Ordering Inductive Reasoning Tests for Adaptive Knowledge Assessments

An Application of Surmise Relations between Tests

Inauguraldissertation
zur Erlangung des Grades einer Dr. rer. nat.
an der Naturwissenschaftlichen Fakultät der
Karl–Franzens–Universität Graz

Gudrun Wesiak

Department of Psychology, University of Graz

Graz, June 2003

Surveyors: o. Univ.–Prof. Dr. Dietrich Albert (Supervisor)
o. Univ.–Prof. Dr. Helmuth P. Huber

Acknowledgments

This research started with my work in an interdisciplinary project on ‘Surmise Relations between Tests’ at the Cognitive Science Section at the University of Graz. The project was supported by the Austrian Science Fund (FWF) through the Grant P12726-SOZ to Dietrich Albert and Wilhelm Schappacher.

I gratefully acknowledge the ‘Heerespsychologischer Dienst des Bundesministeriums für Landesverteidigung’ in Vienna and the ‘Psychologischer Dienst der Bundeswehr der Bundesrepublik Deutschland’ in Bonn for providing the tests and data sets for two investigations.

This research strongly benefited from discussions with, comments and suggestions from my colleagues Silke Brandt, Cord Hockemeyer, Susanne Pötzi, and Ali Ünli in the FWF project as well as my colleagues in the scientific seminar of the Cognitive Science Section.

Special thanks go to my supervisor Dietrich Albert for his continuous support, his comments, and his advice during the times of research and of preparing this report.

Furthermore, I want to thank Gisela Dösinger for her help in conducting the third investigation, Gabriele Janzen for her comments on a preliminary version of this report, and Cord Hockemeyer for his TEXnical and EDP support.

Contents

Summary **1**

1 Introduction **3**

2 Inductive Reasoning **7**

2.1 General introduction 8

2.2 Inductive reasoning tasks 10

2.2.1 Verbal and geometric analogies 11

2.2.2 Series completions 17

2.2.3 Geometric matrices 20

2.2.4 Communalities and differences 24

2.3 Inductive reasoning models 27

2.3.1 Klauer's model of inductive reasoning 27

2.3.2 R.J. Sternberg's cognitive components approach 31

2.4 Psychometric approach to inductive reasoning and intelligence 35

2.4.1 Structure of intelligence models 37

2.4.2 Tests assessing the ability of inductive reasoning 40

2.5 Summary of Chapter 2 46

3 The Theory of Knowledge Spaces **49**

3.1 Surmise relations between items 51

3.2 Surmise relations between tests 55

3.3 Generation of knowledge structures 60

3.3.1 Componentwise ordering of product sets 62

3.4 Validation of knowledge structures 65

3.4.1 Validation of hypotheses via the surmise relation 67

3.4.2 Validation of hypotheses via the knowledge space 70

3.5 Knowledge spaces for inductive reasoning 76
3.5.1 Number series completions 76
3.5.2 Geometric matrices 78
3.6 Summary 79

4 Purpose and Scientific Questions 81

5 Method and Results 85
5.1 Derivation of hypotheses 85
5.1.1 Specification and ordering of components 86
5.1.2 Specification and ordering of attributes 87
5.1.3 Model for the surmise relation between item classes 88
5.1.4 Model for the surmise relation between tests 89
5.1.5 Empirical predictions 90
5.2 Investigation I: Relating a pair of tests: First application 93
5.2.1 Hypothesis 93
5.2.2 Method 94
5.2.3 Results 99
5.2.4 Discussion 108
5.3 Investigation II: Relating a pair of tests: Reconsidered 109
5.3.1 Hypothesis 110
5.3.2 Method 110
5.3.3 Results 116
5.3.4 Discussion 123
5.4 Investigation III: Extending the model to four tests 124
5.4.1 Hypothesis 125
5.4.2 Method 126
5.4.3 Results 133
5.4.4 Discussion 139

6 Adaptive Testing 143
6.1 Approaches to adaptive knowledge assessment 144
6.2 Adaptive assessment algorithms 145
6.2.1 Deterministic assessment algorithm 146

6.2.2 Non–deterministic assessment algorithms 147
6.3 Adaptive assessment for Investigations I–III 149
6.3.1 Adaptive testing procedure 149
6.3.2 Adaptive testing results and discussion 150

7 General Discussion 155

References 165

A Mathematical Basics 179

B Material 181
B.1 Semantic rationales 181
B.2 Items presented in Investigation III 183

C Procedure for Investigation III 185
C.1 Instructions 185
C.2 Randomization 188

D List of Programs 189

E Hypotheses 191
E.1 Relation files 191
E.2 Base files 195

F Results 201
F.1 Relative solution frequencies 201
F.2 χ^2 Statistics for reversed solution frequencies 202
F.3 Distance distributions 203
F.4 Invalidity of items 206
F.5 χ^2 Statistics for distance distributions 209

List of Figures

2.1 Reaction times and error rates for geometric analogies 16

2.2 Sequence of information–processing during the solution of series completion problems 19

2.3 Definition of inductive reasoning by Klauer (1994) 28

2.4 Analytical strategy used for the solution of inductive reasoning problems 30

2.5 Sequence of information–processing during the solution of inductive reasoning problems 33

2.6 Guilford's Structure–of–Intellect model 38

2.7 The Berlin Structure of Intelligence model 39

3.1 Matrix and Hasse diagram for the surmise relation in Example 3.1 . . . 52

3.2 Possible response patterns and assumed prerequisite relation for a pair of items 53

3.3 Derived knowledge structure for Example 3.1 53

3.4 Two tests A and B are in surmise relation from A to B ($B \; \dot{\mathcal{S}} \; A$) 56

3.5 Left–covering surmise relation from test A to test B ($B \; \dot{\mathcal{S}}_l \; A$) 57

3.6 Right–covering surmise relation from test A to test B ($B \; \dot{\mathcal{S}}_r \; A$) 57

3.7 Subsets contained in the Cartesian product for the items of two tests A and B 58

3.8 Cartesian product of two attribute sets A and B (Example 3.3) 63

3.9 Problem structure for Example 3.3 64

3.10 Hypothetical surmise relation and corresponding knowledge space for Example 3.4 66

3.11 Percentage of correct solutions for Example 3.4 68

5.1 Hasse diagrams for the downsets of the five components and their attributes 87

5.2 Cartesian product of the attribute sets for components A, B, and C . . 88

5.3 Derived order for item classes defined by components A, B, and C . . . 89

5.4 Derived surmise relation between two tests AN and SC 91

5.5 Example for a matrix item presented in Investigation I 95

5.6 Example for an analogy item presented in Investigation I 96

5.7 Hasse diagram for the postulated surmise relation between items and relative solution frequencies in Investigation I 101

5.8 Distance distribution of the empirical data set compared to the distributions of random and probability simulations 106

5.9 Example for a matrix item presented in Investigation II 111

5.10 Example for an analogy item presented in Investigation II 112

5.11 Hasse diagram for the postulated surmise relation between items and relative solution frequencies in Investigation II 117

5.12 Distance distribution of the empirical data set compared to the distributions of random and probability simulations 121

5.13 Examples for each problem type presented in Investigation III 128

5.14 Hypothesis and relative solution frequencies for the surmise relation within and between tests in Investigation III 134

5.15 Distance distribution of the empirical data set compared to the distributions of random and probability simulations 138

6.1 Illustration of the adaptive assessment procedure 147

7.1 Mapping sentence for inductive reasoning problems 163

List of Tables

2.1 Taxonomy of semantic relations 14

2.2 Rules for the solution of geometric matrices 23

2.3 Communalities and differences for inductive reasoning problems 26

2.4 The six variants of inductive reasoning problems 29

2.5 ISA problem types and their assignment to ability components 41

2.6 Problem types measuring the ability dimensions of the BIS model . . . 43

3.1 Base for Example 3.1 54

3.2 Relation and base files for Example 3.4 66

3.3 Data set for four items and 10 response patterns (Example 3.4) 67

3.4 Contingency tables for Example 3.4 68

3.5 Symmetric distances between the data set and the hypothesized knowledge space for Example 3.4 72

3.6 Distance distributions and χ^2 values for Example 3.4 73

3.7 Distance distribution for the powerset and DA for Example 3.4 73

3.8 Average distances and DA coefficients for simulated data sets 75

3.9 Examples for the problem construction of number series completions . . 77

5.1 Components, attributes, and assumed downsets for inductive reasoning problems 86

5.2 Item descriptions and downsets for the matrix test in Investigation I . . 96

5.3 Item descriptions and downsets for the analogy test in Investigation I . 98

5.4 Surmise relation across tests in Investigation I 99

5.5 Original and reduced data sets in Investigation I 101

5.6 VC and γ for the surmise relation between items and its subsets (N = 572) 103

5.7 Symmetric distances for the test knowledge space, its substructures, and their powersets (N = 572) 105

5.8 Symmetric distances for the test knowledge space, its substructures, and simulated data sets (N = 572) 106

5.9 Comparison of the empirical and simulated symmetric distances for the test knowledge space and its substructures (N = 572) 107

5.10 Results of the frequency simulations for the test knowledge space and its substructures (N = 572) 108

5.11 Item descriptions and downsets for the matrix test in Investigation II . 113

5.12 Item descriptions and downsets for the analogy test in Investigation II . 114

5.13 Surmise relation across tests in Investigation II 115

5.14 Original and reduced data sets in Investigation II 117

5.15 VC and γ for the surmise relation between items and its subsets (N = 2628) 119

5.16 Symmetric distances for the test knowledge space, its substructures, and their powersets (N = 2628) 120

5.17 Symmetric distances for the test knowledge space, its substructures, and simulated data sets (N = 2628) 121

5.18 Comparison of the empirical and simulated symmetric distances for the test knowledge space and its substructures (N = 2628) 122

5.19 Selected items and tests for Investigation III 127

5.20 Item descriptions and downsets for the four tests in Investigation III . . 130

5.21 Prerequisite relationships across the tests in Investigation III 132

5.22 VC and γ for the surmise relation between items and its subsets (N = 121) 136

5.23 Symmetric distances for the test knowledge space, its substructures, and their powersets (N = 121) 137

5.24 Symmetric distances for the test knowledge space, its substructures, and simulated data sets (N = 121) 137

5.25 Comparison of the empirical and simulated symmetric distances for the test knowledge space and its substructures (N = 121) 139

5.26 Results of the frequency simulations for the test knowledge space and its substructures (N = 121) 140

6.1 Results for the adaptive testing procedures 151

B.1 Semantic rationales for the analogy items in Investigation I 181

B.2 Semantic rationales for the analogy items in Investigation II 182

B.3 Semantic rationales for the verbal analogy items in Investigation III . . 182

B.4 Items and modified answer formats for Investigation III 183

F.1 Relative solution frequencies for items in Investigation I 201

F.2 Relative solution frequencies for items in Investigation II 201

F.3 Relative solution frequencies for items in Investigation III 201

F.4 χ^2 Statistics for reversed solution frequencies in Investigation I 202

F.5 χ^2 Statistics for reversed solution frequencies in Investigation II 202

F.6 χ^2 Statistics for reversed solution frequencies in Investigation III 202

F.7 Distance distributions for the surmise relations between items and across tests in Investigation I 203

F.8 Distance distributions for the surmise relations within tests in Investigation I 203

F.9 Distance distributions for the surmise relations between items and across tests in Investigation II 204

F.10 Distance distributions for the surmise relations within tests in Investigation II 204

F.11 Distance distributions for the surmise relations between items and across tests in Investigation III 205

F.12 Distance distributions for the surmise relations within tests in Investigation III 205

F.13 Invalidity of items in Investigation I 206

F.14 Invalidity of items in Investigation II 207

F.15 Invalidity of items in Investigation III 208

Summary

A series of psychological tests assess the ability of inductive reasoning. Examples are analogies, series completions, or matrices with materials varying across verbal, geometric–figural, or numerical content. The purpose of this study is to establish a prerequisite order on different inductive reasoning tests and to provide a basis for an adaptive assessment instrument that covers various types of inductive reasoning problems.

To generate a hypothesis on the structure of inductive reasoning tests, I selected the non–numerical knowledge space theory, because in this approach the dependencies among items are interpreted as surmise or prerequisite relations, which establish a partial order on the set of items. A recent generalization of the theory to surmise relations between tests renders similar interpretations with respect to sets of tests instead of items. Thereby, the response patterns in a subset of the tests are inferred from the given responses in some other test. For the establishment of a surmise relation between inductive reasoning tests, five common components for the various problem types were extracted from earlier psychological findings. The principle of componentwise ordering of product sets was applied to define difficulty orders on the components' attributes and to establish a test knowledge space (i. e. the set of all predicted solution patterns).

For the empirical validation of the postulated model, three investigations were conducted. For Investigations I ($N = 572$ corporal and officer candidates) and II ($N = 2628$ draftees) the set of related tests comprised a verbal analogy and a geometric matrix test (with 30 and 40 items for Investigations I and II respectively). For Investigation III ($N = 121$ students of both sexes) the set of tests was extended to four problem types with 5 items each, viz. verbal and geometric analogies, number series completions, and geometric matrices. The validation of the derived models is based on procedures via the surmise relation and via the knowledge space. The results of both procedures indicate a good fit of the models established for Investigations I and III. For Investigation II, the results derived via the surmise relation support the postulated model, whereas the results derived via the knowledge space reveal significant deviations.

Considering the possibility of unexpectedly high noise rates in Investigation II, the derived models of all three investigations were implemented into adaptive assessment procedures. Using the postulated knowledge spaces, the adaptive algorithms estimate the complete response patterns from a subset of the given answers. A comparison of the estimated and the empirical patterns showed that the average error rate arising from the adaptive assessments amounts to less than one item per pattern. Moreover, the adaptive algorithms lead to savings of 43.96% to 72.23% of the posed questions. Thus, the presented study shows that ordering inductive reasoning tests on the basis of knowledge space theory provides a good foundation for an integrative and efficient diagnostic instrument in this domain. Further research should focus on the construction of an item and test pool that covers all possible item classes and can subsequently be implemented into a comprehensive adaptive assessment system.

Zusammenfassung

Eine Reihe psychologischer Tests beschäftigt sich mit der Diagnose induktiven Denkens, wobei den Testpersonen verschiedene Aufgabentypen, wie Analogien, Reihenfortsetzen oder Matrizen anhand verbaler, geometrisch–figuraler oder numerischer Materialen vorgegeben werden. Ziel der vorliegenden Arbeit ist es, induktive Denktests derart zu strukturieren, daß unterschiedliche Aufgabentypen in einem umfassenden, adaptiven Testsystem integriert werden können.

Für die Strukturierung induktiver Denktests wurde die nicht–numerische Wissensraumtheorie gewählt, die es erlaubt, auf der Menge der Items eine partielle Ordnung zu etablieren und Abhängigkeiten zwischen Testitems als Vermutungs- oder Voraussetzungsbeziehungen zu interpretieren. In einer Verallgemeinerung der Theorie zu Vermutungsrelationen zwischen Tests wird von den Antwortmustern aus einem Test auf die Antwortmuster in weiteren Tests geschlossen. Zur Konstruktion des Wissensraumes (d. h. der Menge aller postulierten Antwortmuster) bzw. der Vermutungsrelation zwischen induktiven Denktests wurden fünf Komponenten, durch die verschiedene Aufgabentypen beschreibbar sind, abgeleitet. Die Voraussetzungsbeziehungen zwischen Items und Tests basieren auf dem Prinzip der komponentenweisen Ordnung, wonach eine Menge von Aufgaben bezüglich ihrer Schwierigkeiten strukturiert wird.

Die empirische Validierung der erhaltenen Aufgaben– und Teststrukturen erfolgte mittels drei Untersuchungen. In den Untersuchungen I ($N = 572$ Offiziers– und Unteroffiziersanwärter) und II ($N = 2628$ Rekruten) wurden je zwei Tests, bestehend aus verbalen Analogien und geometrischen Matrizen, bearbeitet (mit 30 bzw. 40 Items für die Untersuchungen I bzw. II), während in Untersuchung III ($N = 121$ SchülerInnen und StudentInnen) die vier Aufgabentypen verbale und geometrische Analogien, Zahlenfolgen und geometrische Matrizen durch jeweils 5 Items präsentiert wurden. Die Validierung der drei Modelle erfolgte für die postulierten Vermutungsrelationen und die zugehörigen Wissensräume jeweils getrennt. Die Ergebnisse der Untersuchungen I und III zeigen eine gute Anpassung der Modelle an die empirischen Daten, während für Untersuchung II zwar die Ergebnisse der Vermutungsrelation den Hypothesen entsprechen, die Ergebnisse des Wissensraums jedoch Modellabweichungen aufweisen.

Hinsichtlich der Möglichkeit unerwartet hoher Fehler– und Ratewahrscheinlichkeiten in Untersuchung II, wurden die Modelle aller Untersuchungen in wissensraumbasierte adaptive Testverfahren implementiert. Ein Vergleich der geschätzten mit den empirischen Antwortmustern zeigte, daß die durch die adaptive Diagnose bedingten Abweichungen im Mittel weniger als eine Aufgabe pro Antwortmuster betragen, während die Anzahl der präsentierten Items um bis zu 72.23% reduziert wurde. Die vorliegende Arbeit zeigt, daß die Strukturierung induktiver Denktests auf Basis der Wissensraumtheorie eine gute Grundlage für ein effizientes, adaptives Diagnoseverfahren für diesen Bereich bietet. Für weiterführende Forschungsarbeiten erscheint vor allem die Konstruktion eines Aufgaben– und Testpools, der alle möglichen Aufgabenklassen umfaßt und in ein adaptives Diagnosesystem implementiert werden kann, von Bedeutung.

1 Introduction

This study is about the diagnosis of inductive reasoning abilities. Induction is the process of reasoning from particular instances to reach a general conclusion or find a general rule governing these instances. Inductive inferences are also the means for predicting future instances and for handling new situations by applying already stored knowledge of past events. Thus, inductive reasoning abilities and reasoning in general are fundamental to human intelligence.

The ability to solve intellectual problems differs among individuals. In order to measure these differences a large number of intelligence and aptitude tests has been developed. In many of the diagnostic procedures inductive reasoning skills are tested by presenting at least one subtest containing problems of inducing structure (e. g., analogies or series completion problems). One of the primary motivations for the widespread use of these tests is that inductive reasoning is a central indicator of general intelligence. The interest in this field is documented by the substantial amount of research devoted to either inductive reasoning in general or the study of specific problem types such as analogies.

Inductive reasoning involves forming and testing hypotheses. The testee is given a series of instances, from which he or she must induce a rule that relates the instances to each other. These rule induction skills are usually measured by means of several problem types, including analogies, classifications, series completions, and matrices with materials varying across pictorial, verbal, numerical, and geometric–figural content. For each test or subtest, the testee is assigned a numerical score denoting his or her level of ability on the respective problem type. Generally, the scores derived from psychometric tests consider, whether or not participants answered a question correctly, but do not account for the specific problem requirements that are met by a participant. By this, we are able to differentiate between participants of varying ability but remain uninformed about the problem requirements that are not met and therefore need to be trained on.

The purpose of this research is the development of a new approach to the diagnosis of inductive reasoning abilities. The two foci of the study are (a) on the integration of various problem types into one common classification scheme and (b) on the development of an efficient assessment instrument that provides precise information on problem demands and person abilities. For (a) the integrative representation of inductive reasoning tests, the results of earlier research are taken into account. Therefore, I will provide a review of some prominent findings on inductive reasoning in Chapter 2. I will start out with a general introduction into the field (Section 2.1), which is followed by an overview of the tasks belonging to the domain of inductive reasoning (Section 2.2). This overview includes an outline of the components and attributes inherent in single problem types and an integration of the findings with respect to the problem types' communalities and differences. Since I am interested in a comprehensive classification scheme, I will also report two theories of inductive reasoning, which cover various problem types (Section 2.3). With regard to (b), the psychometric approach to inductive reasoning and intelligence is outlined in Section 2.4. This includes a short description of some intelligence models as well as a selection of some tests measuring inductive reasoning abilities. The chapter on inductive reasoning will conclude with a summary of the most important findings with respect to this study (Section 2.5).

Knowledge of the common features inherent in different types of inductive reasoning problems is fundamental for an integrative diagnostic system. However, the implementation of adaptive assessment algorithms requires additional information. In order to infer the responses to a subset of the items and tests under investigation, a prerequisite order has to be defined on the set of items and tests. The theory of knowledge spaces has been selected as methodological framework for this purpose. This non–numerical approach was originally developed for the representation and efficient diagnosis of knowledge in a given domain. In all knowledge domains or psychometric tests, the set of items varies with respect to difficulty. By considering these implicit dependencies (the so called surmise relation) among a set of problems, the correct or incorrect solutions to a subset of items can be inferred from previously obtained responses. These inferences reduce the number of questions posed to a testee. If the dependencies among items are specified by varying problem demands, it is furthermore possible to obtain precise information on the testee's knowledge state, i. e. of the problem requirements he or she is able to meet. Chapter 3 provides an outline of the knowledge space theory. After an introduction to the basic idea of a surmise or prerequisite relation between items (Section 3.1), I will report a recent generalization of this concept (Section 3.2). The new approach of surmise relations between tests is cardinal to this research, because it permits the establishment of prerequisite relationships between sets of items or, as for this study, between inductive reasoning tests. The specification of the difficulty order on items and tests is based on the components and attributes identified in Chapter 2. Methods for the generation of testable hypotheses will be discussed in Section 3.3, with special focus on a component based approach that allows the establishment of a theoretically founded knowledge structure. In Section 3.4, I will introduce several validation methods that are either based on the surmise relation or on the knowledge space. Finally, in Section 3.5, I will report two studies, in which single inductive reasoning tests have been structured on the basis of knowledge space theory. A short summary will conclude the chapter (Section 3.6).

The theoretical part of this report is followed by a short overview of the study's purpose and conceived scientific questions (Chapter 4). Subsequently, the empirical part (Chapter 5) is entered with a detailed description of how the overall hypothesis on a surmise relation between inductive reasoning tests has been derived (Section 5.1). For the evaluation of the derived model, three closely related investigations have been conducted, which are presented together with a short discussion each in Sections 5.2, 5.3, and 5.4.

The derived hypotheses allow predictions on participants' solution behavior and are, after they have proved to be valid representations of the knowledge domain, applicable to adaptive testing procedures. Since the results of the investigations show that there is no obstacle to applying the derived test knowledge spaces (or corresponding surmise relations) to adaptive assessment algorithms, I will address this issue in Chapter 6. After a general introduction to the adaptive assessment of knowledge and a short outline of traditional approaches (Section 6.1), I will introduce a deterministic and a non–deterministic assessment algorithm, which are both based on the concepts of knowledge space theory (Section 6.2). In Section 6.3, the three postulated models of Investigations I through III are implemented into the algorithms and the empirically obtained answer patterns are compared to the estimated knowledge states.

Finally, in Chapter 7, the empirical findings of the three investigations and the adaptive assessment are reviewed and integrated. A short outlook on further research with respect to an efficient and comprehensive diagnostic instrument for inductive reasoning concludes this report.

2 Inductive Reasoning

Inductive reasoning abilities have been central in theories of human thinking already in the early stages of intelligence research. One reason is the assumption that inductive reasoning is highly associated with the general intelligence factor.

Within the factor–analytic tradition Spearman (as cited in Brody, 1992) believed that his factor g of general intelligence was mainly determined by inductive reasoning processes and Thurstone (1931) considered inductive reasoning as one of the primary mental ability factors. This view was also sustained in the hierarchical model by Cattell (1963; Horn and Cattell, 1966), who divided Spearman's g into the factors g_c of crystallized intelligence and g_f of fluid intelligence. Cattell found that g_f was defined by measurements of an individual's biological capacity to acquire knowledge and was highly determined by the factors inductive and spatial reasoning. Cattell's Culture–Fair tests, which were designed to assess g_f, are again constituted by inductive reasoning problems, such as analogies, matrices, classifications, or series continuations. The importance of inductive reasoning with regard to general intelligence was also confirmed by newer investigations in which also methods like LISREL or multidimensional scalings have been used (e. g., Gustafsson, 1984; Marshalek, Lohman, and Snow, 1983; Shye, 1988; Snow, Kyllonen, and Marshalek, 1984; Tziner and Rimmer, 1984; Undheim and Gustafsson, 1987).

In this chapter, I will first give a short introduction into the domain of inductive reasoning (Section 2.1), which is followed by a description of some typical inductive reasoning tasks (Section 2.2), including analogies (2.2.1), series completions (2.2.2), and matrices (2.2.3). The section on inductive reasoning tasks concludes with a discussion of the communalities and differences among the various problem types (2.2.4). In Section 2.3, I will introduce two prominent models for inductive reasoning, namely Klauer's cognitive training approach (2.3.1) and the cognitive components approach by R.J. Sternberg (2.3.2). The last part of this chapter (Section 2.4) deals with the psychometric approach to inductive reasoning and intelligence. It includes structure of intelligence models (2.4.1) as well as a selection of tests assessing the ability of inductive reasoning either as a subtest within a general test of intelligence or as specific test measuring only inductive reasoning abilities (Section 2.4.2). Furthermore, the most important findings with respect to this study will be reviewed in a short summary (Section 2.5).

2.1 General introduction to inductive reasoning

Reasoning, in general, involves inferences that are drawn from principles and from evidence, whereby the reasoner either infers new conclusions or evaluates proposed conclusions from what is already known (Johnson-Laird, Byrne, and Shaeken, 1992; Johnson-Laird and Byrne, 1993; Rips, 1990; Shye, 1988; Wason and Johnson-Laird, 1972). There are two main types of reasoning, namely deductive and inductive reasoning.

Whereas deductive reasoning denotes the process of reasoning from a set of general premises to reach a logically valid conclusion, inductive reasoning is the process of reasoning from specific premises or observations to reach a general conclusion or overall rule. Deductive inferences therefore draw out conclusions which are implicit in the given information, while inductive inferences add information (Bisanz, Bisanz, and Korpan, 1994; Klauer, 2001; Mayer, 1992).

Examples for deductive reasoning are judging the validity of propositional, categorical, or linear syllogisms, conditional reasoning like the Wason selection task, or mathematical deduction, in which the consequences are deduced from a set of axioms. Examples for inductive reasoning are language acquisition (inducing the rules of a grammar from a set of sentences), scientific induction (e. g., inducing a molecular structure or a formula from a set of numerical data), mathematical induction (e. g., the proof that a rule is valid for all natural numbers), or intelligence test tasks like classifications, analogies, series completions, or matrices (Goertzel, 1993; Greeno, 1978; Greeno and Simon, 1988; Schaefer, 1985).

While mathematical induction contains information about all instances in a class (e. g., the class of all positive integers) and therefore concludes with certainty, psychological induction usually refers to the given instances and does therefore reach conclusions that are not necessarily valid for all possible instances (Klauer, 2001; Schaefer, 1985). Thus, the inductive reasoner can only use probable conclusions to predict further instances (Evans, Newstead, and Byrne, 1993; Sternberg, 1999). A well known example for induction by Karl Popper (1972) is that after observing several instances of white swans the inductive reasoner draws the conclusion that all swans are white. Thus, the process of drawing inductive conclusions about general laws starts with single observations which are combined with the strength of previous observations in order to arrive at a conclusion. However, the derived conclusion is not necessarily accurate or logically valid as can be seen in this example. Nevertheless, in many cases the inductive inferences are valid and provide an important basis for the understanding of regularities. Regularities as well as uniformities are the basis for the generation of concepts and categories, which play a fundamental role in our every–day life (Klauer and Phye, 1994). Rips (1990) argued that we should therefore concentrate on the strength of the inductive conclusion rather than on the validity.

The development of a shared set of concepts is essential for the mutual understanding of human beings. These concepts vary between very concrete ideas, such as what we identify as a table, and more abstract ideas like truth or justice. From an analysis of the similarities and differences between specific experiences, we gather the defining attributes of objects and situations. Thereafter, we can refine and modify these generalizations by applying them to new objects and situations. By this, the derived concepts become part of our permanent knowledge base (Pellegrino, 1985).

The research within the field of inductive reasoning is very broad, ranging from reconstructing the mental processes involved in inductive problem solving (e. g., Dörner, 1976; Greeno, 1978; Holzman, Pellegrino, and Glaser, 1982, 1983; Mulholland, Pellegrino, and Glaser, 1980; Sternberg, 1977a,b) over the construction of computer programs in artificial intelligence (e. g., Carpenter, Just, and Shell, 1990; Ernst and Newell, 1969; Holland, Holyoak, Nisbett, and Thagard, 1986; Kotovsky and Simon, 1973; Michalski, 1983) to investigations on the effects of training on reasoning (e. g., Büchel and Scharnhorst, 1993; Klauer, 2001; Klauer and Phye, 1994; Lehman, Lempert, and Nisbett, 1988; Lehman and Nisbett, 1990).

At the same time the terms induction as well as reasoning are interpreted in a variety of ways (see e. g., Klauer, 2001, for distinctive definitions of the terms induction, inductive reasoning, inductive thinking, inductive inference, or problem of inducing structure). Just looking at the examples for inductive reasoning given above, it becomes clear that the meaning of induction differs depending on the respective area of research. In scientific research induction is used to either generate or to confirm hypotheses (Breuer, 1977). In both cases empirical data form the basis for the inductive processes. The inductive generation of hypotheses starts with the collection of empirical data, from which regularities and theoretical predictions are derived (e. g. by explorative methods such as factor or cluster analysis). The validity of an inductive hypothesis is strengthened by each collected data set,which confirms the hypothesis. An important characteristic of scientific induction is that the validity of the derived hypotheses can only become more probable, i. e. there always remains uncertainty about the unobserved instances. Thus inductive hypotheses are only falsifiable but not verifiable. Mathematical inductions, on the other hand, conclude with certainty. Starting with a mathematical statement or rule about a natural number and the proof, that the rule is true for a small sample of numbers, mathematical induction is a technique to show that the rule is true for the infinite class of natural numbers. A simple example (taken from Schaefer, 1985) is the inductive proof that the sum of the first n natural numbers ≥ 1 equals $\frac{n(n+1)}{2}$. As first step of the inductive process it is shown that the rule is true for the minimal case $n = 1$ ($1 = \frac{1\times 2}{2}$). In a second step, it is assumed that the rule is true for $n - 1$ ($\sum_{i=1}^{n-1} = \frac{(n-1)n}{2}$) and then shown that it also holds for n ($\sum_{i=1}^{n} = \frac{(n-1)n}{2} + n = \frac{(n-1)n+2n}{2} = \frac{n(n+1)}{2}$). Together, the two steps imply that the rule is true for all possible cases, i. e. the minimal case and all its successors. Thus, mathematical induction also generalizes to a whole class from a smaller sample, but as opposed to scientific induction, it gives information about every member of the class. In psychometric inductive reasoning, the testee's task is to extract a rule from a sequence of presented instances (e. g. symbols or numbers). In this case, the extracted rule applies only to the finite class of given instances and does not include generalizations for all possible cases. Since the rule is constricted to the observable cases, the application of the rule should also lead to a single correct response (by for example completing a given pattern by one or two further instances). Thus the different types of induction vary with respect to the size of the class, for which a rule is induced and with respect to the certainty of the conclusion. However, they have in common, that they all involve a generalization process from single instances to an overall rule. Thus, induction can be viewed as drawing conclusions from a smaller sample in order to reach a general rule governing all instances in the class under consideration.

For my research, the focus is on inductive reasoning as it is required in typical intelligence test problems (e. g., analogies, series completion problems, or matrices). Therefore, I refer to inductive reasoning as the process of reasoning from particular instances to reach a general conclusion, i. e. an overall rule that governs the relationships among the single instances.

With respect to inductive reasoning problems, as they are found in psychological aptitude and intelligence tests, the reasoner's task is to discover the pattern of relations among several elements in a given item. The skills required to solve such problems are apprehension of the presented relations and generating an integrated representation or overall rule of the pattern.

2.2 Inductive reasoning tasks

Individuals differ in their ability to solve inductive reasoning problems. In order to assess these differences and to study the underlying factors contributing to individual differences, a variety of tasks has been developed.

Besides the less restrictively structured problem solving contexts, such as Gick and Holyoak's (1980; 1983) research on building analogies between, e. g., 'The General Story' and Duncker's radiation problem (Gick and Holyoak, 1980, 1983, see Box 2.1) or Gentner's (1983; Gentner and Toupin, 1986) investigations on analogies between, e. g., the solar system and an atom, there are several types of inductive reasoning problems, which can be found in nearly every intelligence or aptitude test. Examples are classification problems, analogy problems, series completion problems, or matrices.

All the tasks found in psychometric tests have a common property. After the presentation of a set of stimuli, the testee has to infer the rules or pattern structure for the item and generate or select an appropriate completion or continuation of the pattern.

In the following sections, I will introduce several types of inductive reasoning problems and describe the problems' main features and their solution requirements. Facing the wide selection of proposed models in the cognitive psychology literature, I will concentrate on those problem types and respective models, which are relevant to my own research. The selection of problem types is based on the materials used in the three investigations conducted for this study (see Chapter 5). For the selection of models describing the problem types, my criterion was that the problems are described by components or attributes and that the models provide information on how the items of each problem type vary in difficulty.

The problem types included in this section are verbal and geometric analogies, number series completions, and geometric matrices. For other types of problems, as for example, pictorial or number analogies, pictorial, word, or letter series, verbal matrices, or various types of classification problems, the reader is referred to, e. g., Alderton, Goldman, and Pellegrino (1985), Holzman et al. (1982), Jacobs and Vandeventer (1972),

Box 2.1: Analogical transfer between stories

The General Story:
A general wished to capture a fortress located in the center of a country. Many roads radiated outward from the fortress, but these were mined so that although small groups could pass over them safely, any large group would detonate the mines. Yet the general needed to get his entire large army to the fortress in order to launch a successful attack. The general, however, knew just what to do. He divided his men into small groups and dispatched them simultaneously down multiple roads to converge the fortress. (Holland et al., 1986, p.291)

The Radiation Problem:
Suppose you are a doctor faced with a patient who has a malignant tumor in his stomach. It is impossible to operate on the patient, but unless the tumor is destroyed, the patient will die. There is a kind of ray that at a sufficiently high intensity can be used to destroy the tumor. Unfortunately, at this intensity the healthy tissue that the rays pass through on the way to the tumor will also be destroyed. At lower intensities the rays are harmless to healthy tissue but will not affect the tumor either. How can the rays be used to destroy the tumor without injuring the healthy tissue? (Holland et al., 1986, p.290)

Analogy:

Goal	Use force to overcome a central target.
Resources	Sufficiently great force.
Constraint	Unable to apply full force along one path.
Solution plan	Apply weak forces along multiple paths simultaneously.
Outcome	Central target overcome by force.

(from Gick and Holyoak, 1983)

Participant's task was to detect the analogy between the two stories and find a solution to the radiation problem after hearing the general story.

Klauer (2001), Mayer (1992), Scharroo and Leeuwenberg (2000), Schrepp (1999), and Sternberg and Gardner (1983).

2.2.1 Verbal and geometric analogies

Analogies are the most frequently and most intensively studied type of inductive reasoning problems and can be found in a large number of intelligence tests.

In general, analogy problems are of the form A is to B as C is to D ($A : B :: C : D$). Mostly, they are presented in a forced choice format with a three–term stem ($A : B :: C : ?$) and a set of answer alternatives D_i (standard format). Other variants of psychometric analogy problems are the presentation of the terms $A : B$ as stem pair and a set of alternative pairs $C_i : D_i$ to choose from (see e. g., Bejar, Chaffin, and

Embretson, 1991) or presentations in a true–false format (see e. g., Mulholland et al., 1980). Box 2.2 shows some example analogies of different content and format.

The participant's task is to infer the relation between the terms A and B, to apply this relation to term C, and to choose the correct alternative D_i, so that D has the same relation to C as B has to A (standard item format). The solution process involves the search in a space of relations to find a relation that can be applied to both $A : B$ and $C : D$. For the two most frequent types of analogy problems, viz. verbal and geometric ones, the relations are based on semantic memory and on feature analysis respectively (Greeno and Simon, 1988). Verbal analogies require the consideration of semantic relations and nuances of word meanings, whereas geometric analogies require an analysis of the terms' features and spatial transformations (Pellegrino and Glaser, 1979).

Dependent on the kind of presentation, the solution process furthermore involves either the selection of the correct answer alternative D_i (standard format), the judgment whether the terms $A : B$ and $C : D$ are related by the same rule (true–false format), or the selection of the correct pair $C_i : D_i$.

As I am interested in the difficulty structure of problems, I will next outline the main components contributing to item difficulty. Other research on analogy problems includes, among others, Pellegrino and Glaser's (1982) conceptual and interactive models, Evan's (as cited in Bejar et al., 1991) artificial intelligence model, Embretson's latent trait models (published as Whitely, 1980, 1981), Rumelhart and Abrahamson's (1973) semantic distance model, or Gentner's (1983) structural mapping theory on analogies.

2.2.1.1 Task requirements for verbal analogies

The basis for solving verbal analogy items is the knowledge of word meanings and of semantic relations between the words an item is composed of. Given general knowledge about word meanings and semantic relations, the problem solver has the task to identify the relevant semantic relation between the terms of the stem pair ($A : B$) and to apply the relation to the third term C. Then term D is either generated immediately or each answer alternative D_i is evaluated in order to choose the D_i term which matches the $A : B$ relation best (given the standard forced choice format).

Item difficulty can be described by three major factors, namely operation difficulty, rationale complexity, and degree of constraint (Bejar et al., 1991; Pellegrino and Glaser, 1980). *Operation difficulty* refers to the type of semantic relation connecting the pairs, *rationale complexity* is based on the number of relevant elements or concepts in the given relation, and *constraint* refers to the detectability of the relevant relation. Other factors contributing to item difficulty are word frequency, the use of concrete versus abstract words, or the semantic distance between the terms (Bejar et al., 1991; Rumelhart and Abrahamson, 1973).

Based on taxonomies by Chaffin and Herrmann (1984) and Whitely (1977), Bejar et al. (1991) developed a taxonomy of semantic relations (or *operations*) which consists of 10 families. Each of the families has between 5 and 10 members. Table 2.1 depicts

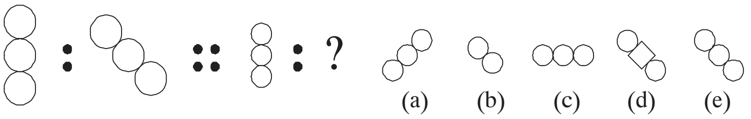
Box 2.2: Examples for psychometric analogy items (from Bejar et al., 1991, Pellegrino, 1985, and Pellegrino and Glaser, 1982)

Verbal analogies:

- sugar : sweet :: lemon : ?
  (a) yellow (b) sour (c) fruit (d) squeeze (e) tea correct answer: (b)
- concert : audience ::
  (a) restaurant : waiter
  (b) orchestra : musicians
  (c) game : spectators
  (d) school : cheerleaders
  (e) zoo : keepers correct answer: (c)

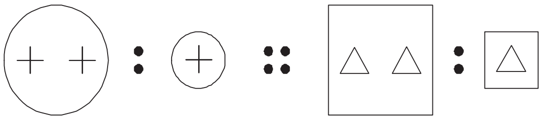
Geometric analogies

- standard format



correct answer: (e)

- true-false format



correct answer: true

The testee's task is to identify the relation between the first two terms, to apply it to the third term, and to either select the correct answer alternative or to judge, whether the analogy is true or false (last example).

the 10 families (1st column) with a short description each (2nd column), some exemplary members (3rd column) for each family, and an example (4th column) for each member. Bejar et al. (1991) analyzed a set of data for 179 GRE (Graduate Records Examination) analogy items, which they first assigned to one of the 10 families of semantic relations (each family contains between 7 and 30 items). By calculating the item difficulty parameter Δ[1] for each family, they found that the relation families *class inclusion* and *similar* are the most difficult ones to solve (in the given order $\Delta = 14.2$ and 14.0 compared to $10.85 \leq \Delta \leq 13.9$ for the remaining families). This result was also confirmed by Klix (1978, 1992) who differentiated between seven types of relations (attribute/quality, cause–purpose, class inclusion, comparative (similar), contrast, coordinates, and location).

Box 2.3 shows an example for *rationale complexity.* A verbal analogy item is expected

[1] Δ is a measure of item difficulty, which is derived from the percentage of correct solutions. It is defined in terms of a normal distribution with a mean of 13 and a standard deviation of 4.

Table 2.1: Taxonomy of semantic relations (adapted from Bejar et al., 1991)

Families	Descriptions	Members	Examples
Class inclusion	One word names a class that includes the entity named by the other word.	taxonomic functional	flower:tulip weapon:knife
Part–whole	One word names a part of the entity named by the other word.	mass:portion object:stuff	water:drop glacier:ice
Similar	One word names a different degree or form of the quality, object, or action represented by the other word.	conversion comparative coordinate	grape:wine breeze:gale son:daughter
Contrast	One word names an opposite or incompatible of the other word.	contrary reverse	old:young love:hate
Attribute	One word names a quality, property, or action of the other word.	item:attribute object:action	beggar:poor glass:break
Nonattribute	One word names a quality that is not an attribute of the other word.	attribute:nonstate item:nonattribute	immortal:death bulwark:flimsy
Case–relation	One word names an action that the other word is usually involved in.	agent:object action:object	tailor:suit plow:earth
Cause–purpose	One word represents the cause, purpose, or goal of the other word.	cause:effect agent:goal	joke:laughter pilgrim:shrine
Space-time	One word names an entity that is associated with a particular location or time named by the other word.	item:location contiguity sequence	arsenal:weapon coast:ocean coda:symphony
Representation	One word names something that is an expression or representation of, or a plan or design for the other word.	expression representation plan	hug:affection person:portrait recipe:cake

to become more difficult, the more elements are contained in the rationale. Bejar et al. (1991) calculated the mean complexity (mean number of elements in the rationale) for each family and compared the derived scores to the difficulty parameters Δ. They found no significant relationship between complexity and difficulty. However, the analysis does not include a direct comparison of the number of elements and difficulty, i. e. the mean Δ for each level of complexity. The range of the number of elements within one family varies from zero to five (e. g., the family *class inclusion* contains only items with 2 elements, while the family *similar* contains items with 1, 2, 3, 4, and 6 elements).

