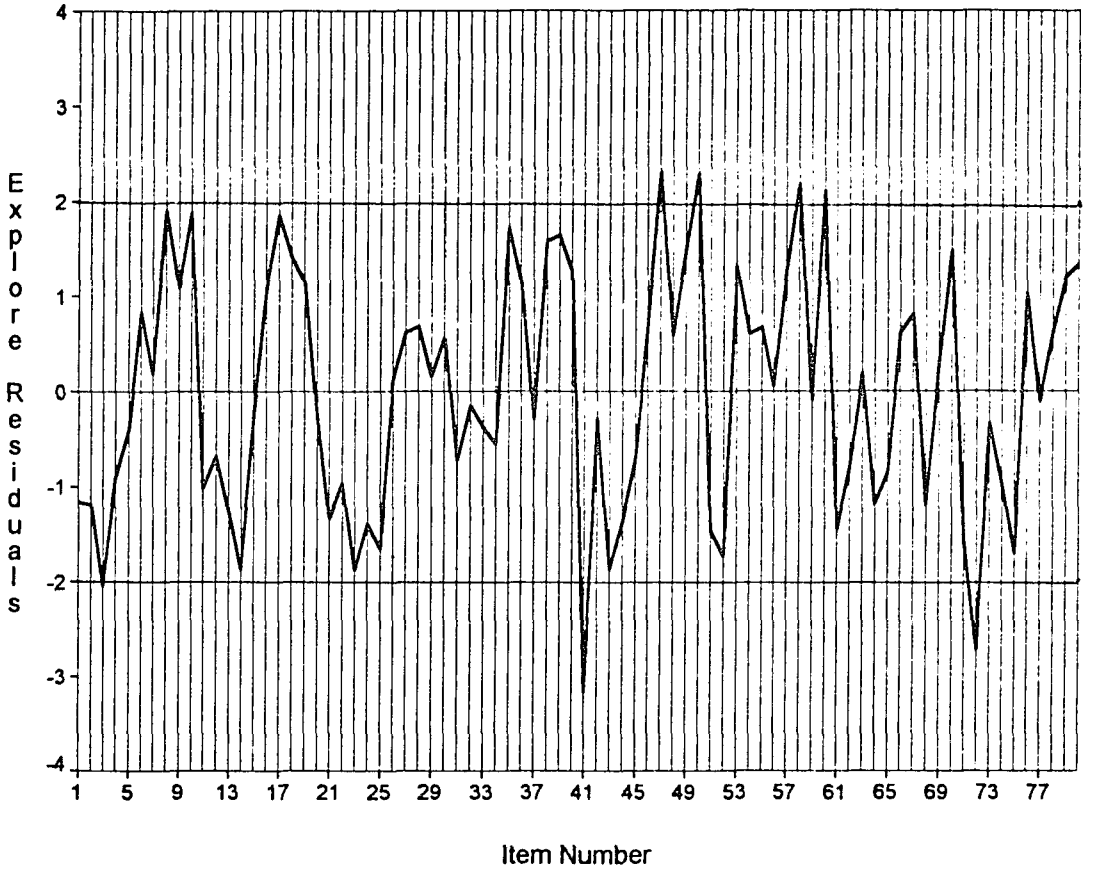
In addition, the residuals, Rasch difficulty minus LLTM predicted difficulty, were plotted for F2 (See Figure 6.1). Six items had values greater than | 2.1 | . Perusing the actual text of these items, however, was informative. For example, item 3 (OPHTHALMOLOGIST:EYE::CARDIOLOGIST:HEART) was empirically an easy item. However, according to the codings, it should have been a difficult item: ophthalmologist rarely appears in text and it had a complex rationale to propositionalize -- it was categorized as a Level2 relation. Here text frequency was a misleading indicator of everyday kinds of knowledge. Examining the actual codings, from the F matrix, offered a few other possible reasons for the discrepancies: (1) of the three difficult items, two were coded as Level 0 relations (Items 47 and 58) and all received middle to easy Stem Rationale Difficulty ratings (Items 47, 50 and 58) and (2) two of the easy items were coded as Level 2 relations (Items 3 and 72) and all three items had key pairs composed of uncommon words (3, 41 and 72).

Confirmatory Phase

The exploratory phase resulted in a set of final F models: (1) F1, the semantic taxonomy, remained the same, (2) four F2 variables prevailed, Levels of Relation, Key Frequency and Stem Rationale Difficulty, and (3) Level of Abstraction was the sole survivor in F3. Therefore, these LLTM models were fit on the confirmatory data set -- the

Figure 6.1

Exploratory Plot



balance of the original sample. The culminating fit indices were virtually identical with the exploratory results (See Table 6.14).

Table 6.14

Final Exploratory and Confirmatory Fit Indices

Model	Explore	Confirm	Difference
F1	.2841	.2862	-.0021
F2	.4062	.4049	.0013
F3	.2888	.2921	-.0033

As shown in Table 6.15, component difficulties were very similar across the two phases. The largest difference between components was -.014 for Mixed.

Table 6.15

Final Component Difficulties for F2 and F3

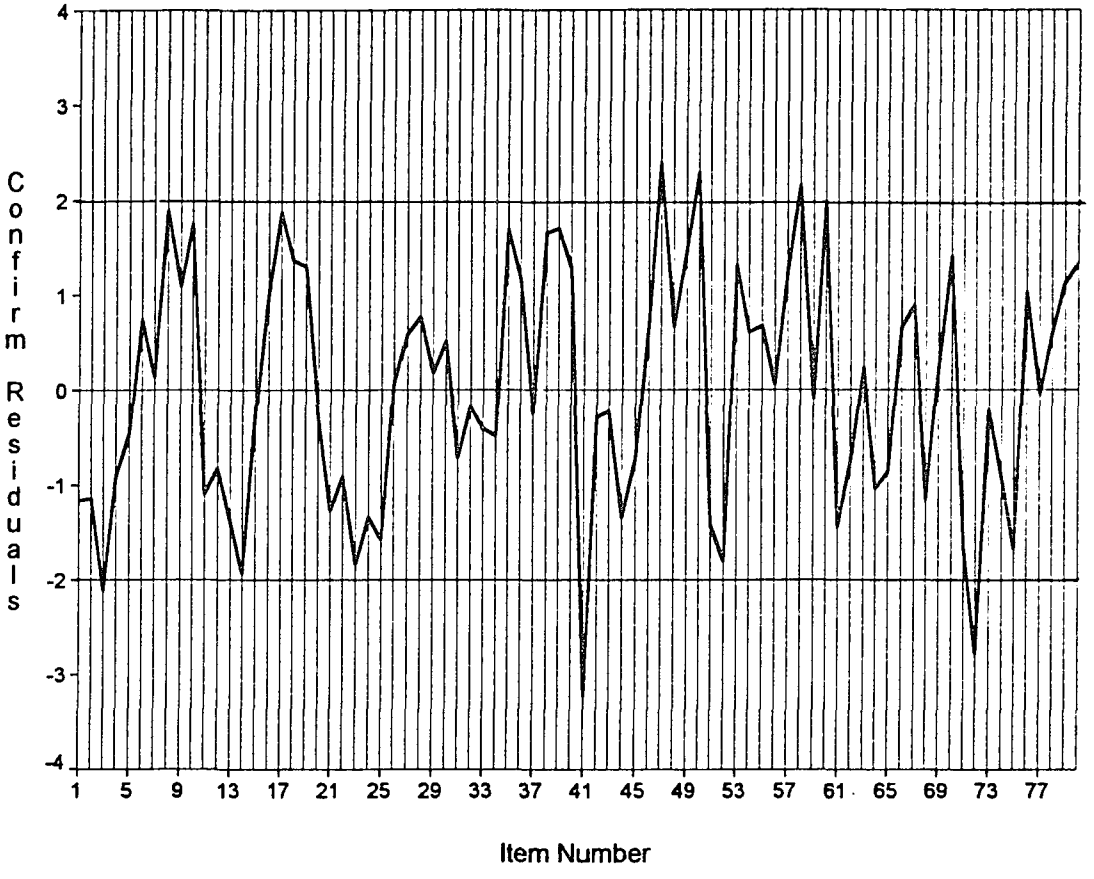
Model	Variable	Explore	Confirm	Difference
F2				
	Level1	.6622	.6742	-.012
	Level2	.9818	.9921	-.0103
	Key Freq	-.0395	-.0390	-.0005
	Stem Diff	-.4922	-.4866	-.0056
F3				
	Mixed	1.0094	.9956	.0138
	Abstract	1.9408	1.9427	-.0019

Likewise, the correlations between LLTM predicted item difficulties and Rasch difficulties, for each F matrix, were .55, .63 and .60 respectively; these values almost exactly coincided with the exploratory correlations. Further, the same six items were flagged when the Rasch minus LLTM residuals were plotted (Figure 6.2). In this study, the high degree of replication between the exploratory and confirmatory phases highlights the benefits of a large sample size.

The consequences of these results will be discussed in the next chapter.

Figure 6.2

Confirmatory Plot



Chapter 7: Discussion and Implications

The goal of this study was to integrate cognitive and psychometric theory. On the psychometric side, a specialized Item Response Theory (IRT) model, the linear logistic test model (LLTM) was implemented. The characteristics of IRT models, especially LLTM, were described; their many strengths were discussed, as well as the difficulties inherent in the LLTM's stringent assumptions and constraints. The F matrix, embedded within LLTM, provided entry for the cognitive side. Thus, the cognitive literature on analogical reasoning and problem-solving was exhaustively surveyed, important variables were noted and an appropriate Information Processing Approach (IPA) model was constructed. The most compelling variables were selected, linked to the IPA model and operationalized. Hence, the Rasch item difficulty parameter could be parsed into components reflecting this IPA model of analogy problem-solving. Given the theoretical nature of the cognitive components, research hypotheses were formulated and tested.

Results

Initial analyses entailed verifying adherence to the model's strict assumptions. Despite the paucity of research and methods in the domain of model fit, support was found for unidimensionality, and by implication independence. Further, a case was made for conformity of the data to the

Rasch Model. These outcomes were essential: the power of a Rasch model was gained, interpretations were simplified and implications for construct validity clarified. Consequently, for this study, a fairly simple testing model could be implemented.

Next, three cognitive models, the F matrices, were tested for fit. The proportion of variance in the baseline Rasch Model explained by using a LLTM was calculated using Embretson's fit index (Embretson & Wetzel, 1987). This fit index was useful in comparing non-nested models, in remaining unaffected by sample size and by functioning analogously to an R^2, thereby enabling analyses similar to backwards regression and semi-partial correlation.

Surprisingly, the taxonomy of semantic relations was supported. In addition, models F2 and F3 also fit reasonably, but several variables could be dropped without significantly reducing the fit index. Thus, in total five variables were important in explaining item difficulty. There were four theoretical variables: the first was a structure-mapping variable, Level of Relation; the others were Semantic Relations, Stem Rationale Difficulty, and Vocabulary Knowledge of the key analogy pair. The last was an Educational Testing Service (ETS) taxonomic variable, Level of Abstraction.

Of the theoretical variables, two semantic memory variables were retained. The first, a taxonomy of semantic

relations, did, as a whole, explain the difficulty of SAT analogy items. This outcome was unexpected, even given a history of conflicting results (See Chapter 4.). Recall that the taxonomy was empirically based and reflected a limited set of basic relations -- frequently accessed, prototypical kinds of semantic relations. It seems, however, that these SAT analogies did measure basic relations and despite the familiarity of the relations, the classes of relations also differentially predicted item difficulty (i.e. the component weights were different as shown in Table 6.14). There were therefore differences in processing different kinds of relations. To return to Bejar et al. (1991, p. 56), "This suggests the possibility that different relations may be represented and processed differently." Further, these results may reaffirm the importance of these relations in portraying memory structures (Chaffin & Pierce, 1987; Rumelhart, 1989). Here, psychometrics has provided feedback to cognitive theory.

Vocabulary knowledge of the least frequent word in the key word pair, as represented by Key Text Frequency, was the other semantic variable. Even though the component weight for this text frequency was small (-.04), this was a reflection of the comparatively large values and variances associated with text frequencies relative to the values of the other model variables. Dropping vocabulary from the final F2 showed that it was indeed an important variable.

Thus, a part of analogy problem-solving relies on previously acquired knowledge bases, or crystallized intelligence. Not knowing word meanings or not being able to retrieve words from Long Term Memory (LTM) could hamper performance on SAT analogy items. As reiterated by the College Board, being a well-prepared student remains an important way to prepare for SAT analogies. Intriguingly, the vocabulary of the stem and alternative pairs was not predictive of item difficulty. Or, alternatively, why was the vocabulary of the key word pair so singularly important? One reason for this could be the mutual intercorrelations between all the text frequency variables. Yet, this is unlikely -- multicollinearity had not been a problem.