The third factor contributing to item difficulty describes the degree of *constraint* on the set of possible answers for an item, i. e. the easiness to detect the relevant relation (Pellegrino and Glaser, 1980). For easy items (high constraint), the semantic relation can be specified immediately and a potential completion term is easily generated. The solution process follows a working–forward strategy, i. e. hypotheses are generated and tested and the processing of answer alternatives only involves a search for the hypothesized answer. An example for high constraint is the analogy ‘wolf:dog::tiger:?’.

Box 2.3: Example for rationale complexity in verbal analogies (adapted from Bejar et al., 1991)

Rationale	No. and kind of elements	Example
A is a member of *B*	1: membership	robin:bird
A is a verbal expression of *B*	2: expression, verbal	scream:fear
A is a device through which the flow of *B* is regulated	3: device, flow regulation	dam:water

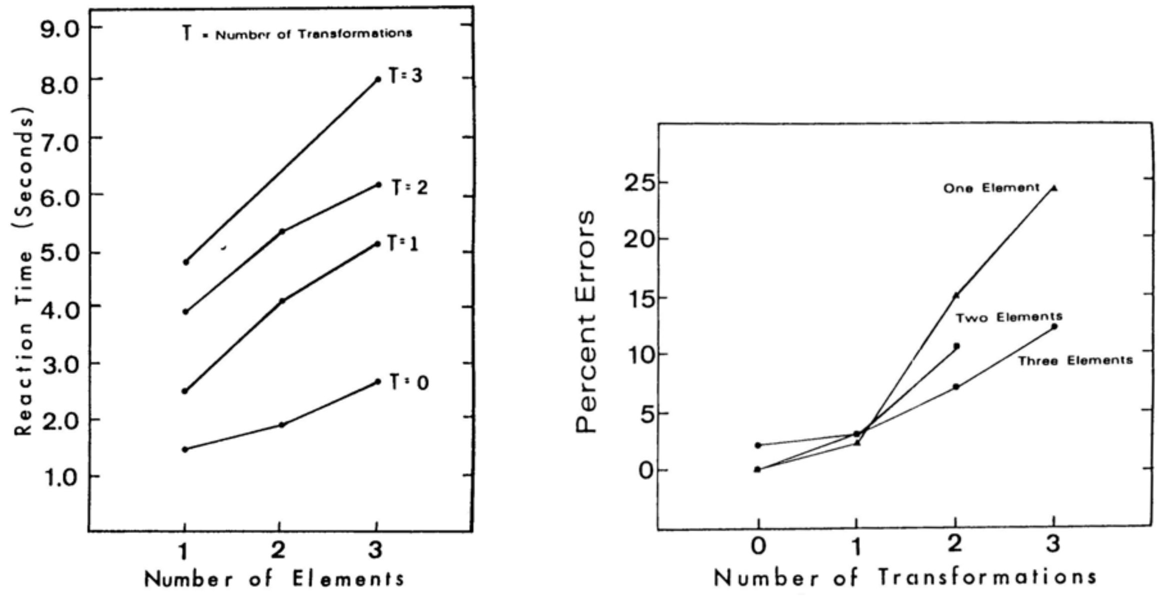
In items of high difficulty (low constraint) the relationship between the terms of the stem pair is not well specified and the item elicits more than one answer. The solution process is partially or completely guided by the set of answer alternatives, i. e. the problem solver uses a working–backward strategy. An example for low constraint is the analogy ‘city:village::army:?’. Presenting the two mentioned analogies in an open answer format, Pellegrino and Glaser (1980) found that the high constraint analogy elicited seven different responses, with 74% agreement on the answer ‘cat’. The low constraint analogy elicited 27 different responses with only 17% agreement on the most frequent answer ‘platoon’.

2.2.1.2 Task requirements for geometric analogies

The correct solution of geometric analogy items (see Box 2.2) requires processes to (a) decompose geometric figures into their constituent elements and (b) to identify specific transformations, which link the terms (Pellegrino and Glaser, 1980). Thus, the problem solver needs two types of declarative knowledge, namely (a) knowledge of the constituent elements used to construct the individual terms and (b) knowledge of the transformations that relate the terms. In general, the relations are found by examining the features of the terms. Looking, for example, at the true–false analogy in Box 2.2, terms *A* and *B* are decomposed into circles and plus signs and terms *C* and *D* into squares and triangles. Then, a transformation in size is applied to the circle and the square in terms *A* and *C*, and a transformation in number is applied to the terms’ plus signs and triangles.

Mulholland et al. (1980) constructed 460 true–false analogies with varying numbers of elements and transformations. The number of elements per term varied between one and three, the number of transformations between zero and three. Figure 2.1 depicts the obtained reaction times (a) and error rates (b) as a function of the number of elements and transformations for 240 true analogies.

The latency data in Figure 2.1a show that the solution time is a direct function of the number of elements and the number of transformations. Each additional element and each additional transformation results in an increase in solution time. This indicates that individuals decompose the patterns of an analogy item sequentially by



(a)

(b)

Figure 2.1: Reaction times (a) and error rates (b) for geometric analogies as a function of the number of elements and transformations (from Pellegrino and Glaser, 1980)

isolating the constituent elements one by one, as well as by performing the transformations in a serial manner. With regard to the error data, Figure 2.1b shows that only the number of transformations but not the number of elements influences the percentage of errors (a repeated ANOVA resulted in $F(2, 240) = 1.63, p > .05$ for the elements and in $F(3, 240) = 129.4, p < .001$ for the transformations). Furthermore, Mulholland et al. (1980) found that the the number of elements and transformations interacted ($F(4, 240) = 5.9, p < .05$), with the major portion of interaction variance being associated with the different trend for transformations in the one– versus two and three–element conditions. The most rapid increase in error rate occurred for items, in which several different transformations had to be performed on a single element. Thus, difficulties seem to arise from retaining and operating on the intermediate products of the transformations. Mulholland et al. (1980) conclude that with an increasing number of elements and transformations, it becomes more difficult to keep all of the performed steps in working memory, whereby the number of required transformations contributes more to item difficulty than the number of basic elements involved. Individual differences in the ability to solve geometric analogy problems is then related to differences in working memory capacity.

Besides the number of elements and transformations, geometric analogy items are mostly varied with geometric figures (e. g., triangles, circles, or squares) and background textures (patterns or shadings) as constituent elements and various types of basic transformations, such as removing, adding, rotating, reflecting, displacing, size changes, and variations in shading, form, shape, or number. The difficulty of the transformations varies in dependence of the elements' features. Features, that are directly perceptible, such as size changes or variations in shading or form are easier to transform than features that require a more abstract analysis, such as counting components

(Hunt, 1974). Furthermore, the application of spatial transformations was shown to contribute to item difficulty, because they are not as salient and are assumed to be not stored visually but acoustically (Posner and Mitchell, 1967).

A further factor contributing to item difficulty is that of constraint. As for verbal analogies, it refers to the easiness to detect the relevant relation or (given a forced choice format) to the likelihood that the stem terms A, B, and C elicit more than one answer D (see Section 2.2.1.1 for a more detailed description of constraint).

Summarized, item difficulty of geometric analogy items depends on the number of constituent elements of the terms, the number and type of required transformations, and the degree of constraint.

2.2.2 Series completions

Series completion problems are also part of many aptitude and intelligence tests, although they are not as frequently encountered as analogy problems. The types of serial patterns which are usually found in psychometric tests are primarily composed of letters or numbers (Pellegrino, 1985). Box 2.4 shows two examples each for number and letter series completions, of which one is presented in an open answer format, the other in a forced choice format. The general item structure is characterized by a set of elements (numbers or letters), which are ordered by one or more relations between the elements. The problem solver's task is to infer the relationship(s) and either generate the next item(s) in the series or select a correct completion of the series from a set of answer alternatives.

Simon and Kotovsky (1963; Kotovsky and Simon, 1973) proposed four basic components that are involved in the solution process for letter series completion problems and Holzman et al. (1983) showed that the same four components are also applicable to number series completions. In the following, a short description of each component

Box 2.4: Examples for number series completion items

Number series:

- 32 11 33 15 34 19 35 _ _ correct answer: 23 36
- 25 23 20 18 15 _
  (a) 12 (b) 17 (c) 13 (d) 14 correct answer: (c)

Letter series:

- j k q r k l r s l m s _ _ _ _ correct answer: m n t t
- a z b y c x d w _
  (a) e (b) x (c) v (d) c correct answer: (a)

The testee's task is to infer the relationship(s) between the elements of the series and to either generate or to select a correct completion of the series.

is given together with an example. The example refers to the first number series in Box 2.4. For letter series completion problems, the same solution steps apply, but numbers have to be replaced by letters and the relations refer to the distance of two letters within the alphabet (e .g same letter, next letter, predecessor, double–, or triple–next letter). The flow chart in Figure 2.2 illustrates the suggested processes involved in the solution of series completion problems. Performance is primarily determined by the following four processes:

1. *Relation detection:* Scanning of the series and generation of a hypothesis about the relationship among two or more elements of the series. In the first example in Box 2.4 scanning the series might first lead to the hypothesis that the elements 32 33 34 ... are related by the rule +1.
2. *Discovery of periodicity:* Using the information about the inferred relationship, the period length of the series is extracted (i. e. the number of elements that constitute one complete cycle of the pattern). The period length can be determined by checking, whether the relation is repeated at regular intervals (or for adjacent elements, such as 12 12 13 13 14 14, whether the relation is interrupted at regular intervals). In the above example, every other number is related by the rule +1, thus the period length is set to 2. Whenever the initially discovered relation is not repeated at regular intervals, the rule is discarded and a search for a new relation starts.
3. *Completion of pattern description:* Identification of rules which relate the remaining elements within the period (by applying the knowledge of the series' periodicity) and definition of a higher order rule for the full sequence. In the example, the remaining numbers are 11 15 19 which are related by the rule +4. The higher order rule in the relation can now be determined as adding 1 to location M_1 $[+1(M_1)]$ and adding 4 to location M_2 $[+4(M_2)]$. For the complete pattern description Holzman et al. (1983) use the notation $[M_1, +1(M_1), M_2, +4(M_2)]$.
4. *Extrapolation:* Completion of the series based upon the pattern description. For each answer blank the relevant rule is isolated and applied to generate the completion term. In the example, the first answer blank belongs to the cycle 11 15 19, and therefore the rule +4 is applied and the number 23 is generated. For the second answer blank the rule +1 is applied to the cycle 32 33 34 35 and the number 36 is generated to complete the series.

The difficulty of series completion problems is influenced by several components, including the type of relation that has to be detected, the period length of the series, and the number of elements in the pattern description. For the following more detailed report of the components and some corresponding empirical results, I will refer to number series completions, since this type of problem is also part of my study and therefore included in the classification scheme for inductive reasoning developed in Chapter 5.

According to Holzman et al. (1983), there are several components, which influence the difficulty of number series completion problems. Firstly, the *types of relations* involve a variety of arithmetic operations, such as addition, subtraction, multiplication,

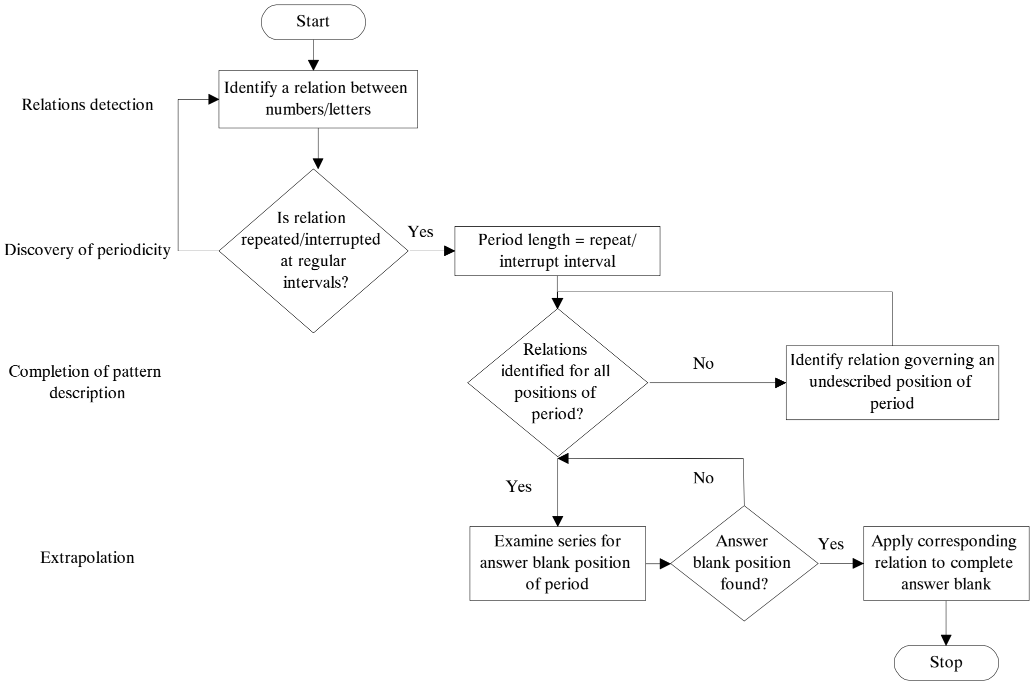

Start
Relations detection
Identify a relation between numbers/letters
Discovery of periodicity
Is relation repeated/interrupted at regular intervals?
Yes
Period length = repeat/ interrupt interval
Completion of pattern description
Relations identified for all positions of period?
No
Identify relation governing an undescribed position of period
No
Yes
Extrapolation
Examine series for answer blank position of period
Answer blank position found?
Yes
Apply corresponding relation to complete answer blank
Stop

Figure 2.2: Flowchart representing the sequence of information–processing during the solution of series completion problems (modified from Holzman et al., 1983)

or division. Secondly, the *magnitude of the arithmetic operation* varies, by which the elements are related (e. g. , +4 vs. +14). The third component contributing to item difficulty is the *complexity of the operation*, which can be increased by applying hierarchical sequences to the operations. As an example, the operation relating the elements of the problem 12 13 16 25 52 is an addition, but the magnitude of the addend is multiplied by three for each step in the sequence. The pattern description for this sequence is $[M_1, +N_1(M_1), x3(N_1)]$, with N denoting the placekeeper for the hierarchical operation. The *length of the period* constitutes the fourth component.

The complexity of the operation and the periodicity of the serial pattern influence the amount of information that must be held and coordinated in working memory, i. e. the number of required *working memory placekeepers* (WMPs). The number of WMPs per item results from the number of operations (M_i in the pattern description) and the number of hierarchical sequences (N_i in the pattern description).

Holzman et al. (1983) developed a set of number series completion items that varied, among others, in the type, magnitude, and complexity of the required operation, in the number of WMPs (0–3), and in period length (1–3). Holzman et al. presented 90 test items to three groups of different ability level (university students, high and average ability children with N = 18 each). Performing ANOVAs for the various components, they found significant effects for the type of operation ($F[1, 1887] = 169.85, p < .001$ with additions and subtractions being easier than multiplications and divisions), for the operation's magnitude ($F[1, 1887] = 67.78, p < .001$ with low magnitude operators being easier than high magnitude operators), for the presence vs. absence of hierarchical sequences ($F[1, 1887] = 20.43, p < .001$), and for the number of WMPs ($F[1, 1887] = 276.89/599.06/48.85, p < .001$ for comparisons between 0 and 1-

3/1 and 2-3/2 and 3 WMPs respectively). The component period length was evaluated within a multiple regression analysis on the proportion of correct solutions. Holzman et al. found no significant effect arising from the length of a period ($t = .32/.99/1.10, n.s.$ for the average, high, and adult ability groups respectively). Overall, the number of WMPs contributed most to the differences in item difficulty. Individual differences in the ability to solve number series completions can thus be related to differences in working memory capacity.

One problem often encountered in series completion problems is that of ambiguity, which occurs when a given sequence can be extrapolated by more than one rule. Korossy (1998), who took a formal approach to the description of rules applied in number series problems, emphasized two consequences that follow from different solution formulae. In one case, different solution formulae lead to identical extensions of the number series, but the solution times differ. Although the solutions are correct no matter which rule is applied, the test scores are biased. Especially in the usually administered speed–power tests, higher solution times per item will lead to lower test scores. In the second case, different solution formulae lead to different extensions of the series. This case is even more critical, because the extensions which were not intended by the test constructors will be judged as incorrect. Box 2.5 shows an example for each of the two cases. In problem (a), the number series can either be solved by the formula $a_{i-1} + 2$ or by the formula $2a_{i-1} - a_{i-2}$ (a_i is the number to be found, a_{i-1} the preceding number, etc.). Both formulae yield the same result, but the second formula takes more time to process all of the elements and to generate the correct response. Problem (b) in Box 2.5 depicts a number series, for which the application of different solution rules leads to different extensions of the series, and therefore to answers that are scored as incorrect. One way to deal with this problem is the use of a set of answer alternatives the participant has to choose from. By this procedure it is possible to rule out several alternative rules by excluding the respective answers from the set of alternatives. With respect to item difficulty, it is necessary to differentiate between varying degrees of constraint when using multiple choice formats.

Other approaches to the solution and construction of series completion problems have been developed within the framework of knowledge space theory (see Chapter 3). Schrepp (1995, 1999) derived a quasi ordinal knowledge space for letter series completion problems which is based on the four described solution processes and Wriessnegger, Janzen, and Albert (2002) developed an eye movement model, which assigns eye movements to the first three solution processes (relation detection, discovery of periodicity, and completion of pattern description). Both investigations support the process model by Simon and Kotovsky (1963) and can be used to predict the difficulty and interdependency of letter series completion problems. Number series completions have been investigated by Albert and Held (1994, 1999) and Ptucha (1994). I will outline the knowledge space based approach by Albert and Held in Section 3.5.1.

2.2.3 Geometric matrices

The most prominent example for matrix problems are probably Raven's Progressive Matrices tests (Raven, 1958, 1965, 1976, see also Section 2.4.2.3), which have been

Box 2.5: Examples for ambiguity in number series completions (adapted from Korossy, 1998)

- Two solution formulae lead to identical extensions:

  (a) 1 3 5 7 9 11

  - formula 1: $a_{i-1} + 2 \longrightarrow$ solution: 13 15
  - formula 2: $2a_{i-1} - a_{i-2} \longrightarrow$ solution: 13 15

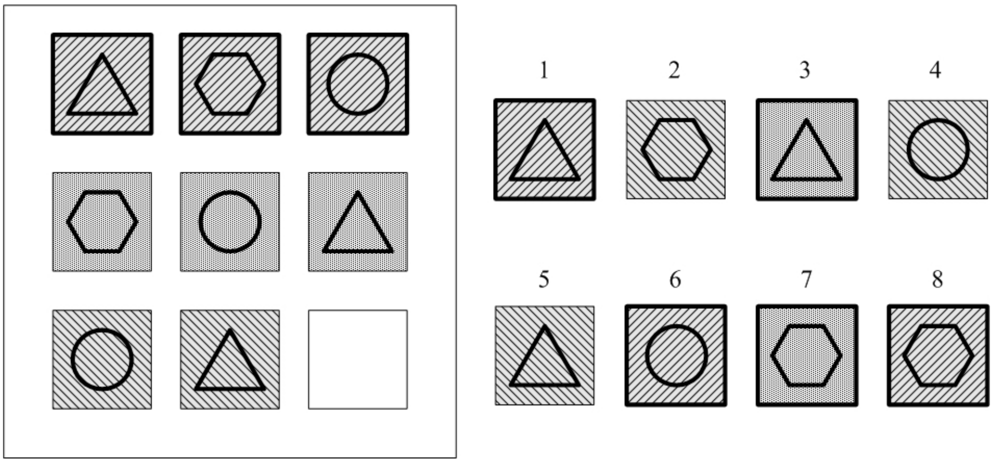
- Two solution formulae lead to different extensions:

  (b) 2 2 2 4 6 10

  - formula 1: $a_{i-1} + a_{i-2} + a_{i-3} - 2 \longrightarrow$ solution: 18 32
  - formula 2: $a_{i-1} + a_{i-2} - a_{i-3} + 2 \longrightarrow$ solution: 14 20

widely applied in both practice and research settings. Geometric matrix items are also often found in intelligence tests. Because of their high loadings on the general intelligence factor g (see e. g., Carroll, 1993; Marshalek et al., 1983; Paul, 1986; Tziner and Rimmer, 1984), there are also several tests that consist only of geometric matrix items, such as the Figure Reasoning Test (FRT by Daniels, 1993), the Vienna Matrices Test (WMT by Formann and Piswanger, 1979), or the already mentioned Coloured, Standard, and Advanced Progressive Matrices (CPM, SPM, and APM by Raven, 1976, 1958, 1965).

Box 2.6: Example for a geometric matrix item



correct answer: alternative 2

The testee's task is to infer the relationships among the figures in the matrix and to select the answer alternative which correctly completes the matrix.

Box 2.6 provides an example for a geometric matrix item (the item is invented, but similar to the items found in Raven's Progressive Matrices or in the WMT). The matrices are usually composed of a diagram with several figures arranged in rows and columns with one part missing. Items can be presented as, e. g., 2 by 3, 2 by 4, or 3 by 3 matrices. The problem solver's task is to induce the relationships among the figures and either to generate the missing element or to select the correct answer from a set of alternatives to complete the diagram. Variations in the entries of the matrix are based on different figural elements (such as triangles, circles, or squares) and different background textures (such as patterns or shadings), which can vary in form, number, spatial orientation, color, etc.

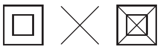
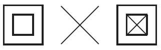
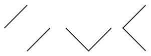
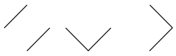
Carpenter et al. (1990) as well as, for example, Hunt (1974) developed algorithms to simulate the processes involved in the performance on Raven's Advanced Progressive Matrices (APM). In order to explain the processing characteristics responsible for individual differences in performance, Carpenter et al. (1990) developed two models, viz. Fairraven and Betterraven. Fairraven models the performance processes of moderately skilled university students, whereas Betterraven is an enhanced version representing the solution processes of highly skilled students. The relatively good match of the computer simulations' and the students' error profiles (i. e. the number of errors per problem) as well as eye–movement data, and self–reports lead to the following explanation of what APM actually assess and how students solve the problems. Carpenter et al. (1990) found that the APM assess the common ability to decompose the problem into manageable units of processing and to iterate through the emerged subgoals one at a time. Besides this common ability, the more difficult problems also require the differential ability to manage the hierarchy of (sub)goals and to form higher level abstractions. The distinction between moderately and highly skilled students is the more or less successful goal management, i. e. the ability of better students to generate and manage the problem solving goals and subgoals in working memory.

Verguts, De Boeck, and Maris (1999) suggested that a rule generation process plays a crucial role in solving geometric matrix items. This generation process can either result in qualitative or in quantitative differences. In their work, Verguts et al. concentrated on the generation speed or response fluency, which they found to be an important factor in solving the problems. The consequence is that testees with higher generation speed have a higher probability to solve the items. Especially in the case of speed–power tests, lower test scores do therefore not necessarily imply that the testees are not able to solve the items when given enough time.

With respect to item difficulty, Carpenter et al. (1990) proposed three factors contributing to the complexity of APM, namely the number of sub–problems, the type of rule governing the variations among the elements, and the difficulty in correspondence finding.

The *number of sub–problems* refers to the number of rules that are required to solve a matrix problem correctly. Usually, the number of sub–problems varies between one and four rules. In one of their experiments, Carpenter et al. (1990) found that the error rate ranges between 16% of errors for the application of only one rule up to 59% of errors for the application of three or four rules. Carpenter et al. attributed the increasing number of errors with an increasing number of rules to the difficulty of keeping track

Table 2.2: Rules for the solution of geometric matrices (from Carpenter et al., 1990; Kinder and Lachnit, 1994; Musch and Albert, 2003, see also text)

Rule	Description	Example
Constant in a row (CR)	The same attribute value occurs throughout a row, but changes down a column.	a a a b b b c c ?
Constant in a column (CC)	The same attribute value occurs throughout a column, but changes along a row.	a b c a b c a b ?
Quantitative pairwise progression (PP)	Increment or decrement between adjacent entries in an attribute such as number, size, or position.	a b c b c d c d ?
Distribution of 2 values (D2)	Two attribute values are distributed through a row, the third differs.	a a b a b a b a ?
Distribution of 3 values (D3)	Permutation of three attribute values through a row.	a b c b c a c a ?
Figure addition (FA)	A figure from one column is added to another figure to produce the third.	
Figure subtraction (FS)	A figure from one column is superimposed to another one to produce the third.	
Exclusive–OR (XO)	Two attributes are combined in the exclusive-or fashion of Boolean algebra[a].	
Boolean AND (BA)	Two attributes are combined in the AND fashion of Boolean algebra[b].	

Note. [a]If an attribute value occurs in exactly one of the first two entries it also occurs in the third entry, if it occurs in both entries it does not occur in the third entry, which leads to the following truth table: 11 ⇒ 0, 01 ⇒ 1, 10 ⇒ 1, 00 ⇒ 0 (1 = attribute present, 0 = attribute absent). [b]If an attribute value occurs in both or neither of the first two entries it also occurs in the third entry, if it occurs in only one of the first two entries it does not occur in the third entry, which leads to the following truth table: 11 ⇒ 1, 01 ⇒ 0, 10 ⇒ 0, 00 ⇒ 1.

of the already inferred rules while inducing the third or fourth rule. This means that each rule imposes additional load on working memory.

Table 2.2 depicts an extended list of the *types of rules* that Carpenter et al. (1990) specified for the solution of geometric matrices. The list of rules covers the findings by Carpenter et al. (1990), as well as Hornke and Habon (1984), Hunt (1974), Jacobs and Vandeventer (1972), Kinder and Lachnit (1994), Klix (1978), Musch and Albert (2003), and Vodegel Matzen, van der Molen, and Dudink (1994). Vodegel Matzen et al. (1994) constructed 52 Raven–like geometric matrices based on the five solution rules from Carpenter et al. (1990), viz. the rules constant in a row (CC), quantitative pairwise progression (PP), distribution of two (D2) and three (D3) values, and figure addition (FA)/figure subtraction (FS). Performing an error analysis, they found that most errors were due to omitting one or more rules. Vodegel Matzen et al. (1994) constructed the eight distractors per item in such a way that the types of omitted rules could be inferred from the incorrectly chosen answer alternative. They found a linear increase in error rate for the type of required rule. The lowest mean omission score occurred for the rule CC ($M = 2.8$ errors, $SD = 2.85$) and the highest score for the rule D2 ($M = 6.2$ errors, $SD = 3.7$). The error rates for the remaining rules are in between CC and D2, with increasing rates from PP over FA/FS to D3. The two Boolean rules exclusive–OR (XO) and Boolean AND (BA), which are expected to exceed all other rules in difficulty (Haygood and Bourne, 1965; Kinder and Lachnit, 1994), were not included in the analyses of Carpenter et al. (1990) or Vodegel Matzen et al. (1994).

Finally, the difficulty in *correspondence finding* refers to the easiness of detecting which elements are governed by the same rule. This last component corresponds to the degree of constraint (see Sections 2.2.1 and 2.2.2) which determines, whether a working–forward or a working–backward strategy is applied. The difficulty in correspondence finding is on the one hand influenced by the number of figural elements or attributes that vary across a row (Carpenter et al., 1990). On the other hand, also the salience of material attributes influences the difficulty in correspondence finding (Musch and Albert, 2003; Posner and Mitchell, 1967), with the attribute spatial order being more difficult to detect than other attributes such as variations in geometric figures, shadings, or patterns.

Other approaches to the investigation of matrices are the two–step theory (variation and retention) of how people solve series of items by Verguts (Verguts, Van Nijlen, and De Boeck, 1999; Verguts et al., 1999), the knowledge space approach by Musch and Albert (2003, see Section 3.5.2), or linear logistic models. Formann (1973), for example, developed the Vienna Matrices Test ('Wiener Matrizen Test', short WMT; Formann and Piswanger, 1979), which is based on a well-balanced construction of items and a linear logistic analysis for Rasch-homogeneity (see Section 2.4.2.4).

2.2.4 Communalities and differences

Looking at the descriptions of the four discussed problem types, there are several components which are common to all of them. With respect to my own research, I want to concentrate on the common components that influence item difficulty.

According to Klauer (2001), inductive reasoning problems require the ability to detect similarities and/or dissimilarities of attributes or relations (see Section 2.3.1). Similarly, Pellegrino and Glaser (1979, 1980) attribute inductive reasoning performance to the abilities to extract relations among the elements of a problem, to assemble those relations into a rule governing the entire problem, and to maintain and update the accumulated results in working memory. Item difficulty then increases with an increasing number of different operators that must be represented in working memory and with lower constraint on the possible rules.

Reviewing the components found for the four presented problem types, there are three major components (or factors) contributing to item difficulty. These are (a) operation difficulty, (b) relational complexity, and (c) the degree of constraint.

Variations in these components have been described for all problem types mentioned in the last three sections (2.2.1, 2.2.2, and 2.2.3), although some of the components' labels differed. Table 2.3 gives an overview of the components, including a short description and the respective labels for each of the discussed problem types. Additionally, the reported components that are specific to each problem type (i. e. the components in which the problem types differ) are listed in Table 2.3. The components that are specific to each problem type (as, e. g., word frequency for verbal material or the number of constituent elements for geometric material) require different abilities in the problem solving process and therefore, should also influence item difficulty. For this first attempt to predict the interdependencies between items of various inductive reasoning tests, I chose a more general model for the classification of problems, in which only the components that are common to all problem types are considered (see Chapter 7, classification scheme, for a discussion of this issue).

The component (a) *operation difficulty* refers to the type of operation that needs to be extracted. For each type of problem there are one or two types which are assumed to be more difficult to extract than all other types of possible operations. The more difficult operations are class inclusion and similar/comparative for verbal analogies, variations in space and number for geometric analogies, hierarchical sequences for number series completion problems[2], and the exclusive–OR and Boolean AND rules for geometric matrices. Component (b) *relational complexity* refers to the number of operations governing the entire problem. For verbal analogies, this component was referred to as rationale complexity (which denotes the number of elements in the semantic rationale), for number series completion problems as period length (i. e. the interval in which a relation is repeated and thus, the number of relations contained in the series). Describing geometric analogies and matrices, I simply spoke of the number of transformations and rules or subproblems. The last component (c) *constraint* is identical for all problem types. For matrix problems it was also referred to as difficulty in correspondence finding. It refers to the easiness of extracting the relevant relation. This last component differentiates between high degrees of constraint (or low salience), which allow a working–forward strategy and low degrees of constraint (or high salience), which require a working–backward strategy.

In addition to the three mentioned components, which can be specified for each single

[2] As reported in Section 2.2.2, the results of Holzman et al. (1983) indicate that the number of WMPs contributes most to item difficulty, whereas the presence of hierarchical sequences and the period length (number of operations) did not always show a significant effect. However, since the number of WMPs is composed of the number of the hierarchical sequences and operations, I will consider these two components separately.

Table 2.3: Communalities and differences for inductive reasoning problems

		Label			
Component	Description	Verbal analogies	Geometric analogies	Number series	Geometric matrices
operation difficulty	variations in the type of applied rules	operation difficulty (type of semantic relation)	difficulty (type) of transformation	operational complexity (hierarchical sequences)	type of rule
relational complexity	variations in the number of applied rules	rationale complexity (number of elements in the semantic rationale)	number of transformations	period length	number of sub–problems (rules)
constraint	easiness of detecting the relevant relation	constraint	constraint	constraint	correspondence finding
type specific components		word frequency, semantic distance, concrete vs. abstract words	number of constituent elements	magnitude and type of the arithmetic operation	material attributes

problem type, the varying types of content are also a factor contributing to item difficulty. Sternberg and Gardner (1983) found that the processing of geometric–figural material leads to longer solution times than the processing of verbal material (mean latencies for geometric material = 5.43 sec compared to 3.5 sec for verbal material; latencies are averaged over analogy, series completion, and classification problems). Whereas verbal stimuli are composed of only one information unit (a word), geometric stimuli are composed of several figural elements, which need to be decomposed before the relevant rules can be inferred (Pellegrino, 1985). Since numerical stimuli are also composed of only one information unit (a number), the processing of numerical material should also be faster than the processing of geometric material.

Under consideration of my own research aims, the mentioned communalities are of special importance, because they permit the establishment of a common structure for the items of the various problem types. Thus, summarized, inductive reasoning problems can be described by three general components (operation difficulty, relational complexity or number of operations, and constraint), which can be applied to all mentioned types of inductive reasoning problems. In addition, the applied material or content influences the difficulty of the problems.

2.3 Inductive reasoning models

In most studies only one or two types of inductive reasoning problems are investigated, as for example classifications (Alderton et al., 1985; Posner and Mitchell, 1967), analogies (Alderton et al., 1985; Bejar et al., 1991; Mulholland et al., 1980; Whitely and Barnes, 1979), series completions (Albert and Held, 1999; Egan and Greeno, 1974; Holzman et al., 1982, 1983; Klahr and Wallace, 1970; Scharroo and Leeuwenberg, 2000; Schrepp, 1999; Simon and Kotovsky, 1963), or matrices (Carpenter et al., 1990; Hornke and Habon, 1984; Hunt, 1974; Musch and Albert, 2003; Raven, 2000; Verguts and De Boeck, 1999). Since the aim of my own research is to integrate various problem types into a common structure of inductive reasoning problems, it is necessary to consider comprehensive models of inductive reasoning, which account for several problem types. Therefore, I will next describe two prominent models, which cover various types of inductive reasoning problems from different perspectives. First, I will outline Klauer's (1997; 2001) model, which is based on a paradigmatic training approach to inductive reasoning (Section 2.3.1). This is followed by Sternberg's (1977a,b; Sternberg and Gardner, 1983) information–processing approach to inductive reasoning (Section 2.3.2), which elicited a high number of investigations on different problem types.

2.3.1 Klauer's model of inductive reasoning

Klauer's (1996; 1997; 2001) model of inductive reasoning was developed within a paradigmatic training approach. On the one hand, it incorporates an exact definition that delimits inductive reasoning problems from other types of problems (e. g. deductive). On the other hand, it specifies the cognitive strategies for solving inductive reasoning problems.

Figure 2.3 presents Klauer's definition of inductive reasoning together with the processes that are necessary to solve inductive reasoning problems. The first part (i. e. the preceding sentence) renders a general definition of inductive reasoning, viz. the detection of regularities and irregularities. It also signifies the end product of the inductive reasoning process, namely the discovery of a generalization or the disproving of an assumed generalization. The second part in Figure 2.3, which is given in form of a mapping sentence[3], describes the strategy to be used to solve inductive reasoning problems. It contains three facets A, B, and C with three, two and five elements respectively. It is therefore possible to develop 30 (3 x 2 x 5) variants of inductive reasoning problems. Central to the model are facets A and B, which display the six basic types of inductive reasoning tasks (see below, Table 2.4).

Inductive reasoning consists of detecting regularities and irregularities by finding out

$$\overset{A}{\left\{\begin{array}{l} a_1 \text{ similarity} \\ a_2 \text{ dissimilarity} \\ a_3 \text{ similarity and} \\ \quad\;\, \text{dissimilarity} \end{array}\right\}} \quad \text{of} \quad \overset{B}{\left\{\begin{array}{l} b_1 \text{ attributes} \\ b_2 \text{ relations} \end{array}\right\}}$$

$$\text{with} \quad \overset{C}{\left\{\begin{array}{l} c_1 \text{ verbal} \\ c_2 \text{ pictorial} \\ c_3 \text{ geometric–figural} \\ c_4 \text{ numerical} \\ c_5 \text{ other} \end{array}\right\}} \quad \text{materials.}$$

Figure 2.3: Definition of inductive reasoning by Klauer (1994)

Facet A determines whether similarities or dissimilarities have to be detected. A class formation task, for example, requires the detection of common features among the elements, whereas a disturbed series requires the detection of a dissimilar element in the series.

[3] Mapping sentences were developed as a basic technique within Guttman's *Facet theory*. A mapping sentence is a formal method to classify a topic by specifying its content as a disjunctive and exhaustive set containing the concepts or *facets* of interest. Each facet consists of several elements that are combined by building the Cartesian product of the facets. The resulting set of all possible element combinations yields the products under investigation. To give another example, Tziner and Rimmer (1984) used a mapping sentence for the description of ability tests, were the facets are (a) the language of presentation (with elements such as verbal, numerical, or figural) and (b) the mental operation required by the test (with elements such as rule inference or variants of rule application). The products are the various ability tests (e. g., analogies, vocabulary, or arithmetic problems).

Table 2.4: The six variants of inductive reasoning problems (adapted from Klauer, 1994)

Processing classes	facet identification	problem types
Generalization	a_1b_1	class formation, class expansion
Discrimination	a_2b_1	identifying irregularities
Cross Classification	a_3b_1	4–, 6–, 9–fold scheme
Recognizing Relationships	a_1b_2	series completion, ordered series, analogy
Differentiating Relationships	a_2b_2	disturbed series
System Construction	a_3b_2	matrices

Facet B determines which elements have to be compared. The model distinguishes between attributes and relations, which are predicates with one or two and more arguments respectively. Similar attributes are, for example, identical colours or geometric forms of the elements in a class formation task, whereas similar relations are, for example, part–whole relations between the terms $A : B$ and $C : D$ in a verbal analogy task.

The comparisons which are presented by facets A and B are abstract and occur analytically. This means, that first it is necessary to consider individual attributes or relations by disregarding irrelevant elements. Secondly, the comparisons do not occur globally but attribute by attribute or relation by relation. The last facet C identifies the materials, i. e. the content of a given problem, such as verbal or geometric–figural material.

The six variants of inductive reasoning problems, which are given by the combination of facets A and B, constitute the basic classes of inductive reasoning processes (e. g. detecting similarity (a_1) of attributes (b_1) by comparing the problem's elements). Table 2.4 shows the resulting processing classes, some of their respective problem types, and the cognitive processes or operations required to solve the problems (facet identification). As an example, *generalization* is defined as the process of recognizing similarities (a_1) of objects or events by comparing their attributes (b_1), whereas *discrimination* is defined as the process of recognizing differences (a_2) among objects or events by comparing their attributes (b_1).

The six basic types of problems are related by the core strategy for solving inductive reasoning problems, namely the process of comparing. Coming from a training approach to inductive reasoning, the goal is to teach individuals solution strategies that enable them to solve all types of inductive reasoning problems. Therefore, Klauer (2001) developed an analytic and a heuristic strategy, both of which share the core process of comparing. The *analytical* strategy (see Figure 2.4) compares objects with respect to their common attributes (one place predicates) or relations (two place predicates). After evaluating all objects regarding the similarities and dissimilarities of all attributes or relations, the problem solver will discover the rule, and consequently the solution. The strategy assumes that the problem solver is able to recognize all at-

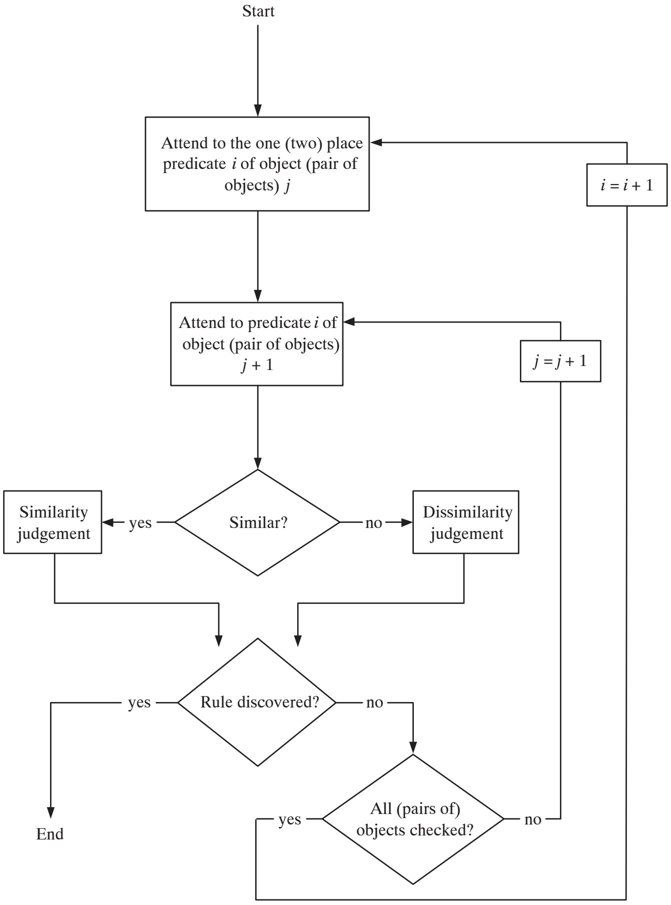

Start
Attend to the one (two) place predicate *i* of object (pair of objects) *j*
i = *i* + 1
Attend to predicate *i* of object (pair of objects) *j* + 1
j = *j* + 1
Similarity judgement
yes
Similar?
no
Dissimilarity judgement
yes
Rule discovered?
no
End
yes
All (pairs of) objects checked?
no

Figure 2.4: Flowchart representing the analytical strategy used for the solution of inductive reasoning problems (adapted from Klauer 2001; Klauer & Phye, 1994)

tributes or relations inherent in the problem. The strategy yields six solution variants dependent on the type of problem to be solved (i. e. whether attributes or relations are involved and whether similarities, dissimilarities, or both have to be detected). The flowchart in Figure 2.4 models the involved processes. The second strategy is based on the assumption that problem solvers often start with rather global hypotheses on the correct solution. They use a *heuristic*, hypothesis–guided strategy starting with global comparisons of the objects in order to test the generated hypothesis. Only when individuals are not successful in establishing a reasonable hypothesis on the problem structure, the more arduous analytic strategy must be employed.