Two working memory (WM) processing variables, Stem Rationale Difficulty and the Level of Relation part of structure-mapping remained in the model. WM variables were critical in that they embodied the reasoning part of analogy problem-solving. Stem Rationale Difficulty was the most "important" variable -- again, the fit of the final model dropped substantially with this variable omitted. Hence, an item's difficulty increased as the difficulty of inferring the stem's rationale increased. However, the operationalization of this variable relied upon the ratings of two expert test developers. In one sense, the relevance of this variable affirms the expertise, knowledge and intuitions of the item writers, a replication of Bejar et

al's findings (1991). Notwithstanding, this variable needs to be "unpacked". What is it that these raters are evaluating? Perhaps protocol analyses would be an appropriate component of future work. As mentioned in Chapter 4, this variable has constantly appeared in the literature, but operationalizations have been notoriously intangible.

Note that Level of Relation was based on Gentner's principle of systematicity. Gentner (1983, 1989) had described several possible levels for structure-mapping to occur between base and target in a Problem Approach kind of problem (e.g. the fortress problem). For realistic kinds of problems, higher-order inferences were psychologically preferred. SAT rationales, however, did occur at all levels, from superficial attributes (Level 0) to the higher-order relations (Level 2). In effect, examinees were required by the task demands, to map between stem and alternatives at all levels; given the nature of the task, theory could not entirely be supported. Nevertheless, as the processing demands increased, the difficulty of the items likewise increased, offering in fact some empirical support for Gentner's theory. It is theoretically meaningful that this variable was supported -- although, of the retained variables, it was the least influential. Other implications will be discussed below.

Out of the "other" variables, Level of Abstraction was

significant in explaining item difficulty. While it makes intuitive sense that concrete items would be easier than abstract, how would this variable fit into the IPA model? Perhaps this variable was a partial manifestation of certain aspects of the Intensional/Pragmatic (I/P) distinction -- concrete relations being pragmatic or perceptually sensed and abstract relations reflecting decontextualized concepts. A mixed relation would include aspects of both types of relations, offering a more realistic world view; in reality, dichotomies rarely exist.

Unfortunately, several theoretic variables did not contribute to the explanation of item difficulty. Most of the problems with these variables were a theoretic. For example, all alternative choice variables could be improved. Independent/overlapping has always been a troublesome variable to categorize. In addition, there was virtually no consensus between raters for the context variable (Kappa=.07). Further, using the ETS raters in the operationalization of alternative similarity was a mistake. In forming rationales, the raters used a common format across all pairs and all items. Consequently, "The first lives in the second" was the form for STUDENT:DORMITORY's rationale. Therefore, when the values for alternative similarity were actually calculated, the differences in number of common elements between the key and the alternatives were tremendously reduced. This was also true

for the structure-mapping variable, number of common elements between the stem and key relation rationales. A better way to have operationalized these variables would have been to enlist high school students instead of ETS test developers. Researchers should persist in trying to operationalize this variable using propositionalization (See Chapter 4).

As mentioned above, the semantic variable I/P may also have had problems with operationalization. While using WORDNET, a psychologically oriented lexical compendium, was less of a qualitative methodology than the ratings previously employed (Diones et al, in preparation) and more independent of the semantic taxonomy (Bejar et al, 1991), the method actually implemented may still have not been reflective of the I/P distinction. Another procedure needs to evolve conjoining quantitative and qualitative methodologies, which could be augmented by the Level of Abstraction taxonomy. Interestingly, however, if as in the past the semantic taxonomy had been used to categorize I/P, the intensional classes would have been ranked the five most difficult, perhaps supporting Bejar et al's (1991) previous contentions (See also Chapter 4.). An exploration of possibilities will be examined in future work.

Given that several theoretical variables were not supported, that .72, .59 and .71 percent of variance remained unexplained for the three LLTM's (1-Embretson's Fit

Index) and that there were only moderate correlations (.54 to .65) between the Rasch difficulty estimates and LLTM predicted difficulty estimates, the need for additional work in explaining analogy item difficulty is clearly indicated. A complete portrait of the sources of difficulty continues to be elusive.

Implications

Construct Validity

As demonstrated in Chapter 2, construct validation, also termed construct representation (Embretson, 1983) and internal validation (Sternberg, 1977a), or verifying that the test measures the intended construct, is a major goal of LLTM research. Thus, since the unidimensionality assumption was upheld, only one construct was considered. Next, recall that LLTM's combine person and component item characteristics to determine the probability of a correct response. For this study, the person ability, θ, or measured construct, was analogical reasoning or inductive reasoning behaviors utilizing selective encoding and selective comparisons with the intention of inferring relations (Sternberg, 1986), reasoning with them and making relational similarity judgments (Goswami, 1991). Further, as discussed in Chapter 1, the tie between cognitive and psychometric research relies upon item difficulty. Hence, once a processing model has been extracted from the cognitive literature, the F matrices specify the proposed

theoretical mechanisms underlying task performance or the task structure (i.e. the cognitive side) in terms of item characteristics (Fischer, 1973; Embretson, 1985). In effect, the coded variables represent the items' stimulus characteristics, which in turn determine each component's processing demand (Embretson, 1985; Embretson & Wetzel, 1987; Whitely & Schneider, 1981); more difficult stimulus characteristics catalyze greater processing demands. Additionally, returning full circle, more difficult items demand more of the construct (Stenner, Smith and Burdick, 1983). So finally, Θ could be interpreted in terms of the LLTM outcomes (Mislevy, 1981).

This of course relates back to the notion of item difficulty defining the measured variable (Figure 2.4). As discussed, the β's and Θ's lie on the same scale/continuum. In the case of LLTM, the component difficulties determine each item's predicted β estimate. So for the confirmatory group, the F2 prediction equation was

$$\beta^{*}=\text{intercept}+.67f_{\text{level1}}+.99_{\text{level2}}-.04_{\text{keyfreq}}-.49\text{stemdiff}$$

F2 specified each item's characteristics or the actual values for the variables. Thus, an examinee's place on the ability continuum (defined as the Θ value where the examinee would have a 50% probability of correctly answering item i), matching the predicted β, says something about the meaning of this examinee's ability in terms of a particular set of

known components.

Further, though unidimensionality was supported, as Scheuneman and Steinhaus (1987) pointed out, a data set that is quantitatively singular need not be theoretically singular. Indeed, analogical reasoning incorporated a constellation of activities between the LTM and WM processing requirements. Consequently, for LTM, vocabulary knowledge and semantic relations comprised a part of analogical reasoning ability as did the WM variables, Stem Rationale Difficulty and Level of Relation. Stem Rationale Difficulty expert ratings of the difficulty of inferring the stem rationale. Level of Relation comprised a part of structure-mapping activities between the stem (base) and the alternatives (target). Accessing and referring to prototypical semantic relations facilitated aspects of the problem-solving processes. Vocabulary knowledge also contributed to item difficulty. As always construct validation is an ongoing process. This study was part of this effort and did successfully identify some meaningful components of the IPA model.

The Processing Model Revisited

Yet, how do these results affect the IPA model (Figure 3.7)? Since facets of four of the six original theoretic variables linked to the model were empirically supported, they will remain as they were in the model. Additionally, since the problems with the semantic I/P dichotomy, the

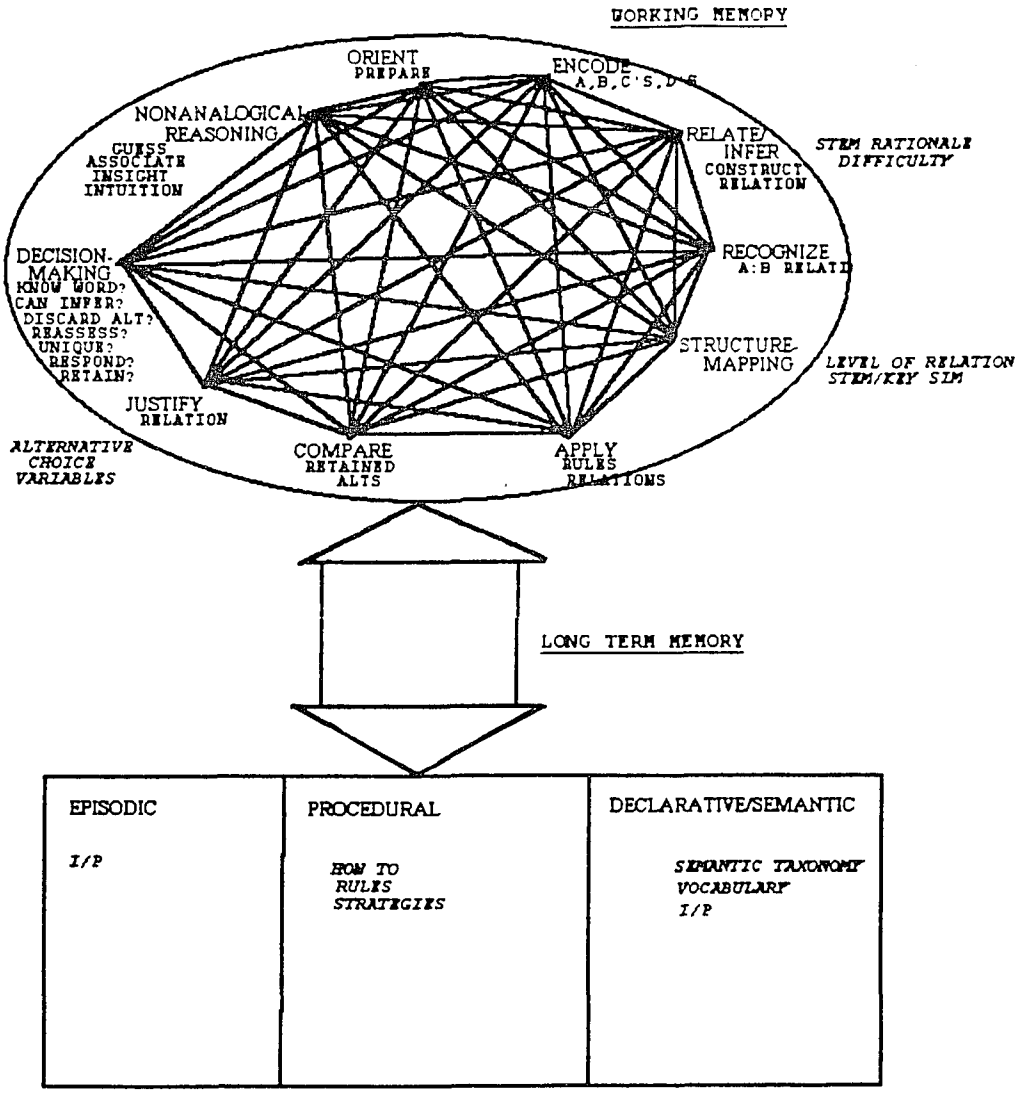
structure-mapping's key/stem similarity and independent/overlapping and the alternative choice variables mainly involved issues of operationalization, these variables will also be retained, especially given their strong emphasis in the literature, but only after the above-mentioned methodological issues have been addressed. Thus, the original model has essentially been preserved. This was, however, to be expected. An LLTM study can not impact the actual structure of a model -- only an experimental cognitive-processing study can do that.

Further, as more was learned about analogy problem-solving, the processing model became more cumbersome and complex, making the IPA model difficult to read and interpret (See Chapter 3). At the same time, while a LLTM study did need to reflect psychology's most current thinking, the processing model did not need to include every detail. In fact, a simpler, heuristic model may be even more useful. Added to this, was the need to incorporate a connectionist portrayal of memory.

Therefore, in reconsidering this study, a different version of the model emerged (See Figure 7.1). Note that all the processing components have been retained, but they are now circularly oriented. The lines between the components reflect the nonlinear, connectionist nature of analogy problem-solving as the examinee jumps from component to component and from WM to LTM, depending on the problem's

Figure 7.1

A Reconceptualized Processing Model



demands. Additionally, if each component is conceptualized as a spoke in a wheel, then each spoke can be represented in a LLTM. At present, only a few of the spokes are represented (See italicized words in Figure 7.1). Thus, a researcher glancing at this model can easily identify areas for future development. Further, the ties between Level of Relation, prototypical semantic relations and I/P relations should be examined. Also, the nature of Level of Abstraction's role in the model should be further investigated.