The goal of Klauer's training approach is that the trainees become experts in inductive reasoning, i. e. that they acquire the competences necessary to solve inductive reasoning problems. Therefore, the trainees have to learn to recognize identical problem structures that differ only in their kind of presentation and content. The transfer of the learned solution strategies to new problems requires the identification of the problem type (generalization, discrimination, etc.) and the selection of the adequate solution strategy (detection of similarities or dissimilarities of attributes or relations).

Klauer (2001) performed a meta–analysis of 61 experiments to evaluate the effect of the training on the performance on inductive reasoning tests (mostly the Culture Fair Tests by Cattell and Weiss or Raven's Progressive Matrices). He found a positive training effect for all but one of the experiments with the effect sizes d ranging between -0.05 and 1.3, $M_d = 0.6$, $SD = .32$ ($d = 1$ indicates that a trained child performed by an average of one standard deviation better than an untrained child). An estimation of the effect size Δ across all 61 experiments resulted into $\Delta = .54, p < .001, N = 2632$. Further analyses to the long–term effect of the training, based on 17 experiments, also showed a significant training effect. The mean effect size right after the training amounted to $M_d = .77 (SD = .22)$ and stayed almost unchanged for the retests after 3-12 month, with $M_d = .7 (SD = .31)$.

2.3.2 R.J. Sternberg's cognitive components approach

Whereas Klauer (2001) classified various inductive reasoning problems and their corresponding solution strategies, Sternberg performed a detailed analysis of the cognitive processes underlying the solution of inductive reasoning problems.

Sternberg's (1977a,b; Sternberg and Gardner, 1983) cognitive components approach belongs to the broader field of information–processing. Within the information–processing framework, researchers try to identify the processes that underly performance. The goal is to answer questions about what the basic psychological processes are that are involved in solving psychometric tasks and which mental activities contribute to the interindividual differences, as they are assessed in psychometric tests. Mostly reaction times are used to investigate and separate the assumed cognitive processes. There are two main approaches, namely the cognitive correlates approach and the cognitive components approach.

The *cognitive correlates* approach has the aim to specify the information processing abilities that are related to different levels of aptitude. Questions are of the form: "What does it mean to be high in some ability?" To answer these questions, psychometric tests are used to identify subgroups of different ability, which are then compared on cognitive processing characteristics, such as decoding or holding information in long– and short–term memory respectively (Brocke and Beauducel, 2001; Mayer, 1992; Pellegrino and Glaser, 1979). These processes constitute the cognitive correlates of the respective psychometric scores. Representatives of the cognitive correlates approach include, among others, Beauducel and Brocke (1993), Hunt (1978, 1985), Neubauer (1995), or Schweizer (1995).

The aim of the second major approach, viz. the *cognitive components approach* is to break a task down into its underlying cognitive subprocesses (components[4]) and to assess individual differences in each of the component processes. In this task–analytic approach, questions are of the form: "What do intelligence tests measure?" To answer these questions, a problem is broken down into a list of component processes. The goal is to discover processes that individuals use in problem solving, from the time

[4] Note that within the cognitive components approach the term component refers to non–observable cognitive processes, such as encoding or comparing, while in Section 2.2 it was used to describe observable problem components (e. g. as type or number of operations) that influence the difficulty to solve an item correctly.

the problem is presented to the time an answer is given. In order to identify plausible models, performance measures such as latency or accuracy data and verbal self–reports are used (Bisanz et al., 1994; Mayer, 1992; Pellegrino and Glaser, 1979; Sternberg, 1977a,b). Furthermore, in connection with knowledge space theory (see Chapter 3), the component processes involved in letter series completion problems (see Section 2.2.2) have been investigated by means of predictable answer and latency patterns (Schrepp, 1995, 1999) as well as eye tracking data (Wriessnegger, 2000; Wriessnegger et al., 2002).

Whereas the majority of research within the cognitive components approach investigated the components of only one type of problem (e. g., Holzman et al., 1982, 1983; Mulholland et al., 1980; Sternberg, 1977a,b; Whitely and Barnes, 1979), Sternberg and Gardner (1983) conducted comparisons across several tasks to identify their common components. They started out with Sternberg's (1977a,b) model for analogy solution and extended their investigations to series completion and classification problems (including verbal, pictorial, and geometric material).

With regard to the four typical psychometric inductive reasoning tasks (classifications, analogies, series completions, matrices) Sternberg and Gardner (1983) and Pellegrino (1985) showed that the following five components can be used to describe the processes underlying the performance on all four tasks. For an example of each component process, I will refer to the verbal classification problem "(a) Furnace, Stove (b) Refrigerator, Air Conditioner – classify Oven" from Sternberg and Gardner (1983).

1. *Encoding processes* to create a mental representation of the given stimuli. In the above example, the testee has to encode the terms Furnace, Stove, Refrigerator, Air Conditioner, and Oven.

2. *Inference processes* to infer rules which relate the individual stimuli to each other. In the above example, the testee has to infer that the two terms in class (a) are objects used for heating and that the two terms in class (b) are objects used for cooling.

3. *Comparison processes* to compare the consistency of the inferred rules and to create a relational structure in memory. In the above example, the testee compares the term 'Oven' to the inferred concepts for (a) and (b).

4. *Justification processes* to select the best answer from a set of alternatives which does not contain the ideal answer or contains several possible answers; this component is only applicable for ambiguous items in multiple–choice formats. If the term 'Oven' does neither belong to the concept inferred for class (a) nor to the concept inferred for class (b), the testee has to justify one of the options as being closer to the ideal solution.

5. *Response processes* to give the chosen answer, i. e. overtly indicate which answer alternative (or in this case which class) is selected. In the example, the correct answer is (a).

Besides those five common processes, Sternberg and Gardner (1983) found two further processes which are required for the solution of analogies. The first one is *mapping* a

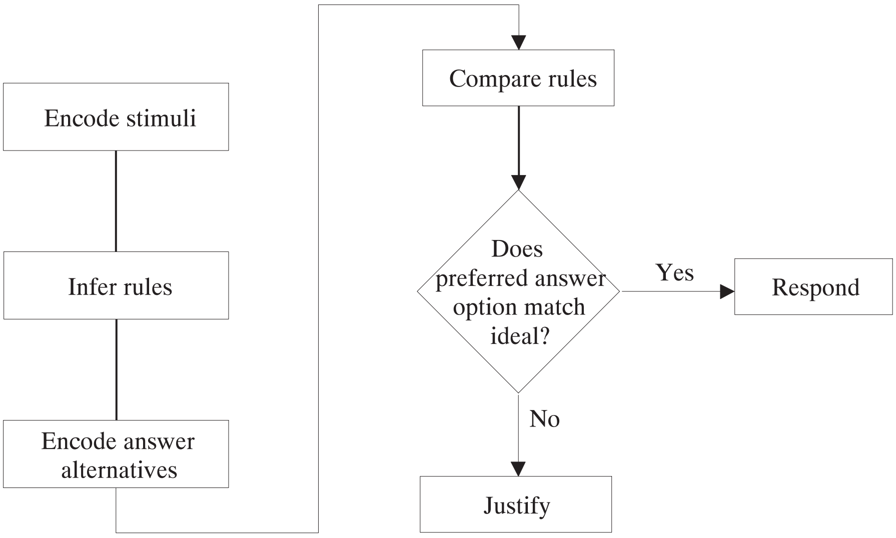

Encode stimuli
Infer rules
Encode answer alternatives
Compare rules
Does preferred answer option match ideal?
Yes
Respond
No
Justify

Figure 2.5: Simplified flowchart representing the sequence of information–processing during the solution of inductive reasoning problems (adapted from Sternberg and Gardner, 1983)

higher–order rule between the two parts of the analogy and the second is *applying* this rule to generate an ideal solution. The latter process is also required for the solution of series completion problems (in this case, the comparison process is followed by an application process, in which the inferred rule is applied to the relevant element of the series in order to generate an ideal completion). Matrices, as they constitute an extended form of analogies, require the same processes as analogies do. Generally, the processes are executed sequentially, but often the problem solver needs to perform multiple cycles through the various processes until a response can be made. Figure 2.5 shows a simplified flowchart which represents the sequence of information–processing during the solution of inductive reasoning problems.

In order to test the model of componential processing, Sternberg (1977a,b) introduced an experimental method called partial cueing. Showing the participants a variable number of terms (problem elements) before the entire problem (the three stem terms plus the answer alternatives) is presented, it is possible to extract the latencies for each component process. Box 2.7 shows an example analogy, the cues presented in the four precueing conditions, and the equations of Sternberg's basic solution model.

In the four–term analogy shown in Box 2.7, showing the participant the first term of the problem (Washington) should decrease the latency for encoding (parameter a in the model) by one term or 25% (assuming an additive model and sequential processing). By showing the first two terms of the analogy (Washington and 1), the latency for encoding should be cut in half and the time used for inference processes (parameter x in the model) should drop out.

In one of their experiments, Sternberg and Gardner (1983) presented nine types of task–content combinations (classifications, analogies, series completion and verbal, geometric, and pictorial material) to estimate the fit of the processing model. The correlations between the predicted and the observed values were significant for each of the nine task–content combinations ($.70 \leq r \leq .97, p < .05, Mdn = .82$). Furthermore, intercorrelations between tasks (collapsed over contents) and between contents (collapsed over tasks) supported the assumption of a single information–processing model for the different types of problems (r ranged between .96 and .97 for tasks and between .72 and .91 for contents).

The analysis of the components across tasks or contents (as pairwise comparisons) yielded the highest correlations for the comparison component (.40 – .80), followed by the reasoning component (including the three reasoning processes inference, mapping, and application; .10 – .69), and the justification component (.19 – .43). For the encoding component, all but one of the correlations were nonsignificant (-.08 – .43), indicating that there are different encoding processes involved in the tasks and contents. Comparing the component processes with psychometric tests measuring reasoning abilities, Sternberg and Gardner (1983) found the highest correlations between the psychometric scores and the component processes for reasoning (.50 – .79) and comparison (.61 – .75).

Comparing Sternberg's componential approach to Klauer's training approach (see Section 2.3.1), both models indicate that various types of inductive reasoning problems can be described by the following set of underlying cognitive processes. The two flowcharts in Figures 2.4 and 2.5, which illustrate the solution process for inductive reasoning problems, both involve the processes of encoding, inference, and comparison. In Klauer's model (Figure 2.4), encoding is given by the two entries "Attend to predicate i ...", inferences are necessary for similarity and dissimilarity judgments,

Box 2.7: Experimental design and basic additive model of Sternberg's (1977a) cognitive component approach

Example:
Washington : 1 :: Lincoln : ? (a) 10 (b) 5 correct answer: (b)

Terms presented in the first part of the precueing condition (with 0–3 cues):
0:
1: Washington
2: Washington : 1
3: Washington : 1 :: Lincoln

Equations for the basic additive model:
0: $ST_0 = 4a + fx + gy + fz + c$
1: $ST_1 = 3a + fx + gy + fz + c$
2: $ST_2 = 2a + gy + fz + c$
3: $ST_3 = 1a + fz + c$

ST = solution time, a = encoding time, x = inference time, y = mapping time, z = application time, c = constant response time, f = number of values changed from A to B, g = number of values changed from A to C.

and the comparison process for deciding, whether or not the overall "rule is discovered". Differences are found in the model specifications. Klauer's model yields a more detailed description of the differences among various problem types. Whereas Sternberg differentiates between different contents and tasks as well as the sets of necessary cognitive components per task, Klauer also specifies the types of elements that have to be encoded, inferred, and compared. More exactly, with facets A and B he differentiates between problems that require the detection of similarities and/or dissimilarities as well as between the detection of attributes or relations. However, Klauer's aim was to give a definition of inductive reasoning that also differentiates between different types of problems, whereas Sternberg's intention was to identify their common cognitive components. With respect to the evaluation of the two models, Klauer (2001), who was mainly interested in the training of inductive reasoning, validated his model by measuring the overall training effect of the proposed strategies. Sternberg and Gardner (1983), on the other hand, validated their model by investigating the adequacy of each of the assumed component processes by means of the described precueing procedure.

Essential for my own research is the result that various inductive reasoning problems can be subsumed within a single model and that the main components required for the solution of the problems are common to all problem types.

2.4 Psychometric approach to inductive reasoning and intelligence

Besides the cognitive training approach taken by Klauer (2001; see Section 2.3.1) and the information–processing approach taken by Sternberg (1977a,b; see Section 2.3.2), the psychometric approach to inductive reasoning is also of importance for this research. Klauer and Sternberg both showed that various types of inductive problems can be described by common components. This result is taken up for my own work, in which different problem types are structured within a common model of inductive reasoning tasks. This common structure is intended to form the basis for an adaptive testing system in the domain of inductive reasoning (see also Chapter 6). Thus, the psychometric approach, which is concerned with the development of tests has to be considered.

I will start out with a short overview of intelligence models, with the focus on structure models. Structure models of intelligence are closely related to my own work, but located on a higher level of classification. While structure of intelligence models identify the major components of intelligence and classify various problem types within their models, I want to identify the major components of inductive reasoning tests only and classify the single problems. After introducing some of the models used to describe intelligence and the placement of inductive reasoning within these models (see Section 2.4.1), I will describe some intelligence tests in order to show the different principles applied for the test developments (see Section 2.4.2).

Within the psychometric approach to the study of intelligence and interindividual differences the basic mental abilities are determined and then tests are developed to assess these abilities. The general idea is that humans are equipped with a set of cognitive factors and that individuals differ along these factors. The interindividual differences are reflected by the differences shown in intellectual performance on the tests. One main question within the psychometric approach is, how many cognitive factors are involved in intellectual performance (Mayer, 1992; Kail and Pellegrino, 1988).

Starting with Spearman's (as cited in Brody, 1992) two–factor–theory of intelligence, psychometric intelligence models based on factor analysis dominated the research in this area for several decades. Therefore, most of the traditional intelligence tests are based on factor analytic models but they vary with respect to the number and hierarchical order of the assumed factors. Spearman related performance on an intellectual task to only two factors (or factor sets), namely to a general intellectual ability factor g which is common to all tasks and to a set of specific factors s, each of which is specific to a single task. Spearman's theory implies that intelligence is best represented as an aggregate of diverse ability measures. Thurstone (1931), on the other hand, assumed several unrelated factors. He extracted seven independent primary mental abilities, namely perceptual speed, verbal comprehension, word fluency, induction/reasoning, memory, number, and space. A person's intelligence level should therefore be represented by one ability score on each of the seven primary factors as opposed to an aggregated score.

In Cattell's (1963, Horn and Cattell, 1966) hierarchical model of intelligence, Spearman's g is divided into the two factors g_c of crystallized and g_f of fluid general ability. Factor g_c measures the influence of schooling and acculturation, whereas g_f measures the ability to acquire knowledge and to adapt to new situations. Within each of the two major factors, there are several, more specific factors, from which the verbal ability factor (assessed, e. g. by vocabulary tests) loads high on g_c and the reasoning factor loads high on g_f (assessed, e. g. by Cattell's Culture–Fair tests). Later on Horn and Cattell (1966) elaborated the model by adding the factors general visualization g_v, general fluency f, and general speediness g_s to g_f and g_c.

Carroll (1993, 1994) analyzed 461 factor–analytic data sets and, thereafter, proposed his three–stratum theory of intelligence, which subsumes the factors found by Spearman, Thurstone, and Cattell. At the lowest level, Stratum I, there are over 40 rather specific abilities, as for example, lexical knowledge, memory span, perceptual speed, numerical facility, or induction. The eight second–order factors at Stratum II include, among others, fluid and crystallized intelligence, general memory, visualization capacity, or general cognitive speed. Finally, Stratum III constitutes the highest level with general intelligence g as the only factor.

Inductive reasoning abilities show high loadings on Spearman's g, Cattell's g_f, and Thurstone's primary ability induction/reasoning. In Carroll's model, induction is located at Stratum I as one of the abilities constituting the second–order factor fluid intelligence.

2.4.1 Structure of intelligence models

A different approach to the notion of intelligence was taken with the structure of intelligence models. Prominent models within this approach are Guilford's (1965; 1981) Structure–of–Intellect model and the Berlin Structure of Intelligence (BIS) model by Jäger (1982, 1984).

2.4.1.1 Guilford's Structure–of–Intellect model

Guilford (1965, 1981) based his model on a three–dimensional taxonomy of intellectual abilities or tasks. Figure 2.6 illustrates the model, which can be understood in terms of a cube that represents the intersection of three dimensions. Each ability and its corresponding tasks can be described by the three aspects kind of mental *operation*, kind of informational *content*, and kind of *product* information. Operations are mental processes (e. g., cognition C, memory M, or evaluation E), which are applied to the contents, i. e. the types of materials that appear in a problem (e. g., visual V, symbolic S, or semantic M). The results of applying an operation to a content are described in terms of products, which constitute the required responses (e. g., units U, classes C, or relations R). The possible combinations of five operations, five contents, and six products lead to 5 x 5 x 6 = 150 types of intellectual abilities, to each of which one or more types of mental tasks are assigned. Inductive reasoning problems assess, for example, the cognition of semantic and visual relations (CMR and CVR assessed by verbal and figural analogies or matrices), the cognition of symbolic systems (CSS assessed by letter or number series), or the cognition of semantic and visual classes (CMC and CVC assessed by verbal and figural classifications).

In the original formulation of his model, Guilford assumed the resulting combinations as independent primary factors. Later on (Guilford, 1981) he revised this view in favor of a hierarchical model, in which second–order factors are characterized by the combination of only two dimensions (e. g., CM for cognition of semantic content) and third–order factors by considering only one dimensions (e. g., cognition C).

2.4.1.2 The Berlin Structure of Intelligence model (BIS)

The BIS model (Jäger, 1982, 1984) is a descriptive system of intellectual abilities, which is based on three basic assumptions, namely (1) the multi–factorial conditionality, (2) the multi–modality principle, and (3) the hierarchical structure of abilities. With respect to (1) the model postulates that each intellectual achievement is determined by all intellectual abilities but that the weight of each ability varies. Assumption (2) states that intellectual achievements and abilities can be classified by different modalities or components. The model specifies a bimodal classification in the modalities operations and contents. The operation modality is divided into four ability components, the content modality into three components (see below and Figure 2.7). However, Jäger (1984) points out that the current model is expandable in the number of modalities as well as in the number of components per modality. Assumption (3) states that intellectual abilities are structured hierarchically, that is they can be differentiated on

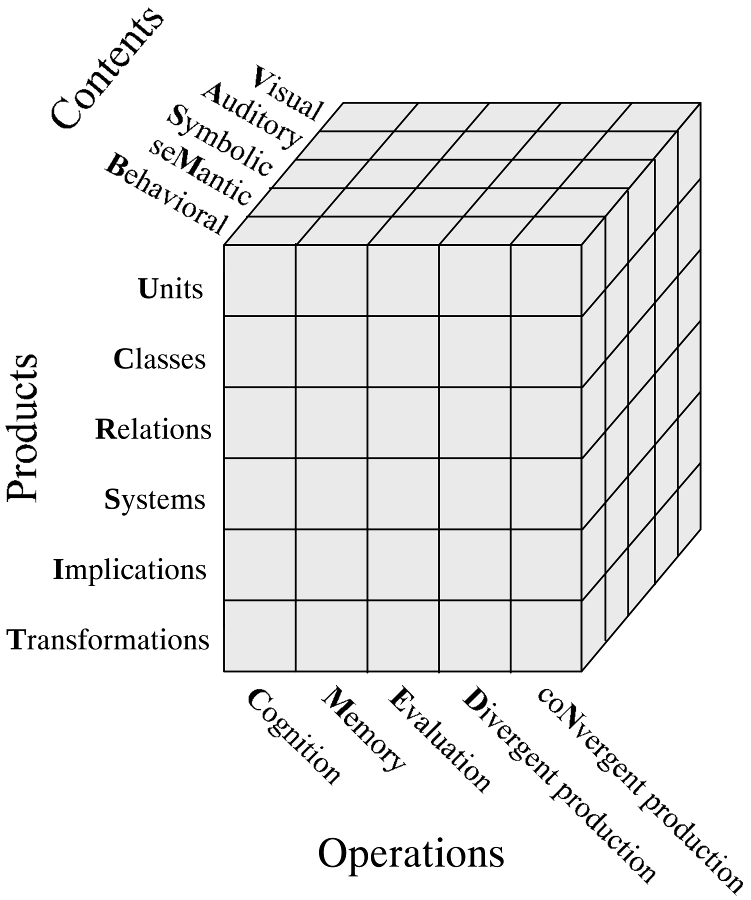


Figure 2.6: Guilford's Structure–of–Intellect model (adapted from Guilford, 1981 and Mayer 1992)

several levels of generality. Figure 2.7 illustrates the structure of the BIS model. On the most general level, the factor g_{BIS} of general intelligence is assumed as the integral part of all abilities (similar to Spearman's g or Carroll's Stratum III, see above). The seven ability components, which are arranged on the second level of generality, are comparable to second order factors (like Carroll's Stratum II). The combination of the four operational components and the three content related components yields 3 x 4 = 12 cells, to which the various intelligence tasks can be assigned. The 12 cells constitute the third level of generality. However, as opposed to Guilford's Structure–of–Intellect model (see above), the cells of the BIS model do not constitute primary factors but abilities that are determined by multiple factors (assumption (2) of the BIS model).

As shown in Figure 2.7 the four operations comprise the components reasoning (R), processing speed (S), memory (M), and creativity (C), which refer to the following abilities. Component R refers to the ability to process information in complex tasks. It includes inductive, deductive, and spatial reasoning (a more detailed description of the inductive reasoning tasks is given in Section 2.4.2.2. Component S refers to the ability to process simple tasks quickly but accurately. The memory component M refers to the ability to encode and recall or recognize sets of items and the creativity component C refers to the ability to produce ideas fluently and to look at things from different

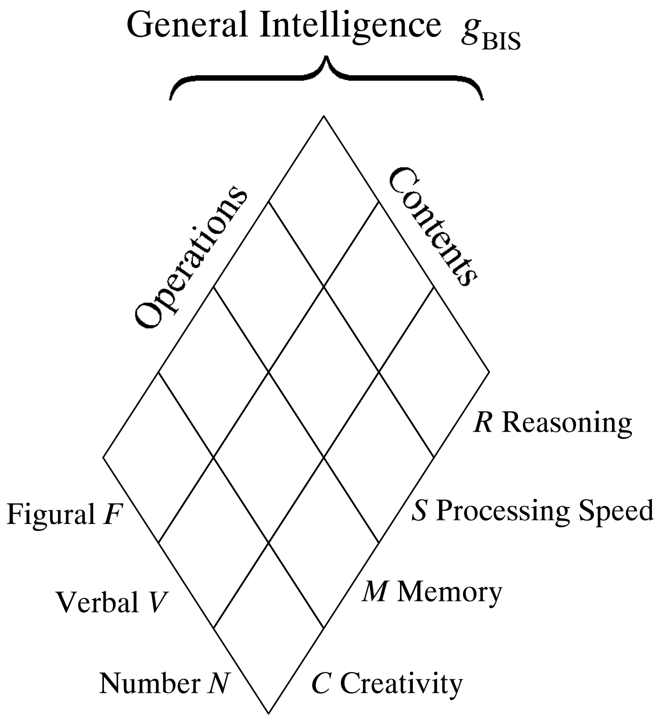


Figure 2.7: The Berlin Structure of Intelligence model (adapted from Jäger, 1984)

points of view. The second modality differentiates between three types of contents, namely spatial–figural (F), verbal (V), and numerical (N) content. The three content components refer to geometric–figural and spatial sense (F), and the acquisition and availability of the language (V) and number (N) systems.

The structure of the BIS model is based on a representative sample of intelligence test problems found in the literature to intelligence and creativity research. Jäger and his group analyzed 2000 different problem types, which they reduced to a set of 98 problem types. The selection criteria for the remaining 98 problem types included to maintain diversity and to maintain representations of the marking variables for other intelligence models, such as the models by Spearman, Thurstone, Guilford, or Cattell (see above). From these 98 problem types, 45 have finally been used for the BIS test (Jäger, Süß, and Beauducel, 1997, see Section 2.4.2.2). The selection criteria for these 45 problem types included, among others, an equal distribution of problem types per content component, how well the problems predict the operational and content related components, the elimination of bottom and ceiling effects, and adequate processing times. Jäger and Tesch-Römer (1988) could also replicate the BIS model with other sets of items, such as the 'Kit of Reference Test for Cognitive Factors' (French, Ekstrom, & Price, 1963, as cited in Jäger & Tesch–Römer, 1988) and Schmidt (1984) confirmed the content and operation related ability dimensions using LISREL.

Comparing the BIS model to Guilford's Structure–of–Intellect model (1965; 1981; Section 2.4.1.1), both models are based on a multi–modality concept, whereby the BIS model includes only a subset of Guilford's modalities (contents and operations, but not products). Differences between the two models are primarily found in the assumptions on the factor hierarchy. While Guilford originally assumed that the combinations of operations, contents, and products are independent primary factors (later on first–order factors), the BIS model conceives the combinations of contents and operations as multifactorial conditioned achievements.

2.4.2 Tests assessing the ability of inductive reasoning

Inductive reasoning tests belong to the group of intelligence or aptitude tests. Regarding the large but still growing number of existing intelligence tests (Brickenkamp's test compendium includes 57 published intelligence tests for the German speaking area), I will concentrate on a small selection of tests, which either assess inductive reasoning abilities by means of subtests within a general intelligence model or as a specific test including only one type of inductive reasoning problem, but constituting a good indicator for the factor g of general intelligence. For an overview of the prevalent tests available in the German-speaking area I refer the reader to Brähler, Holling, Leutner, and Petermann (2002).

The four selected tests are the Intelligence Structure Analysis (ISA), the Berlin Structure of Intelligence test (BIS test), Raven's Advanced Progressive Matrices (APM), and the Vienna Matrices Test (WMT). While the ISA and the BIS test are general intelligence tests, which consist of several subtests, the APM and the WMT are specific intelligence tests, which consist of only one type of inductive reasoning task (geometric matrices). The BIS test and the WMT are the two tests, from which I selected the material for my third investigation. I chose the ISA and APM for comparison, because they contain similar materials but are based on different principles regarding the test development. All four of the tests constitute traditional intelligence tests with fixed numbers of items. In Chapter 6, I will also discuss different approaches to adaptive testing.

With respect to the content of my own work, I will only report the basic concepts the tests are build on and focus on the problem types used to assess inductive reasoning abilities. For information on the tests' quality criteria (objectivity, reliability, and validity), assessment and evaluation procedures, norms, and areas of application, I refer the reader to the respective test manuals.

2.4.2.1 Intelligence Structure Analysis (ISA)

The ISA (Fay, Trost, and Gittler, 1998) was developed to assess primary cognitive abilities on a differentiated level (comparable to Thurstone's intelligence model, see above) and to provide a measure for general intelligence by aggregating the scores of the single ability dimensions. The authors of the ISA based their concepts and test materials on Amthauer's IST–70 (1973). However, the items of the ISA were newly constructed in order to overcome some weaknesses of the IST–70, as for example different language uses in Germany, Switzerland, and Austria or inconsistent item sequences (mostly but not always according to item difficulty).

Table 2.5: ISA problem types and their assignment to ability components (from Fay et al., 1998)

Problem Type	ISA	Thurstone	BIS[a]
SE Sätze ergänzen (sentence completion)	verbal	verbal comprehension	*RV*
GF Gemeinsamkeiten finden (detecting similarities)	verbal	verbal comprehension	*RV*
WM Waren merken (memorizing products)	memory	memory	*MV*
ZF Zahlenreihen fortsetzen (number series completion)	numerical	induction/reasoning (number)	*RN*
BE Beziehungen erschliessen (inducing relationships)	verbal	verbal comprehension	*RV*
WE Würfel erkennen (recognizing cubes)	figural–spatial	space	*RF*
PR Praktisches Rechnen (word problems)	numerical	reasoning (number)	*RN*
BB Begriffe bilden (forming concepts)	verbal	verbal comprehension	*RV*
FZ Figuren zusammensetzen (assembling figures)	figural–spatial	space	*RF*

Note. [a]R = reasoning, M = memory, V = verbal, N = numerical, F = figural.

The test construction is based on classical test theory, starting with the construction of an item pool and a stepwise selection of items based on difficulty and discriminatory power indices.

The ISA comprises nine problem types which can be assigned to the four factors or ability domains verbal intelligence (4 problem types), numerical intelligence (2 problem types), figural-spatial intelligence (2 problem types), and memory (one problem type). The nine item types cover five of Thurstone's seven primary abilities, the three content related components of the BIS model (see Section 2.4.1.2) as well as two of the four operational components specified in the BIS model. Table 2.5 depicts the nine problem types and their assignments to the four ISA factors, Thurstone's primary ability factors, and the BIS components. Although there are high intercorrelations among the subtests (r = .41 − .80) and all of the subtests correlate highly with the overall test score (i. e. with the general intelligence factor; r = .69 − .89), the test authors could reallocate the nine problem types within their respective factors by applying a confirmatory factor analysis.

Each problem type comprises 20 items (except for memorizing products with 17 items) and is preceded by instructions and warming up items. Inductive reasoning abilities are assessed by the three problem types GF (detecting commonalities), ZF (number series completion), and BE (inducing relationships). The items of the problem type GF consist of five words, four of which share a superordinate concept (e. g. fruit). The testee's task is to find the word that does not share a common concept with the remaining words. The number series completion items (problem type ZF) are presented in an open answer format and consist of seven elements each. The testee's task is to complete the series by generating the eighth element. The items of the last problem type measuring inductive reasoning abilities are verbal analogy problems (BE) presented in a standard item format ($A : B = C : ?$). The testee has to select one out of five answer alternatives to complete the analogy.

For the test results, the number of correctly solved items is recorded for each problem type and for the complete test. The raw scores are standardized by percentile ranks and T-scores. The standard presentation of the results is a table, but the test authors point out that a presentation as ability profile is also possible.

2.4.2.2 Berlin Structure of Intelligence Test (BIS test)

The BIS test (Jäger et al., 1997) is an instrument for measuring the intellectual abilities specified in the BIS model (see Section 2.4.1.2). The test scores derived from the BIS test can therefore be interpreted on the basis of the BIS model.

The test development is based on structural test theory, an extension of classical test theory, which allows a controlled portion of heterogeneity per subscale. The selection of problem types is based on the investigations for the BIS model, which started with the 2000 problem types the authors found in the intelligence and creativity literature. The final BIS test consists of 45 problem types (see Section 2.4.1.2 for the selection criteria) which can be located in the model as follows. For the operation component reasoning (R) there are five problem types per content component (i. e. spatial–figural F, verbal V, and numerical N), for the components speed (S) and memory (M) there are three problem types per content component each, and for the component creativity (C) there are four problem types per content component. Each problem type contributes to the overall score of three ability scales, namely to one of the operational scales, to one of the content related scales, and to the general intelligence scale (g_{BIS}). Table 2.6 depicts one example problem type for each component combination.

With respect to construct validity, a confirmatory factor analysis showed a high correspondence between the BIS test and the BIS model. The test authors report that the model structure was replicated without any allocation errors. The correlations between the abilities and their respective task bundles (problem types measuring the same ability scale) range between .62 and .72 for the operation components and between .6 and .73 for the content related components. The correlations between the single task bundels and the general factor g_{BIS} vary between .87 and .88.

With regard to inductive reasoning, there are seven problem types measuring this ability, viz. figural and verbal analogies (RF and RV), figural, number, and letter series completion problems (RF and twice RN), and figural and verbal classification

Table 2.6: Examples for problem types measuring the ability dimensions of the BIS model (adapted from Jäger et al., 1997)

g_{BIS}	Spatial–figural F	Verbal V	Numerical N
Speed S	number–symbol test (assigning symbols to given numbers according to a set of specified number–symbol pairs)	classifying words (crossing out words that name a plant)	arithmetic operators (inserting arithmetic operators in simple equations)
Memory M	figures (remembering the borders of company logos)	phantasy words (remembering pairs consisting of a german and a phantasy word)	number pairs (remembering pairs of numbers)
Creativity C	design of layouts (designing company logos for a small shop)	naming attributes (naming attributes people in a certain profession should not have)	phone numbers (constructing phone numbers with n digits that are easy to remember)
Reasoning R	geometric analogies	verbal analogies	number series

problems (RF and RV). The test items of each problem type are preceded by two practice items. The two types of analogy problems (figural and verbal) comprise eight test items each and are presented in a standard item format ($A : B = C : ?$) with five answer alternatives. The series completion problems comprise six (figural), eight (letter), and nine (number) items and are presented in an open answer format. The figural items consist of four elements each, the letter items of 8–15 elements, and the numerical items of five or seven elements. The testee's task is to complete the series by either two figures, two letters, or one number. Finally, the two types of classification problems comprise five (figural) and nine (verbal) items. The figural classification problems are presented as two sets A and B with six elements each. On the right–hand side three further elements are depicted, each of which the testee has to assign to either set A or set B. The verbal classification items consist of four words, of which three share a common concept (e. g. furniture). The testee's task is to find the word that does not share this concept.

For the analysis of the results, the BIS test includes a computer program, which converts the raw scores into percentiles, z–scores, and standardized scores with $M = 100$, $SD = 10$. Furthermore, the results are presented as ability profiles, with scores for each operational scale (based on 9 to 15 problem types), each content related scale (based on 15 problem types each), and a score for the general intelligence factor (based on all 45 problem types). The feedback for the participants includes a ranking of their abilities, but no comparisons with group norms.

As compared to the ISA (Section 2.4.2.1) with only nine problem types that assess four different ability dimensions (with one to four problem types each), the BIS test covers seven ability dimensions with 9 to 15 problem types each (the processing time for the ISA takes about 110 minutes, for the BIS about 130 minutes). In addition, the BIS test is based on a theoretically well founded model with a series of preliminary investigations (see Section 2.4.1.2). A further special characteristic of the BIS test is that, on the one hand, each problem type contributes to multiple ability scales, and on the other hand, that for each ability scale a relatively large number (9 to 15) and diversity of independent problem types contribute to the resulting ability measure.

2.4.2.3 Advanced Progressive Matrices (APM)

The APM are a speech–free intelligence test (based on classical test theory) for the assessment of reasoning abilities. Raven's intent when developing the progressive matrices (PM) CPM (Coloured PM, 1976), SPM (Standard PM, 1958), and APM (Advanced PM, 1965) was to create tests that are easy to administer and to interpret in a clear and theoretically relevant way. The main component assessed by Raven's tests is eductive or analytical reasoning ability, which refers to the ability to extract rules or educe relations (Carpenter et al., 1990; Raven, 2000). Raven (1948) himself pointed out that the matrices are not a test for general intelligence. However, Marshalek et al. (1983), as well as Tziner and Rimmer (1984), who applied multidimensional scaling techniques, allocated Raven's matrices at the center of the derived Radex structures, i. e. the tests show high g loadings and are therefore a good indicator for general intelligence (see also, e. g., Bisanz et al., 1994; Carroll, 1993; Neubauer, 1995; Paul, 1986). This might also be the reason why the Raven tests have often been used as models for similar tests, such as the WMT (see below) or the FRT (Figure Reasoning Test by Daniels, 1993). They also inspired the development of other speech free intelligence tests, such as number series or paper–folding tests.

Being the most difficult of Raven's tests, the APM was developed to differentiate between the more able, i. e. people above average. Raven constructed the set of items in order to avoid ceiling effects. For the selection and sequencing of the items, discrimination and difficulty parameters were calculated and the distractor alternatives were analyzed (based on classical test theory).

The APM consist of 48 geometric matrices (12 in Set I, 36 in Set II). To solve the problems it is necessary to form comparisons and reason by analogy, i. e. the solutions require analytic and integrating operations (exceptions are the first four problems in Set I, which can be solved by applying Gestalt principles).

Each item is depicted as a three by three matrix, in which the bottom right–hand cell is empty. Below the matrix eight answer alternatives are depicted (see Box 2.6 for an example of an APM–like matrix). Participants' task is to select the answer alternative that completes the matrix correctly. The cells of the matrix contain figural elements, such as lines, geometric figures, or background textures, which are related by one to four rules (Carpenter et al., 1990; Musch and Albert, 2003). Participants' are instructed to look for rules governing the first and the second row (or column) of the matrix, and then to apply the found rule(s) to the third row (column) in order to complete the matrix.

For the results, the number of correctly solved items is added up and the raw scores can be converted to percentiles, IQs, and T–scores.

With respect to item difficulty, Raven designed the items on intuition and did not explicitly report the types of rules used for the APM. However, there are several analyses of the items by other authors (e. g., Carpenter et al., 1990; Hunt, 1974; Jacobs and Vandeventer, 1972; Musch and Albert, 2003; Vodegel Matzen et al., 1994). The extracted rules are listed in Section 2.2, Table 2.2.

Vodegel Matzen et al. (1994) also pointed out that the PM were not constructed with the intent to provide information on how a particular test score is achieved. The interpretation of the test score takes only account of the number of correctly solved items but not of the underlying cognitive processes (see Section 2.2.3 for research on the requirements of geometric matrices).

2.4.2.4 Vienna Matrix Test (WMT)

The WMT (Formann and Piswanger, 1979) is also a speech–free geometric matrix test, which strongly builds upon Raven's progressive matrices. It was developed to assess a testee's reasoning ability associated with abstract symbols, but also shows high correlations with measures of general intelligence ($r_{(\text{WMT,IST})} = .85$). While the WMT shares the intention and basic concepts of the APM, the construction and selection of items differ. As compared to the APM, which were designed on intuition, the item set of the WMT was developed according to construction principles (see below). The set of principles was used to create an item universe that defines all possible item classes. From this item universe, Formann (1973) constructed a subset of 42 items and analyzed the item set with Rasch's probabilistic test model (Fischer and Molenaar, 1995). From the set of 42 items, nine items had to be eliminated because they did not fit the Rasch model. In addition, four items were excluded, because their parameters deviated from the preferred hypothesis on item difficulty (see below) and one item was eliminated due to the high solution probability of 98.5%. From the remaining 28 items, one serves as instructional item, three as practice items, and the final 24 items constitute the actual set of test items (two of the 24 items are taken from the SPM, one from the APM).

With regard to the presentation of the problems, Formann (1973) based the appearence of the items on the APM, i. e. each item is presented as a three by three matrix with eight anwer alternatives (see Box 2.6 for a WMT–like matrix). However, for the specific design of each item, Formann specified three components, which determine the construction principle of the items and can be used to describe the items. The three components are (1) effective rules, (2) relevant material attributes, and (3) the direction of the rule. The first component (1) consists of the three rules continuation, variation, and superimposition. The applied material attributes (2) are form, pattern, number, and spatial orientation, the rule directions (3) are vertical, horizontal, or both. As for the APM, the testee's task is to find a correct completion of a matrix by applying one of the rules to the elements' material attributes in either of the directions.

In order to determine the relationship between the structure of the items (given by the components' attributes) and item difficulty, Formann (1973) examined several hypotheses by means of the linear logistic Rasch model. Generally, he found that item difficulty cannot be completely explained by the item structure. The best approximation was that the three components influence item difficulty independent of each other, with the component effective rule being most important and the component material attributes least important. With respect to the attributes' difficulty, he found that the rule superimposition is most difficult, followed by the rule variation, whereas the rule continuation is least difficult. With regard to the material attributes, spatial order and number are more difficult than form or pattern. For the component direction, vertical rules are most difficult for the particpants, while horizontal rules are least difficult.

With respect to the diagnostic results of the WMT, the number of correct responses is added up and the raw scores are converted to percentiles, z–scores, IQs, and verbal descriptions of the ability levels.

The WMT and the APM both assess the ability of analytic reasoning, more specifically to educe relations in abstract symbols. As compared to the APM, the development of the WMT is based on predefined construction principles, which allows an anlysis of the items with respect to the demands or components contributing to item difficulty. Furthermore, the selection of items is based on the Rasch model, which has the advantage that the test assesses exactly one ability dimension that is the same for all persons independent of their educational background.

However, inspite of the precise formulation of the underlying model and the specification of an item universe by construction principles, Formann had to exclude several items that did not meet the model requirements. Other investigations on geometric matrices did also not succeed in predicting item difficulty as a linear ordered combination of the problem components and had to eliminate some items with non–fitting item parameters (see e. g., Hornke and Habon, 1984; Hornke, Küppers, and Etzel, 2000; Nährer, 1980). In Section 3.5.2, I will introduce an alternative approach to the modelling of geometric matrices by Musch and Albert (2003), which is based on knowledge space theory (see Chapter 3) and assumes a partial order on the items' components.