Psychometric Concerns

Ultimately, however, LLTM outcomes should return full circle to re-inform test development. For instance, in test construction, writers could use LLTM derived information to write items of varying difficulties. This would be especially useful in writing difficult items that discriminate -- always a challenging task. In addition, residual analyses, similar to those in Chapter 6, could help identify poorly constructed items or outliers. Also, LLTM components could provide a basis for algorithms needed to produce a computerized adaptive test item bank. Furthermore, Mislevy's (1981) extension of LLTM could be used to make a cognitive diagnosis of group differences (i.e. differential item functioning). Moreover, using LLTM components to equate tests, while not a novel idea, should be attempted. Lastly, as analogy problem-solving is further

decomposed and understood, students could be directly taught the thinking skills needed for successful analogical reasoning (e.g. Sternberg, 1979a).

However, LLTM outcomes could additionally be used to qualitatively alter the nature of the analogy task. For example, attention should turn towards levels of relations, focusing particularly on higher level relations. Perhaps more analogies like NUCLEUS:ELECTRONS::SUN:PLANETS, with an over-arching analogy ATOM:SOLAR SYSTEM, can be written. Or, different item format approaches could be investigated. For example, the test could require students to provide a word pair analogous to the stem or write out each item's actual stem rationale.

Conclusions

An LLTM study was successfully pursued through its many steps. Thus the current cognitive theory on analogical reasoning for A:B::?:? problems was integrated with a simple IRT model to test relevant hypotheses. Several variables were supported; others were not. Both events were informative. First, cognitive theory was informed. The importance of semantic relations was reaffirmed. A structure-mapping variable, Level of Relation, was advanced. A stem rationale rated as difficult to infer created a difficult item. Second, several variables need to be further refined and developed. Additionally, future cognitive research can further develop the processing model

towards a better, deeper understanding of analogy problem-solving.

Rater _____
Item# _____

Appendix A

The Decontextualized Stem Rationale

Analogy Stem: _____:_____

Please provide an as-complete-as-possible relation rationale for this analogy stem. Your rationale should be in the form of a declarative sentence .

_____
_____
_____

ETS Specialists

Rater _____
Item# _____

Full Item Rationales

Total Analogy: Stem: _____ : _____

A. : _____
_____

B. : _____
_____

C. : _____
_____

D. : _____
_____

E. : _____
_____

1. If necessary, revise your stem rationale.

_____
_____
_____

2. Please write in the space next to each alternative an as-complete-as-possible relation rationale, in the form of declarative sentences.

3. Circle the letter of the alternative you would have selected as the key.

Context Effect

4. Would you judge that this rationale is contextualized by the stem? That is, does the final stem rationale depend on observing and considering the choice alternatives?

yes____
no ____

ETS Specialists

Rater _____
Item# _____

Consensus on Rationales and Context Effect

Further, after your independent appraisal of 1, 2 and 4, if your answer differs from your colleague's answer, please resolve any differences through discussion. Indicate any changes below:

stem rationale for :

_____
_____
_____

alternative rationales:

A. : _____
_____

B. : _____
_____

C. : _____
_____

D. : _____
_____

E. : _____
_____

context effect:

yes____
no ____

ETS Specialists

Rater _____
Item# _____

Stem Rationale Difficulty

Given this final stem rationale, _____
_____,
please rate the following statement by circling the number most representative of the statement:

How difficult was it or would it have been to formulate this stem rationale.

1	2	3	4	5	6	7
very difficult			average difficulty			very easy

ETS Specialists

Rater _____
Item# _____

Semantic Taxonomy

All stem item pairs need to be classified into the mutually exclusive semantic taxonomy shown below. However, it may be difficult to classify some items into only one class. In that case, check off the item rationale's **predominant** kind of relation. (Note that these classes were defined by Bejar et al. (1991, p. 58).)

Thus, given this final stem rationale, _____
_____
_____
_____,
categorize it into one of the following classes by putting a check in the appropriate box. As an example, the analogy pair *engine:car* has the rationale "An engine is part of a car.". Therefore, a check mark should be placed in the Part-Whole Check Off box (see Table A.1).

Table A.1

The Semantic Taxonomy Check Off Form

Relation	Check Off	Rationale
Class Inclusion		A is a member of the class B
Part-Whole		A is a part of B
Similar		A is a more intense B
Contrast		A is the opposite of B
Attribute		A is an attribute of B
Non-Attribute		A is not an attribute of B
Case Relation		A works on B
Cause/Purpose		A is the cause of B
Space/Time		A can be found in B B occurs during A
Representation		A is representation of B

Rater _____
Item# _____

Semantic Taxonomy Continued

Further, when comparing your determination of this item's classification with your colleagues' judgments, resolve any differences, if any exist, through discussion. Please indicate any changes on an additional form. The form should be labelled RESOLVED CLASSIFICATION.

Graduate Students

Rater _____
Item# _____

Choice Alternative Similarity

Below are several propositional maps, one for the key pair and four more for the choice alternatives. The rater must first count the number of elements in the key. Then she must count the number of elements in common with the key, for all alternatives. Thus, in the example below, the key has 6 elements; each element is numbered. The numbers over the alternative elements correspond to the elements in common with the key's elements. So for alternative (a) there are two elements in common with the key. The digits next to each alternative represents the difference score. Again for alternative (a), the number of elements in the key minus the number of common elements is 6-2=4. These digits were then summed across the four alternatives and to standardize across items, divided by the number of elements in the key. Here the final computation was 14/6.

Example:
Key Pair

A soldier lives in a barracks.

(location: in (live, S: soldier), barracks)

Alternative Pairs

(a) A cook uses a stove.

(use, A: cook, O: stove) 4

(b) A nurse works at a clinic.

(location: at (work, A: nurse), clinic) 2

(c) An accountant works with numbers.

(work, A: accountant, I: numbers) 4

(d) An athlete performs a workout.

(perform, A: athlete, G: workout) 4

Researcher and Expert Graduate Student

Rater _____
Item# _____

Syntactic Order

Given the word order of this stem pair _____:_____ and the word order of this final stem rationale:_____ _____ _____, it was necessary, when formulating this rationale to:

(1) reverse the order of the words in the analogy pair_____

or

(2) retain the order of the words in the analogy pair_____.

For example, the pair *engine:car*, with a rationale "An engine is part of a car.", would be categorized as retain the order of the stem pair rationale while the pair *paint:artist*, with the rationale "An artist paints.", would be categorized as reverse the order of the stem pair rationale.

Graduate Students

References

Alderton, D. L., Goldman, S. R. & Pellegrino, J. W. (1985). Individual differences in process outcomes for verbal analogy and classification solution. Intelligence, 9, 69-85.

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

Anderson, J. R. (1985). Cognitive psychology and its implications. New York: W. H. Freeman & Co.

Andrich, D. (1988). Rasch models for measurement. Newbury Park, Ca.: Sage.

Baddeley, A. (1992). Working memory. Science, 255, 556-559.

Baker, F. B. (1992). Item response theory: parameter estimation techniques. New York: Marcel Dekker.

Barnes, G. M. (1980). A concurrent model for solving well- and ill-structured verbal analogies. Unpublished doctoral dissertation, University of Kansas.

Barnes, G. M. & Whitely, S. E. (1980). Problem restructuring processes for ill structured verbal analogies. (Technical Report Number NIE-80-2). University of Kansas.

Becker, B. J. (1990). Item characteristics and gender differences on the SAT-M for mathematically able youths. American Educational research Journal, 27(1), 65-87.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). Cognitive and psychometric analysis of analogical problem solving. N.J.:Springer-Verlag.

Bejar, I. I., Embretson, S., & Mayer, R. (1987). Cognitive psychology and the SAT: A review of some implications. ETS Research Report-87-28.

Bock, R.D., Gibbons, R. & Muraki, E. (1988). Full information factor analysis. Applied Psychological Measurement, 12(3), 261-280.

Bovair, S. & Kieras, D. E. (1985). A guide to propositional analysis. In B. K. Britton & J. B. Black (Eds.), Understanding expository text: A theoretical and practical handbook for analyzing explanatory text. Hillsdale, NJ: Lawrence Erlbaum.

Breland, H. M., Jones, R. J. & Jenkins, L. (1994). Alphabetical word list for the College Board study. CBR-94-4.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models. Newbury Park, CA: Sage.

Carpenter, P. A., Just, M. A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. Psychological Review, 97(3), 404-431.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect". In L. B. Resnick (Ed.), The nature of intelligence. Hillsdale, NJ:Lawrence Erlbaum.

Carroll, J. B. (1979). Measurement of abilities constructs. In Construct validity in psychological measurement: Proceedings of a colloquium on theory and application in education an employment (pp.43-52). Princeton, NJ: Educational Testing Service.

Chaffin, R. & Herrmann, D. J. (1987). Relation element theory: A new account of the representation and processing of semantic relations. In D. Gorfein & R. Hofman (Eds.), Learning and memory: The Ebbinghaus Centennial Conference. Hillsdale, N.J.: Erlbaum.

Chaffin, R. & Herrmann, D. J. (1988). The nature of semantic relations: A comparison of two approaches. In M. Evens (Ed,), Relational models of the lexicon. New York: Cambridge University Press.

Chaffin, R. & Pierce, L. (1987, April). Types of verbal analogy relations and academic skills. Paper presented at National Council of Measurement in Education, Washington, D.C.

Chaffin, R. & Pierce, L. (draft). A taxonomy of semantic relations for the classification of GRE analogy items and an algorithm for the generation of GRE-type analogies.

Chaffin, R. & Pierce, L. (draft). A taxonomy of semantic relations for the classification of GRE analogy items. ETS-RR-50.

Clement, C. A. & Gentner, D. (1991). Systematicity as a selection constraint on analogical mapping. Cognitive

Science, 15, 89-132.

Collins, A. M. & Loftus, E. F. (1975). A spreading theory of semantic processing. Psychological Review, 82(6), 407-428.

Dawis, R. V. & Siojo, L. T. (1972). Analogical reasoning: A review of the literature. Tech Report No. 1

Dierbach, C. (1990). Abstractional concept mapping: A formal basis for analogical reasoning. Unpublished doctoral dissertation, University of Delaware.

Dillon, W. R. & Goldstein, M. (1984). Multivariate analysis: Methods and applications. New York: John Wiley & Sons.

Diones, R., Bejar, I. & Chaffin, R. (in preparation). Dimensionality issues for SAT analogy items.

Donlon, T. F. (Ed.) (1984). The College Board technical handbook for the scholastic aptitude test and achievement tests. New York: College Board Publications.

Embretson, S. E. (1982). Component latent trait models for test design. ED 264 272.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93(1), 179-197.

Embretson, S. E. (1984). A general latent trait model for response processes. Psychometrika, 49(2), 175-186.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In S. E. Embretson (Ed.), Test design: Developments in psychology and psychometrics. Orlando: Academic Press.

Embretson, S. E. (1989). Program LINLOG: An extension of Fischer and Formann's program for the linear logistic trait model.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. Psychometrika, 56(3), 495-515.

Embretson, S. E. (1992a). Measuring and validating cognitive modifiability as an ability: A study in the spatial domain. Journal of Educational Measurement, 29(1), 25-50.

Embretson, S. E. (1992b). Psychometric models for learning and cognitive processes. In (Eds.), N. Frederiksen, R. J. Mislevy, & I. Bejar, Test theory for a new generation of tests. Hillsdale, N.J.: Lawrence Erlbaum.

Embretson, S. E. & Curtright, C. (1980). Performance and stimulus complexity norms for ability test analogies. (Technical Report Number NIE-80-4). Kansas: University of Kansas.

Embretson, S. E. & Curtright, C. (1982). Problem structure and response format in solving verbal analogies. (Technical Report Number NIE-82-2). Kansas: University of Kansas.

Embretson, S. E. & Schneider, L. M. (1989). Cognitive component models for psychometric analogies: Conceptually driven versus interactive process models. Learning and individual differences, 1, 155-178.

Embretson, S. E. & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Embretson, S. E., Schneider, L. M. & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. Journal of Educational Measurement, 23(1), 13-32.

Emmerich, W. (1989). Appraising the cognitive features of subject tests. ETS-RR-89-53.

Enright, M. K., Duran, R. P. & Pierce, L. P. (1986). Strategies and processes in solution of GRE analogies. Paper presented at the American Educational Research Association, San Francisco, CA.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. Acta Psychological, 37, 359-374.

Fischer, G. H. (1972). Some probabilistic models for measuring change, In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), Advances in psychological and educational measurement. New York: Wiley.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. Psychometrika, 48(1), 3-26.