2.5 Summary of Chapter 2

Inductive reasoning can be defined as the process of drawing conclusions from single instances or observations in order to reach a general rule that governs all instances under consideration (Section 2.1). Inductive reasoning problems can be found in many intelligence or aptitude tests, where they are mostly presented as analogy, series completion, classification, or matrix items. In Section 2.2, I outlined the demands inherent in those problem types, which are relevant to my own research (verbal and geometric analogies, number series, geometric matrices) and compared the found components to each other. The aim of this research is to order the inductive reasoning problems by their difficulty so that an efficient basis for an adaptive testing system can be developed. For the establishment of a difficulty structure on various problem types it is necessary to define comparable features among the problems. In Sections 2.3.1 and 2.3.2, I introduced two models by Klauer (2001) and Sternberg (1977a,b; Sternberg and Gardner, 1983), which are both designed for various problem types. Whereas Sternberg's approach concentrates on the identification of common cognitive processes underlying different inductive reasoning problems, Klauer's model is based on a definition of inductive reasoning, which gives detailed problem specifications, such as the problems' content or the detection of (dis)similarities in attributes or relations. Both models provide important assumptions on the components that are shared by different types of problems as well as on the differences among the various problem types. Knowing the demands or components of inductive reasoning problems, it is still necessary to find an appropriate test model for the intended adaptive assessment system. Focusing on structure models, Section 2.4 dealt with the psychometric approach to inductive reasoning, including a selection of tests measuring this ability. Still missing is an integrative test model of inductive reasoning abilities, that constitutes the basis for efficient testing procedures. In the next chapter, I will outline a method that is suitable for the establishment and validation of a difficulty structure on various problem types and can be applied for adaptive knowledge assessments.

3 The Theory of Knowledge Spaces

In Chapter 2, I outlined several models on the cognitive requirements of inductive reasoning problems. Furthermore, I discussed the communalities and differences among various types of inductive reasoning problems and gave some examples for tests assessing the ability of inductive reasoning.

What is still needed, though, is a precise and comprehensive description of the properties of various types of inductive reasoning problems, which at the same time can be used to describe individual differences in performance. After defining common components and properties of inductive reasoning problems, it is possible to develop means for an exact diagnosis of a persons knowledge in the domain. The currently available instruments (most of which are based on classical test theory) describe a person's ability level by a numerical test score, which refers to the performance on a subtest or on the entire test (see also Section 2.4.2). The test scores are usually standardized, so that an individual's ability level can be located within the population of reference. Probabilistic test theory additionally provides information on the difficulty and ability ratios between items and persons respectively. Moreover, the estimated item and person parameters are independent of each other (see also Section 6.1). One requirement for the application of probabilistic test theory is that the items are unidimensional, i. e. that they assess only one ability dimension. Item as well as person parameters are based on interval scales and the established linear orders on the items can be used for adaptive testing procedures (see also Chapter 6).

Another approach is to define a person's knowledge state by the set of problems the person is able to solve. If the demands of each problem are known, a person's missing knowledge can be derived directly from the diagnosed knowledge state. Such an approach has the advantage that psychological theories on the problem requirements or cognitive demands are directly related to the empirical solution patterns. Furthermore, the diagnostic instrument can simultaneously be used as a basis for adaptive testing procedures as well as for tutorial systems.

In this chapter, I want to discuss a methodological framework, which permits the diagnosis of a person's knowledge state in a specified domain as well as the prediction of observable behavior. In order to describe problem demands and persons abilities not only by numerical parameters but by the specific requirements inherent in a problem or met by the person, non–numerical test theory seems to be the appropriate instrument of measurement.

The behavioral *Knowledge Space Theory*, originally developed by Doignon and Falmagne (1985; 1999; Falmagne, Koppen, Villano, Doignon, and Johannesen, 1990), is a major constituent within the non–numerical test theory. Some of the main advantages of this approach are its low scale requirements and non–linearity combined with the possibility of empirical validation. Furthermore, the mathematical modeling underlying the theory shows a high degree of formal precision, while starting out with fairly simple psychological assumptions.

The primary idea behind the theory of knowledge spaces was the development of a model that permits an efficient estimation of a person's knowledge state in a given domain. Generally, it is assumed that a person's proficiency level in a specified field of information can be assessed by testing the person on a set of problems from this field. One method to carry out the testing is to present all available problems. In this case the set of problems solved correctly represents the knowledge state of the person. However, with a large set of problems and participants the method would prove itself as rather impracticable and uneconomical in time and expense.

The knowledge space theory provides a framework for more efficient testing procedures, which diagnose a person's knowledge state by specifying the problem demands the person is able to meet and/or the underlying skills the person possesses. A further objective is to reduce the number of presented problems by taking advantage of the dependencies among a set of problems. In this case only a subset of the test items needs to be presented, while the solution of the remaining items can be surmised from the correct or incorrect solution of the previously observed responses. As an example, one might imagine a person who is capable of multiplying fractions. Assuming that the same person will also be capable of multiplying real numbers, it would be inefficient to present problems containing this type of task. Hence, by taking advantage of the implicit structure relating a set of problems, it is possible to reduce the number of items presented in a test and simultaneously obtain precise information on the person's knowledge state.

With this issue in mind Doignon and Falmagne (1985; 1999; Falmagne et al., 1990) developed a mathematical model in which the representation and diagnosis of knowledge are directly related. The notational framework of the knowledge space theory is defined in terms of behavioral data structures. Furthermore, the theory allows computer aided knowledge assessments and yields a basis for the development of computerized tutoring systems.

The following sections give an overview of the main concepts and ideas which are relevant for the presented research. I will start with a description of the main ideas for surmise relations between items (Section 3.1) and tests (Section 3.2) and proceed with different methods for the generation (Section 3.3) and validation (Section 3.4) of knowledge spaces. This Chapter will conclude with two examples of how knowledge space theory has already been applied to single inductive reasoning tests (Section 3.5).

For a comprehensive review of the theory, I refer the reader to Doignon and Falmagne (1999) and for theoretical extensions and empirical applications to Albert and Lukas (1999). Considering the model's formal approach, the mathematical basics and used terminology are outlined in Appendix A for a better understanding of the concepts.

3.1 Surmise relations between items

Consider a set $Q = \{q_1, q_2, \ldots, q_n\}$ of test items which are related implicitly by their difficulty. For example q_1 is at least as difficult as q_2, which again is at least as difficult as q_3, and so on. In this case, all items can be ordered linearly along one dimension, which is known as Guttman scale. A Guttman scale is a special case of a quasi order, where all items are connected. The theory of knowledge spaces generalizes the linear order to a partial order by including independence between items. The related concept is called a *surmise* or *prerequisite relation*. More specifically, a surmise relation is a binary relation on a set of problems or test items with the following interpretation:

> "Whenever a person masters an item $x \in Q$ and we can surmise that this person is also able to master item $y \in Q$ we say that the pair (y, x) is in a *surmise relation*."

In other words, from the mastery of item x we can surmise or assume the mastery of item y, i. e. item y is a prerequisite for item x. Surmise relations are reflexive and transitive but not necessarily connex, i. e. they are quasi orders. Doignon and Falmagne (1985, 1999) defined this dependency between test items as follows:

Definition 3.1 Let Q be a set of problems and $S \subseteq Q \times Q$ a reflexive and transitive binary relation on Q. Then the quasi order S is called a *surmise relation*. □

Common notations for a surmise relation between two items x and y are ySx, $(y, x) \in S$, $S = \{(y, x)\}$, $y \preceq x$, or $(y, x) \in \preceq$. I will use the notations ySx or $S = \{(y, x)\}$.

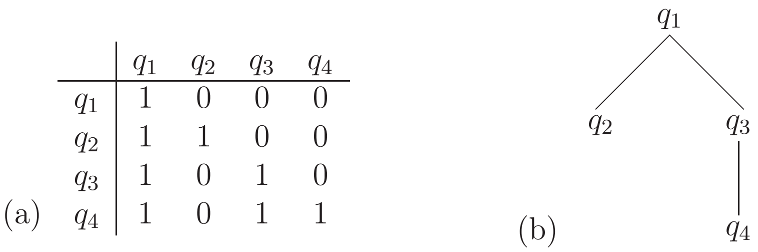
If the pairs (y, x) and (x, y) are both elements of a surmise relation S (i. e. ySx and xSy) then the items x and y are called *equivalent*.

Example 3.1 Imagine a set $Q = \{q_1, q_2, q_3, q_4\}$ of three items and a surmise relation $S = \{(q_1, q_1), (q_2, q_2), (q_3, q_3), (q_4, q_4), (q_2, q_1), (q_3, q_1), (q_4, q_1), (q_4, q_3)\}$ on Q. With regard to S we assume that the items q_2 and q_3 are prerequisites for item q_1 and that the item q_4 is prerequisite for items q_1 and q_3. Or, in other words, from a correct solution to item q_1 we can surmise a correct solution to items q_2, q_3, and q_4 (q_2Sq_1, q_3Sq_1, and q_4Sq_1) and from a correct solution to item q_3 we can surmise a correct solution to item q_4 (q_4Sq_3). The two items q_2 and q_3 as well as the pair q_2 and q_4 are independent ($q_2Sq_3 \notin S$, $q_3Sq_2 \notin S$, $q_2Sq_4 \notin S$, and $q_4Sq_2 \notin S$). □

A surmise relation can be depicted as matrix or as Hasse diagram. In the matrix, a '1' in column x and row y indicates that the pair (y, x) is element of the surmise relation (ySx). In the Hasse diagram a descending line signifies that a correct solution of the lower item is surmisable from a correct solution of the upper item. Figure 3.1 illustrates the surmise or prerequisite relation described in Example 3.1 as matrix (3.1a) and as Hasse diagram (3.1b).

The set of all elements of Q that are surmisable from item x is also called downset of x (D_x). Formally,

$$D_x = \{y \mid ySx\} \quad \forall x, y \in Q. \tag{3.1}$$



(a)
$$\begin{array}{c|cccc} & q_1 & q_2 & q_3 & q_4 \\ \hline q_1 & 1 & 0 & 0 & 0 \\ q_2 & 1 & 1 & 0 & 0 \\ q_3 & 1 & 0 & 1 & 0 \\ q_4 & 1 & 0 & 1 & 1 \end{array}$$

Figure 3.1: Matrix (a) and Hasse diagram (b) for the surmise relation in Example 3.1

In other words, a downset D is a downset of x, whenever x is the greatest element of D (Davey and Priestley, 1990). For Example 3.1 the following downsets can be derived:

$$\begin{aligned} D_{q_1} &= \{q_1, q_2, q_3, q_4\} \\ D_{q_2} &= \{q_2\} \\ D_{q_3} &= \{q_3, q_4\} \\ D_{q_4} &= \{q_4\} \end{aligned}$$

The surmise relation in Example 3.1 is a partial order, which has the properties of transitivity, reflexivity, and antisymmetry. Surmise relations that are additionally connected are called linear ordered surmise relations. In a linear ordered surmise relation, there are no independent item pairs.

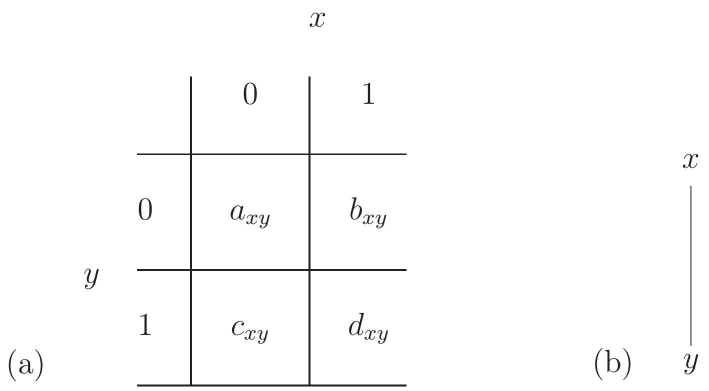
With regard to the expected item combinations or solution patterns, a surmise relation reduces the set of all possible solution patterns (the powerset P with $2^{|Q|}$ elements) to a subset of admissible solution patterns. Imagine a surmise relationship between two items x and y (ySx), where item y is prerequisite for item x. The possible response patterns (see Figure 3.2a) are that both items are solved correctly, which is denoted $\langle 1, 1\rangle$ (cell d_{xy}), that only item x is answered correctly ($\langle 1, 0\rangle$, cell b_{xy}), that only item y is answered correctly ($\langle 0, 1\rangle$, cell c_{xy}) or that neither of the items is answered correctly ($\langle 0, 0\rangle$, cell a_{xy}). The admissible solution patterns for the items include only cells a_{xy}, c_{xy} and d_{xy}. This means, that under the assumption ySx, x should only be mastered in combination with a correct solution to y. The solution pattern $\langle 1, 0\rangle$ is possible, but not admissible. Figure 3.2 shows (a) the contingency table for a pair of items x and y and (b) the respective prerequisite relationship as Hasse diagram. The variables a_{xy}, b_{xy}, c_{xy}, and d_{xy} denote the frequencies for the four possible answer vectors. The frequency for cell b_{xy} should equal zero.

Each admissible item combination resulting from a given surmise relation is called a *quasi ordinal knowledge state*. Formally, a quasi ordinal knowledge state K is defined by

$$K \subseteq Q \Leftrightarrow (\forall x, y \in Q, ySx \wedge x \in K \Rightarrow y \in K) \tag{3.2}$$

Correspondingly, Doignon and Falmagne (1985) defined the *knowledge state* of a person as the set of problems in a specified domain Q this individual is able to master under ideal conditions. The set of all knowledge states is called *knowledge structure* ($\mathcal{K}$).

Definition 3.2 A *knowledge structure* is a pair $(Q, \mathcal{K})$, in which Q is a nonempty set, and $\mathcal{K}$ is a family of subsets of Q, containing at least the complete set of items Q and



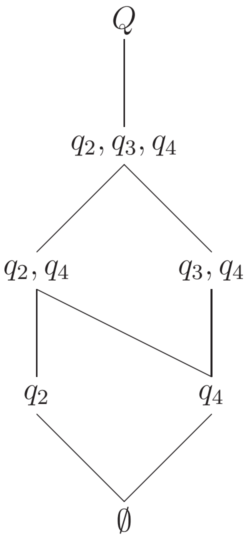
(a)

		x	
		0	1
y	0	a_{xy}	b_{xy}
	1	c_{xy}	d_{xy}

Figure 3.2: Possible response patterns and assumed prerequisite relation for a pair of items with $b_{xy} = 0$

the empty set $\emptyset$. The elements of Q are the problems or test items of a knowledge domain and the subsets in the family $\mathcal{K}$ are called *knowledge states*. □

In Example 3.1 the number of possible solution patterns $2^{|Q|} = 16$ is reduced to seven admissible solution patterns or knowledge states. The resulting knowledge structure is depicted in Figure 3.3.



$$\mathcal{K} = \{\emptyset, \{q_2\}, \{q_4\}, \{q_2, q_4\}, \{q_3, q_4\}, \{q_2, q_3, q_4\}, Q\}$$

Figure 3.3: Derived knowledge structure for Example 3.1

Knowing that only a specified family of subsets of Q is accepted as knowledge states, one can presume that a knowledge structure itself can be described by its special characteristics. For a knowledge structure $\mathcal{K}$ that is derived from a surmise relation and fulfills the requirements of a quasi order, two properties are always satisfied.

$$\begin{aligned} (i) \quad & K, L \in \mathcal{K} \Rightarrow K \cup L \in \mathcal{K} \\ (ii) \quad & K, L \in \mathcal{K} \Rightarrow K \cap L \in \mathcal{K} \end{aligned} \tag{3.3}$$

Knowledge structures which fulfill property (i) are called *knowledge spaces*. Knowledge structures fulfilling both properties (i) and (ii), i. e. they are closed under union $\cup$ and intersection $\cap$, are called *quasi ordinal knowledge spaces*. This means that for any two knowledge states K and L, their union $\cup$ and their intersection $\cap$ are also knowledge states.

According to a theorem by Birkhoff (1937), quasi orders on a set of items establish a one–to–one correspondence between a surmise relation and its corresponding knowledge space. Thus, the quasi ordinal relation can be directly inferred from the quasi ordinal space, and vice versa.

Theorem 3.1 The **Birkhoff–Theorem** defines two equivalences:

$$ySx \;\Leftrightarrow\; (\forall K \in \mathcal{K}, x \in K \Rightarrow y \in K) \tag{3.4}$$

$$K \in \mathcal{K} \;\Leftrightarrow\; (\forall x, y \in Q \text{ with } ySx, x \in K \Rightarrow y \in K) \tag{3.5}$$

□

Another important concept within the knowledge space theory is the *base* of a knowledge space. The base is the minimal subfamily $\mathcal{B}$ of $\mathcal{K}$, from which each knowledge state can be restored by building the closure under union of its elements. Hence, the base constitutes an economic storage for all knowledge states. Doignon and Falmagne (1999) defined the base as follows:

Definition 3.3 A *base* for a knowledge structure $(Q, \mathcal{K})$ is a minimal family $\mathcal{B}$ of states spanning $\mathcal{K}$. A knowledge structure has a base only if it is a knowledge space. A knowledge state K belonging to some base $\mathcal{B}$ of $\mathcal{K}$ cannot be the union of other elements of $\mathcal{B}$ or other states of $\mathcal{K}$. □

The base can also be seen as the family of downsets D_x of Q. Hence, the base for the knowledge space in Example 3.1 contains the minimal states $\mathcal{B} = \{\{q_2\}, \{q_4\}, \{q_3, q_4\}, \{q_1, q_2, q_3, q_4\}\}$. Figure 3.1 depicts the base for Example 3.1 as a matrix. In the matrix, a '1' in row i indicates that the base element is minimal for item i. A '2' in row i indicates that the item in column j is prerequisite for item i and therefore contained in the base element that is minimal for i.

Table 3.1: Base for Example 3.1

	q_1	q_2	q_3	q_3
q_1	1	2	2	2
q_2	0	1	0	0
q_3	0	0	1	2
q_4	0	0	0	1

In general, the advantage of organizing knowledge according to surmise or prerequisite relations is that the number of possible response vectors, i. e. the powerset P (with $2^{|Q|}$ elements) of all problems, can be reduced to a subset $\mathcal{K} \in 2^Q$ of knowledge states.

The basic concepts and ideas of the theory of knowledge spaces elicited several new theoretical approaches. Examples are the generalization to surmise systems or surmise functions, which allow alternative prerequisites (Doignon and Falmagne, 1985, 1999), the inclusion of skills (Doignon, 1994a; Düntsch and Gediga, 1995), competencies (Korossy, 1997), or process models (Schrepp, 1995, 1999), the assessment of misconceptions (Lukas, 1997; Körner, 1998), the extension to probabilistic knowledge structures (Cosyn and Thiéry, 2000; Doignon and Falmagne, 1999; Falmagne, 1989a,b, 1994; Lakshminarayan and Gilson, 1998; Villano, 1991), and the development of techniques to obtain a knowledge space (Albert and Held, 1994, 1999; Dowling, 1993b; Held, 1999; Held and Korossy, 1998; Kambouri, Koppen, Villano, and Falmagne, 1994; Koppen, 1994, 1993; Koppen and Doignon, 1990). A theory based approach to obtain a knowledge space developed by Albert and Held (1994) will be outlined in Section 3.3. Furthermore, the concept has been extended from relationships between items to relationships between tests (Albert, 1995; Brandt, Albert, and Hockemeyer, 1999, 2003; Albert, Brandt, Hockemeyer, Ünlü, and Schappacher, 2003), which will be discussed next (Section 3.2).

3.2 Surmise relations between tests

So far, the theory of knowledge spaces referred to single tests. However, in common psychological assessment procedures we often deal with a set $\mathcal{T}$ of different tests that are usually related. Examples are educational or cognitive psychology where prerequisite relationships between different educational stages or cognitive functions are investigated.

The classical conception of the relations between tests is based on correlations. Usually, the strength of relationships between two tests is investigated, but not the direction of the relationship. Here, the non–numerical test theory can be employed to investigate directed relationships between tests. The idea is that the possession of one, e. g. cognitive ability might be prerequisite for some other ability. Or, with regard to educational psychology, the mastery of one course within a curriculum might be prerequisite for the mastery of another course.

On the background of Doignon and Falmagne's (1985; 1999) framework, Albert and his group (1995; Albert et al., 2003; Brandt et al., 1999; 2003) extended the concept of the non–symmetric surmise relation between items (i. e. *within* tests) to surmise relations *between* tests. The interpretation of a surmise or prerequisite relation between tests is as follows:

> "Whenever a person masters a given set of items in test A and we can surmise that this person is also able to master a particular non-empty subset of items in test B, we say that the two tests $A, B \in \mathcal{T}$ are in surmise relation from A to B."

Figure 3.4 illustrates the described concept. In this example mastering item b_3 in test B is a prerequisite for the mastery of item a_1 in test A. Formally, the surmise relation between tests is defined as follows:

Definition 3.4 Let $\mathcal{T}$ be a set of tests. Then the relation $\dot{\mathcal{S}} \subseteq \mathcal{T}$ x $\mathcal{T}$, defined by

$$B\,\dot{\mathcal{S}}\,A \Leftrightarrow \exists a \in A : B_a \neq \emptyset \quad \forall A, B \in \mathcal{T} \tag{3.6}$$

is called *surmise relation between tests*. When $B\ \dot{\mathcal{S}}\ A$ holds we say that the tests A and B are in surmise relation from A to B. □

Common notations for a surmise relation between two tests A and B are $B\ \dot{\mathcal{S}}\ A$, $(B, A) \in \dot{\mathcal{S}}$, $\dot{\mathcal{S}} \subseteq \mathcal{T} \times \mathcal{T}$, or $\dot{\mathcal{S}}_{\mathcal{T}x\mathcal{T}}$.

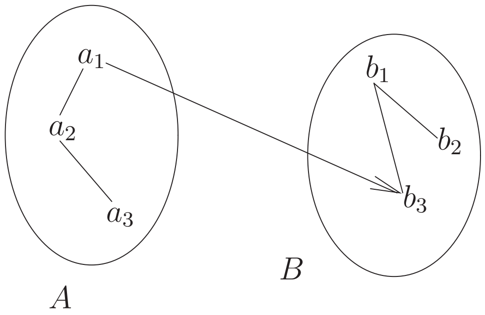


Figure 3.4: Two tests A and B are in surmise relation from A to B ($B\ \dot{\mathcal{S}}\ A$)

Regarding the properties of a surmise relation between tests, one important point is that, unlike the surmise relation between items, it is in general not a quasi order. More specifically, for a set of tests $\mathcal{T} = \{A, B, C, \ldots\}$, a surmise relation between tests has the property of reflexivity but not necessarily transitivity. However, there are special cases for which transitivity holds, namely left–, right–, and total–covering surmise relations. Analogous to items (see Birkhoff–Theorem 3.1), the advantage of a quasi order on a set of tests is that transitive surmise relations between tests can be inferred from the corresponding test knowledge structure (see below, Definitions 3.8 and 3.9). However, the reverse inference is not valid for a set of tests (Albert et al., 2003).

The interpretation of a *left–covering surmise relation* ($B\ \dot{\mathcal{S}}_l\ A$) is that for each item $a \in A$ there exists a nonempty subset of prerequisites in test B. This means that a person who doesn't solve any item in B will not be able to solve any item in A, either. There is no need to test this person on test A (see Figure 3.5).

Definition 3.5 Let $\mathcal{T}$ be a set of tests. Then the relation $\dot{\mathcal{S}}_l \subseteq \mathcal{T}$ x $\mathcal{T}$, defined by

$$B\,\dot{\mathcal{S}}_l\,A \Leftrightarrow \forall a \in A : B_a \neq \emptyset \qquad \forall A, B \in \mathcal{T} \tag{3.7}$$

is called *left–covering surmise relation.* When $B\ \dot{\mathcal{S}}_l\ A$ holds we say the two tests A and $B \in \mathcal{T}$ are in a left–covering surmise relation from test A to test B. □

The interpretation of a *right–covering surmise relation* ($B\ \dot{\mathcal{S}}_r\ A$) is that for each item $b \in B$, there exists at least one item $a \in A$ for which b is a prerequisite. This means that failing to solve a given item in test B implies a failure on a subset of items in test A. In other words, a person who solves all items in test A is also able to solve all items in test B. Hence, there is no further need to test this person on test B (see Figure 3.6).

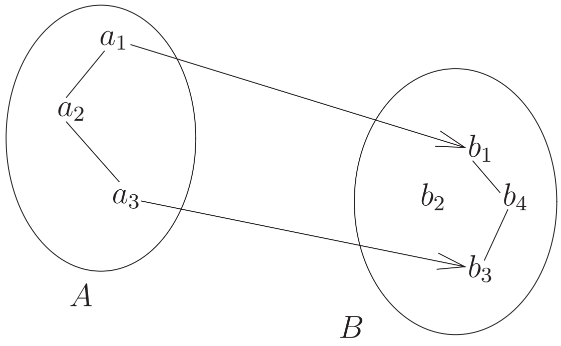


Figure 3.5: Left–covering surmise relation from test A to test B ($B\ \dot{\mathcal{S}}_l\ A$)

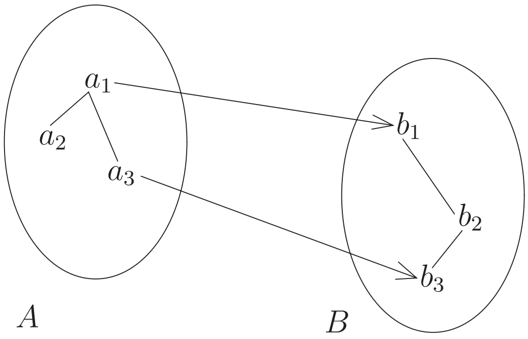


Figure 3.6: Right–covering surmise relation from test A to test B ($B\ \dot{\mathcal{S}}_r\ A$)

Definition 3.6 Let $\mathcal{T}$ be a set of tests. Then the relation $\dot{\mathcal{S}}_r \subseteq \mathcal{T}$ x $\mathcal{T}$, defined by

$$B\ \dot{\mathcal{S}}_r\ A \Leftrightarrow \bigcup_{a \in A} B_a = B \qquad \forall A, B \in \mathcal{T} \tag{3.8}$$

is called *right–covering surmise relation*. When $B\ \dot{\mathcal{S}}_r\ A$ holds we say the two tests A and $B \in \mathcal{T}$ are in a right–covering surmise relation from test A to test B. □

Finally, we speak of a *total–covering surmise relation* ($B\ \dot{\mathcal{S}}_t\ A$), if all items in test A have a prerequisite $b \in B$ and all items in test B are prerequisite for some item $a \in A$, i. e. the surmise relation is left– as well as right–covering. For surmise relations between tests which are neither left– nor right–covering, I will refer to as *general surmise relations* between tests.

Aside of the surmise relation between tests, it is necessary to differentiate between various subsets of the surmise relation on the entire set Q of items. S_{QxQ} denotes the surmise relation on the whole set of items and is referred to as the surmise relation *between items (SRbI)*. The disjoint subsets of the surmise relation S_{QxQ} on two tests A and B are denoted S_{AxA}, S_{BxB}, S_{AxB}, and S_{BxA}. The sets S_{AxA} and S_{BxB} are called surmise relations *within tests (SRwT)*, the sets S_{AxB} and S_{BxA} surmise relations *across tests (SRxT)*. Each subset is defined as follows:

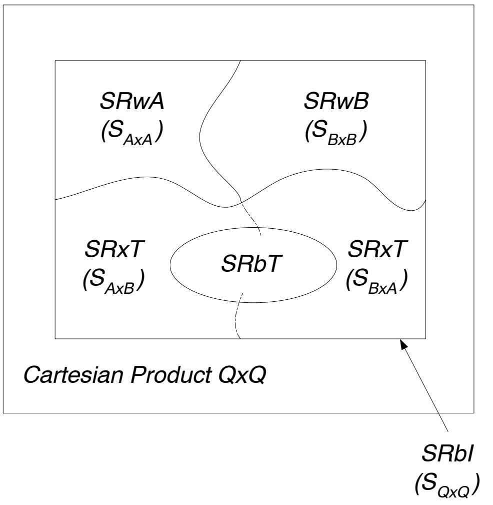


Figure 3.7: Illustration of the different subsets contained in the Cartesian product for the items of two tests A and B

$$\begin{aligned}
S_{QxQ} &= \{(y,x) \,|\, x,y \in Q \wedge ySx\} \\
S_{AxA} &= \{(a_i,a_j) \,|\, a_i,a_j \in A \wedge a_iSa_j\} \\
S_{BxB} &= \{(b_i,b_j) \,|\, b_i,b_j \in B \wedge b_iSb_j\} \\
S_{AxB} &= \{(a,b) \,|\, a \in A, b \in B \wedge aSb\} \\
S_{BxA} &= \{(b,a) \,|\, a \in A, b \in B \wedge bSa\} \qquad (3.9)
\end{aligned}$$

If a surmise relation fulfills either the condition S_{AxB} or S_{BxA}, we speak of a surmise relation between tests ($\dot{\mathcal{S}}_{\mathcal{T}x\mathcal{T}}$ or $SRbT$) as it is defined in terms of Definition 3.4. If S_{AxB} or S_{BxA} fulfill the conditions specified in Definition 3.5 or 3.6, we speak of a left– respectively right–covering surmise relation. Surmise relations between tests which fulfill both conditions are called total–covering. Figure 3.7 depicts the item pairs contained in the Cartesian product of the set Q of items contained in both tests, the $SRbI$ (S_{QxQ}) as subset of the Cartesian product QxQ, and the partitioning of the $SRbI$, including the set of item pairs contained in the $SRbT$.

To illustrate the concepts of surmise relations between items, within, and across tests, I will use the surmise relation depicted in Figure 3.6 and refer to it as Example 3.2.

Example 3.2 Imagine a surmise relation between two tests $A = \{a_1, a_2, a_3\}$ and $B = \{b_1, b_2, b_3\}$ as depicted in Figure 3.6. Then we can derive the following subsets of pairs for the surmise relation between items (S_{QxQ}), within tests (S_{AxA} and S_{BxB}), and across tests (S_{AxB} and S_{BxA}).

$$\begin{aligned}
S_{QxQ} &= \{(a_1,a_1),(a_2,a_2),(a_3,a_3),(b_1,b_1),(b_2,b_2),(b_3,b_3),(a_2,a_1),(a_3,a_1), \\
&\quad (b_2,b_1),(b_3,b_1),(b_3,b_2),(b_1,a_1),(b_2,a_1),(b_3,a_1),(b_3,a_3)\} \\
S_{AxA} &= \{(a_1,a_1),(a_2,a_2),(a_3,a_3),(a_2,a_1),(a_3,a_1)\} \\
S_{BxB} &= \{(b_1,b_1),(b_2,b_2),(b_3,b_3),(b_2,b_1),(b_3,b_1),(b_3,b_2)\}
\end{aligned}$$

$$\begin{aligned} S_{AxB} &= \{\} \\ S_{BxA} &= \{(b_1,a_1),(b_2,a_1),(b_3,a_1),(b_3,a_3)\} \end{aligned}$$

□

As for the generalization of the surmise relation between items to the surmise relation between tests, Albert and his group (Albert et al., 2003; Brandt et al., 1999; 2003) also extended the concepts of a knowledge state, a knowledge structure, and a knowledge space. I will illustrate the concepts by means of Example 3.2.

The *test knowledge state* $\dot{K}_i$ of a person i is defined as the combination of item subsets per test this person is capable of mastering. Accordingly, the collection of all test knowledge states $\dot{\mathcal{K}}$ is called the *test knowledge structure*.

Definition 3.7 Let $K_i \in \mathcal{K}$ be a knowledge state. Then for this knowledge state K_i the n–tuple $\dot{K}_i = (A_i, B_i, \ldots)$ is called *test knowledge state*, with A_i being the subset of items mastered in test A, B_i the subset of items mastered in test B, and so on. □

Definition 3.8 Let $\dot{\mathcal{K}}$ denote the set of all test knowledge states. Then the pair $(\mathcal{T}, \dot{\mathcal{K}})$ is called *test knowledge structure*, with $\mathcal{T}$ denoting the set of tests $\{A, B, C, \ldots\}$. □

For a test knowledge structure $\dot{\mathcal{K}}$ the combinations of the knowledge states from each single test are restricted to the set of test knowledge states. The set of test knowledge states can be derived from the surmise relation between tests.

Regarding Example 3.2 the following test knowledge structure can be derived from the surmise relation between the items of the two tests A and B.

$$\begin{aligned} \dot{\mathcal{K}}_{\{A,B\}} = \; & \{(\emptyset,\emptyset),(\emptyset,\{b_3\}),(\emptyset,\{b_2,b_3\}),(\emptyset,B),(\{a_2\},\emptyset),(\{a_2\},\{b_3\}), \\ & (\{a_2\},\{b_2,b_3\}),(\{a_2\},B),(\{a_3\},\{b_3\}),(\{a_3\},\{b_2,b_3\}), \\ & (\{a_3\},B),(\{a_2,a_3\},\{b_3\}),(\{a_2,a_3\},\{b_2,b_3\}), \\ & (\{a_2,a_3\},B),(A,B)\}. \end{aligned}$$

As mentioned above, a surmise relation between tests and, therefore, its corresponding test knowledge structure are not necessarily quasi ordinal. However, the properties of closure under union $\cup$ and intersection $\cup$ can also be defined for test knowledge structures (Brandt et al., 2003).

$$\text{For } \dot{K}_i = (A_i, B_i, \ldots) \text{ and } \dot{K}_j = (A_j, B_j, \ldots):$$

$$\begin{aligned} (i) \quad & \dot{K}_i \,\dot{\cup}\, \dot{K}_j := (A_i \cup A_j, B_i \cup B_j, \ldots) \\ (ii) \quad & \dot{K}_i \,\dot{\cap}\, \dot{K}_j := (A_i \cap A_j, B_i \cap B_j, \ldots) \end{aligned} \tag{3.10}$$

Test knowledge structures which fulfill property (i) are called *test knowledge spaces.* If a test knowledge structure fulfills both properties (i) and (ii) we speak of a *quasi ordinal test knowledge space.*

Definition 3.9 A test knowledge structure $(\mathcal{T}, \dot{\mathcal{K}})$ is a *test knowledge space*, if $\dot{\mathcal{K}}$ is closed under union $\cup$. $(\mathcal{T}, \dot{\mathcal{K}})$ is a *quasi ordinal test knowledge space* if $\dot{\mathcal{K}}$ is closed under union $\cup$ and intersection $\cap$. □

Regarding the correspondence between a knowledge space on the items of different tests and a test knowledge space, the test knowledge structure $\dot{\mathcal{K}}$ is a (quasi ordinal) test knowledge space, if its corresponding knowledge structure $\mathcal{K}$ is a (quasi ordinal) knowledge space.

Finally, the set of test knowledge states can be stored in the *base* $\dot{\mathcal{B}}$ of the corresponding test knowledge space, i. e. as a subset of $\dot{\mathcal{K}}$. The base is defined as follows:

Definition 3.10 A subset $\dot{\mathcal{B}} \subseteq \dot{\mathcal{K}}$ of test knowledge states is called the *base* of a test knowledge structure, if the following condition holds:

$$\begin{aligned} \dot{\mathcal{B}} &= \{(A_i, B_i, \ldots), (A_j, B_j, \ldots), \ldots\} \text{ is the base of } \dot{\mathcal{K}} \Leftrightarrow \\ \mathcal{B} &= \{A_i \cup B_i \ldots, A_j \cup B_j \ldots, \ldots\} \text{ is the base of } \mathcal{K}. \end{aligned} \tag{3.11}$$

□

By means of the base $\dot{\mathcal{B}}$, it is possible to reestablish the test knowledge space $\dot{\mathcal{K}}$, the knowledge space $\mathcal{K}$ on the set of items and its substructures, the surmise relation between items and its subsets, the surmise relation between tests, and its properties (e. g., transitivity, left,– and right coveringness).

For the test knowledge structure of Example 3.2, the base consists of the following test knowledge states:

$$\dot{\mathcal{B}} = \{(\emptyset, \{b_3\}), (\emptyset, \{b_2, b_3\}), (\emptyset, B), (\{a_2\}, \emptyset), (\{a_3\}, \{b_3\}), (A, B)\}.$$

Analogous to the surmise relation, I will refer to the test knowledge space as TKS, to the knowledge space between the items of all tests as $KSbI$, to its substructures within tests as $KSwT$ and across tests as $KSxT$.

The advantage of the concept of surmise relations between tests is that it is possible to specify prerequisite relations not only between single items but between subsets of items. Such subsets can be psychological tests or, for example, cognitive or developmental stages, where the possession of one ability is prerequisite for the acquisition of some other ability. Regarding the area of diagnostics, the number of employed tests can be reduced and adaptive testing systems can be developed, which are more economical and more informative by covering a wide range of problems. Further potential applications include the fields of educational psychology (e. g., curriculum development; Hockemeyer, Albert, and Brandt, 1998; Albert and Hockemeyer, 1999) or computer sciences (e. g., structuring hypertext documents; Albert and Hockemeyer, 1997).

3.3 Generation of knowledge structures

The theory of knowledge spaces, as it was outlined in Sections 3.1 and 3.2, describes how a structure or a relation on a set of problems can be represented in a formal model.

The question now is, how the relational dependencies between problems and their corresponding knowledge states can be determined. The derivation of well founded knowledge structures is a crucial requirement for the application of diagnostic procedures. Especially, for the use of adaptive testing procedures, the knowledge spaces have to provide a proper representation of the knowledge domain by simultaneously reducing the number of knowledge states to a subset of the powerset. The efficiency of a diagnostic procedure is directly related to the number of states in a knowledge space or correspondingly to the number of pairs in a surmise relation. With a decreasing number of knowledge states and an increasing number of pairs in the surmise relation, also the number of questions decreases that need to be asked to determine a person's knowledge state. Several approaches have been developed to generate a knowledge structure.

One way to establish a surmise relation on a set of items is the analysis of binary data matrices. Van Leeuwe (1974, see also Held & Korossy, 1998), for example, describes a deterministic method called *item tree analysis* (ITA). In this approach, the surmise relation is estimated directly from an analysis of empirical contingency tables for each pair of items. Similar deterministic methods with a more general approach to binary data analysis can be found in Airasian and Bart (1973), Bart and Krus (1973), and van Buggenhaut and Degreef (1987). Probabilistic approaches were suggested by, for example, Cosyn and Thiéry (2000), Doignon and Falmagne (1999), Falmagne (1994, 1989a,b) and Villano (1991). However, the application of data analytic approaches is limited, since the resulting knowledge structure depends strongly on the sample of participants. Probabilistic procedures furthermore require the existence of very large data sets in order to estimate the models' parameter values.

A second method to generate a knowledge structure is to consult experts of the relevant knowledge domain. Dowling (1993a,b), Koppen (1993, 1994), and Koppen and Doignon (1990) have developed elaborated algorithms for computerized query procedures. In accordance with the behavioral approach of the knowledge space theory, the querying procedures refer directly to the given set of items. In the course of a query, the expert judges implications on the relationships among subsets of items. Additionally, inferences on further pairs in the relation can be drawn. Due to frequent inconsistencies in the experts' ratings (Baumunk and Dowling, 1997; Wesiak, 1998), a knowledge space established by means of a query can only be considered as a first sketch. The consultation of experts is especially suitable for fields of information which are not highly formalized. In these cases, a theoretical, content based analysis of the domain would not yield sufficient information about the set of reasonable knowledge states. One example is the field of psychopathology, which was investigated by means of querying experts by Riegler (1999) and Wesiak (1998). Among others, Dowling, Koch, and Quante (1996), Kambouri et al. (1994), and Koppen (1993, 1994) improved and refined the querying methods. One example is the development of a user–friendly interface, which includes visualizations of the derived structure and yields not only more efficient procedures but also more consistent results (Dowling et al., 1996).

In a third approach, the knowledge spaces are built from theoretically founded item hierarchies. Starting from psychological findings, surmise relations and their corresponding knowledge spaces are established. Based on task analyses or a cognitive theory on the solution processes, the relationships between items are explicitly formulated, (Albert and Held, 1994, 1999; Albert, Schrepp, and Held, 1994; Held, 1999; Schrepp, 1995, 1999; Schrepp, Held, and Albert, 1999). Thereby, the problems of a specified domain are broken down into a set of components or into several steps of the solution process. Then, a surmise relation and the corresponding knowledge space can be derived by applying various ordering principles (e. g. set–inclusion, sequence inclusion, multiset inclusion, or componentwise product formation). For an overview of the ordering principles, I refer the reader to Lukas and Albert (1993) and Albert and Lukas (1999).

The main requirement for theoretically founded problem structures is a well defined set of problem components. In one approach the components, i. e. sets of demands or competencies, are assigned to the items and an order is established by way of set–inclusion (see Albert and Held, 1994, 1999; Albert et al., 1994; Schrepp et al., 1999, for applications to the domain of chess). This means that items defined by a given set of demands have all those items as prerequisites, which are defined by a subset of these demands. Going beyond this model, where demands are either present or not, items are described by the same set of components but with variable attributes on each component. In this case difficulty orders are defined on the attributes of each component and the problems are ordered by a componentwise comparison of their respective attributes (Albert and Held, 1994, 1999; Held, 1999, see Section 3.3.1). Finally, the most specific way of structuring a set of problems is based on process models, according to which individual steps of the problem solving process are analyzed and structured (Schrepp, 1995, 1999).

For a detailed discussion, comparison, and integrated application of the three approaches to generate a knowledge structure see Held, Schrepp, and Fries (1995).