Fischer, G. H. & Formann, A. K. (1972). An algorithm and a FORTRAN program for estimating the item parameters of the linear logistic test model. Research Bulletin No.

11. Vienna: University of Vienna.

Fischer, G. H. & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. Applied Psychological Measurement, 6(4), 397-416.

Fischer, G. H., Formann, A. K. & Wild B. (1988). Programm LLTM.

Fleiss, J. L. (1973). Statistical methods for rates and proportions. New York: Wiley.

Freedle, R. & Kostin, I. (1991). Semantic and structural factors affecting the performance of matched black and white examinees on analogy items from the scholastic aptitude test. ETS-RR-91-28.

Freedle, R., Kostin, I. & Schwartz, L. M. (1987). A comparison of strategies used by black and white students in solving SAT verbal analogies using a thinking aloud method and a matched percentage-correct design. ETS-RR-87-48.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning. New York: Cambridge University Press.

Gitomer, D. H., Curtis, M. E., Glaser, R. & Lensky, D. B. (1987). Processing differences as a function of item difficulty in verbal analogy performance. Journal of Educational Psychology, 79(3), 212-219.

Glaser, R. & Pellegrino, J. W. (1978). Uniting cognitive process theory and differential psychology: Back home from the wars. Intelligence, 2, 305-319.

Glaser, R. & Pellegrino, J. W. (1979). Cognitive process analysis of aptitude: The nature of inductive reasoning tasks. Pittsburgh, PA: Learning Research and Development Center. (ED 198 175).

Goldman, S. R. & Pellegrino, J. W. (1984). Deductions about induction: Analyses of developmental and individual differences. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence Vol.2. Hillsdale, NJ: Lawrence Erlbaum.

Goswami, U. (1989). Relational complexity and the development of analogical reasoning. Cognitive Development, 4, 251-268.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. Child Development, 62, 1-22.

Grundin, J. (1980). Processes in verbal analogies. Journal of Experimental Psychology: Human Perception and Performance, 6(1), 67-74.

Hall, W. S., Nagy, W. E. & Linn, R. (1984). Spoken words: Effects of situation and social group on oral word usage and frequency. Hillsdale, N.J.: Lawrence Erlbaum.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer•Nijhoff Publishing.

Hasher, L. & Zacks, R. (1979). Automatic and effortful processes in memory. Journal of Experimental Psychology: General, 108(3), 356-388.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9(2), 139-164.

Henley, T. B. (1989). The primacy of relations. Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Holyoak, K. J. (1984). Analogical thinking and human intelligence. In R. J. Sternberg (Ed.), 199-230.

Holyoak, K. J. & Nisbett, R. E. (1988). Induction. In R. J. Sternberg & E. E. Smith, (Eds.), The psychology of human thought (pp.50-91). New York: Cambridge University Press.

Holyoak, K. J. & Thagard, P. (1989a). Analogical mapping by constraint satisfaction. Cognitive Science, 13, 295-355.

Holyoak, K. J. & Thagard, P. (1989b). A computational model of analogical problem solving. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning. New York: Cambridge University Press.

Hornke, L. F. & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. Applied Psychological Measurement, 10(4),

369-380.

Ingram, A. & Pellegrino, J. W. (1977). Response generation norms for verbal analogies. ED 144 046.

Iran-Nejad, A., Wittrock, M. C. & Hidi, S. (Eds.). (1992). Brain and education [Special issue]. Educational Psychologist, 27(4).

Johnson-Laird, P. N., Herrmann, D. J. & Chaffin, R. (1984). Only connections: A critique of semantic networks. Psychological Bulletin, 96, 292-315.

Just, M. A. & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. Psychological Review, 99(1), 122-149.

Kintsch, W. & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Kirsch, I. S. & Mosenthal, P. B. (1988). Understanding document literacy: Variables underlying the performance of young adults. RR-88-62. Princeton, NJ: Educational Testing Service.

Klix, F. (1980). On structure and function of semantic memory. In F. Klix & J. Hoffmann (Eds.), Cognition and memory. New York: North-Holland.

Kramer, G. A. & Smith, R. M. (1990). An investigation of components influencing the difficulty of form-development items. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, Ma.

Kramer, G. A., Smith, R. M. & Kubiak, A. T. (1989). A cross validation of components influencing the difficulty of cube items. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, Ca.

Lane, S. (1991). Use of restricted item response models for examining item difficulty ordering and slope uniformity. Journal of Educational Measurement, 28(4), 295-309.

Liu, I. (1981). Common and specific features in pictorial analogies. Memory and Cognition, 9, 515-523.

Markowitz, J. A., Nutter, J. T. & Evens, M. W. (1992). Beyond is-a and part-whole: More semantic network

links. Computers Mathematical Applications, 23(6-9), 377-390.

Masters, G. N. & Mislevy, R. J. (March, 1991). New views of student learning: Implications for educational measurement. ETS-RR-91-24-ONR.

McClelland, J. L. & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. Journal of Experimental Psychology: General, 114(2), 159-188.

McKinley, R. (1989). Confirmatory analysis of test structure using multidimensional item response theory. RR-89-31.

Medin, D. L. (1989). Concepts and conceptual structure. American Psychologist, 44(12), 1469-1481.

Miller, G. A. (1990). WordNet: An on-line lexical database. International Journal of Lexicography, 3(4).

Millsap, R. E. & Everson, H. T. (1993). Methodology Review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17(4), 297-334.

Millsap, R. E. & Meredith, W. (1993). Inferential conditions in the statistical detection of measurement bias. Applied Psychological Measurement.

Mislevy, R. J. (1981). A general linear model for the analysis of Rasch item threshold estimates. Unpublished doctoral dissertation, University of Chicago, Illinois.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. Applied Psychological Measurement, 12(3), 281-296.

Mislevy, R. J., Sheehan, K. M. & Wingersky, M. (1993). How to equate tests with little or no data. Journal of Educational Measurement, 30(1), 1-24.

Mislevy, R. J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. Psychometrika, 55(2), 195-215.

Mitchell, K. J. (1983). Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Battery. ED 260 122.

Muthén, B. O., Kao, C. & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of

new IRT-based detection technique to mathematics achievement test items. Journal of Educational Measurement, 28(1), 1-22.

Olsen, J. F., Scheuneman, J. & Grima, A. (1989). Statistical approaches to the study of item difficulty. ETS-RR-89-21.

Pellegrino, J. W. (1985). Mental models and mental tests. In S. E. Embretson, (Ed.), Test design.

Pellegrino, J. W. & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. Intelligence, 3, 187-214.

Pellegrino, J. W. & Glaser, R. (1980). Components of inductive reasoning. In R. E. Snow, P-A. Frederico & W. E. Montague (Eds.), Aptitude, learning and instruction: Cognitive process analyses of aptitude (Vol 1). Hillsdale, NJ: Lawrence Erlbaum.

Raaijmakers, J. G. W. & Shiffrin, R. M. (1992). Models for recall and recognition. Annual Review of Psychology, 43, 205-234.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Ramsey, J. O. (December, 1991). TESTGRAF: A program for the graphical analysis of multiple choice test data.

Ramsey, J. O. Kernel smoothing approaches to nonparametric item characteristic curve estimation.

Rumelhart, D. E. (1989). Toward a microstructural account of human reasoning. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning. New York: Cambridge University Press.

Rumelhart, D. E. & Abrahamson, A. A. (1973). A model for analogical reasoning. Cognitive Psychology, 5, 1-28.

Scheiblechner, H. (1972). Das lernen und lösen komplexer denkaufgaben. Zeitschrift für Experimentelle und Angewandte Psychologie, 19, 476-506.

Scheuneman, J. D., Gerritz, K. & Embretson, S. E. (1991). Effects of prose complexity on achievement test item difficulty. ETS RR-91-43.

Scheuneman, J. D. & Steinhaus, K. S. (1987). A theoretical

framework for the study of item difficulty and discrimination. ETS Research Report 87-44.

Schmitt, A. P. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27(1), 67-81.

Schmitt, A. P. & Bleistein, C. A. (1987). Factors affecting differential item functioning for black examinees on scholastic aptitude test analogy items. Educational Testing Service RR-87-23.

Schmitt, A. P. & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27(1), 67-81.

Sheard, C. & Readence, J. E. (1988). An investigation of the inference and mapping processes of the componential theory of analogical reasoning. Journal of Educational Research, 81(6), 347-353.

Sheehan, K. & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. Journal of Educational Measurement, 27(3), 255-272.

Shinn, Hong Shik. (1989). A unified approach to analogical reasoning. Unpublished doctoral dissertation, Georgia Institute of Technology, Georgia.

Simpson, J. A. & Weiner, E. S. C. (1989). The Oxford English Dictionary. New York: Oxford University Press.

Smith, E. E. (1988). Concepts and thought. In R. J. Sternberg & E. E. Smith, (Eds.), The psychology of human thought (pp.50-91). New York: Cambridge University Press.

Smith, R. M. & Green, K. E. (1985). Components of difficulty in paper-folding tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Ill.

Smith, R. M. & Kramer, G. A. (1988). Component analysis of the factors influencing the difficulty of perceptual ability items. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Lo.

Smith, R. M., Kramer, G. A. & Kubiak, A. T. (in press). Components of difficulty in spatial ability test items.

Spada, H. (1978). Understanding proportionality: A

comparison of different models of cognitive development. International Journal of Behavior Development, 1, 363-376.

Spada, H. (1982). The assessment of learning effects by means of linear logistic test models. Paper presented at the annual meeting of the National Council on Measurement in Education,

Spearman, C. (1923). The nature of 'intelligence' and the principles of cognition. New York: Macmillan.

Stenner, A. J., Smith III, M. & Burdick, D. S. (1983). Toward a theory of construct definition. Journal of Educational Measurement, 20(4), 305-316.

Sternberg, R. J. (1977a). Intelligence, information processing and analogical reasoning. Hillsdale, NJ: Lawrence Erlbaum.

Sternberg, R. J. (1977b). Component processes in analogical reasoning. Psychological Review, 84(4), 353-378.

Sternberg, R. J. (1981). Intelligence and nonentrenchment. Journal of Educational Psychology, 73, 1-16.

Sternberg, R. J. (1986). Toward a unified theory of human reasoning. Intelligence, 10, 281-314.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52(4), 589-617.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredericksen, Glaser, Lesgold & Shatto, (Eds.), Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, N.J.: Lawrence Erlbaum.

Thissen, D., Steinberg, L. & Fitzpatrik, A. R. (1989). Multiple-choice models: The distractors are part also part of the item. Journal of Educational Measurement, 26(2), 161-176.

Turner, A. & Green, E. (1978). Construction and use of a propositional text base. JSAS Catalog of Selected Documents in Psychology, 3, 58. (Ms. No. 1713)

Warren, H. C. (1967). History of association psychology. New York: Charles Scribner's Sons.

Whitely, S. E. (1976). Solving verbal analogies: Some

cognitive components of intelligence test items. Journal of Educational Psychology, 68(2), 234-242.

Whitely, S. E. (1977). Information-processing on intelligence test items: Some response components. Applied Psycholcgical Measurement, 1(4), 465-476.

Whitely, S. E. (1977b). Relations in analogy items: A semantic component of a psychometric task. Educational and Psychological Measurement, 37, 725-739.

Whitely, S. E. (1980). Latent trait models in the study of intelligence. Intelligence, 4, 97-132.

Whitely, S. E. (1981). Construct validation from latent trait models for aptitude processes. Technical Report NIE-81-1. University of Kansas.

Whitely, S. E. & Schneider, L. M. (1981). Information structure for geometric analogies. Applied Psychological Measurement, 5(3), 383-397.

Wilson, D. T., Wood, R. & Gibbons, R. (1987). TESTFACT: Test scoring, item statistics, and item factor analysis. Mooresville, IN: Scientific Software.

Wright, B. D. & Linacre, J. M. (1992). A user's guide to BIGSTEPS. Chicago: Mesa Press.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D. & Stone, M. H. (1979). Best test design. Chicago: Mesa Press.

Yamamoto, K. (1989). Hybrid model of IRT and latent class models. ETS-RR-89-41.

Yamamoto, K. (1990). Modeling the mixture of IRT and patterned responses by a modified hybrid model. Unpublished manuscript.

Yamamoto, K. & Gitomer, D. H. (1992). Application of a HYBRID model to a test of cognitive skill representation. In N. Frederiksen (Ed.), Test Theory for a New Generation of Tests. Edison, N.J.: Lawrence Erlbaum.