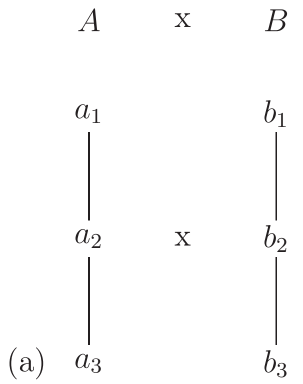
In the following subsection, I will outline the method for generating hypotheses which is relevant to my research in more detail. I chose the theory based method of componentwise ordering of product sets. As already outlined in Section 2.2, inductive reasoning problems are often described by the components that contribute to item difficulty and influence the solution process of the respective items. Also Klauer's (2001, see Section 2.3.1) definition of inductive reasoning is based on components and their attributes. Klauer refers to the components as facets and uses a mapping sentence to derive the set of all possible inductive reasoning problems. This approach basically corresponds to forming the Cartesian product of the facets (or components). Thus, the selected approach has the advantage that the existing research on inductive reasoning problems can be integrated in a straight forward way. Moreover, the applicability of the component based approach to single inductive reasoning tests has already been shown in several studies (cf. Section 3.5).

3.3.1 Componentwise ordering of product sets

The principle of componentwise ordering of product sets (Albert and Held, 1994, 1999) assumes that every item in a set Q of problems is described by the same set of components $C = \{A, B, C, \ldots\}$. Each component consists of a set of attributes $(A = \{a_1, a_2, \ldots\}, B = \{b_1, b_2, \ldots\}, C = \{c_1, c_2, \ldots\})$, on which order relations are defined. It is assumed that the attributes belonging to the same component cannot be combined with each other. This means that every item has exactly one attribute a_i, b_i, c_i, ..., whereby an attribute can also be defined as empty set. For example, when describing number series completion problems, one of the components could be the type of mathematical operation involved in the series. Then the attributes of this component could be addition, subtraction, multiplication, division, and no operation (i.e. the empty set).

Forming the Cartesian product of the components results in the set of all possible attribute combinations, which can be used as a basis for constructing problems or for a demand analysis of a given set of problems. Items that are described by the same set of attributes form an *item class*.

Example 3.3 Let us imagine a set of two components ($C = \{A, B\}$), with three attributes each. On each set of attributes ($A = \{a_1, a_2, a_3\}, B = \{b_1, b_2, b_3\}$) a linear difficulty order is defined, with i_1 being the most difficult attribute and i_3 the least difficult attribute. Figure 3.8a illustrates the linear orders on the components A and B, Figure 3.8b depicts the set of pairs derived by forming the Cartesian product of the two components A and B. □



$$A \text{x} B = \begin{bmatrix} (a_1, b_1) & (a_1, b_2) & (a_1, b_3) \\ (a_2, b_1) & (a_2, b_2) & (a_2, b_3) \\ (a_3, b_1) & (a_3, b_2) & (a_3, b_3) \end{bmatrix}$$

(b)

Figure 3.8: Cartesian product of two attribute sets A and B (Example 3.3)

The surmise relation on a set of problems is established by a pairwise comparison of the problems with respect to the components' attributes. More exactly, the ordering rule assumes that a problem y is prerequisite for a problem x, if the demands of all attributes $(a_i, b_i, c_i, \ldots)$ of y are equal or less difficult than the corresponding attributes $(a_i, b_i, c_i, \ldots)$ of x. This principle is known as coordinatewise order (see Davey and Priestley, 1990), which also corresponds to a choice heuristic from decision theory, viz. the dominance rule. For Example 3.3, the problem structure derived through a pairwise comparison of the components' attributes a_i and b_i is shown in Figure 3.9a. In this example, forming the product of the two components A and B results in nine possible attribute combinations or item classes (see also Fig. 3.8b). Because of the two linear attribute orders (see Fig. 3.8a) the obtained surmise relation contains 36 item pairs (including reflexive pairs).

Of course, it is also possible to define a partial order on the attributes of one or more components. In this case the resulting surmise relation will include less pairs than are derived by the linear ordered attributes. Defining, for example, a partial order on both components A and B with the attributes i_2 and i_3 being prerequisites for the attributes i_1, but independent of each other, results in a surmise relation with 25 pairs.

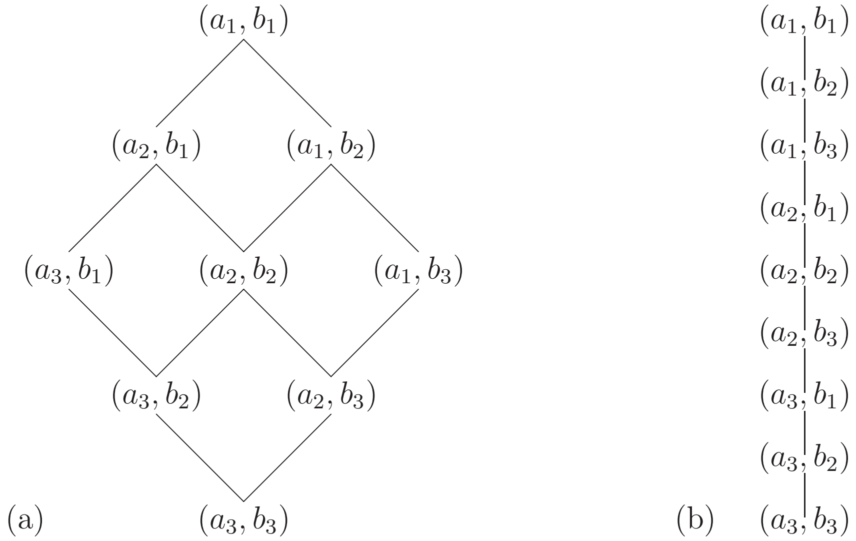


Figure 3.9: Problem structure for Example 3.3 with linear attribute orders and (a) incomparable components versus (b) lexicographic ordering

In Figure 3.9a an antichain is defined on the two components A and B, i. e. the components are incomparable. In more specific cases, the components themselves can be ordered lexicographically, meaning that one component A is assumed to be more important than another component B. In this case, the attributes of component A are first compared. Attributes of component B are only considered for problems equipped with the same attribute a_i. By this, the resulting surmise relation becomes a linear order (provided that the attributes of each component are also ordered linearly). Figure 3.9b shows a lexicographic order on the components of Example 3.3. With the linear order on the two sets of attributes, the surmise relation contains 45 pairs (including reflexive pairs).

In a similar approach a partial order can be defined on the components, meaning that some of the components are assumed to be incomparable with respect to importance. An order on a set of three components $\mathcal{C} = \{A, B, C\}$ could, for example, define components A and B as incomparable but both as more important than component C.

The principle of componentwise ordering of product sets has already been applied successfully for the domains of elementary stochastics (Held, 1993, 1999), and various types of inductive reasoning problems (Albert and Held, 1994, 1999; Musch and Albert, 2003; Albert and Wesiak, 2002, see Section 3.5).

3.4 Validation of knowledge structures

Strictly viewed, the goodness of fit of a given knowledge structure to a set of data can be defined by the number of response patterns which can be assigned exactly to one of the hypothesized states. However, the set of items a person is able to solve at a certain point in time during a testing procedure does not necessarily reflect the true knowledge state of this person. It has to be accounted for various influences such as careless errors (e. g., in computations or in transcribing the answer) while responding to some of the items. Additionally, restricted answer formats like multiple choice tests provide opportunities for lucky guesses on items the person could otherwise not answer correctly. Hence, the knowledge states in a deterministic model describe the latent knowledge of persons or the response patterns observed under ideal conditions. Thus, it is necessary to apply validation methods that are able to account for noise in the data.

In the following, I will introduce different methods for validating a knowledge space and its respective surmise relation under consideration of eventual discrepancies between a person's true knowledge state and his or her empirical response pattern. First, I will describe procedures that measure the agreement of a hypothesized surmise relation to a set of data. The second approach is based on the knowledge space, i. e. the correspondence of hypothetical knowledge states and empirical answer patterns is evaluated. All of the discussed procedures can be applied to surmise relations and knowledge spaces between items ($SRbI/KSbI$), within tests ($SRwT/KSwT$), and across tests ($SRxT/KSxT$). In this case, the set of answer patterns and the postulated pairs in the surmise relation or the knowledge states in the knowledge space are reduced to the respective subsets of the $SRbI/KSbI$ (see Section 3.2).

In order to compare the validity of the different subsets, it is necessary to account for the varying number of pairs in the relations as well as the sizes of the knowledge spaces. Generally, with a higher number of pairs in the relation or a lower number of states in the knowledge space, the hypothetical model contains more assumptions on the dependencies among items and tests. Because each additional assumption might lead to additional contradictions of the observed response patterns, knowledge structures of varying sizes should not be directly compared with regard to their validity. Therefore, I will calculate various indices that account for these discrepancies by relativizing the number of observed contradictions to the number of pairs in the relation or to the size of the space. Concerning the surmise relation between tests ($SRbT$), the presence of a right–, left–, or total–covering $SRbT$ is derived from the $SRbI$.

For an easier understanding of the discussed procedures, I will illustrate the methods by means of Example 3.4, which refers to a single test. As mentioned above, for the validation of a set of tests only the relevant substructures and the corresponding parts of the answer vectors are considered.

Example 3.4 Let $Q = \{a, b, c, d\}$ be a set of four items and $S \subseteq Q$ x Q a hypothetical surmise relation on Q with $S = \{(a, a), (b, b), (c, c), (d, d), (b, a), (c, a), (d, a), (c, b), (d, b)\}$.

Then the corresponding quasi ordinal knowledge space $\mathcal{K}$ contains the elements
$\mathcal{K} = \{\emptyset, \{c\}, \{d\}, \{c, d\}, \{b, c, d\}, \{a, b, c, d\}\}$ and is described by the base

Table 3.2: Relation and base files for Example 3.4

Relation file	Base file
	4
4	4
1000	1222
1100	0122
1110	0010
1101	0001

$\mathcal{B} = \{\{c\}, \{d\}, \{b, c, d\}, \{a, b, c, d\}\}$.

Figure 3.10 shows the Hasse diagram for the resulting surmise relation (Fig. 3.10a) and the corresponding knowledge space (Fig. 3.10b). Table 3.2 depicts the respective relation file with four items and the base file with four items and four base elements (see Section 3.1 for the interpretation of the two matrices).

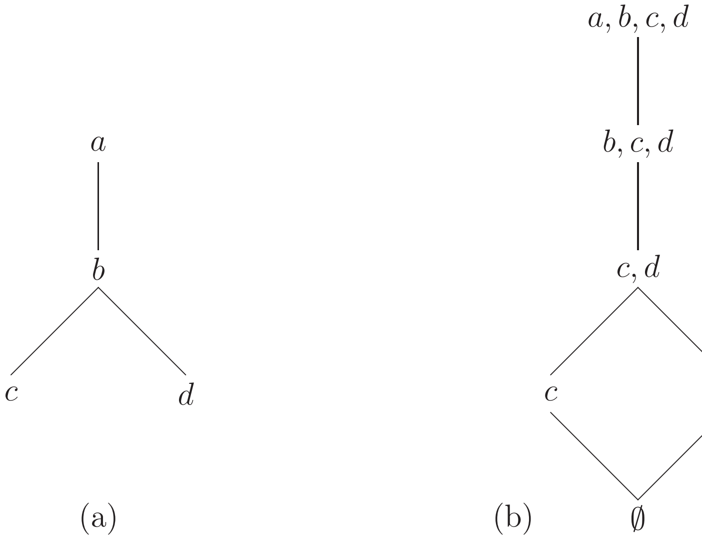


Figure 3.10: Hypothetical surmise relation (a) and its corresponding knowledge space (b) on a set $Q = \{a, b, c, d\}$ of four items (Example 3.4)

For the validation of the hypothetical surmise relation and knowledge space in this example a set of 10 fictitious response patterns will be used (see Table 3.3). Correct responses are coded '1', incorrect responses are coded '0'. Hence, the empirical response patterns are represented by binary response vectors as, for example, $\langle 0010 \rangle$ (pattern 1 in Table 3.3). With regard to the pairs contained in the surmise relation, Table 3.4 depicts the derived contingency tables for the set of response patterns. □

Before introducing the various validation methods, it has to be noted that the empirical validation of knowledge space hypotheses requires complete response patterns, i. e. each item has to be processed by all participants. Otherwise, it is not possible to decide whether the empirical states correspond to one of the hypothetical states or not. Furthermore, the solution frequencies of the items should yield no floor of ceiling effects. Trivial response patterns, in which either all or none of the items are solved correctly, do not contribute to the validation of a knowledge space hypothesis, because the empty and the full set are always knowledge states, i. e. element of $\mathcal{K}$ (see Definition 3.2). On the level of item pairs, response vectors with either both or neither of the items solved correctly do not give any evidence on the model's validity. Thus, for some of the validation methods, which include trivial response vectors or item pairs, floor and ceiling effects can lead to an overestimation of the model's fit.

Finally, it should be mentioned that all computer programs (Hockemeyer, 2001; Hockemeyer and Pötzi, 2001; Pötzi, 2001; Pötzi and Wesiak, 2001) necessary for the discussed validation procedures are available at the Cognitive Science Section at the University of Graz. For a list and short description of the used programs see Appendix D.

3.4.1 Validation of hypotheses via the surmise relation

The fit of a surmise relation to a set of data can be estimated by calculating the relative frequencies of correct solutions for each item or by computing various indices, which measure the overall fit of a hypothesized surmise relation to a set of data. As mentioned above, the validation methods will be demonstrated for a surmise relation on a single test. For a set of tests, the same methods are applied to the various subsets of the relation. For the *SRwT* the pairs within each test are considered separately, for the *SRxT* only the pairs between items of two different tests are considered, and for the *SRbI* all pairs are taken into account. The validity of the *SRbT* is estimated by means of the relative solution frequencies (see below), by checking whether the requirements

Table 3.3: Data set for four items and 10 response patterns (Example 3.4)

pattern	item a	item b	item c	item d	Σ
1	0	0	1	0	1
2	0	0	0	1	1
3	0	0	1	1	2
4	0	1	1	1	3
5	0	1	1	1	3
6	0	0	1	1	2
7	0	0	0	1	1
8	0	1	1	0	2
9	0	1	0	0	1
10	1	1	1	0	3
Σ	1	5	7	6	19

Note. 1 = correct response, 0 = incorrect response; Σ = number of correct solutions per pattern and item.

Table 3.4: Contingency tables for Example 3.4

		b	
		0	1
c	0	2	1
	1	3	4

(a)

		a	
		0	1
c	0	3	0
	1	6	1

(b)

		b	
		0	1
d	0	1	3
	1	4	2

(c)

		a	
		0	1
d	0	3	1
	1	6	0

(d)

		a	
		0	1
b	0	5	0
	1	4	1

(e)

for an eventual general, left–, right–, or total–covering surmise relation are fulfilled (see Section 3.2, Definitions 3.4, 3.5 and 3.6).

Percentage of correct solutions Imagine a surmise relationship between two items x and y, where item y is prerequisite for item x (ySx). The admissible response vectors for the items are that both items are solved correctly $\langle 1,1 \rangle$, neither of the items is solved correctly $\langle 0,0 \rangle$, or only item y is solved correctly $\langle 0,1 \rangle$ (see also Section 3.1). Since x should only be solved in combination with a correct solution to y, it is expected that the solution frequency for item y is equal or higher than the solution frequency for item x (Albert and Held, 1994). In the Hasse diagram the lower item is supposed to have the higher solution frequency. In the case of reversed solution frequencies (x has a higher solution frequency than y) a one–dimensional χ^2 statistic is calculated to decide whether the difference is significant or not.

Regarding Example 3.4 the relative solution frequencies for items a, b, c, and d are 10%, 50%, 70%, and 60% respectively. Figure 3.11 depicts the hypothetical surmise relation together with the relative solution frequency for each item.

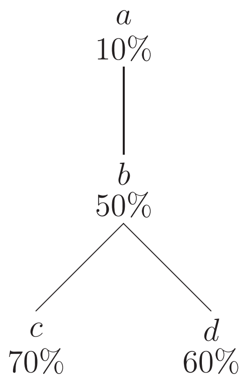


Figure 3.11: Percentage of correct solutions for Example 3.4

As can be seen in the Hasse diagram, the hypothesis is verified by all values (lower items always show higher solution frequencies than the items they are prerequisite for). However, from the percentage of correct solutions it is not possible to infer, whether the individual response vectors for each item pair are in accordance with the surmise relation (i. e. whether the responses only reflect pairs which are elements of the relation). Therefore, the percentage of correct solutions has to be viewed as a quick method to get a general impression of the hypothesis' validity.

Indices for the fit of a surmise relation In order to evaluate the hypothesis on an individual level, I will apply two indices which measure the fit of a hypothetical surmise relation to a set of data.

The first index is called *gamma–index* (γ). It was initially proposed by Goodman and Kruskal (1972) and is based on the item easiness index by Scheiblechner (1997; see also Körner, 2000). In Section 3.1, I described the possible correct/incorrect response patterns for an item pair. Assuming a prerequisite relationship between two items x and y, where item y is prerequisite for item x, it was pointed out, that the response pattern $\langle 1, 0\rangle$ contradicts the hypothesized relationship (item x is solved correctly, its prerequisite y incorrectly). These cases are called discordant pairs (cell b_{xy} in Figure 3.2). The pattern $\langle 0, 1\rangle$, on the other hand, confirms the hypothesis, because some of the participants are expected to solve item y but not the more difficult item x. These cases are called concordant pairs (cell c_{xy} in Figure 3.2). The cases in which either both or none of the items are solved (cells d_{xy} and a_{xy} in Figure 3.2) are neither contradicting nor confirming the hypothesis. Thus, it is not possible to draw a conclusion about the predicted relationship between the items.

Since the cases a_{xy} and d_{xy} are ambiguous, the γ–index only uses the number of concordant and discordant answer patterns to calculate the validity of each pair in a surmise relation. Furthermore, a global index (γ_G) is computed to evaluate the overall fit of a model by accumulating the frequencies of concordant and discordant cases over all pairs of items in the surmise relation. Formally, the γ–index is defined by

$$\gamma_{xy} = \frac{N_c - N_d}{N_c + N_d}, \tag{3.12}$$

with N_c being the number of concordant pairs (cell c_{xy}) and N_d the number of discordant pairs (cell b_{xy}) over all response patterns. The γ–index varies between -1 and $+1$ ($\gamma \in [-1, 1]$), with $+1$ indicating a perfect fit (no contradictions at all). Item pairs that are incomparable with regard to a given hypothesis [$(x, y) \notin S$ and $(y, x) \notin S$; within the Hasse diagram, there is no direct or indirect line connecting the two items] are not taken into account, because they are neither confirming nor contradicting the hypothesis. In order to decide, whether the derived γ value supports the hypothesis, a McNemar χ^2 test is calculated to compare the number of concordant pairs to the number of discordant pairs.

For the five pairs of the surmise relation in Example 3.4, the contingency tables depicted in Table 3.4 show the frequencies for the concordant (N_c) and discordant (N_d) pairs in the hypothesized relation.

Looking, for example, at Table 3.4a, the γ–index is computed as follows:

$$\gamma_{(c,b)} = \frac{N_c - N_d}{N_c + N_d} = \frac{3 - 1}{4} = 0.5$$

The indices for the remaining item pairs amount to $\gamma_{(d,b)} = .14$, $\gamma_{(d,a)} = .71$, and $\gamma_{(b,a),(c,a)} = 1$. Thus, all of the five pairs in the relation show a positive γ value, which indicates that N_c is always larger than N_d. For the computation of the global index (γ_G) the frequencies for N_c and N_d are added over all pairs in the relation.

$$\gamma_G = \frac{N_c - N_d}{N_c + N_d} = \frac{23 - 5}{28} = 0.64$$

With $\chi^2(1, N = 28) = 10.94, p < 0.005$ (after Yate's continuity correction) the number of concordant pairs is significantly larger than the number of discordant pairs, which supports the hypothetical surmise relation.

The second index used in the validation process is called *violational coefficient VC* (Schrepp et al., 1999). The index compares the fit of a surmise relation to a given data set by counting the number of violations or contradictions for each pair ySx in a surmise relation S. Violations are defined as those cases, in which a person mastered item x but failed in mastering item y. Formally, the VC is defined as follows:

$$VC = \frac{1}{n(|S| - m)} \sum_{x,y} v_{xy} \tag{3.13}$$

with n denoting the number response vectors, m the number of items, $|S|$ the number of pairs in the relation, and v_{xy} the number of violations for all pairs in S. Thus, the VC value denotes the averaged number of violations of the item pairs ySx contained in a surmise relation S (with $x \neq y$). The index varies within the limits of 0 and 1 ($VC \in [0, 1]$), with 0 denoting a perfect fit (no violations at all). For Example 3.4 the VC value is calculated as follows (the sum of v_{xy} can be inferred from Table 3.4 by adding the frequencies of the cells b_{xy} over all item pairs).

$$VC = \frac{1}{10(9-4)} \times 5 = 0.1$$

The obtained VC value indicates that 10% of the empirical response vectors contradict the item pairs contained in the hypothetical surmise relation. Note, that contrary to the γ–index, VC also includes the response vectors $\langle 0, 0\rangle$ and $\langle 1, 1\rangle$ and represents therefore a weaker test of the hypothesis.

With regard to the interpretation and/or applicability of the indices, it has to be considered that they are pragmatical approaches to test the fit of a surmise relation and that they are primarily used to compare different models. There is also no statistical test available to judge the significance of VC. However the indices can be used to compare the fit of the *SRbI*, the *SRwT*, and the *SRxT* to each other. Such a comparison is valuable in order to uncover which part(s) of the *SRbT* has (have) to be refined.

3.4.2 Validation of hypotheses via the knowledge space

The general idea of validation procedures via the knowledge space is to compare the hypothetical knowledge states to a set of empirical response patterns. First the averaged minimal symmetric distance between a knowledge space and a set of empirical response patterns is calculated (this distance will be referred to as *ddat*). In a second step the *distance agreement coefficient DA* is calculated, which estimates the fit between a knowledge structure and a binary data matrix by taking account of the structure's size (i. e. the number of knowledge states $|\mathcal{K}|$). This is necessary in order to compare the distances for knowledge spaces of varying sizes (the distance *ddat* decreases with an increasing number of knowledge states). Since the symmetric distances yield a frequency distribution for the distances between the empirical response patterns and the postulated knowledge states, it is also possible to estimate the fit of the knowledge space by comparisons with simulated data sets.

In the case of a set of tests, the same methods are applied to the various substructures of the test knowledge space (see Section 3.2). Furthermore, the knowledge spaces induced by the $KSbI$, the $KSwT$, and the $KSxT$ are compared to each other in order to decide which part of the test knowledge structure deviates most from the empirical response patterns. Analogous to the $SRbI$ and its subsets, the $KSbI$ and its substructures are validated by considering only the relevant parts of the postulated test knowledge states and the empirical response patterns. For the $KSwT$, only the part of the knowledge states referring to the items in a single test as well as the corresponding part of the response patterns are considered. For the $KSxT$, only the parts denoting prerequisites between items of different tests are considered and for the $KSbI$ the whole test knowledge space is taken into account. The $KSxT$ and the $KSbI$ require the complete response patterns for the validation procedures.

Symmetric distances Symmetric distances between a knowledge structure and a binary data matrix denote the averaged minimal distance or number of deviations between each person's response pattern and the nearest hypothesized knowledge state (e. g., Garnier and Taylor, 1992; Albert et al., 1994; Kambouri et al., 1994). Generally, the *distance* d between two sets A and B is defined as the number of elements contained in the symmetric set difference of A and B and abbreviated by $d(A, B)$.

$$d(A, B) = |A \bigtriangleup B|, \text{where } A \bigtriangleup B = (A \setminus B) \cup (B \setminus A). \tag{3.14}$$

Hence, the symmetric distance of two sets equals the number of elements that are contained in one set but not in the other.

The *minimal distance* between a response pattern $r \in R$ and a knowledge space $\mathcal{K}$ is defined as the distance to the nearest knowledge state $K \in \mathcal{K}$ and abbreviated by $dmin(r, \mathcal{K})$.

$$dmin(r, \mathcal{K}) = min\{d(r, K)|K \in \mathcal{K}\} \tag{3.15}$$

Now, the averaged minimal distance (which will be called $ddat$) between a set of response patterns R and a knowledge space $\mathcal{K}$ can be defined as follows:

$$dmin(R, \mathcal{K}) = \frac{\sum\{dmin(r, \mathcal{K})|r \in R\}}{n} = ddat, \tag{3.16}$$

with n denoting the number of response patterns in R. For two sets, there is also a distance's theoretical minimum ($dmin$) and maximum ($dmax$). The theoretical minimum corresponds to perfect results, i. e. there are no deviations at all ($dmin$ always equals 0). The theoretical maximum denotes the greatest possible distance with regard to the number of items, which is equivalent to the greatest integer k that is smaller or equal $\frac{|Q|}{2}$ ($dmax = \left\lfloor \frac{|Q|}{2} \right\rfloor = \frac{n}{2}$ or $\frac{(n-1)}{2}$ for even and odd numbers respectively).

Table 3.5 shows the minimal symmetric distances for Example 3.4. With $|Q| = 5$ items the maximal distance $dmax$ equals two. The observed distances show a frequency of seven for $dmin(r, \mathcal{K}) = 0$ and a frequency of three for $dmin(r, \mathcal{K}) = 1$. Thus, the distance distribution is positively skewed and the averaged minimal distance ($ddat = .3$) is much smaller than the theoretical maximum. Regarding a distribution's skewness, the fit of a knowledge space to a data set is better the more positive the distribution's skew (for a perfect fit all distances equal zero). It should be noted that the distance for trivial response patterns (all or none of the items solved correctly) always equals zero.

Table 3.5: Symmetric distances between the data set and the hypothesized knowledge space for Example 3.4

$dmin(r, \mathcal{K})$	frequency f	$dmin$	$dmax$	$ddat$ (SD)	Mdn
0	7	0	2	0.3 (0.46)	0
1	3				

For a statistical test of the knowledge structure's empirical validity, Heller (2001) suggests to use the frequency distribution of the symmetric distances. Starting with the null hypothesis that there is no structure inherent in the empirical data, the validity of the postulated knowledge structure is estimated by means of a one–dimensional χ^2 statistic. The distance distribution of the items' powerset P is used as basic model for the test (a more detailed description of P is given in the next subsection). The question is, whether the distance distribution for the observed response patterns differs significantly (is more positively skewed) from the distribution for the expected patterns (generated with the powerset P with $2^{|Q|}$ elements). Whenever the null hypothesis is accepted, it has to be assumed that there is no structure in the empirical data and that the data therefore contradict the hypothetical knowledge space.

Table 3.6 shows the observed and expected distance distributions as well as the derived χ^2 values for Example 3.4. To obtain the expected frequencies, the relative frequencies for each distance $dmin(., \mathcal{K})$ are multiplied with the number of empirical response patterns $N = 10$. With $k - 1 = 1$ degree of freedom and at a 5% level of significance ($\chi^2_{crit} = 3.84$), the null hypothesis is rejected, because of $\chi^2_{obs}(1, N = 10) = 4.51$. Thus, the set of response patterns reflects the hypothesized knowledge structure in the example better than the powerset does.

Assuming a valid test knowledge space, it still has to be considered that for the knowledge spaces between items, within, and across tests the number of knowledge states varies. Therefore, I will next apply a validation method which accounts for the structures' sizes, that is the number of knowledge states $|\mathcal{K}|$.

Distance agreement coefficient As I am mainly interested in the validity of the test knowledge space, it is necessary to find out, whether the distance of the entire structure ($KSbI$) is primarily due to the distances within or across tests. Regarding the varying sizes of the hypothetical substructures, the $KSbI$, $KSwT$, and $KSxT$ can

Table 3.6: Distance distributions and χ^2 values for Example 3.4

	powerset $(P, \mathcal{K})$		data set $(R, \mathcal{K})$		
$dmin(., \mathcal{K})$	absolute f	relative f	observed f	expected f	χ^2_{obs}
0	6	0.3750	7	3.75	2.82
1	9	0.5625	3	6.25	1.69
2	1	0.0625	0	0	0
$\sum$	16	1.0000	10	10.00	4.51

only be compared by taking into account the number of knowledge states within the respective powersets. Hence, I will calculate the *distance agreement coefficient* DA (Schrepp, 1993; Schrepp et al., 1999), which measures the fit between a knowledge space and a data set under consideration of the structure's size. The index indicates the validity of a knowledge space $\mathcal{K}$ by relativizing the empirical distance *ddat* to the mean distance of the space's powerset P (*dpot*). Formally DA is defined by

$$DA = \frac{ddat}{dpot}, \tag{3.17}$$

with *dpot*[1] denoting the averaged minimal distance between a knowledge space $\mathcal{K}$ and its powerset P. The value *dpot* yields the expected distance for random response patterns, i. e. if $\mathcal{K}$ cannot account for the behavior of participants. Generally, *dpot* is an inverse function of the number of states contained in $\mathcal{K}$. The term *ddat* is defined in Equation 3.16. DA varies between the limits of 0 and $\frac{dmax}{dpot}$ ($DA \in [0, \frac{dmax}{dpot}]$). A lower value of the distance agreement coefficient indicates a better fit of a knowledge structure to a given set of data. In the case of a small *ddat* value that is due to a large number of knowledge states, DA compensates the small value by a small *dpot* value.

For Example 3.4, the powerset P contains $2^4 = 16$ elements. The distance distribution, *dpot*, and the coefficient DA are given in Table 3.7.

Table 3.7: Distance distribution for the power set (*dpot*) and DA for Example 3.4

$dmin(p, \mathcal{K})$	frequency f	$dpot$ (SD)	Mdn	DA
0	6	0.688 (0.58)	1	0.44
1	9			
2	1			

As mentioned above, the index DA is mainly applied to compare alternative hypotheses or models. To evaluate a single knowledge space and its respective distance distribution by means of different data sets, I will simulate two types of data sets for a comparison with the empirically observed response patterns.

[1] $dpot = \frac{\sum\{dmin(p,\mathcal{K})|p \in P\}}{2^{|Q|}}$, with $|Q|$ denoting the number of items.

Simulations In order to compare the goodness of the empirical distance ($ddat$) and the DA value with other data sets, I will compute simulations with different degrees of specificity. The least specific type is the computation of random patterns for the respective number of items. In this case knowledge states are randomly drawn from the states contained in the powerset. In the following, this type of simulation will be called *random simulation*. For various sets of random data the mean symmetric distances between the random data sets and the respective knowledge structure are calculated ($dsim_r$). Afterwards, a one–dimensional χ^2 test is applied to estimate, whether the fit between the empirical data set and the knowledge structure is better than chance.

In a second approach, which will be referred to as *probability simulation*, patterns are simulated on the hypothetical knowledge structure under consideration of the probabilities for careless errors (β) and lucky guesses (η). In case of multiple–choice items the probabilities for lucky guesses results from the number of answer alternatives (e. g. 8 alternatives allow 12.5% guesses), whereas the probability for careless errors is unknown. Therefore, I will vary the probability for careless errors (with $0.05 < \beta \leq 0.15$) and simulate sets of response patterns for each β value. As for the random simulations, the mean distances ($dsim_p$) and DA values are calculated and compared to the empirical results. However, the interpretation of the results differs.

While it is expected that the empirical data fit the hypothetical knowledge structure significantly better than random simulations, for probability simulations it is expected that there is no significant difference between the two types of data. In the first case, the aim is to show that there are differences between the empirical and the simulated data, i. e. that the agreement of the empirical data set with the postulated model is not put forth by random. Simulations on the hypothetical structure, on the other hand, reflect response patterns under the assumption of a correct model, but permitting a certain amount of noise in the data. Hence, if the empirical set of data does not deviate significantly from the simulated data sets (or even shows smaller distances), deviations can be attributed to noise in the data and it can be assumed that the model is correct.

Finally, the most specific type of simulation, which will be referred to as *frequency simulation*, is to compute data matrices under consideration of the solution frequencies for items and persons. This means that the marginal frequencies of the simulated matrices are equivalent to those of the original data matrix, whereas the distributions of '0's and '1's differ. The used algorithm[2] (Ponocny and Waldherr, 2002) randomly selects two rows and two columns of the given matrix. If the four selected entries show either the pattern $\begin{smallmatrix}1&0\\0&1\end{smallmatrix}$ or $\begin{smallmatrix}0&1\\1&0\end{smallmatrix}$, the '0's and '1's are exchanged. Otherwise, the algorithm keeps the pattern and selects another pair of rows and columns. The total number of potentially exchanged entries amounts to $n \times m + 3000$, with n denoting the number of response patterns and m the number of items. As for the other two types of simulations, various data matrices will be simulated and their symmetric distances ($dsim_f$) to the corresponding knowledge spaces as well as the DA values will be calculated. For the interpretation of the results, it has to be considered that the postulated model assumes that a partial order reflects the dependencies among items better than a linear order. This means, that not only the solution frequencies of the items (and persons) but also

[2] I thank Ivo Ponocny for providing the algorithm, which was adapted by C. Hockemeyer.

the specific patterns of the responses are relevant for the prediction of testee's solution behavior. Thus, for the frequency simulations, the expectation is, that the empirical data matrix significantly differs from the simulated matrices. Being the most specific type of simulation and therefore the strictest test of the hypothesis, the frequency simulation will only be applied for data sets that confirm the hypothesis according to the random and probability simulations.

For an illustration of these approaches, I simulated various data sets for Example 3.4. Table 3.8 shows the distance distributions and DA values for the empirical response patterns ($ddat$), sets of random data ($dsim_r$), data sets simulated under consideration of the knowledge space hypothesis ($dsim_p$), or the solution frequencies ($dsim_f$). For each type of simulation, 10 different sets of data with 10 response patterns each were generated. For $dsim_r$ the response patterns were drawn randomly from the power set, which contains $2^4 = 16$ states. For $dsim_p$ the probabilities for lucky guesses (η) and careless errors (β) were set to $\eta = 0.125$ and $0.05 < \beta \leq 0.15$ and for $dsim_f$ the original data matrix was permuted as described above (with 10 response patterns and 4 items, there are $10 \times 4 + 3000 = 3040$ potential exchanges of entries). The numbers in Table 3.8 denote the averaged values of the 10 data sets for each type of simulation.

Table 3.8: Average distances and DA coefficients for simulated data sets

$dmin(.,\mathcal{K})$	$ddat$	$dsim_r$	$dsim_p$	$dsim_f$
0	7	4.2	8.0	6.6
1	3	5.1	1.9	3.4
2	0	0.7	0.1	0
M (SD)	0.30(0.46)	0.65 (0.57)	0.21 (0.38)	0.34 (0.05)
DA	0.44	0.95	0.31	0.49
z		-1.84	0.66	-0.76
$\chi^2(df = 1)$		3.22	0.63	0.07

Note. The values for $dsim_r$, $dsim_p$, and $dsim_f$ are averaged over 10 sets of data á 10 response patterns for each type of simulation.

To assure the differences between the empirical and simulated data, I standardized the empirical values (obtaining z–scores) and calculated one–dimensional χ^2 statistics to estimate the differences with regard to the distribution of the symmetric distances (for larger item and data sets a Mann–Whitney U test will be calculated to estimate differences regarding the central tendency; in this example there are too many tied ranks). At a 5% level of significance ($z_{crit} = 1.96$; $\chi_{crit} = 3.84$) there is no difference between the empirical distribution and the three simulated distributions. Thus, with the small sample in this example, the hypothesis is neither contradicted nor confirmed and the results are therefore ambiguous.

3.5 Knowledge spaces for inductive reasoning

In this last section of the theoretical part of this report, I want to show how the theory of knowledge spaces has already been applied for the domain of inductive reasoning. Up to now, there are several investigations, in which the items of single problem types have been ordered on the basis of knowledge space theory. The investigated problem types include letter series completions (Schrepp, 1995, 1999; Wriessnegger, 2000), number series completions (Albert and Held, 1994, 1999; Ptucha, 1994), and geometric matrices (Musch and Albert, 2003). Since letter series completion problems are not part of my own research, I will demonstrate the approach by the two problem types number series completions and geometric matrices. For number series completion problems, the approach taken by Albert and Held (1994, 1999) will be reported. Both Albert and Held's (1994; 1999) as well as Musch and Albert's (2003) approaches are based on the principle of componentwise ordering of product sets (see Section 3.3.1), which also constitutes the method for hypothesis generation in my own work.

3.5.1 Number series completions

Albert and Held (1994, 1999) investigated the solution of number series completion problems (see also Section 2.2.2) as one empirical example for the component based construction of problems and the application of ordering principles, such as the componentwise ordering of product sets (see Section 3.3.1). Albert and Held started with the assumption that two cognitive demands are required for the solution of number series completions, namely (1) the recognition of the properties and regularities of a sequence and (2) the establishment, application, and testing of a hypothesis on the rule that governs the sequence. The problems themselves are described by solution formulae, such as $x_n = x_{n-1} + 2^n$ as rule for the number series 30 32 36 44 60 (x_n denotes the number to be found, x_{n-1} its immediate predecessor).

For the component based construction of problems, Albert and Held specified three distinct components in order to construct 12 classes of number series completions with varying difficulty.

The first component, the *level of recursion* (M_1), is based on the number of immediate predecessors used for the solution (see also Krause, 1985). For example, the problem descriptions $x_n = x_{n-1}$ and $x_n = x_{n-1} + x_{n-2}$ have recursion levels of '1' and '2' respectively. Albert and Held specified three attributes on component M_1 ($M_1 = \{a_1, a_2, a_3\}$), namely the levels of recursion 1, 2, and 3 (1 corresponds to attribute a_3, etc.). The attributes are ordered linearly, with a_1 being the most difficult and a_3 being the least difficult attribute.

The second and third component are the *multiplicative* (M_2) and *additive factor* (M_3). The two factors specify, whether or not it is necessary to multiply one of the predecessors by some factor and/or to use an addend in the pattern description. Examples for problem descriptions with and without the multiplicative factor are $x_n = 2x_{n-1}$ versus $x_n = x_{n-1}$, and for the description with and without the additive factor are $x_n = x_{n-1} + 4$ and $x_n = x_{n-1}$. The components multiplicative and additive factor consist of two attributes each, namely presence versus absence of a multiplicand or an addend ($M_2 = \{b_1, b_2\}$, $M_3 = \{c_1, c_2\}$). The attributes b_1 and c_1 denote that a multiplicative factor > 1 and an additive factor > 0 are present, the attributes b_2 and c_2 denote that the multiplicative factor equals 1 and that the additive factor equals 0. The attributes b_1 and c_1 are assumed to be more difficult than the attributes b_2 and c_2 respectively.

Forming the Cartesian product of the three components results in 3 x 2 x 2 = 12 classes of number series completion problems with varying difficulty. Accordingly, Albert and Held constructed 12 solution formulae and one number series completion problem per formula. Table 3.9 shows some examples of the developed formulae and items.

Table 3.9: Examples for the problem construction of number series completions (from Albert and Held, 1999)

Attributes	Solution formula	Item	Solution
a_1, b_1, c_1	$x_n = 2x_{n-3} + x_{n-2} + x_{n-1} + 4$	1 5 9 20 43 85	172
a_1, b_2, c_2	$x_n = x_{n-3} + x_{n-2} + x_{n-1}$	26 34 41 101 176 318	595
a_2, b_1, c_1	$x_n = x_{n-2} + 2x_{n-1} + 2$	1 4 11 28 69 168	407
a_2, b_2, c_1	$x_n = x_{n-2} + x_{n-1} + 5$	12 17 34 56 95 156	256
a_3, b_1, c_2	$x_n = 2x_{n-1}$	4 8 16 32 64 128	256
a_3, b_2, c_2	$x_n = x_{n-1}$	113 113 113 113 113	113

For the construction of a knowledge space, Albert and Held applied the coordinatewise as well as the lexicographic ordering principle (see Section 3.3.1), which resulted in knowledges spaces with 50 and 13 knowledge states respectively (the powerset for 12 problems contains $2^{12} = 4096$ elements). For the coordinatewise order all three components were viewed as independent of each other, for the lexicographic order component M_1 (level of recursion) was assumed to be most important, component M_2 (multiplicative factor) second most important, and component M_3 (additive factor) least important.

For an empirical validation of the two knowledge spaces, Albert and Held conducted two investigations with altogether 48 response patterns. The trivial problem (a_3, b_2, c_2) was excluded from the analysis, because in one investigation it was solved by 100% of the participants, in the other it was not presented. Regarding the remaining 11 problems, the response patterns of the participants yielded a mean symmetric distance (see Section 3.4.2) of 0.44 for the coordinatewise order and a mean distance of 1.23 for the lexicographic order. In the first case 34 response patterns showed a distance of zero, in the second case 10 response patterns (however, since the lexicographic order is much more restrictive, a comparison of the two values is not possible).

Regarding the small distances derived by the coordinatewise ordered knowledge space, the results of Albert and Held's investigations clearly show that item difficulty is increased by higher levels of recursion and by the application of the multiplicative as well as the additive factor.

3.5.2 Geometric matrices

The knowledge space approach to the problem type geometric matrices (see also Section 2.2.3) was taken up by Musch and Albert (2003), who established a complete item taxonomy for the Sets I and II of Raven's APM (1965, see Section 2.4.2.3). Their aim was to develop a formal model for the description of item characteristics and individual differences in performance on the APM.

Applying the principle of componentwise ordering of product sets (see Section 3.3.1), Musch and Albert defined three distinct components with varying numbers of attributes each. The specification of the three components and their attributes is based on the results of previous investigation on the APM, especially Carpenter et al. (1990), Jacobs and Vandeventer (1972), Kinder and Lachnit (1994), and Vodegel Matzen et al. (1994). The derived components are (1) number of rules, (2) types of rules, and (3) material attributes.

The first component (1) refers to the *number of rules* (N) or operations necessary to solve the problem. For the 46 analyzed matrices of the APM, the number of rules ranges between one and four. Thus, component N has four attributes, N = {1,2,3,4}. It is assumed that a higher number of rules leads to higher item difficulty.

Component (2) refers to the *types of rules* (T) that govern the variations among the items' elements. In order to account for all items of the APM, Musch and Albert (2003) extended the five rules specified by Carpenter et al. (1990, see Section 2.2.3) to a set of seven rules (see Table 2.2 for a description of the rules). The rules are divided into two attribute classes, namely the application of the more difficult exclusive–OR rule and rules of other types, viz. constant in a row, pairwise progression, distribution of two or three values, figure addition, and figure subtraction. Thus, component T has two attributes, T = {X,O}, with X denoting the presence of the exclusive–OR rule and O denoting that only rules of other types are necessary to solve a given item.

The rules operate on materials of different detectability. The last component (3) considers these *material attributes (M)*, which influence the difficulty in the correspondence finding process. The two attribute classes of this component are spatial order, which is supposed to be less salient (high demand on correspondence finding process, low detectability), and materials making low demands on the corresponding finding process due to higher salience (high detectability). The latter attribute class includes variations in geometric figures, patterns, and number. With the two attribute classes high demands (H) and low demands (L), component M = {H,L}.

Regarding the item taxonomy, Musch and Albert (2003) classified the set of APM items according to the number (N) and type (T) of rules involved, as well as the difficulty of the material attributes (M). As an example, a matrix item with two rules of the types constant in a row (CR) and distribution of three values (D3) operating on the geometric form and shading of the elements would therefore be described by (2,O,L). Looking, for example, at the geometric matrix in Figure 2.6, the testee's task is to detect and apply the CC rule to the inner part of the figures and the D3 rule to the outer part of the figures. Since neither of the two rules is applied to spatial order, the detection of the relevant elements is supposed to be low in difficulty (low demand).

A combination of the attributes of all three components results in 4 x 2 x 2 = 16 possible item classes, of which 10 are actually realized in the APM. The number of items per class varies between one and 13. In order to establish a knowledge space on the 16 possible item classes by the componentwise ordering principle, it had to be decided, whether the components should be ordered coordinatewise, lexicographicly, or as a mixture of both (see Section 3.3.1). Furthermore, it had to be decided, whether all three components or only a subset of the three components should be taken into account. Since there was no evidence on the importance of the single components, Musch and Albert (2003) specified all possible models on the set of items, which are either based on a coordinatewise or a lexicographic order on all three components or on a subset of one or two components. Altogether, they derived 19 different models (7 based on a coordinatewise order, 12 on a lexicographic order), which they first evaluated by means of the relative solution frequencies (see Section 3.4.1) from a data set containing 1015 response patterns. After this first estimation of the models' fit, 14 models were excluded from further validation procedures. The remaining five models (4 are based on a coordinatewise order, one on a lexicographic order) were validated by two data sets of 41 and 44 response patterns. The number of realized item classes per model varied between 3 and 10. The results derived via the mean symmetric distances and the distance agreement coefficient DA (see Section 3.4.2) show for both samples that the models based on coordinatewise ordering fit the empirical data better than the lexicographic model (for the two samples, DA ranged between 0.00 and 0.40 for the coordinatewise ordered models and amounted to $DA = 0.24$ and 0.68 for the lexicographic model; the reported values refer to a threshold of 75% with regard to the percentage of items solved within an item class).

Musch and Albert (2003) concluded that a partial order on the components is more suitable than a lexicographic order to model item difficulty and individual performance on the APM. Furthermore, the results indicate that the three specified components type and number of rule, as well as material attributes influence item difficulty.

3.6 Summary of Chapter 3

Knowledge space theory is a non–numerical approach, which can be used for both the representation of a knowledge domain and the diagnosis of an individual's performance state in this domain. Starting out with the idea of prerequisite relationships or surmise relations between items, Doignon and Falmagne (1985; 1999; Falmagne et al., 1990) developed a mathematical model, in which the structure of a given knowledge domain as well as people's knowledge states are formally defined (Section 3.1). Albert and his group (1995; Albert et al., 2003, Brandt et al., 1999; 2003) generalized the model to surmise relations between tests, where sets of items or tests are the basic units (Section 3.2). There are several methods to construct a surmise relation or knowledge space (Section 3.3), of which the principle of componentwise ordering of product sets by Albert and Held (1994, 1999) is of most importance to my research. As outlined in Section 3.3.1, the principle of componentwise ordering assumes that all items in a given knowledge domain can be described by the same set of components. Furthermore, each component consists of a set of attributes, on which order relations are defined.

Forming the Cartesian product of the components results in all possible attribute combinations (item classes), which are then compared to specify an order on the set of item classes. The principle of componentwise ordering has already been successfully applied to single inductive reasoning tests (see Section 3.5). Applying the principle to sets of tests requires the specification of components, which account for all occurring types of problems. Thereafter, it is possible to establish a surmise relation between tests by following the same procedure as for items. The surmise relation between tests has the advantage that items of different problem domains can be ordered in a common test knowledge structure and later be used for adaptive testing procedures. Before implementing a (test) knowledge structure into an adaptive testing system it is of course necessary to validate the structure. In Section 3.4 several validation methods are introduced which are either based on the surmise relation (Section 3.4.1) or the knowledge space (Section 3.4.2). In this work both approaches are applied.

4 Purpose and Scientific Questions

The purpose of the presented study is to establish an integrative structure for different types of inductive reasoning problems in order to lay the basis for an adaptive testing instrument in this domain. In Section 2.2, I described several types of inductive reasoning problems including the demands required to solve these problems. The presentation of the problems followed the customs in common literature on this topic, namely separately for each problem type. After discussing the components that influence item difficulty of the various tests, I compared the components of different problem types with respect to corresponding elements, and thereby arrived at a set of components all types of problems have in common (see Section 2.2.4). That various problem types and their underlying cognitive demands can be described by a set of common components, has already been shown within the two inductive reasoning models by Sternberg and Gardner (1983) and Klauer (2001), which are outlined in Section 2.3. Both models focus on the similarities and differences of various problem types. My goal is to integrate the findings on the common components of different problem types and the findings on the varying demands required by the single items that represent a problem type. Thus, the general objective is to integrate various types of problems into one common classification scheme that specifies the problem requirements and is able to relate items of different tests.

For the establishment of such a classification scheme, a structural approach will be taken, which is comparable to some of the existing models of intelligence and inductive reasoning. The BIS model (see Section 2.4.1.2) covers a set of four operational and a set of three content related dimensions of intelligence. Forming the product of the two sets results in 4 x 3 = 12 ability combinations (similarly, Guilford's Structure–of–Intellect model specifies 150 ability dimensions by the combination of five operations, five contents, and six products, see Section 2.4.1.1). Going a step further to Klauer's model of inductive reasoning (see Section 2.3.1), the problem types are also described by a set of content related demands (with 5 elements), but the operational abilities are specified with respect to inductive reasoning. For this specification, Klauer (2001) differentiates between (A) the detection of similarities, dissimilarities, or both and (B) the comparison of elements or attributes. Forming the product of the three sets results in 5 x 3 x 2 = 30 types of inductive reasoning problems. On the next level, the items of single inductive reasoning problem types are analyzed to identify the components contributing to the difficulty of individual items. Examples for such analyses are given within the descriptions of inductive reasoning tasks in Section 2.2. Furthermore, Formann and Piswanger (1979) used construction principles for the WMT (see Section 2.4.2.4), that are based on three components with varying numbers of attributes. The result are 36 classes of geometric matrices, which are composed of three types of rules, four material attributes, and three rule directions. A similar approach was taken by some of the applications of knowledge space theory for the structuring of inductive reasoning problems. In Sections 3.5.1 and 3.5.2, I outlined the principles for the componentwise ordering of number series completion problems and Raven's APM, which are also based on the specification of a set of components with varying attributes.

What is still missing, though, is the specification of components for various problem types on the level of single items. A first attempt to the establishment of such a classification scheme is made in Section 5.1, where I will specify a set of common components and their respective attributes for the four types of inductive reasoning problems used in the succeeding investigations.

Since the obtained classification scheme should be applicable as a basis for an adaptive testing system of inductive reasoning, it is also necessary to establish an order on the set of items and problem types (see Chapter 6 for adaptive testing applications). Thus, another question concerns the development of a test model, in which the established classification system as well as the derived order can be validated.

In Section 2.4.2, I presented a selection of tests measuring inductive reasoning abilities either as a subtest within a test assessing several dimensions of intelligence (ISA and BIS test, see Sections 2.4.2.1 and 2.4.2.2) or as specific intelligence test assessing the ability of analytic reasoning by means of geometric matrices (APM and WMT, see Sections 2.4.2.3 and 2.4.2.4). As of now, several tests or subtests are presented to assess a person's abilities in solving analogies, series completions, or matrices, as well as for the assessment of how capable a person is in processing different content types such as verbal, numerical, or geometric–figural material. As diagnostic result the testees are described by a standardized numerical test score or an ability parameter for each (sub)test and/or for an ability scale. Even if construction principles are used for the development of the test items (as e. g., for the WMT), the test results are usually not intended to provide information on the demands a testee is (un)able to fulfill.

Obtaining more detailed information on a testee's performance has the advantage that the ability level of the testee is not only given by an ability parameter (based on the number of correctly solved items), but by the problem requirements the testee is able to meet. This additional information allows for an exact feedback regarding the lacking skills and can be employed for tutoring purposes. The information can also be used in terms of item difficulty characteristics, which are the basis for more efficient, adaptive testing procedures. Dealing with various types of inductive reasoning problems, the established test model should also be able to predict from a person's performance in one of the tests the response behavior in one or more of the other tests.

In order to combine the request for detailed item descriptions and for precise information on the set of problem demands that are met by a person, the theory of knowledge spaces provides an appropriate framework.

The knowledge space theory, as outlined in Chapter 3, has been developed as a methodological framework for precise and efficient knowledge assessments. Considering the recent developments of the theory, such as the generalization of surmise relations between items to surmise relations between tests, it provides a promising approach for the establishment of an integrative diagnostic system. The general reasoning is that a mathematical model permits a precise representation of information in a given knowledge domain and additionally adapts easily to computerized systems.

Foremost, it is necessary to construct a plausible test knowledge structure, which contains all relevant prerequisite relationships within, across, and between inductive reasoning tests. Considering the methods for the establishment of knowledge structures (see Section 3.3), the principle of componentwise ordering of product sets seems to be most suitable for the investigated domain. Inductive reasoning tests have been studied intensively. Thus, the derivation of hypotheses should definitely make use of earlier research and incorporate the psychological findings. Data–driven approaches and querying methods yield information on the prerequisite relationships among items (and consequentially tests), but not on specific problem demands. For these reasons, the classification scheme is developed by defining components and attributes based on the findings reported in Sections 2.2 and 3.5. The establishment of surmise relations between the actually presented sets of items and tests (see Chapter 5) follows a task analysis. In order to evaluate the applicability of the models to an adaptive diagnostic instrument, the derived knowledge spaces are implemented into a deterministic and a non–deterministic adaptive assessment algorithm (see Chapter 6).

Remembering the objectives of this study, namely the establishment of a common structure for various inductive reasoning problems, which can be used as a basis for an adaptive assessment system, the following scientific problems are conceived.

(i) Is the approach of surmise relations between tests suitable to order a set of inductive reasoning tests?

  - Is it possible to define meaningful relationships between the items of different tests, such that the solution behavior in one test allows predictions on the solution behavior in one or more other tests?
  - Do the postulated relationships between tests accurately predict participants' solution behavior?

(ii) Can the principle of componentwise ordering of product sets be applied to a set of inductive reasoning tests?

  - Is it possible to describe various types of inductive reasoning problems by a set of common components and attributes?
  - Are the specified components, their attributes, and the postulated order on the components and attributes a valid representation of participants' solution behavior?

(iii) Are the derived models valid representations of the knowledge domain when implemented into adaptive assessment algorithms?

  - Are the estimations of participants' knowledge states an accurate representation of their empirical response patterns?
  - Does the application of adaptive assessment algorithms reduce the number of presented items without a substantial loss in accuracy?

5 Method and Results

The aim of my research is to establish a structure on the items of various inductive reasoning tests, which can be used as a basis for an adaptive testing system (see Chapter 6). Since I am dealing with a set of tests, the mathematical approach to surmise relations between tests (see Section 3.2) is applied. In order to specify the surmise relation on the set of items and tests, I have chosen the principle of componentwise ordering of product sets (see Section 3.3.1).

A detailed description of how the hypotheses are derived is given in the next section (Section 5.1). The basic assumptions for the surmise relations between items, within, across, and between tests are equivalent for all three of the conducted investigations. Differences in the specific hypotheses are due to variations in the presented materials. I will discuss the methods and results for the three investigations in the succeeding sections.

Investigation I (Section 5.2) serves as a first examination of the surmise relations between tests approach to the domain of inductive reasoning and the developed classification scheme (Section 5.1). The set of data and the two inductive reasoning tests used in this investigation have been provided for reanalysis by the psychological service of the Austrian military (see Section 5.2 for detailed information). For methodological reasons a second reanalysis of a larger data set and computer–aided tests provided by the psychological service of the German military is conducted for Investigation II (Section 5.3). For a final investigation (Investigation III, Section 5.4), the set of data has been collected by myself in accordance with the standards of knowledge space research and the set of tests has been increased to four different problem types.

5.1 Derivation of hypotheses

In order to establish the surmise relation between tests by means of the componentwise ordering principle (see Section 3.3.1), items of various inductive reasoning tests have been analyzed and described by components and attributes that are applicable to all types of problems. On this basis the items of all analyzed tests can be compared and possible surmise relations between items and tests can be defined.

The specification of common components and attributes for various inductive reasoning tests is based on the problem requirements described in Sections 2.2 and 3.5. With regard to the materials used in the following three investigations (see Sections 5.2, 5.3, and 5.4), a general classification scheme has been developed, in which items of the types

Table 5.1: Components, attributes, and assumed downsets for inductive reasoning problems

Components	Downsets for components	Attributes	Downsets for attributes
A: Operation difficulty[a]	$\{A, D, E\}$	a_1 O other (less difficult)	$\{a_1\}$
		a_2 D difficult	$\{a_1$,$a_2\}$
B: Number of operations	$\{B, D, E\}$	b_i = 1 - 4	$\{b_{\leq i}\}$[b]
C: Constraint	$\{C, D, E\}$	c_1 L low demand	$\{c_1\}$
		c_2 H high demand	$\{c_1$,$c_2\}$
D: Material	$\{D\}$	d_1 V verbal	$\{d_1\}$
		d_2 N numerical	$\{d_2\}$
		d_3 G geometric-figural	$\{d_1$,d_2,$d_3\}$
E: Number of answer alternatives	$\{E\}$	e_1 4 alternatives	$\{e_1\}$
		e_2 5 alternatives	$\{e_1$,$e_2\}$
		e_3 8 alternatives	$\{e_1$,e_2,$e_3\}$

Note. [a]Types of operations (A) vary with respect to component D (e. g., rotation is only applicable to geometric, class inclusion to verbal material). [b]The smaller the number of operations the easier the attribute.

analogy, series completion, and matrix problem can be integrated. The specification of common components follows the findings outlined in Section 2.2.4, while attributes that differ with respect to problem types are based on the findings outlined in Sections 2.2.1, 2.2.2, 2.2.3, and 3.5. Table 5.1 depicts the derived classification scheme including the assumed downsets for each component and attribute. Figure 5.1 illustrates the assumed downsets by means of Hasse diagrams for each component.

5.1.1 Specification and ordering of components

Column one in Table 5.1 depicts five components, which can be applied to all problem types under consideration. Components that are specific to one of the problem types (e. g., word frequency for verbal analogies, the magnitude of an arithmetic operation for number series completions, or the number of constituent elements in geometric analogies or matrices, see Table 2.3) have not been included in the classification scheme (see Chapter 7, classification scheme, for a discussion of this issue). As already discussed in Section 2.2.4, the components *(A) operation difficulty, (B) relational complexity* or *number of operations*, and *(C) constraint* are the three major factors contributing to the difficulty of problem demands. Furthermore, the influence of the items' content was pointed out, for which component *(D) material* is introduced. Finally, component E is added to account for the varying *number of answer alternatives* used as answer formats in the investigated tests.

The assumed order of importance (see Section 3.3.1) on the five components is presented in column two in Table 5.1 and Figure 5.1f. The three major components A, B,

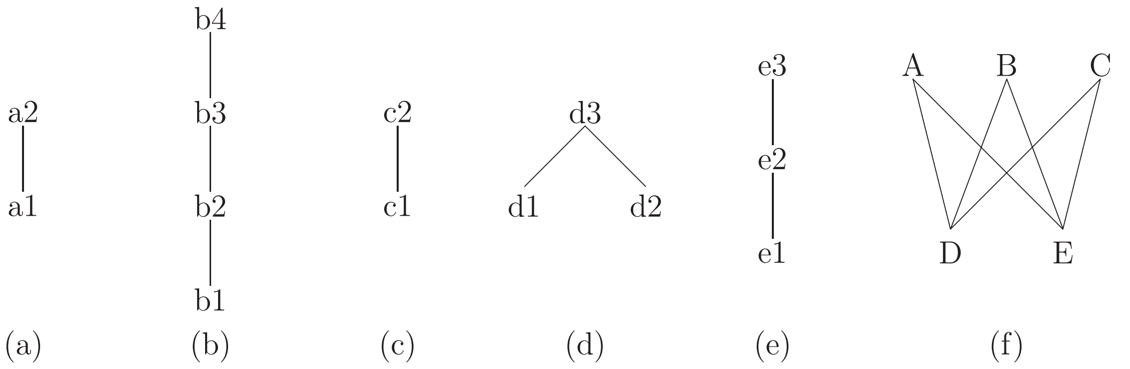


Figure 5.1: Hasse diagrams for the attributes of the five components operation difficulty (a), number of operations (b), constraint (c), material (d), and number of answer alternatives (e) and order of importance on the components A through E (f)

and C are defined as being more important than components D and E. The reasoning is that in common inductive reasoning tests the first three components vary within single tests and can therefore also distinguish between items of the same problem type. For the latter two components, the items within one test are usually described by the same attributes d_i and e_i. Assigning the same degree of importance to all five components would result in a rather unreasonable test knowledge structure, in which a whole test A is prerequisite for a test B because of a single attribute, such as five vs. eight answer alternatives or geometric vs. verbal material. In other word, items with geometric material or a higher number of answer alternatives would never be prerequisite for items with other materials or fewer answer alternatives (see Section 5.1.2 for a description of the components' attributes). Regarding for example component E, the mastery of an easy geometric analogy item with eight answer alternatives but only one operation of less difficulty and low demand in constraint could not be surmised from the mastery of a more difficult geometric matrix item with five answer alternatives, three necessary operations, and high demand in constraint.

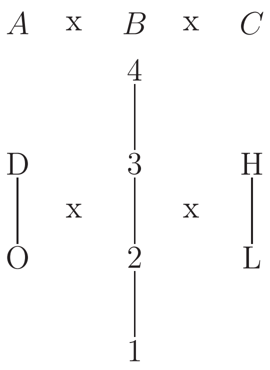
5.1.2 Specification and ordering of attributes

Columns three and four in Table 5.1 show the components' attributes and their assumed downsets. The attribute orders on each component are based on the findings outlined in Section 2.2. Component A (operation difficulty) has two attributes, viz. *difficult operations* (a_2: D) and *less difficult operations* (a_1: O), with 'O' being prerequisite for 'D' (see Fig. 5.1a). The kinds of operations vary with regard to the problems' material. *Difficult operations* (D), as they are pointed out in Section 2.2, are semantic relations of the types class inclusion and similar/comparative, numerical operations that include hierarchical sequences, and geometric transformations in number and space, as well as the Boolean AND and exclusive-OR operators. *Other operations* (O) include, for example, the semantic relations part–whole, contrast, or attribute, numerical operations without hierarchical sequences, and geometric transformations in form, shading, or size, as well as the rules constant in row, pairwise progression, or distribution of three values (see Section 2.2 for a description of the whole set of operations). For component B (number of operations), it is assumed that an item is more difficult the more operations are necessary to solve the item (see Fig. 5.1b). With regard to the tests analyzed for this study, there are four attributes presenting the presence of *one* (b_1: 1), *two* (b_2: 2), *three* (b_3: 3), or *four* (b_4: 4) *necessary operations*. The two attributes of component C (constraint) are high constraint or *low demand* on correspondence finding processes (c_1: L) and low constraint or *high demand* on correspondence finding processes (c_2: H), with 'L' being prerequisite for 'H' (see Fig. 5.1c). For component D (material), the specified types of material are *verbal* content (d_1: V) and *numerical* content (d_2: N), which are both prerequisite for the third attribute *geometric–figural* content (d_3: G), as depicted in Figure 5.1d. Other materials used for the presentation of inductive reasoning problems, such as letters or pictorial material, are not included in this scheme, because they do not occur in the tests under investigation and can therefore not be validated. Finally, the last component E specifies the number of answer alternatives. It is assumed that the smaller the number of answer alternatives the easier the problem demand on this component. A higher number of answer alternatives reduces the possibility to find the correct answer by way of exclusion and decreases the probability to guess the correct answer. The attributes can vary between two and n answer alternatives, with sets of *four*, *five*, and *eight alternatives* being realized in the three presented investigations. Therefore, there is a linear order on the three attributes, with e_1 (four alternatives) defined as the easiest and e_3 (eight alternatives) defined as the most difficult attribute (see Fig. 5.1e).

5.1.3 Model for the surmise relation between item classes

According to the presented classification scheme, the test items can be described by their respective attributes on each component. Forming the Cartesian product of the five sets of attributes results in 2 x 4 x 2 x 3 x 3 = 144 item classes, i. e. items described by the same set of attributes. The item classes can then be ordered by comparing the attributes of each item class. Figure 5.2 exemplifies the derivation of item classes by forming the Cartesian product of the three major components A, B, and C. The product of their attributes results in 2 x 4 x 2 = 16 item classes.

Figure 5.3 shows the derived order for the three major components A, B, and C (for



$$A\text{x}B\text{x}C = \begin{bmatrix} (\text{O},1,\text{L}) & (\text{O},2,\text{L}) & (\text{O},3,\text{L}) & (\text{O},4,\text{L}) \\ (\text{D},1,\text{L}) & (\text{D},2,\text{L}) & (\text{D},3,\text{L}) & (\text{D},4,\text{L}) \\ (\text{O},1,\text{H}) & (\text{O},2,\text{H}) & (\text{O},3,\text{H}) & (\text{O},4,\text{H}) \\ (\text{D},1,\text{H}) & (\text{D},2,\text{H}) & (\text{D},3,\text{H}) & (\text{D},4,\text{H}) \end{bmatrix}$$

Figure 5.2: Cartesian product of the attribute sets for components A, B, and C

reasons of representation components D and E are not included). In order to obtain the set of all 144 possible attribute combinations, the Cartesian product is extended by components D and E. Because of the order of importance specified in Table 5.1 and Figure 5.1, the components D and E are disregarded for item pairs with different attributes a_i, b_i, and c_i, e. g. for the item classes (O,1,L,V,5) and (O,2,L,G,5). With respect to the order on all 144 item classes, components D and E have to be considered only for item classes with the same attributes a_i, b_i, and c_i, e. g. for the item classes (O,1,L,V,5) and (O,1,L,G,5).

5.1.4 Model for the surmise relation between tests

In Figure 5.3 the derived item classes (for components A, B, C) are depicted in a common structure. For the establishment of a surmise relation between tests, the item classes are partitioned into two or more sets of item classes. Thereby, the differentiation of problem types (e. g., analogy vs. matrix problems), as it is usually found in inductive reasoning tests, can be retained. Furthermore, the partitioning allows both the consideration of the entire set of item classes and the consideration of single problem types. This distinction has the advantage that, according to the diagnostic demands, it is possible to assess either the ability of inductive reasoning in general or the more specific abilities associated with only one problem type (e. g., the ability to solve verbal analogy problems).

Figure 5.4 depicts an example for surmise relations within and between two tests AN and SC. For both tests all possible attribute combinations for components A through C are shown. Regarding the components D and E, test AN is assumed to consist of

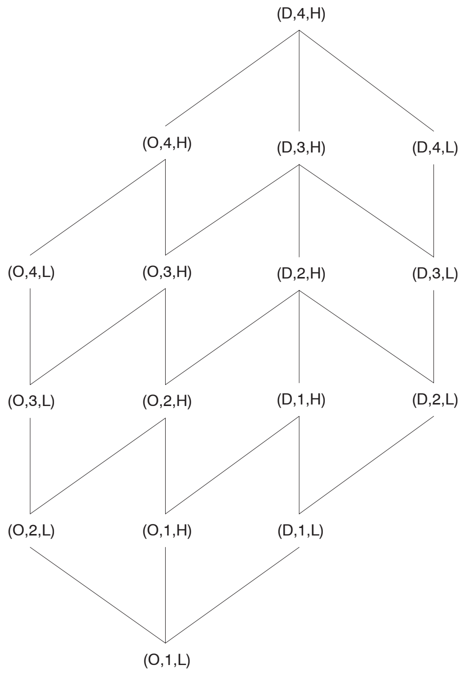


Figure 5.3: Derived order for item classes defined by components A, B, and C

verbal analogy items with eight answer alternatives each (V,8) and test SC is assumed to consist of numerical series completion items with five answer alternatives each (N,5). In this example component D does not contribute to the established order, because the materials verbal and numerical content are not comparable. Due to component E (number of answer alternatives), for item classes of different tests but with the same attributes a_i, b_i, and c_i, the series completion item class will always be prerequisite for the analogy item class. For reasons of more clarity, the prerequisite relationships between the item classes of different tests (i. e. the $SRxT$) are not shown in detail. However, it is easy to imagine a line from each AN item class to its respective SC item class with the same attributes a_i, b_i, and c_i. For example, item class (O,1,L,N,5) is prerequisite for item class (O,1,L,V,8). Therefore, the two tests are in a total–covering surmise relation from the analogy test to the series completion test ($SC \; \dot{\mathcal{S}}_t \; AN$). This means that there is a left–, as well as a right–covering surmise relation (cf. Definitions 3.5 and 3.6) from AN to SC. Each item class in AN has a prerequisite in SC and each item class in SC is prerequisite for some item class in AN. Therefore, it is possible to surmise that a person who solves all items in test AN will also solve the entire test SC and that a person who fails on all items in SC will also fail on the entire test AN. Furthermore, there is a general surmise relation (cf. Definition 3.4) from the series completion test to the analogy test ($AN \; \dot{\mathcal{S}} \; SC$), which means that at least one item in AN is prerequisite for some item in SC. Because components D and E are defined as being less important than components A through C (see Tabel 5.1), analogy items with equal or less difficult attributes on components A, B, and C are prerequisite for the respective series completion items. For example, item class (O,1,L,V,8) is prerequisite for item class (O,2,L,N,5).

In order to specify the surmise relation between item classes and its subsets, only the relationships contained in the respective relation are taken into account (see Equation 3.9 and Fig. 3.7 for the definition and an illustration of the subsets). For the surmise relation between items ($SRbI$) all pairs of item classes are considered, for the surmise relation across tests ($SRxT$) only the relationhips between item classes of different tests, and for each surmise relation within tests ($SRwT$) only the relationships between item classes contained in the same test are considered.

5.1.5 Empirical predictions

The hypotheses for each of the following three investigations (see Sections 5.2, 5.3, and 5.4) are derived by analyzing each item with respect to the five components listed in Table 5.1 and by assigning each item to its respective item class. Depending on the item classes realized in each of the analyzed tests, surmise relations between items, within, across, and between tests are established according to the componentwise ordering principle illustrated in Figures 5.3 and 5.4.

The postulated surmise relations between items of the sets of tests render a difficulty order with the following interpretation:

> “Whenever the demands of the attributes of a problem y are equal or less difficult than the corresponding attributes of a problem x, we assume that

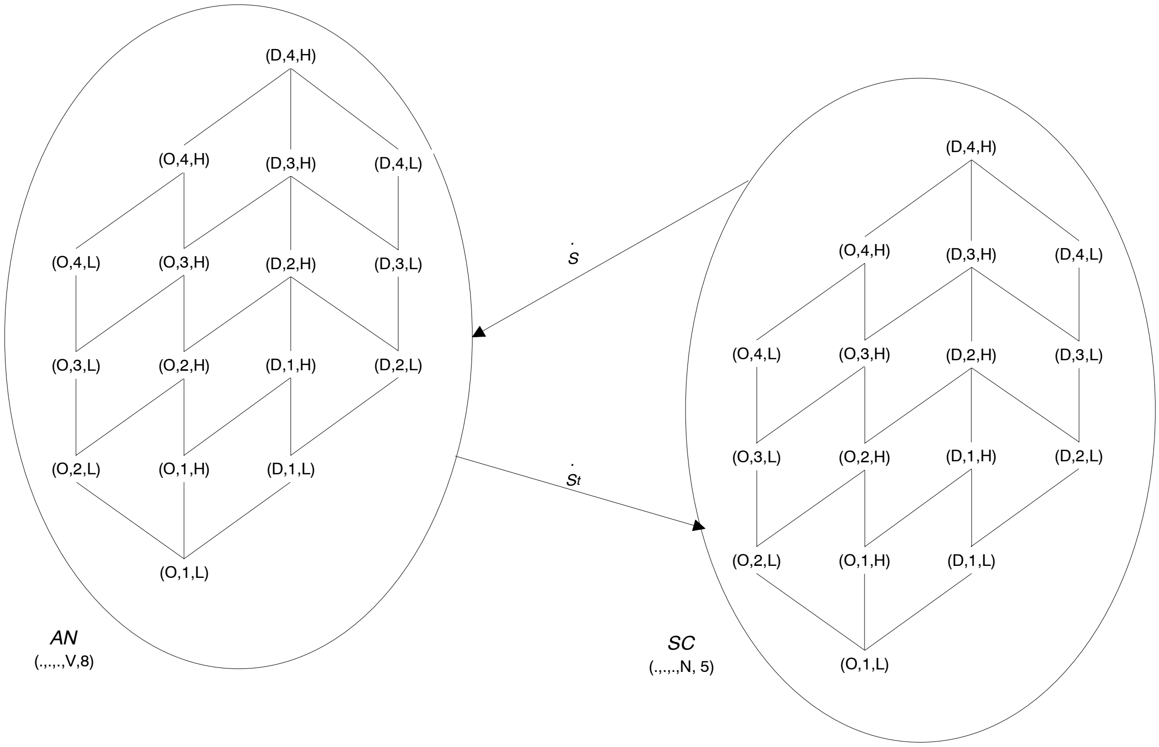


Figure 5.4: Derived surmise relation between two tests AN and SC with $SC\ \dot{\mathcal{S}}_t\ AN$ and $AN\ \dot{\mathcal{S}}\ SC$

y is prerequisite for problem x."

In order to account for the various substructures for each set of tests and the different validation methods via the surmise relation and the knowledge space, several empirical predictions have been derived.

For the presented sets of tests, the difficulty order leads to the following predictions on the surmise relation between tests ($SRbT$):

Ia For each postulated general $SRbT$ ($\mathcal{T}_i\ \dot{\mathcal{S}}\ \mathcal{T}_j$), it is expected that at least one item class in test $\mathcal{T}_j$ has a prerequisite and therefore an equal or lower solution frequency than one of the item classes in test $\mathcal{T}_i$.

Ib For each postulated left–covering $SRbT$ ($\mathcal{T}_i\ \dot{\mathcal{S}}_l\ \mathcal{T}_j$), it is expected that each item class in test $\mathcal{T}_j$ has a prerequisite and therefore an equal or lower solution frequency than one of the item classes in test $\mathcal{T}_i$.

Ic For each postulated right–covering $SRbT$ ($\mathcal{T}_i\ \dot{\mathcal{S}}_r\ \mathcal{T}_j$), it is expected that each item class in test $\mathcal{T}_i$ is prerequisite for one of the item classes in test $\mathcal{T}_j$ and has therefore an equal or higher solution frequency.

Id For each postulated total–covering $SRbT$ ($\mathcal{T}_i\ \dot{\mathcal{S}}_t\ \mathcal{T}_j$), it is expected that the predictions Ib and Ic apply.

For the presented set of items, the difficulty order leads to the following predictions on the surmise relation between items ($SRbI$), across tests ($SRxT$), and within tests ($SRwT$):

IIa For each postulated $SRbI$, $SRxT$, and $SRwT$, it is expected that the percentage of correct solutions for item classes described by attributes of equal or less difficulty is equal or higher than for the item classes they are prerequisite for.

IIb For each postulated $SRbI$, $SRxT$, and $SRwT$, it is expected that the γ–index yields a significantly higher number of concordant pairs than discordant pairs.

For the presented set of items, the difficulty order leads to the following predictions on the knowledge space between items ($KSbI$), across tests ($KSxT$), and within tests ($KSwT$):

IIIa For each postulated $KSbI$, $KSxT$, and $KSwT$, it is expected that the mean symmetric distance between the knowledge space and the empirical data set is significantly lower than the mean symmetric distance between the knowledge space and its powerset.

IIIb For each postulated $KSbI$, $KSxT$, and $KSwT$, it is expected that the mean symmetric distance between the knowledge space and the empirical data set is significantly lower than the mean symmetric distance between the knowledge space and random data sets (random simulations).

IIIc For each postulated $KSbI$, $KSxT$, and $KSwT$, it is expected that the mean symmetric distance between the knowledge space and the empirical data set is not significantly higher than the mean symmetric distance between the knowledge space and data sets simulated on the hypothesis (probability simulations).

IIId For each postulated $KSbI$, $KSxT$, and $KSwT$, it is expected that the mean symmetric distance between the knowledge space and the empirical data set are significantly lower than the mean symmetric distance between the knowledge space and data sets simulated under consideration of the data matrices' marginal frequencies (frequency simulations).

Summarized, prediction I refers to the postulated relationships between tests. It assumes the validity of the hypothesized $SRbT$ and its properties (left–, right–, or total–covering). Prediction II was derived in order to validate the postulated surmise relations on the various substructures, while prediction III refers to the corresponding knowledge spaces. The differentiation between the surmise relation and the knowledge space also allows for a comparison of the different validation methods (see Chapter 7, validation methods, for a disscussion of this issue).

5.2 Investigation I

Relating a pair of inductive reasoning tests: First application

The general aim of this work is to develop a structure for various inductive reasoning tests, which contains different problem types. The derived structure should also be applicable to adaptive testing procedures. In Section 5.1, I have introduced a classification scheme for inductive reasoning problems, which is based on the results of earlier investigations in the domain (Chapter 2). By applying the non–numerical knowledge space theory (Chapter 3), it is possible to order the set of problems in such a way that implicit prerequisite relationships between items and tests can be used to draw inferences from previously answered items. In order to implement the derived structure into an adaptive testing system, it is first of all necessary to validate the hypothesized surmise relations. For the first evaluation of the approach of surmise relations between tests (Section 3.2) and the postulated classification scheme (Section 5.1), only a subset of the possible problem types is investigated.

Therefore, a set of data and tests provided by the "Heerespsychologischer Dienst des BMLV" (HPD) in Vienna, Austria, has been reanalyzed (see Section 5.2.2). The original data set contained response patterns of 1221 participants, who performed a psychological test for the selection of corporals and officers ("Psychologische Auswahltestung für Offiziers– und Unteroffiziersanwärter"). For the establishment and validation of the hypotheses, two inductive reasoning tests (geometric matrices and verbal analogies) have been selected. The sets of data and items were reduced (see Section 5.2.3) to fulfill the condition of complete answer patterns as it is required for the validation of knowledge space hypotheses (see Section 3.4).

5.2.1 Hypothesis

The hypothetical knowledge structure for the two inductive reasoning tests (geometric matrices and verbal analogies) used in Investigation I was constructed as outlined in Section 5.1. The surmise relation on the set of items and tests was established in two steps. First, I analyzed the given items (see Section 5.2.2) with respect to their attributes on each of the five components (see Table 5.1) and assigned each item to the corresponding item class. In a second step, I ordered the item classes realized in the two tests (i. e. a subset of the 144 possible item classes) according to the structures derived in Sections 5.1.3 and 5.1.4 (see Figures 5.3 and 5.4 for an illustration of the ordering principles).

Expectations regarding the derived surmise relation between tests ($SRbT$), the surmise relation between items ($SRbI$) and its subsets ($SRxT$ and $SRwT$), as well as the corresponding knowledge spaces ($KSbI$, $KSxT$, and $KSwT$) are according to the empirical predictions made in Section 5.1.5.

For the $SRbT$, the item classes realized in the two tests (see Section 5.2.2.2 for details) suggest a total–covering surmise relation from the matrix (MT) test to the analogy (AN) test ($AN\ \dot{\mathcal{S}}_t\ MT$) and a general surmise relation from the analogy test to the matrix test ($MT\ \dot{\mathcal{S}}\ AN$). For ($AN\ \dot{\mathcal{S}}_t\ MT$), it is expected that each item class in the matrix test has a prerequisite item class in the analogy test ($AN\ \dot{\mathcal{S}}_l\ MT$) and that each item class in the analogy test is prerequisite for a matrix item class ($AN\ \dot{\mathcal{S}}_r\ MT$). For ($MT\ \dot{\mathcal{S}}\ AN$), it is expected that at least one item class in the analogy test has a prerequisite item class in the matrix test (see also Section 5.1.5, predictions Ia and Id).

A detailed description of the items and prerequisite relationships contained in the two tests under investigation is given in the next section.

5.2.2 Method

5.2.2.1 Participants

From the original 1221 male participants, 572 response patterns have been analyzed for Investigation I. All participants were corporal and officer candidates of the Austrian military and therefore taking the tests voluntarily. Participants' age ranged between 18 and 35 years with the average located in the early 20s (personal information). For confidentiality reasons more detailed personal data of the participants was not provided.

5.2.2.2 Material

Materials consist of two inductive reasoning tests developed by the HPD. The first test (MT) originally comprised 20 geometric matrix items. Because of the requirement of complete answer patterns (see Section 3.4), the set of items has been reduced to 14 matrix problems.

Each matrix item consists of nine squares arranged in three rows and three columns. The square on the lower right contains a question mark, all other squares are divided into nine cells. Each cell is either a blank or colored in red, green, yellow, dark or light blue, or brown. By analyzing the squares row by row and/or column by column, between one and three relational rules can be induced (see Table 5.2). On the right of the matrix eight answer alternatives (labeled A through H) are depicted. Participants had to decide, which alternative belongs in the square with the question mark. Figure 5.5 shows an example matrix (for confidentiality reasons the item is invented but similar to the original items), which requires the rule 'Figure Addition' as only necessary operation to induce the correct answer 'E'. With one rule of less difficulty, low demands on the correspondence finding process, geometric material, and eight answer alternatives, the corresponding item class is (O,1,L,G,8).

An analysis of the items' attributes on each of the five components (see Table 5.1) results in six item classes with one to four items each. Table 5.2 depicts the attribute combinations realized in the six item classes and a description (item number, operation types, and downsets) for each item.

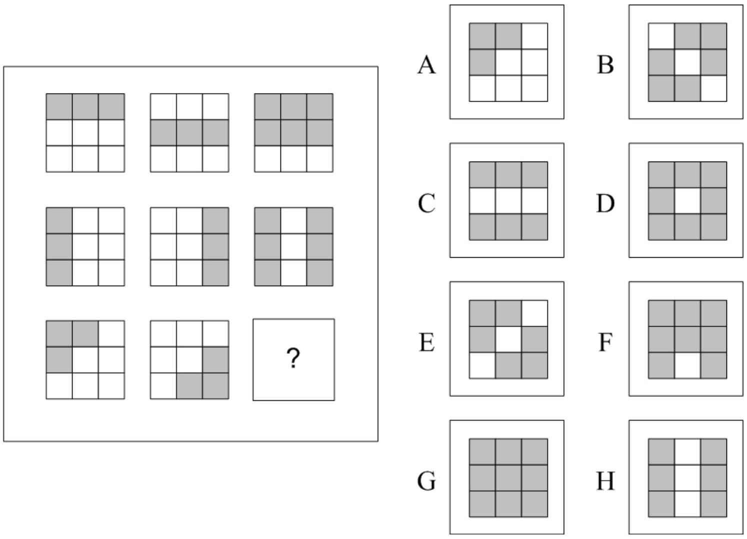


Figure 5.5: Example for a matrix item presented in Investigation I

The first item class (first column in Table 5.2), for example, is described by the attributes (O,1,L,G,8). This means that the included four items (item numbers in the second column) all have **O**ther operation difficulty (component *A*, operation difficulty), **1** necessary operation to solve the problem (component *B*, number of operations), and **L**ow demand on correspondence finding processes (component *C*, constraint). The attributes on components *D* (material) and *E* (number of answer alternatives) are identical for all item classes, viz. **G**eometric–figural material and **8** answer alternatives. The third column in Table 5.2 specifies the types of operations that need to be inferred to solve the items. The operations realized within the six item classes are constant in a row (CR), constant in a column (CC), pairwise progression (PP), figure addition (FA), distribution of two (D2) and three values (D3), and the exclusive-OR rule (XO). The operations are described in detail in Section 2.2.3, Table 2.2. Overall, one out of the six item classes has the attribute **D**ifficult operation and the number of operations varies between one and three. Four item classes have the attribute **H**igh demand on correspondence finding processes and all six item classes are described by **G**eometric material and **8** answer alternatives. Finally, the last two columns in Table 5.2 depict the prerequisites (downsets) for item classes (fourth column) and items (fifth column), which were derived according to Section 5.1, Table 5.1.

The second test (*AN*) originally comprised 25 verbal analogy problems in German, which have been reduced to a set of 16 items (see Section 5.2.3). Each analogy has a three–term stem of the form $A : B = C : ?$. Below the stem terms five answer alternatives (labeled *a* through *e*) are presented. Exceptions are two analogies with only four answer alternatives. Participants had to choose the alternative that completes the analogy best. Stem terms as well as alternatives are nouns (8 items), verbs (4 items), or adjectives (1 item). Three of the analogies are mixed (noun-verb once and noun-adjective twice). Figure 5.6 gives an example analogy (for confidentiality reasons the example is an extended item taken from Bejar et al., 1991, but it is similar to the original items), which can be solved by applying the semantic rationale ‘A is part of

Table 5.2: Item descriptions and downsets for the matrix test in Investigation I

Item classes[a]	Item numbers	Operation types[b]	Downsets for item classes[c]	Downsets for items
(O,1,L,G,8)	1 3 5 7	D3 PP CR FA	(O,1,L,G,8)	1,3,5,7
(O,1,H,G,8)	6	FA	(O,1,L,G,8),(O,1,H,G,8)	1,3,5,6,7
(O,2,L,G,8)	2 4 10 12	CC,PP D3,D3 D2,D2 PP,PP	(O,1,L,G,8),(O,2,L,G,8)	1,2,3,4,5,7, 10,12
(O,2,H,G,8)	8 9	CC,D3 CR,CC	(O,1,L,G,8),(O,1,H,G,8), (O,2,L,G,8),(O,2,H,G,8)	1,2,3,4,5,6,7,8, 9,10,12
(O,3,H,G,8)	11 13	D2,D2,D3 CR,CC,CC	(O,1,L,G,8),(O,1,H,G,8), (O,2,L,G,8),(O,2,H,G,8), (O,3,H,G,8)	1,2,3,4,5,6,7,8, 9,10,11,12,13
(D,2,H,G,8)	14	XO,XO	(O,1,L,G,8),(O,1,H,G,8), (O,2,L,G,8),(O,2,H,G,8), (D,2,H,G,8)	1,2,3,4,5,6,7,8, 9,10,12,14

Note. [a] Components are ordered alphabetically, i. e. *A* through *E*. [b] CR = constant in a row, CC = constant in a column, PP = pairwise progression, FA = figure addition, D2 (D3) = distribution of two (three) values, XO = exclusive-OR. [c] See Table 5.1 for the derivation of downsets.

B' to the third term of the analogy (the correct answer is 'a'). With a rule of less difficulty (part–whole relation), one element in the semantic rationale, low demands on the correspondence finding process, verbal material, and five answer alternatives, the corresponding item class is (O,1,L,V,5).

wheel : car = leg : ?

a) horse b) bicycle c) forest d) bookcase e) snake

Figure 5.6: Example for an analogy item presented in Investigation I

An analysis of the items' attributes on each component (see Table 5.1) results in nine item classes with one to four items each. Table 5.3 depicts the attribute combinations (first column), item numbers (second column), types of operations (third columns), and downsets for item classes (fourth column) and items (last column). In the case of verbal analogies the number of operations (component B) refers to the number of elements in the analogies' rationales (see Section 2.2.1, Box 2.3). In Table 5.3 only the semantic relations between the terms are listed (third column), whereas the complete rationales are given in Appendix B.1, Table B.1.

As can be seen from the attribute combinations (first column), the types of operations (component A; specified in the third column) that need to be inferred include the semantic relations of **O**ther types part–whole (PW), contrast (CO), attribute (AT), cause–purpose (CP), and space–time (ST), as well as the more **D**ifficult relation similar/comparative (SI). The latter is required by three out of nine item classes. The semantic rules are described in detail in Section 2.2.1, Table 2.1. The number of elements in the rationales (component B) varies between one and three. **H**igh demands on correspondence finding processes (component C) are required by four item classes, while the attributes on component D are identical for all items, viz. **V**erbal material. Eight item classes have **5** answer alternatives (component E), one item class only **4** (for item 21 alternative e is missing, for item 26 the alternatives c and d are identical).

Using the order of importance on the components (Tabel 5.1, second column) and the difficulty orders on the attributes (Tabel 5.1, fourth column) a surmise relation was established on the item classes of both tests. Table 5.4 shows the prerequisite relationships for the surmise relation across tests ($S_{MT \times AN}$ and $S_{AN \times MT}$, see Section 3.2). The surmise relation from the matrix test to the analogy test ($S_{AN \times MT}$) is represented by '$\diamond$'. Since each matrix item class has a prerequisite item class in the analogy test and each of the analogy item classes is prerequisite for a matrix item class, a total–covering surmise relation ($AN\ \dot{\mathcal{S}}_t\ MT$, see Definitions 3.5 and 3.6) is postulated. The surmise relation from the analogy test to the matrix test ($S_{MT \times AN}$) is represented by 'x'. In this case, only a subset of the analogy item classes has prerequisites in the matrix test and only a subset of matrix item classes is prerequisite for an analogy item class. Therefore, a general surmise relation ($MT\ \dot{\mathcal{S}}\ AN$, see Definition 3.4) from the analogy test to the matrix test is postulated. The surmise relation between items ($SRbI$), i. e. within ($SRwMT$ and $SRwAN$) and across ($SRxT$) the two tests is depicted as a Hasse diagram in Figure 5.7. The relation files for the $SRbI$, $SRxT$, $SRwMT$, and $SRwAN$ are given in Appendix E.1. With regard to the corresponding knowledge spaces, the base files for the $KSbI$, the $KSxT$, the matrix test ($KSwMT$), and the analogy test ($KSwAN$) are listed in Appendix E.2.

5.2.2.3 Procedure

Participants were run in group sessions with an average of 21 persons each ($SD = 5.2$), ranging between 5 and 32 individuals per group. The data were collected between 1995 and 1997 by members of the HPD. The entire examination lasted 22 hours including an initial testing (first day, 2 pm to 10 pm) and a final testing (second day, 6 am to 12 am), which were separated by an endurance phase with no sleep and physical strain (10 pm to 5 am). The initial and the final testings consisted of 14 and 10 subtests respectively, including tests for concentration, perception, and memory, various intelligence tests, oral presentations, and writing an essay. The two inductive reasoning tests of the initial

Table 5.3: Item descriptions and downsets for the analogy test in Investigation I

Item classes[a]	Item numbers	Operation types[b]	Downsets for item classes[c]	Downsets items
(O,1,L,V,4)	21 26	PW CP	(O,1,L,V,4)	21,26
(O,1,L,V,5)	15 17 25 30	CO CO PW PW	(O,1,L,V,4),(O,1,L,V,5)	15,17,21,25,26, 30
(O,1,H,V,5)	27 29	AT AT	(O,1,L,V,4),(O,1,L,V,5), (O,1,H,V,5)	15,17,21,25,26, 27,29,30
(O,2,L,V,5)	18	ST	(O,1,L,V,4),(O,1,L,V,5), (O,2,L,V,5)	15,17,18,21,25, 26,30
(O,2,H,V,5)	22 24	ST ST	(O,1,L,V,4),(O,1,L,V,5), (O,1,H,V,5),(O,2,L,V,5), (O,2,H,V,5)	15,17,21,22,24, 25,26,27,29,30
(O,3,H,V,5)	20 28	CP CP	(O,1,L,V,4),(O,1,L,V,5), (O,1,H,V,5),(O,2,L,V,5), (O,2,H,V,5),(O,3,H,V,5)	15,17,18,20,21, 22,24,25,26,27, 28,29,30
(D,1,L,V,5)	16	SI	(O,1,L,V,4),(O,1,L,V,5), (D,1,L,V,5)	15,16,17,21,25, 26,30
(D,1,H,V,5)	19	SI	(O,1,L,V,4),(O,1,L,V,5), (O,1,H,V,5),(D,1,L,V,5), (D,1,H,V,5)	15,16,17,19,21, 25,26,27,29,30
(D,2,L,V,5)	23	SI	(O,1,L,V,4),(O,1,L,V,5), (O,2,L,V,5),(D,1,L,V,5), (D,2,L,V,5)	15,16,17,18,21, 23,25,26,30

Note. [a] Components are ordered alphabetically, i. e. *A* through *E*. [b] PW = part–whole, CP = cause purpose, CO = contrast, AT = attribute, ST = space–time, SI = similar/comparative. [c] See Table 5.1 for the derivation of downsets.

testing were selected for this investigation.

The tests were presented as booklet with a separate machine–readable answersheet for the responses. The matrix test was the third test to take, following a concentration test and a verbal activity test (40 items), which took together 13 minutes. The analogy test was presented after the matrix test, but with a break of 15 minutes in between.

Table 5.4: Surmise relation across tests in Investigation I

Analogies	Matrices (D = G)					
(D = V)	(O,1,L,8)	(O,1,H,8)	(O,2,L,8)	(O,2,H,8)	(O,3,H,8)	(D,2,H,8)
(O,1,L,4)	◇	◇	◇	◇	◇	◇
(O,1,L,5)	◇	◇	◇	◇	◇	◇
(O,1,H,5)	x	◇		◇	◇	◇
(O,2,L,5)	x		◇	◇	◇	◇
(O,2,H,5)	x	x	x	◇	◇	◇
(O,3,H,5)	x	x	x	x	◇	
(D,1,L,5)	x					◇
(D,1,H,5)	x	x				◇
(D,2,L,5)	x		x			◇

Note. An x indicates that the matrix item class in column i is prerequisite for the analogy item class in row j (iSj); a ◇ indicates that the analogy item class in row j is prerequisite for the matrix item class in column i (jSi).

The items of both inductive reasoning tests were presented as speed–power tests[1] in paper-pencil form. Participants had 12 minutes to complete the matrix test (20 items) and 5 minutes for the analogy test (25 items). Both tests included two practice items, which were reviewed with the instructor. For the matrix test, the instructions to the paractice items were given only verbally (personal information), whereas the analogy test includes a written explanation of the relevant relation for the first practice item.The given sequence of the items (see item numbers in Tables 5.2 and 5.3) was not binding for the participants.

5.2.3 Results

For the validation of the established hypotheses the methods outlined in Section 3.4 are applied. After a short description of the procedure used for pre–editing data and tests, I will continue with the results regarding the surmise relation. This is followed by the results derived via the knowledge space, which also include comparisons with simulated data sets and statistical analyses.

[1] The presentation of the items as speed–power test lead to incomplete answer patterns, which was the reason for eliminating items and patterns from the original data matrix (see Section 5.2.3)

Pre–editing of data and tests The validation of knowledge space hypotheses requires complete response patterns, i. e. all the items have to be answered by each participant (see Section 3.4). The data are coded as binary strings of '0' and '1' with '0' denoting an incorrect answer and '1' denoting a correct answer. Missing responses are not intended. The presentation of items as a speed–power test generally leads to unprocessed items at the end of the test, which results in a lower percentage of correct responses for these items. Therefore, incomplete answer patterns have to be removed from the data set. Furthermore, the validation procedure requires a reasonable high number of response patterns given by the same participants for both tests. In the original data matrix only 51 persons answered all of the 45 items. In order to handle the trade–off between the total number of response patterns and the number of items answered by all participants, the following procedure was applied.

First, the program `patt-statistics` was applied to count the number of complete response patterns for various numbers of items. The order, in which the items are removed depends on the number of missing responses per item. First, the item with the highest number of missing responses is removed and the number of complete patterns for the remaining items is recounted. For the two tests in Investigation I, the last item presented in the analogy test was removed first, which left 56 persons who answered all of the remaining 44 items. Then the item with the second highest number of missing responses was deleted, the number of complete patterns was recounted, and so on.

In a second step, the postulated knowledge spaces were calculated for various numbers of items and the item set, for which the number of complete response patterns exceeded the number of hypothetical knowledge states, was selected for further analyses. The goal was to achieve a maximal number of items by simultaneously maintaining the number of postulated states smaller than the number of response patterns. The procedure resulted in a reduction to 30 items (14 matrix and 16 analogy items) and 572 response patterns for both tests. The postulated test knowledge space for 30 items contains 293 states.

The reduced data matrix contains 17,160 responses from 572 participants. Table 5.5 gives an overview of the original and the reduced data sets. The minimal number of correctly solved items equals zero (1 response pattern) the maximal number 30 (26 response pattern), which amounts to 4,7% of all response patterns (due to the small percentage, the trivial response patterns have not been excluded from the final data set).

5.2.3.1 Validation of hypotheses via the surmise relation

Percentage of correct solutions All in all 84.66% of the items were solved correctly. This corresponds to a total of 14,528 correct responses. The relative solution frequencies for each single item are provided in Appendix F.1, Table F.1. Figure 5.7 depicts the relative solution frequencies together with the postulated surmise relation between the item classes of both tests. The relative solution frequencies for the nine item classes with more than one item (see Tables 5.2 and 5.3) represent the averaged relative solution frequencies of all items in the respective class. The number of items

Table 5.5: Original (N=1221) and reduced (N=572) data sets in Investigation I

Tests	No. of items	No. of responses	No. of correct responses	No. of incorrect responses	No. of missing responses
MT original	20	24,420	15,853	4,278	4,289
AN original	25	30,525	18,661	4,680	7,184
Total	45	54,945	34,514	8,958	11,473
MT reduced	14	8,008	6,801	1,207	0
AN reduced	16	9,152	7,727	1,425	0
Total	30	17,160	14,528	2,632	0

Note. MT = matrix test, AN = analogy test, total refers to the set of items contained in both tests.

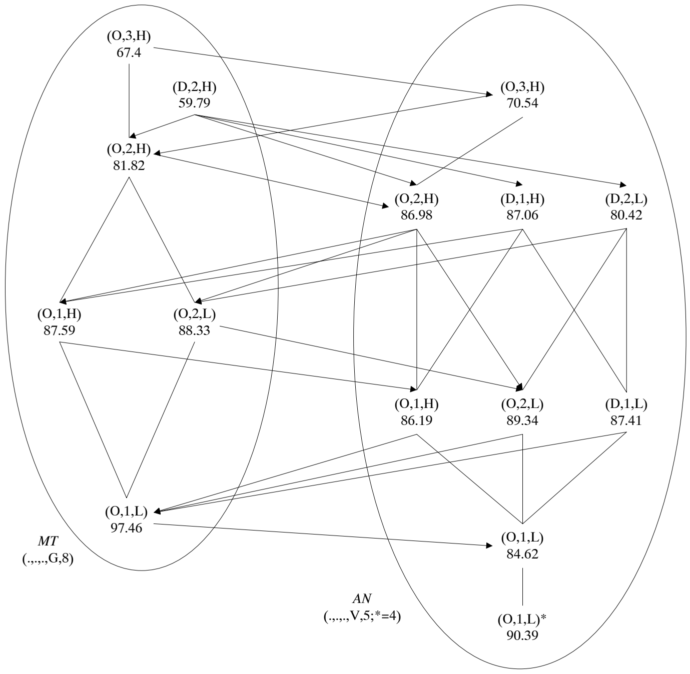


Figure 5.7: Hasse diagram for the postulated surmise relation between items and relative solution frequencies in Investigation I ($AN\ \dot{\mathcal{S}}_t\ MT$ and $MT\ \dot{\mathcal{S}}\ AN$)

contained in one item class varies between one and four. The average maximal difference in solution frequencies of items contained in the same class amounts to 5.89% ($SD = 5.1$). Regarding the item classes, the percentages of correctly solved item classes vary from 59.79% for item class (D,2,H,G,8) up to 97.46% for item class (O,1,L,G,8).

Looking at the two surmise relations within tests, the results for the geometric matrices ($SRwMT$, left ellipse in Figure 5.7) represent a perfect fit of the hypothesized (see Table 5.2) and the empirical surmise relation. For all of the 13 pairs contained in the $SRwMT$, the relative solution frequencies for the postulated prerequisite item classes are higher than those for the item classes they are surmised from. The results for the analogy test ($SRwAN$, right ellipse) show that the relative solution frequencies for item classes (O,1,H,V,5) and (O,1,L,V,5) are lower than those for two respectively five item classes they are prerequisite for (see Table 5.3 for the postulated downsets). More precisely, 17 out of 24 pairs of item classes confirm the hypothetical surmise relation (reflexive pairs are not counted). However, one–dimensional χ^2 tests show that the differences for all of the seven reversed pairs are not significant ($\alpha = .05$;[2] see Appendix F.2, Table F.4 for the exact values). Considering that the hypothesis states that the solution frequency for a prerequisite item class should be higher than or equal to the frequency of the item class it can be surmised from (see prediction IIa in Section 5.1.5), the reversed pairs are still in accordance with the postulated knowledge space.

With regard to the surmise relation across tests (see Figure 5.7), i. e. prerequisite relationships between item classes of different tests ($SRxT$, represented by lines going from one ellipse to the other), 22 out of 27 pairs (see Table 5.4) confirm the surmise relation from the matrix test to the analogy test. Deviations occur for the three item classes (O,1,L,G,8), (O,1,H,G,8), and (O,2,L,G,8) with higher solution frequencies than part of their postulated prerequisites in the analogy test. From the five pairs of item classes with reversed solution frequencies, only the pair (O,1,L,G,8) and (O,1,L,V,5) shows a significant difference ($\chi^2(1, N = 1041) = 5.19, p < .05$; see Appendix F.2, Table F.4 for the remaining values).

Regarding the surmise relation from the analogy test to the matrix test (see Figure 5.7) the solution frequencies for all of the 14 pairs are in accordance with the hypothesis (see Table 5.4).

For the surmise relation between tests, a total–covering surmise relation from the matrix test to the analogy test ($AN\ \dot{\mathcal{S}}_t\ MT$) and a general surmise relation from the analogy test to the matrix test ($MT\ \dot{\mathcal{S}}\ AN$) was postulated in Section 5.2.1. For the total–covering surmise relation, it is expected that each of the six item classes in the matrix test has a prerequisite in the analogy test and that each of the nine analogy item classes is prerequisite for one of the matrix item classes (prediction Id in Section 5.1.5). For the general surmise relation, it is expected that at least one item class in the analogy test has a prerequisite in the matrix test (prediction Ia in Section 5.1.5). The total–covering surmise relation ($AN\ \dot{\mathcal{S}}_t\ MT$) is verified by all but one of the 15 item classes in both test. With 97.46% of correct answers, item class (O,1,L,G,8) shows the highest solution frequency within the set of tests, and does therefore not have a prerequisite in the analogy test (yet the difference between this item class and the class (O,1,L,V,4) is not significant; $\chi^2(1, N = 1075) = 1.54, p < .05$). The general surmise relation ($MT\ \dot{\mathcal{S}}\ AN$) is clearly verified, since not only one but all of the 14 postulated prerequisite relationships from the analogy test to the matrix test are confirmed by the relative solution frequencies. Summing up, the surmise relation between items of both tests (i. e. the pairs within each test and across the two tests) is confirmed by 66 out of 78 postulated pairs of item classes (reflexive pairs are not counted). Of the 12 pairs with reversed solution frequencies only one pair shows a significant difference.

[2] Generally, with the α–adjustment for item classes involved in multiple χ^2 tests, the α–level can be reduced. However, because the model postulates that there are no significant differences, the α–adjustment would lead to a weaker test of the hypothesis and was therefore disregarded.

Table 5.6: VC and γ for the surmise relation between items and its subsets (N = 572)

	No. of items	No. of pairs[a]	VC	γ_G	$\gamma > 0$	χ^2
$SRbI$	30	351	0.08	0.36	270 (76.92%)	6437.43
$SRxT$	30	191	0.08	0.31	136 (71.20%)	2606.61
$SRwMT$	14	71	0.04	0.68	69 (97.18%)	4715.82
$SRwAN$	16	89	0.09	0.20	65 (73.03%)	501.06

Note. $SRbI$ = surmise relation between items of both tests, SRxT = surmise relation across the two tests, $SRwMT$ = surmise relation within the matrix test, $SRwAN$ = surmise relation within the analogy test. [a]Reflexive pairs are not counted.

Summarized, the surmise relation between items is confirmed by 84.62% of the postulated pairs, the surmise relation across tests by 87.81%, and the surmise relations within the matrix and the analogy test by 100% and 70.83% respectively. Regarding only the statistically significant differences, the $SRxT$ is confirmed by 97.56%, the $SRbI$ and the two $SRwT$ by 100% each.

Indices for the fit of a surmise relation To validate the fit of each item pair in the surmise relation, the violational coefficient (VC) and the gamma–index (γ) were calculated by specifying the respective pairs for the $SRbI$ and its subsets, namely the $SRxT$, the $SRwMT$, and the $SRwAN$. Note that the values refer to relationships between single items as compared to item classes.

As outlined in Section 3.4.1, a better fit of a surmise relation to a set of data is indicated by lower VC but higher γ values. As shown in Table 5.6, the indices VC and γ_G both indicate that the $SRwMT$ fits the set of data best, whereas $SRwAN$ deviates most from the data. Regarding VC, the data violate the hypotheses for the various (sub)structures in 4% to 9% of the pairs in the relation. However, the differences between the $SRbI$, the $SRxT$, and the $SRwAN$ are extremely small (8% and 9%). Furthermore, it needs to be considered, that the VC considers all but the reflexive pairs in the relation. Hence, the small number of violations is partly caused by the high percentage of correctly solved items (see above). Trivial response vectors (both items correct or both items incorrect) are always in accordance with the hypothesis. The results from the γ–index are more specific, because the index considers only those item pairs where only one of the items is answered correctly. Table 5.6 also lists the number of pairs with positive γ–indices and the McNemar χ^2 values for concordant and discordant pairs over all items and participants. The percentage of pairs with γ values greater than zero range between 71.2% ($SRxT$) and 97.18% ($SRwMT$). All four of the McNemar χ^2 values are highly significant and thus confirm the hypothesis (see Section 5.1.5, prediction IIb).

The results yielded by the two indices are in accordance with the analysis of relative solution frequencies (see above). This means that also on the item level (instead of the more global view of item classes), the values indicate that the analogy test contributes most to the deviations of the surmise relation between items, while the matrix test shows the best fitting results. Regarding the postulated surmise relation between tests, the indices do not permit a direct conclusion about its validity.

5.2.3.2 Validation of hypotheses via the knowledge space

Symmetric distances and distance agreement coefficient Table 5.7 shows the mean symmetric distances (*ddat*; 5th column in Table 5.7) between the empirical data set and the test knowledge structure ($KSbI$) as well as its substructures across tests ($KSxT$) and within tests ($KSwMT$ and $KSwAN$). The empirical distances for all postulated knowledge spaces are far below the theoretical maxima ($0.84 \leq ddat \leq 2.99$ as compared to $7 \leq dmax \leq 15$; 4th column in Table 5.7). An analysis of the items' invalidity shows that the distances are mainly associated with careless errors, which means that the contradicting response patterns included for the main part incorrect solutions for items that are assumed to be prerequisite for some other correctly solved item(s). The mean relative frequencies for careless errors per item and person amounts to .084 as compared to .016 lucky guesses per item and person[3] (see Appendix F.4, Table F.13 for the invalidities of each item).

A comparison of the (sub)structures' *ddat* values with the distances between the knowledge spaces and their powersets[4] (*dpot*; 7th column in Table 5.7) clearly indicates that the empirical data fit the knowledge spaces better than random response vectors. The distance *dpot* ranges between 4.57 ($KSwMT$) and 10.08 ($KSbI$), the medians for *dpot* (8th column in Table 5.7) between 5 and 10 as compared to 1 and 3 for the medians of the empirical data (6th column in Table 5.7). For all four (sub)structures, χ^2 tests show that the empirical distance distributions are significantly better (more positively skewed) than the distance distributions of the respective powersets ($9,717 \leq \chi^2 \leq 38,446$; see Appendix F.5 for the exact values). These results support the postulated model according to prediction IIIa in Section 5.1.5. The distance distributions for the test knowledge space, its substructures, and the corresponding powersets are provided in Appendix F.3, Tables F.7 and F.8.

As already mentioned in Section 3.4.2, a direct comparison of the various structures' *ddat* values is not reasonable, because the number of items as well as the number of knowledge states varies among the structures. In order to relate the test knowledge space to its substructures, I calculated the distance agreement coefficient (DA; see Section 3.4.2). The interpretation of the coefficient is that the lower DA the better

[3] Note that the analysis of careless errors and lucky guesses is based on the assumption of a correct model, i. e. each contradicting response is interpreted as either careless error or lucky guess (depending on the nearest knowledge state). Furthermore, the reported number of careless errors and lucky guesses only reflects the portion, which contradicts one of the postulated knowledge states. This means that the participants might have made additional careless errors or lucky guesses which are, however, in accordance with one of the knowledge states and are therefore not counted as such.

[4] For the knowledge spaces between items and across tests with 30 items each, the powerset was not computable and therefore, the values for *dpot* were calculated with 20,000 simulated random patterns. Similar computations with smaller knowledge spaces (20 items) showed that the values do not differ up to the 2nd decimal place.

Table 5.7: Symmetric distances for the test knowledge space, its substructures, and their powersets (N = 572)

	m	$\lvert\mathcal{K}\rvert$	$dmax$	$ddat$ (SD)	Mdn	$dpot$ (SD)	Mdn	DA
$KSbI$	30	293	15	2.99 (2.17)	3	10.08 (1.90)	10	0.30
$KSxT$	30	83,219	15	1.82 (1.26)	2	5.98 (1.46)	6	0.30
$KSwMT$	14	57	7	0.84 (0.95)	1	3.77 (1.23)	4	0.22
$KSwAN$	16	71	8	1.42 (1.36)	1	4.57 (1.34)	5	0.31

Note. m denotes the number of items, $\lvert\mathcal{K}\rvert$ the number of knowledge states; $KSbI$ = knowledge space between items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test.

the fit of the knowledge space to a data matrix. As opposed to $ddat$, the highest DA value does not result for the test knowledge structure ($KSbI$) but for the analogy test. Corresponding to the results found for the surmise relation (percentage of correct solutions, VC, and γ_G, see Section 5.2.3.1), the matrix test shows the best (lowest) DA value, whereas the values for the $KSbI$, the $KSxT$, and the $KSwAN$ differ only slightly (see last column in Table 5.7).

Simulations As outlined in Section 3.4.2, I simulated data sets to compare the results of the empirical data set to results derived from random sets of data as well as data sets based on the postulated knowledge spaces. Table 5.8 depicts the mean symmetric distances and DA values for the empirical data ($ddat$) and for the simulated sets of data. The values for random simulations ($dsim_r$) and probability simulations ($dsim_p$) are the averaged mean distances and standard deviations from 1000 data sets each. The number of response patterns per data set corresponds to the number in the empirical data set, i. e. 572 patterns. For the probability simulation, the probability for lucky guesses corresponds to the number of answer alternatives ($\eta = 0.16$ for $KSbI$ and $KSxT$, 0.125 for $KSwMT$, and 0.2 for $KSwAN$). The probability for careless errors was varied in 10 steps with $0.05 < \beta \leq 0.15$, because the true probability of participants making a careless error is unknown. For each probability level, I simulated 100 sets of data. The averaged distance distributions for the simulated data sets are provided in Appendix F.3, Tables F.7 and F.8.

The empirical mean distances for the $KSbI$ and its substructures are all far below the distances resulting from random simulations ($3.76 \leq dsim_r \leq 10.08$) and slightly lower than the distances for the respective probability simulations ($1.07 \leq dsim_p \leq 3.15$). The same is true for the DA values of the various knowledge spaces (see Table 5.8) .

Figure 5.8 depicts the frequency distribution of the empirical distances to the $KSbI$ and the averaged distance distributions derived from random and probability simulations. Regarding the random simulations (left figure), there is only a small overlap of the empirical and the simulated data sets. Furthermore, the empirical data set is located further to the left on the distance scale and has a positive skew as opposed to the

Table 5.8: Symmetric distances for the test knowledge space, its substructures, and simulated data sets (N = 572)

			random simulations			probability simulations		
	$ddat$	DA	$dsim_r$ (SD)	Mdn	DA	$dsim_p$ (SD)	Mdn	DA
$KSbI$	2.99	0.30	10.08 (1.90)	10	1.000	3.15 (1.66)	3	0.31
$KSxT$	1.82	0.30	5.99 (1.46)	6	1.002	1.86 (1.25)	2	0.31
$KSwMT$	0.84	0.22	3.76 (1.23)	4	0.997	1.07 (0.95)	1	0.28
$KSwAN$	1.42	0.31	4.56 (1.34)	5	0.998	1.58 (1.18)	1	0.35

Note. For the simulated data sets SD refers to the mean standard deviations; $KSbI$ = knowledge space between items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test.

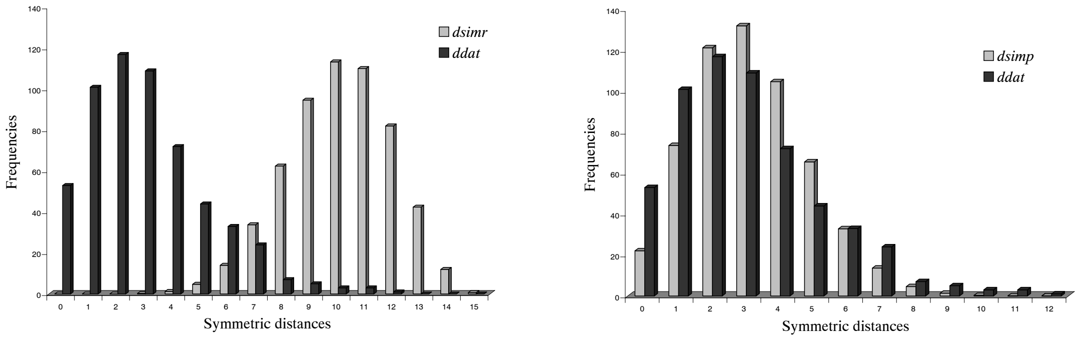


Figure 5.8: Distance distribution of the empirical data set ($ddat$) compared to the distributions of random ($dsim_r$) and probability simulations ($dsim_p$)

random data sets. This result clearly indicates that the postulated test knowledge space fits the empirical data far better than sets of data simulated under the assumption that there is no structure inherent in the data. The distribution resulting from the probability simulations (right figure) strongly overlaps with the empirical distribution. On the lower end of the distance scale ($di = 0$ or 1) the empirical frequencies are higher than the simulated ones, which indicates a slightly better fit of the test knowledge space to the empirical data than to the simulated data sets.

Statistical analyses Regarding the differences between the empirical data set and the distributions of simulated data, I computed several tests to judge, whether or not the differences are statistically significant. In order to allocate the empirical mean distances ($ddat$) within the distributions of simulated data sets, the values have been standardized. Table 5.9 shows the averaged mean symmetric distances ($dsim_r, dsim_p$) of the 1000 simulated data sets each, the distributions' standard deviations (SD), and the standardized z–scores for the empirical mean distances ($ddat$). For the data sets simulated on the postulated knowledge spaces ($dsim_p$), furthermore the results of

Table 5.9: Comparison of the empirical and simulated symmetric distances for the test knowledge space and its substructures (N = 572)

	random simulations			probability simulations				
	$dsim_r$	SD	z	$dsim_p$	SD	z	$U(z)$	χ^2 (df)
$KSbI$	10.08	0.07	-89.75	3.15	0.37	-0.43	-2.91	111.58 (8)
$KSxT$	5.99	0.06	-65.97	1.86	0.33	-0.10	-0.39	3.84 (6)
$KSwMT$	3.76	0.05	-58.72	1.07	0.12	-2.00	-4.55	50.06 (4)
$KSwAN$	4.56	0.05	-59.29	1.58	0.20	-0.76	-3.15	52.55 (5)

Note. SD refers to the distributions' standard deviations; $KSbI$ = knowledge space between items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test; $U(z)$ denotes the standardized U–score for large samples with tied ranks.

Mann–Whitney U and χ^2 tests are given.

Regarding the results for random simulations ($dsim_r$), the empirical z–scores are so far below the mean of the distributions, that further calculations have not been necessary ($-89.75 \leq z \leq -58.72$). According to prediction IIIb (Section 5.1.5), the results support the postulated model.

The standardization of the mean empirical distances ($ddat$) to the distributions of data sets derived from the probability simulations ($dsim_p$) reveals no significant difference for the test knowledge space or its subsets at an alpha–level of 0.01 ($-2.0 \leq z \leq -0.1$). The results of U and χ^2 tests show significant differences in favor of the hypotheses (prediction IIIc in Section 5.1.5) for all but one value. The values for the $KSxT$ are not significant ($U(z) = 0.39, \chi^2(6, N = 1144) = 3.84, p < 0.05$). This means that for the $KSbI$ and the two knowledge spaces within tests, the empirical probabilities for careless errors and/or lucky guesses are smaller than the assumed probabilities. For the $KSbI$, a more differentiated analysis of the distributions obtained by the probability simulations shows that the simulated data sets yield significantly lower mean distances ($z \geq 3.99$) at β levels $\leq .07$ and significantly higher mean distances ($z \leq -3.24$) at β levels $\geq .11$. For $.08 \leq \beta \leq .10$ the differences are not significant, which indicates that the probability for careless errors is located between 8% and 10%.

These results imply that, under the assumptions of $\eta = 0.16$ and $\beta \geq 0.08$, the postulated model and all of its substructures reliably reflect the empirical response patterns.

Since the results of the probability simulations confirm the postulated $KSbI$ and its substructures, I also performed the strictest test of the hypothesis by simulating data based on the solution frequency of items and persons. As for the other types of simulations, 1000 data matrices were simulated. The matrices have the same marginal frequencies as the empirical data matrix, but the distribution of '0's and '1's in the rows and columns of the matrices differ. Table 5.10 shows the mean symmetric distances ($dsim_f$), the DA values, and the results of the statistical analyses for the four substructures.

Table 5.10: Results of the frequency simulations for the test knowledge space and its substructures (N = 572)

	frequency simulations					
	$dsim_f$ (SD)	Mdn	DA	z	$U(z)$	χ^2 (df)
$KSbI$	3.01 (2.16)	3	0.30	-1.06	-0.35	4.43 (8)
$KSxT$	1.82 (1.28)	2	0.30	0.26	0.08	2.49 (5)
$KSwMT$	0.85 (0.93)	1	0.23	-0.93	-0.40	1.22 (4)
$KSwAN$	1.43 (1.34)	1	0.31	-0.53	-0.17	0.35 (5)

Note. $KSbI$ = knowledge space between items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test; $U(z)$ denotes the standardized U–score for large samples with tied ranks.

The results show that the symmetric distances ($0.85 \leq dim_f \leq 3.01$) and the DA values ($0.23 \leq DA \leq 0.31$) are basically equivalent to those of the empirical data. The results from the statistical analyses reveal only one significant value, namely the χ^2 for the $SRbI$ ($\chi^2(8, N = 1144) = 4.43, p < 0.05$). Thus, for the $SRbI$, the exact solution patterns are an important predictor within the postulated model, whereas the solution frequencies seem to suffice for its substructures.

5.2.4 Discussion

The results of Investigation I show that the approach of surmise relations between tests can be applied successfully when structuring a set of inductive reasoning tests by means of the componentwise ordering principle. The differentiated validation of the surmise relation between items ($SRbI$) and its subsets ($SRxT$, $SRwMT$, and $SRwAN$) allows the following assumptions about the way, in which each part of the surmise relation contributes to the results. The various validation methods all imply that the surmise relation established on the matrix test explains the set of data best, whereas the surmise relation on the analogy test deviates most from the empirical data. Thus, refinements of the postulated test knowledge space should primarily concern the hypothesis on the analogy test. However, the results of the simulation studies show that the postulated knowledge spaces fit the empirical data equally well or even better than sets of data that were simulated under the assumption of a correct model but with a certain amount of lucky guesses and careless errors. Therefore, all of the found deviations can be attributed to noise in the data. With regard to the assumed partial order, the frequency simulations showed that the postulated knowledge states are an important predictor for the solution behavior in both tests, whereas the solution behavior within the single tests can also be explained by the frequencies of correct responses per item and person.

Nevertheless, the reported results have to be interpreted with caution. The presentation of the matrix as well as the analogy test as speed–power tests with relatively strict time limits of 12 respectively 5 minutes lead to a high percentage of items that have not been processed by all participants. Therefore, it was necessary to apply the elimination procedure described in Section 5.2.3 (pre–editing of data and tests), which resulted in a reduction to only 30 out of the original 45 items. Additionally, the more difficult items were presented at the end of each session. Hence, the items remaining after the pre–editing of data and tests belong to an easier subset of the entire set of items. Regarding the matrix test, for example, there is only one item left in the finally analyzed set, which shows the attribute difficult type of operation. The resulting ceiling effect becomes obvious when looking at the averaged solution frequency of 84.66% for both tests.

As a consequence of the ceiling effect, the remaining subset of items shows only little variation in the solution frequencies (see Section 5.2.3, percentage of correct solutions). Ranging between 57.14% and 96.73% ($M = 82.58$, $SD = 12.45$) for the matrix items and between 69.02% and 88.4% ($M = 82.1$, $SD = 6.34$) for the analogy items, the relative solution frequencies do not yield the desired range for knowledge space validation procedures. As mentioned in Section 3.4, item pairs with either both or neither of the items solved correctly cannot contribute to the validation of a knowledge space hypothesis. In the case of a large set of high solution frequencies, we can also assume a large number of pairs with both items solved correctly. In fact, for the 351 item pairs in the SRbI, which were processed by the 572 participants, in 72.58% of the cases (i. e. 145,711 vectors of $\langle 1, 1\rangle$) both items were answered correctly. Such item pairs do not contradict the hypothesis. Consequently, the very well fitting knowledge space might be an artifact of the high number of correctly solved items.

5.3 Investigation II

Relating a pair of inductive reasoning tests: Reconsidered

The surmise relation between tests established in Investigation I could clearly be confirmed by means of the applied validation methods. However, as already pointed out in Section 5.2.4, there are several problems concerning the set of data and tests analyzed in Investigation I. The purpose of Investigation II is to carry out a second evaluation of the model and postulated classification scheme using a larger set of data obtained by a computer–aided presentation of items. Because the processing time for the two analyzed tests was restricted, the set of response patterns had to be reduced in order to obtain complete answer patterns. However, because of the large number of original patterns (12,759), a complete set of items (20 per test) was available even after deleting incomplete patterns. The inclusion of items with a broader range of difficulty also yields larger variations in the solution frequencies (in Investigation I, the more difficult items presented at the end of each test have been deleted). Furthermore, the computer–aided presentation of the tests permitted a clear distinction between items that were not proceeded versus items that could not be solved.

The set of data and the tests (see Sections 5.3.2.1 and 5.3.2.2) reanalyzed for Investigation II have been provided by the “Psychologischer Dienst der Bundeswehr der Bundesrepublik Deutschland” (PDB) in Bonn, Germany. The complete set of data consisted of response patterns from 23,802 participants, who worked on two sets of parallel tests (forms A and B). For the establishment and validation of the hypotheses, form A (12,759 patterns) of the inductive reasoning tests (a matrix and an analogy test) has been selected. In order to fulfill the condition of complete response patterns, the set of data was reduced to 2628 patterns (see Section 5.3.3).

5.3.1 Hypothesis

The hypothetical knowledge structure for the two inductive reasoning tests (geometric matrices and verbal analogies) used in Investigation II was constructed as outlined in Section 5.1. As in Investigation I, the surmise relation on the set of items and tests was established in two steps. First, the given items were analyzed with respect to their attributes on each of the five components and assigned to their corresponding item class (see Table 5.1). In a second step, an order on the item classes realized in the two tests was established, i. e. the subset of pairs relevant for the given material was extracted from the relations derived in Sections 5.1.3 and 5.1.4 (see Figures 5.3 and 5.4 for an illustration of the ordering principle).

Expectations regarding the derived surmise relation between tests ($SRbT$), the surmise relation between items ($SRbI$) and its subsets ($SRxT$ and $SRwT$), as well as the corresponding knowledge spaces ($KSbI$, $KSxT$, and $KSwT$) are according to the empirical predictions made in Section 5.1.5.

For the $SRbT$, the item classes realized in the two tests (see Section 5.3.2.2 for details) render a total–covering surmise relation from the matrix test to the analogy test ($AN\ \dot{\mathcal{S}}_t\ MT$) and a general surmise relation from the analogy test to the matrix test ($MT\ \dot{\mathcal{S}}\ AN$). For $AN\ \dot{\mathcal{S}}_t\ MT$, it is expected that each item class in the matrix test has a prerequisite item class in the analogy test ($AN\ \dot{\mathcal{S}}_l\ MT$) and that each item class in the analogy test is prerequisite for an item class in the matrix test ($AN\ \dot{\mathcal{S}}_r\ MT$). For $MT\ \dot{\mathcal{S}}\ AN$ it is expected that at least one item class in the analogy test has a prerequisite item class in the matrix test (see also Section 5.1.5, predictions Ia and Id). A detailed description of the items and prerequisite relationships contained in the two tests under investigation is given in the next section.

It should be noted that in spite of the different sets of item classes realized in Investigations I and II, the $SRbT$ have the same properties in both investigations (viz., a total–covering surmise relation from the matrix to the analogy test and a general surmise relation from the analogy test to the matrix test).

5.3.2 Method

5.3.2.1 Participants

From the original 12,759 male participants, 2628 response patterns were analyzed for Investigation II. Participants were draftees of the German Military who took form

A of the psychological aptitude test ("Psychologische Eignungsuntersuchung und –feststellung") developed by the PDB. Taking the test was obligatory for all participants. For confidentiality reasons personal data of the participants was not provided.

5.3.2.2 Material

Materials consist of two computer–aided inductive reasoning tests developed by the PDB.

The first test (MT) contains 20 geometric matrix items. The items are constructed as 2 by 3, 2 by 4, or 3 by 3 matrices with 3, 2, and 15 items respectively. The items contain different types of geometric forms (squares, circles, etc.) colored in black and a blank square on the lower right. By analyzing the geometric forms row by row and/or column by column, between one and three relational rules can be induced (see Table 5.11). Below each matrix eight answer alternatives (labeled 1 through 8) are depicted. Participants had to decide which alternative belongs in the blank square to complete the matrix correctly. Figure 5.9 depicts an example matrix (for confidentiality reasons the example is modified), which requires the rule 'Figure Addition' as only necessary operation to induce the correct answer '5'. With one rule of less difficulty, low demands on the correspondence finding process, geometric material, and eight answer alternatives, the corresponding item class is (O,1,L,G,8).

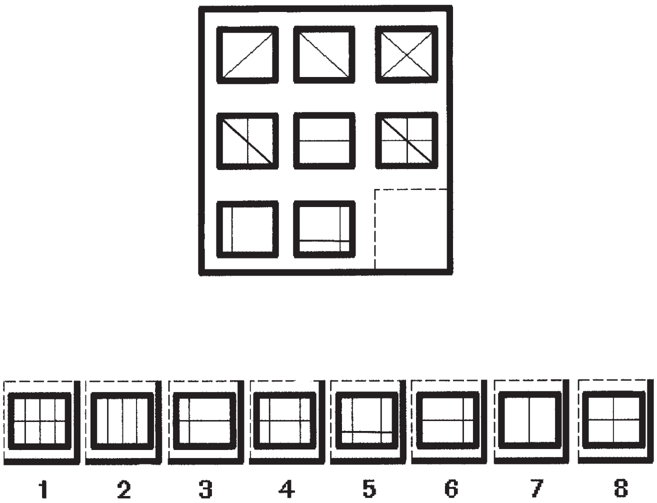


Figure 5.9: Example for a matrix item presented in Investigation II

An analysis of the items' attributes on each of the five components (see Table 5.1) results in seven item classes with one to five items each. Table 5.11 shows the attribute combinations (first column), item numbers (second column), operation types (third column), and downsets for the item classes and items (last two columns).

For component A (operation difficulty), the attribute **D**ifficult type of operation (three item classes) includes the Boolean AND operator (BA) and the exclusive-OR rule (XO), whereas **O**ther types of operations (four item classes) include the rules constant in a row (CR), pairwise progression (PP), figure addition (FA), distribution of two values (D2), and distribution of three values (D3). The rules are listed in Table 5.11 (column three) and described in detail in Section 2.2.3, Table 2.2. The attributes one, two, and three operations necessary to solve the problem are realized for component B (number of operations) with two, four, and one item classes respectively. Component C (constraint) differentiates between four item classes with **H**igh and three item classes with **L**ow demands on correspondence finding processes, while components D (material) and E (number of answer alternatives) show identical attributes for all item classes, viz. **G**eometric–figural material and **8** answer alternatives. The last two columns in Table 5.11 depict the downsets for item classes (fourth column) and items (fifth column), which were derived according to Section 5.1, Table 5.1.

The second test (AN) consists of 20 verbal analogy problems in German. Each analogy has a three–term stem of the form $A : B = C : ?$. Below the stem terms, five answer alternatives (labeled 1 through 5) are presented. Participants had to choose the alternative that completes the analogy best. Stem terms as well as alternatives are nouns (18 items) or adjectives (1 item). One analogy is mixed with nouns for terms A and C and verbs for term B and the five answer alternatives. Figure 5.10 depicts an example analogy (for confidentiality reasons the example is an extended item taken from Bejar et al., 1991, but is similar to the original items), which belongs to the item class (O,1,L,V,5). A more detailed description of the item is given in see Section 5.2.2.2.

wheel : car = leg : ?

1) horse
2) bicycle
3) forest
4) bookcase
5) snake

Figure 5.10: Example for an analogy item presented in Investigation II

An analysis of the items' attributes on each component (see Table 5.1) results in nine item classes with one to four items each. Table 5.12 depicts the attribute combinations (first column), item numbers (second column), operation types (third column), and downsets for the item classes and items (last two columns). As already mentioned in Section 5.2.2.2, for verbal analogies the number of operations (component B) refers to the number of elements in the analogies' rationales (see Section 2.2.1, Box 2.3). Table 5.12 shows the semantic relations between the terms (third column) of each item, whereas the complete rationales are given in Appendix B.1, Table B.2.

With regard to the different attribute combinations (first column), the types of semantic relations (component A, specified in the third column) that are realized in this test are part–whole (PW), contrast (CO), attribute (AT), cause–purpose (CP), and space–time (ST), which constitute the operations of **O**ther types (realized in five item classes), and class inclusion (CI) and similar/comparative (SI) as **D**ifficult operations (four item classes). The semantic rules are described in detail in Section 2.2.1, Table 2.1. The number of elements in the rationales (component B) varies between one (2

Table 5.11: Item descriptions and downsets for the matrix test in Investigation II

Item classes[a]	Item numbers	Operation types[b]	Downsets for item classes[c]	Downsets for items
(O,1,L,G,8)	1 2 7 8	FA FA FA FA	(O,1,L,G,8)	1,2,7,8
(O,1,H,G,8)	3 4 6	PP PP FA	(O,1,L,G,8),(O,1,H,G,8)	1,2,3,4,6,7,8
(O,2,L,G,8)	5 9 10 11 12	PP,PP D2,PP D2,CR PP,PP PP,PP	(O,1,L,G,8),(O,2,L,G,8)	1,2,5,7,8,9,10, 11,12
(O,2,H,G,8)	14 15 16 17 19	D3,FA PP,PP FA,PP D3,PP D3,D3	(O,1,L,G,8),(O,1,H,G,8), (O,2,L,G,8),(O,2,H,G,8)	1,2,3,4,5,6,7,8,9, 10,11,12,14,15, 16,17,19
(D,2,L,G,8)	13	PP,BA	(O,1,L,G,8),(O,2,L,G,8), (D,2,L,G,8)	1,2,5,7,8,9,10, 11,12,13
(D,2,H,G,8)	18	PP,BA	(O,1,L,G,8),(O,1,H,G,8), (O,2,L,G,8),(O,2,H,G,8), (D,2,L,G,8),(D,2,H,G,8)	1,2,3,4,5,6,7,8,9, 10,11,12,13,14, 15,16,17,18,19
(D,3,H,G,8)	20	XO,BA,BA	(O,1,L,G,8),(O,1,H,G,8), (O,2,L,G,8),(O,2,H,G,8), (D,2,L,G,8),(D,2,H,G,8), (D,3,H,G,8)	1,2,3,4,5,6,7,8,9, 10,11,12,13,14, 15,16,17,18,19, 20

Note. [a] Components are ordered alphabetically, i. e. A through E. [b] CR = constant in a row, PP = pairwise progression, FA = figure addition, D2 (D3) = distribution of two (three) values, BA = Boolean AND, XO = exclusive–OR. [c] See Table 5.1 for the derivation of downsets.

item classes) and three (one item class). **H**igh demands on correspondence finding processes (component C) are required by five item classes, **L**ow demands by four item classes. The attributes on components D and E are identical for all items, viz. **V**erbal material and **5** answer alternatives.

A combined analysis of the items of both tests results in the $SRxT$ shown in Table

Table 5.12: Item descriptions and downsets for the analogy test in Investigation II

Item classes[a]	Item numbers	Operation types[b]	Downsets for item classes[c]	Downsets items
(O,1,L,V,5)	21 23	ST CO	(O,1,L,V,5)	21,23
(O,1,H,V,5)	30	AT	(O,1,L,V,5),(O,1,H,V,5)	21,23,30
(O,2,L,V,5)	36	PW	(O,1,L,V,5),(O,2,L,V,5)	21,23,36
(O,2,H,V,5)	35	PW	(O,1,L,V,5),(O,1,H,V,5), (O,2,L,V,5),(O,2,H,V,5)	21,23,30,35,36
(O,3,H,V,5)	39	PW	(O,1,L,V,5),(O,1,H,V,5), (O,2,L,V,5),(O,2,H,V,5), (O,3,H,V,5)	21,23,30,35,36, 39
(D,1,L,V,5)	22 24 27	CI SI SI	(O,1,L,V,5),(D,1,L,V,5)	21,22,23,24,27
(D,1,H,V,5)	25 28 32	SI SI SI	(O,1,L,V,5),(O,1,H,V,5), (D,1,L,V,5),(D,1,H,V,5)	21,22,23,24,25, 27,28,30,32
(D,2,L,V,5)	26 29 34 37	SI SI SI SI	(O,1,L,V,5),(O,2,L,V,5), (D,1,L,V,5),(D,2,L,V,5)	21,22,23,24,26, 27,29,34,36,37
(D,2,H,V,5)	31 33 38 40	SI SI SI SI	(O,1,L,V,5),(O,1,H,V,5), (O,2,L,V,5),(O,2,H,V,5), (D,1,L,V,5),(D,1,H,V,5), (D,2,L,V,5),(D,2,H,V,5)	21,22,23,24,25, 26,27,28,29,30, 31,32,33,34,35, 36,37,38,40

Note. [a] Components are ordered alphabetically, i. e. *A* through *E*. [b] PW = part–whole, CP = cause purpose, CO = contrast, AT = attribute, ST = space–time, SI = similar/comparative. [c] See Table 5.1 for the derivation of downsets.

5.13. The prerequisite relationships were derived by using the order of importance on the components and the difficulty order on the attributes specified in Table 5.1 (column two and four). The item classes realized in the two tests yield a total–covering surmise relation from the matrix test to the analogy test and a general surmise relation from the analogy test to the matrix test. For the total–covering surmise relation ($AN\ \dot{\mathcal{S}}_t\ MT$), Table 5.13 shows that each matrix item class has a prerequisite item class (indicated by '◇') in the analogy test ($AN\ \dot{\mathcal{S}}_l\ MT$, see Definition 3.5) and that each analogy item class is prerequisite for some item class in the matrix test ($AN\ \dot{\mathcal{S}}_r\ MT$, see Definition 3.6). Regarding the general surmise relation ($MT\ \dot{\mathcal{S}}\ AN$, see Definition 3.4), it can be inferred from the prerequisites (indicated by 'x') specified in Table 5.13 that these two conditions are not fulfilled for the surmise relation from the analogy to the matrix test, but that there are prerequisite item classes in the matrix test. The surmise relation between the items of both tests (i. e. within and across the tests) is depicted as Hasse diagram in Figure 5.11. The relation files for the $SRbI$, $SRxT$, $SRwMT$, and $SRwAN$ are given in Appendix E.1, the corresponding base files for the $KSbI$, $KSxT$, $KSwMT$, and $KSwAN$ in Appendix E.2.

Table 5.13: Surmise relation across tests in Investigation II

Analogies	Matrices (D = G, E = 8)						
(D = V, E = 5)	(O,1,L)	(O,1,H)	(O,2,L)	(O,2,H)	(D,2,L)	(D,2,H)	(D,3,H)
(O,1,L)	◇	◇	◇	◇	◇	◇	◇
(O,1,H)	x	◇		◇		◇	◇
(O,2,L)	x		◇	◇	◇	◇	◇
(O,2,H)	x	x	x	◇		◇	◇
(O,3,H)	x	x	x	x			◇
(D,1,L)	x				◇	◇	◇
(D,1,H)	x	x				◇	◇
(D,2,L)	x		x		◇	◇	◇
(D,2,H)	x	x	x	x	x	◇	◇

Note. An x indicates that the matrix item class in column i is prerequisite for the analogy item class in row j (iSj); a ◇ indicates that the analogy item class in row j is prerequisite for the matrix item class in column

5.3.2.3 Procedure

The data were collected between December 1997 and and October 1998 by members of PDB. The application of the aptitute test was carried out via computer and covered various subtests (geometric matrices, verbal analogies, arithmetics, mechanics, electronics). The matrix test was presented first, followed by the analogy test. At the beginning of each test a short description of the tests' objective and several instruction items were presented (seven items for the matrix test and eight items for the analogy test). Each example included a typical item for the respective test, the correct solution, and a comprehensive explanation of the solution (i. e. which relations are relevant to complete a given matrix or analogy problem). After the instructional items, the test items were presented in the same order for all participants (see item numbers in Tables 5.11 and 5.12). The processing time for each test was limited[5] (18 minutes for the matrices, 4.5 minutes for the analogies), starting with the presentation of the first test item. There was no time limit for the processing of single items. Participants had the possibility to skip items and review them at the end of the test (within the given time limit). The items were presented in a multiple choice format, where participants had to type in the correct solution number. For each item the computer program recorded the given response, the solution time[6], and whether an item was skipped.

5.3.3 Results

The validation of the established hypotheses follows the methods outlined in Section 3.4. First, I will report the results derived via the surmise relation, which is followed by the results derived via the knowledge space. The latter will also include comparisons with simulated data sets and statistical analyses.

Pre–editing of data and tests The original data set analyzed in Investigation II contained 12,759 response patterns. In order to obtain a greater variability in item difficulty (than in Investigation I), only incomplete answer patterns were excluded from the data set. Due to the large number of participants, it was not necessary to exclude items. The remaining raw data analyzed in this investigation contain 105,120 responses from 2628 participants for the 20 matrix and the 20 analogy items. Table 5.14 gives an overview of the original and the reduced data sets. The minimal number of correctly solved items equals zero (2 response patterns), the maximal number 37 items (1 response pattern). The two trivial response patterns (0.08% of all patterns) have been kept in the analysed data set.

5.3.3.1 Validation of hypotheses via the surmise relation

Percentage of correct solutions Overall, the participants solved 50.62% of the items, which corresponds to a total of 53,212 correct responses. The frequencies for each single item are provided in Appendix F.1, Table F.2. The postulated surmise relation between items together with the relative solution frequencies for the item classes in Investigation II is depicted in Figure 5.11. The relative solution frequencies for the nine item classes with more than one item (see Tables 5.11 and 5.12) represent the averaged relative solution frequencies of all items in the respective class. The average maximal difference in solution frequencies of items contained in the same class amounts to 14.69% ($SD = 9.51$). With regard to the item classes, the percentages of correct solutions vary from 8.98% for item class (D,3,H,G,8) up to 82.18% for item class (O,1,L,G,8).

[5] The presentation of the items as speed–power test lead to incomplete answer patterns, which was the reason for eliminating patterns from the original data matrix (see Section 5.3.3)

[6] Since the postulated model predicts the empirical response patterns with respect to correct/incorrect answers but not the processing times, latencies are not included in the results.

Table 5.14: Original (N=12,759) and reduced (N=2628) data sets in Investigation II

Test	No. of items	No. of responses	No. of correct responses	No. of incorrect responses	No. of missing responses
MT original	20	255,180	139,375	106,130	9,675
AN original	20	255,180	123,023	83,365	48,792
Total	40	510,360	262,398	189,495	58,467
MT reduced	20	52,560	26,092	26,468	0
AN reduced	20	52,560	27,120	25,440	0
Total	40	105,120	53,212	51,908	0

Note. MT = matrix test, AN = analogy test, total refers to the set of items contained in both tests.

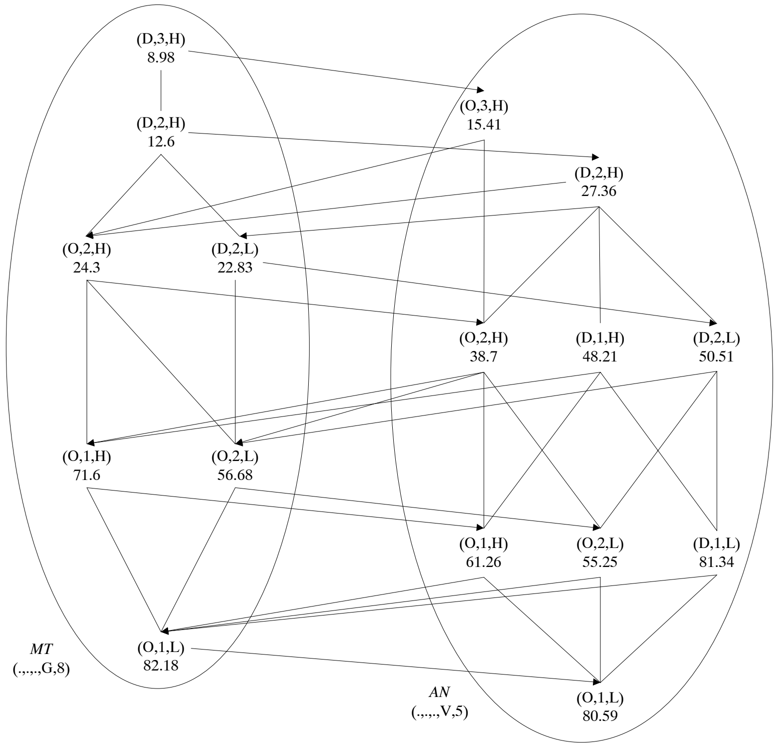


Figure 5.11: Hasse diagram for the postulated surmise relation between items and relative solution frequencies in Investigation II ($AN\ \dot{\mathcal{S}}_t\ MT$ and $MT\ \dot{\mathcal{S}}\ AN$)

In the following, referring to pairs of item classes, does not include reflexive pairs.

The results for the two surmise relations within tests show that all of the 18 postulated pairs of item classes in the geometric matrix test ($SRwMT$, left ellipse) and 22 out of 23 pairs in the verbal analogy test ($SRwAN$, right ellipse) are confirmed by the solution frequencies. Within the analogy test, the percentage of correct solutions for item class (D,1,L,V,5) is higher than that for its postulated prerequisite item class

(O,1,L,V,5), but the difference is not significant ($\chi^2(1, N = 4256) = .09, p < .05$[7]). Since the hypothesis (prediction IIa in Section 5.1.5) states that the solution frequency for a prerequisite item class should be higher than or equal to the frequency of the item class it can be surmised from, the reversed pair is still in accordance with the postulated surmise relation.

Regarding the *SRxT* (see Figure 5.11), i. e. prerequisite relationships between item classes of different tests, 27 out of 30 pairs confirm the postulated relationships from the matrix test to the analogy test and 17 out of 19 pairs the relationships from the analogy to the matrix test (see also Table 5.13). In the first case, the relative solution frequencies for item classes (O,1,L,G,8), (O,1,H,G,8), and (O,2,L,G,8) are higher than their respective prerequisites (O,1,L,V,5), (O,1,H,V,5), and (O,2,L,V,5). A significant difference was only found for the pair (O,1,H,G,8) and (O,1,H,V,5), $\chi^2(1, N = 3492) = 21.19, p < .05$ (see Appendix F.2, Table F.5 for the remaining values). A notable point is, that the three item classes in the matrix test and their respective prerequisites in the analogy test are all described by the same attributes on the three main components A, B, and C. For the relation from the analogy test to the matrix test, deviations occur for the analogy class (D,2,H,V,5) with respect to its postulated prerequisites (O,2,H,G,8) and (D,2,L,G,8). For both pairs the differences are significant, $\chi^2(1, N = 1358) = 4.76$ for (O,2,H,G,8) and $\chi^2(1, N = 1319) = 10.74$ for (D,2,L,G,8), $p < .05$.

With regard to the *SRbT* (predictions Ia and Id in Section 5.1.5), the results clearly confirm the hypothesized general surmise relation from the analogy test to the matrix test ($MT\ \dot{\mathcal{S}}\ AN$), since all eight of the expected analogy item classes have a prerequisite in the matrix test. The total–covering surmise relation from the matrix test to the analogy test ($AN\ \dot{\mathcal{S}}_t\ MT$) is confirmed by 15 out of the expected 16 item classes of both tests. More precisely, all of the nine item classes in the analogy test are prerequisite for some item class in the matrix test, while six out seven item classes in the matrix test have a prerequisite in the analogy test. Item class (O,1,L,G,8) has a higher solution frequency than its postulated prerequisite (O,1,L,V,5) but the difference is not significant ($\chi^2(1, N = 4278) = .41$).

Summing up, the *SRbI* (i. e. pairs in the *SRxT* and in the two *SRwT*) is confirmed by 84 out of 90 postulated pairs of item classes. Of the six pairs with reversed solution frequencies four pairs differ significantly.

Summarized, the surmise relation between item classes of both tests is confirmed by 93.33% of the postulated pairs, the surmise relation across tests by 89.8%, and the surmise relations within the matrix and the analogy test by 100% and 95.65% respectively. Regarding only the statistically significant differences, the *SRbI* is confirmed by 96.66% of the pairs, the *SRxT* by 93.88%, and the two *SRwT* by 100% each.

Indices for the fit of a surmise relation To estimate the fit of each item pair in the surmise relation, I calculated the violational coefficient (VC) and the gamma–index (γ_G) for the *SRbI* and its subsets, viz. the *SRxT* and the two *SRwT*. For both indices, the obtained values refer to relationships between single items (as compared to item classes).

[7] See footnote on page 102.

Table 5.15: VC and γ for the surmise relation between items and its subsets (N = 2628)

	No. of items	No. of pairs[a]	VC	γ_G	$\gamma > 0$	χ^2
SRbI	40	543	0.09	0.64	503 (92.63%)	278,142.75
SRxT	40	284	0.10	0.58	248 (87.32%)	117,034.71
SRwMT	20	138	0.06	0.76	137 (99.28%)	106,044.29
SRwAN	20	121	0.09	0.63	118 (97.52%)	59,211.45

Note. $SRbI$ = surmise relation between the items of both tests, $SRxT$ = surmise relation across the two tests, $SRwMT$ = surmise relation within the matrix test, $SRwAN$ = surmise relation within the analogy test. [a]Reflexive pairs are not counted.

As shown in Table 5.15, the indices signify that the surmise relation on the matrix test fits the set of data best ($\gamma_G = .76$, $VC = .06$), while the surmise relation across tests deviates most from the data ($\gamma_G = .58$, $VC = .1$). With regard to VC, the data violates the postulated surmise relations in 6% to 10% of the respective pairs. The differences between the $SRbI$, the $SRxT$ and the $SRwAN$ are too small to be interpreted.

The results derived from the γ–index show that between 87.32% ($SRxT$) and 99.28% ($SRwMT$) of the postulated item pairs yield a positive γ value, i. e. the number of concordant responses is higher than the number of discordant responses. Furthermore, all four of the McNemar χ^2 values indicate that the differences between the concordant and discordant pairs are highly significant, which supports the postulated models (prediction IIb in Section 5.1.5).

The results derived from the two indices correspond to the more global analysis of relative solution frequencies (see above). Thus, also on an item level the relationships across the two tests contribute most to the deviations in the $SRbI$, while the relationships postulated within the matrix test fit the empirical data best. With regard to the validity of the $SRbT$, the indices do not allow any conclusions.

5.3.3.2 Validation of hypotheses via the knowledge space

Symmetric distances and distance agreement coefficient Table 5.16 depicts the mean symmetric distances (*ddat*), the theoretical maxima (*dmax*), and the distances for the powersets (*dpot*) of the test knowledge structure ($KSbI$) and its substructures ($KSxT$, $KSwMT$, and $KSwAN$). For all four of the postulated knowledge spaces the mean empirical distances ($1.79 \leq ddat \leq 6.08$) are far below their theoretical maxima ($10 \leq dmax \leq 20$) and the distances for the respective powersets[8]

[8] For the $KSbI$ and the $KSxT$ with 40 items each, the powerset was not computable and therefore, the values for *dpot* were calculated with 20,000 simulated random patterns (see also footnote on page 104).

Table 5.16: Symmetric distances for the test knowledge space, its substructures, and their powersets (N = 2628)

	m	$\lvert\mathcal{K}\rvert$	$dmax$	$ddat$ (SD)	Mdn	$dpot$ (SD)	Mdn	DA
$KSbI$	40	7,633	20	6.08 (2.26)	6	12.96 (2.33)	13	0.47
$KSxT$	40	2,278,770	20	4.67 (1.76)	5	8.20 (1.64)	8	0.57
$KSwMT$	20	343	10	1.79 (1.25)	2	5.53 (1.57)	6	0.32
$KSwAN$	20	484	10	2.45 (1.32)	2	5.38 (1.57)	5	0.46

Note. m denotes the number of items, $|\mathcal{K}|$ the number of knowledge states; $KSbI$ = knowledge space between the items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test.

($5.38 \leq dpot \leq 12.96$). χ^2 statistics show that the empirical data fit the test knowledge space as well as its substructures significantly better than their respective powersets ($36,649 \leq \chi^2 \leq 264,596$; see Appendix F.5 for the exact values), which supports the hypothesis according to prediction IIIa (Section 5.1.5). The distance distributions for the $KSbI$, its substructures, and the corresponding powersets are provided in Appendix F.3, Tables F.9 and F.10. Regarding the proportion of careless errors and lucky guesses contributing to the mean empirical distance, an analysis of the items' invalidity (see Appendix F.4, Table F.14) shows no difference between the two measures (on the average, the mean distance is composed of .076 careless errors and .075 lucky guesses per item and person)[9].

For a comparison of the various structures' fit, the distance agreement coefficient (DA) was calculated. The DA values for the test knowledge space and its substructures are comparable to the results derived via the surmise relation (percentage of correct solution, γ_G, and VC, see Section 5.3.3.1). The matrix test shows the best fitting knowledge space ($DA = 0.32$), whereas the $KSxT$ deviates most from the empirical data ($DA = 0.57$).

Simulations In Table 5.17 the mean symmetric distances and the DA values for the empirical data ($ddat$) are compared to the according values for simulated data sets. The values for random simulations ($dsim_r$) and probability simulations ($dsim_p$) are the averaged mean distances and standard deviations from 1000 data sets each. The number of response patterns per data set corresponds to the number in the empirical data set, i. e. 2628 patterns. For $dsim_p$, the probability for lucky guesses corresponds to the number of answer alternatives ($\eta = 0.16$ for $KSbI$ and $KSxT$, 0.125 for $KSwMT$, and 0.2 for $KSwAN$). Since the probability for careless errors is not known, it was varied in 10 steps with $0.05 < \beta \leq 0.15$. For each probability level, 100 sets of data have been simulated. The averaged distance distributions for the simulated data sets are provided in Appendix F.3, Tables F.9 and F.10.

[9] See footnote on page 104.

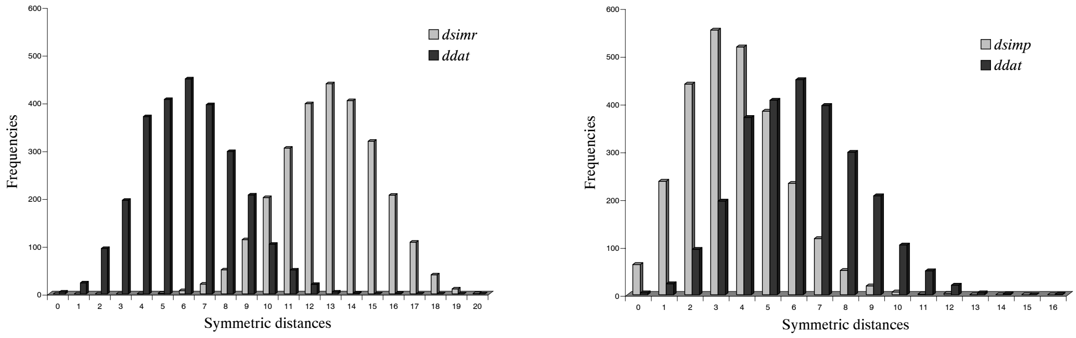


Figure 5.12: Distance distribution of the empirical data set (*ddat*) compared to the distributions of random ($dsim_r$) and probability simulations ($dsim_p$)

The values for random simulations ($dsim_r$) are equivalent to those found for the powersets of the various structures ($5.53 \leq dsim_r \leq 12.95$). Regarding the results for the simulations based on the postulated knowledge spaces ($dsim_p$), the mean symmetric distances as well as the DA values for the $KSbI$ and its substructures are below, i. e. better than those for the empirical data set ($1.44 \leq dsim_p \leq 3.68$).

Figure 5.12 illustrates the differences between the empirical and the simulated distance distributions of the test knowledge space. With regard to the random simulations (left figure) the overlap of the empirical and the simulated distribution is rather small and the empirical data is located further left on the distance scale. Thus, it can be concluded, that the postulated test knowledge space fits the empirical data far better than random response patterns. The distribution obtained from the probability simulations (right figure) shows a greater overlap with the empirical distribution but in this case, the empirical distances are located further right on the distance scale. Thus, the postulated test knowledge space fits the empirical data set less than the data sets simulated on the hypothesis.

Table 5.17: Symmetric distances for the test knowledge space, its substructures, and simulated data sets (N = 2628)

			random simulations			probability simulations		
	$ddat$	DA	$dsim_r$ (SD)	Mdn	DA	$dsim_p$ (SD)	Mdn	DA
$KSbI$	6.08	0.47	12.95 (2.34)	13	0.999	3.68 (1.81)	4	0.28
SxT	4.67	0.57	8.19 (1.63)	8	0.999	2.42 (1.48)	2	0.30
$KSwMT$	1.79	0.32	5.53 (1.57)	6	1.000	1.44 (1.11)	1	0.26
$KSwAN$	2.45	0.46	5.38 (1.57)	5	1.000	1.68 (1.20)	2	0.31

Note. For the simulated data sets SD refers to the mean standard deviations; $KSbI$ = knowledge space between the items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test.

Table 5.18: Comparison of the empirical and simulated symmetric distances for the test knowledge space and its substructures (N = 2628)

	random simulations			probability simulations				
	$dsim_r$	SD	z	$dsim_p$	SD	z	$U(z)$	χ^2 (df)
$KSbI$	12.95	0.05	-142.00	3.68	0.51	4.68	36.70	7481.01 (11)
$KSxT$	8.19	0.03	-114.83	2.42	0.49	4.56	41.34	9794.60 (8)
$KSwMT$	5.53	0.03	-123.86	1.44	0.15	2.33	10.31	268.85 (5)
$KSwAN$	5.38	0.03	-96.14	1.68	0.21	3.67	22.63	1091.69 (6)

Note. SD refers to the distributions' standard deviations; $KSbI$ = knowledge space between the items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test; $U(z)$ denotes the standardized U–score for large samples with tied ranks.

In order to find out, whether or not the obtained differences are significant, the following statistical analyses have been performed.

Statistical analyses As for Investigation I, I computed several tests to judge, whether or not the differences between the empirical and the simulated data sets are statistically significant. With regard to the differences between the empirical data and randomly simulated data sets ($dsim_r$), the z–scores in Table 5.18 show that the empirical distances are located significantly below the distributions of the random simulations ($-142 \leq z \leq -96.14$), which supports the postulated model (see prediction IIIb in Section 5.1.5).

The standardization of the mean empirical distances to the distributions of probability simulations ($dsim_p$) reveals significant differences for the $KSbI$, the $KSxT$, and the $KSwAN$. At an alpha–level of 0.01, the z–score for the matrix test is not significant ($z = 2.33$). However, the results derived from U and χ^2 statistics show significant differences for all four (sub)structures ($U(z) \geq 10.31, \chi^2(N = 5256) \geq 268.85$). Regarding only the highest assumed β level with 15% careless errors, $dsim_p$ for the $KSbI$ amounts to 4.38 (with the distribution's $SD = 0.03$, $z = 50.63$ for the respective 100 data sets). Thus, even with the highest assumed noise level, the model cannot be explained by the empirical data. It therefore has to be concluded, that the deviations of the postulated model to the set of data are not only due to the assumed noise variables and the hypothesis regarding prediction IIIc (Section 5.1.5) has to be rejectet (considering these results, frequency simulations were not computed, since they constitute an even stricter test of the hypothesis). The results obtained for random simulations, on the other hand, imply that the postulated test knowledge space still explains the empirical data better than random response vectors.

5.3.4 Discussion

The investigation reported in this section was conducted in order to overcome some of the methodological problems that emerged in Investigation I (see Section 5.2.4). The reanalysis of a larger set of data (2628 participants) derived by a computer–aided presentation of items yielded the desired range of item difficulty. With a mean of 50.62% of correctly solved items, the percentage of correct solutions ranges between 8.98% and 82.18%. Except for the relationships among the three item classes (O,1,L,G,8), (O,1,L,V,5), and (D,1,L,V,5) and between the two classes (O,2,L,G,8) and (O,2,L,V,5), the solution frequencies differ clearly for each prerequisite relationship in the postulated surmise relation.

As already noted at the end of Section 5.3.3, the results obtained in this investigation imply that the postulated model on the test knowledge space between items (and its substructures) fits the empirical data significantly better than random answer vectors. However, the data sets obtained from the probability simulations are significantly better fitting than the empirical data. This result implies that either the model contains false predictions or that the assumed amount of noise in the data does not suffice to explain the differences between the hypothesized (test) knowledge states and the empirical answer patterns.

Reviewing the results under consideration of the various (sub)structures, all validation procedures indicate most deviations for the surmise relation across tests. A closer look at the solution frequencies shows that the item classes of three out of five contradicting pairs are described by the same attributes on the three main components A, B, and C (operation difficulty, number of operations, and constraint). In all three cases, the postulated prerequisite in the analogy test differs from the corresponding matrix item class only with respect to the components material and number of answer alternatives. Hence, the assumption that geometric–figural material is more difficult to process than verbal material might be too strong. The influence of the number of answer alternatives may also be reduced when the chances for lucky guesses are relatively small in both cases (five vs. eight alternatives).

With regard to the surmise relation between tests, the relative solution frequencies (regarding the statistically significant results) confirm the total–covering surmise relation from the matrix to the analogy test ($AN\ \dot{\mathcal{S}}_t\ MT$) as well as the general surmise relation from the analogy to the matrix test ($MT\ \dot{\mathcal{S}}\ AN$). This indicates, that although there are deviations with respect to the relationships between single item classes, the predictions rendered by the $SRbT$ can be maintained. More exactly, a correct completion of the matrix test implies a correct completion of the analogy test and failing on the analogy tests implies a failure on the matrix test.

Comparing the results of this investigation to those of Investigation I, two issues with respect to the validation methods have to be discussed. The first issue concerns the differences in the results obtained via the surmise relation vs. the knowledge space. Investigation II shows a higher percentage of pairs in the surmise relation that are confirmed by the solution frequencies as well as higher γ_G values and higher percentages of positive γ indices for the surmise relation between items and its subsets (VC yielded negligible smaller values for Investigation I). Hence, looking only at pairs of items (or item classes), the results for this investigation imply a higher validity of the model than for Investigation I. The values obtained via the knowledge space, on the other hand, indicate the direct opposite (DA and simulation studies).

The reason for these contradicting results might be found in the different approaches the two validation methods take. The procedures working with the surmise relation validate each pair in the relation separately, while the knowledge space procedures compare the complete answer vectors with the knowledge states. The high solution frequencies in Investigation I seem to positively influence the validation of the surmise relation to a smaller degree than the validation of the knowledge space. For the percentage of correct solutions, all values increase and the γ–index only accounts for concordant and discordant pairs (pairs with both or neither of the items solved are excluded). The calculation of symmetric distances, on the other hand, is based on the nearest knowledge state. The higher the number of correct responses, the smaller is the distance to the full set. Hence, high solution frequencies will yield relatively small distances.

With this issue in mind, the interpretation of the results derived via the knowledge space should always include a consideration of the data set's properties, viz. the range of solution frequencies and the number of trivial response patterns .

The second issue to be discussed deals with the assumed noise in the data set. While probability simulations lead to even higher distances than the empirical data set in Investigation I, the same simulations yield significantly better values than the data set of Investigation II. One reason is the already mentioned problem of high solution frequencies in Investigation I. Another methodological difference in the two investigations arises from the participants. The data analyzed in Investigation I stems from corporal and officer candidates who had obviously personal interest in doing their best. The participants in Investigation II were draftees, i. e. they were liable for military service. In this case, we cannot be sure that all participants gave their best. What follows, is that the amount of careless errors will increase and eventually be higher than the amount of noise assumed for the probability simulations.

Investigation III should, among others, clarify, whether the knowledge space results of this investigation are to be contributed to properties of the postulated model or to an unexpectedly high amount of noise in the data.

5.4 Investigation III

Extension of the model to a surmise relation between four inductive reasoning tests

Investigation III has two main purposes. The results of the two previous investigations show that there is no obstacle to applying the method of surmise relations between tests (see Section 3.2) and the developed classification system for inductive reasoning tests (see Sections 5.1.1 and 5.1.2) to a set of more than two tests. Hence, the first purpose of Investigation III is to find out, how well the postulated model fits a set of data when a larger set of problem types is presented. Secondly, the methodological issues discussed in Sections 5.2.4 and 5.3.4 need to be accounted for by collecting data in accordance with the requirements of knowledge space research.

For these reasons, I conducted a third investigation, in which four different types of inductive reasoning problems were presented. The problems included verbal and geometric analogies, numerical series completion problems, and geometric matrices with five items each. With regard to the empirical standards of knowledge space theory, the items were presented in random order and participants were given enough time to process all of the presented items. In order to achieve a more consistent form of presentation for the various problem types, I standardized the number of answer alternatives to five alternatives per item. The items originally stem from two intelligence tests, namely the Berlin Structure of Intelligence test (BIS test) by Jäger et al. (1997, see Section 2.4.2.2) and the Vienna Matrices Test (WMT) by Formann and Piswanger (1979, see Section 2.4.2.4), but were adapted to meet the criterion of uniform answer formats[10].

5.4.1 Hypothesis

The hypothetical knowledge structure for the four inductive reasoning tests (verbal and geometric analogies, number series completions, and geometric matrices) used in Investigation III was constructed as outlined in Section 5.1. Analogously to Investigations I and II, the surmise relation on the set of items and tests was established in two steps. First, I analyzed the given items with respect to their attributes on each of the five components (see Table 5.1) and assigned each item to its respective item class. In a second step, the $SRbI$, $SRwT$, $SRxT$, and $SRbT$ were established for the item classes realized in the four tests. The so found surmise relations constitute subsets of the respective relations derived in Sections 5.1.3 and 5.1.4 (see Figures 5.3 and 5.4 for an illustration of the ordering principle).

Expectations regarding the derived surmise relation between tests ($SRbT$), the surmise relation between items ($SRbI$) and its subsets ($SRxT$ and $SRwT$), as well as the corresponding knowledge spaces ($KSbI$, $KSxT$, and $KSwT$) are according to the empirical predictions made in Section 5.1.5.

Regarding the $SRbT$, the derived structure yields the following pairs for the general, left–, and right–covering surmise relation between tests (see Section 5.4.2.2 for details).

- The left–covering surmise relation contains five test pairs ($\mathcal{T}_i\ \dot{\mathcal{S}}_l\ \mathcal{T}_j$). Three of the five pairs are relationships from the verbal analogy test to the remaining three tests, viz. to the geometric analogy test ($AN_G\ \dot{\mathcal{S}}_l\ AN_V$), to the series completion test ($SC_N\ \dot{\mathcal{S}}_l\ AN_V$), and to the matrix test ($MT_G\ \dot{\mathcal{S}}_l\ AN_V$). The

[10] The original set of items presented to the participants also contained five geometric classification problems. However, the structure of these problems does not allow a modification of the answer format without a substantial alteration of the original task. Therefore, the classification problems were not included in further analysis.

remaining two test pairs are relationships from the series completion test to the geometric analogy test ($AN_G\ \dot{\mathcal{S}}_l\ SC_N$) and from the matrix test to the geometric analogy test ($AN_G\ \dot{\mathcal{S}}_l\ MT_G$). Note, that there are two transitive triplets, namely $AN_G\ \dot{\mathcal{S}}_l\ SC_N\ \dot{\mathcal{S}}_l\ AN_V$ and $AN_G\ \dot{\mathcal{S}}_l\ MT_G\ \dot{\mathcal{S}}_l\ AN_V$. For all pairs $\mathcal{T}_i\ \dot{\mathcal{S}}_l\ \mathcal{T}_j$, it is expected that each item class in $\mathcal{T}_j$ has a prerequisite item class in $\mathcal{T}_i$ (see also Section 5.1.5, prediction Ib).

- The right–covering surmise relation contains three test pairs ($\mathcal{T}_i\ \dot{\mathcal{S}}_r\ \mathcal{T}_j$) from the matrix test to the remaining tests, viz. to the verbal analogy test ($AN_V\ \dot{\mathcal{S}}_r\ MT_G$), to the geometric analogy test ($AN_G\ \dot{\mathcal{S}}_r\ MT_G$), and to the series completion test ($SC_N\ \dot{\mathcal{S}}_r\ MT_G$). For all pairs $\mathcal{T}_i\ \dot{\mathcal{S}}_r\ \mathcal{T}_j$, it is expected that each item class in $\mathcal{T}_i$ is prerequisite for some item class in $\mathcal{T}_j$ (see also Section 5.1.5, prediction Ic). Note, that the surmise relation from the matrix test to the geometric analogy test is left– as well as right–covering and thus a total–covering surmise relation ($AN_G\ \dot{\mathcal{S}}_t\ MT_G$, prediction Id in Section 5.1.5).

- The remaining five test pairs are all element of the general surmise relation ($\mathcal{T}_i\ \dot{\mathcal{S}}\ \mathcal{T}_j$). Each test contains item classes that have a prerequisite in all other three tests. However, the conditions for a left– or right– covering surmise relation are not fulfilled. The general surmise relation includes three relationships from the geometric analogy test to the verbal analogy test ($AN_V\ \dot{\mathcal{S}}\ AN_G$), to the series completion test ($SC_N\ \dot{\mathcal{S}}\ AN_G$), and to the matrix test ($MT_G\ \dot{\mathcal{S}}\ AN_G$), as well as two relationships from the series completion test to the verbal analogy test ($AN_V\ \dot{\mathcal{S}}\ SC_N$) and to the matrix test ($MT_G\ \dot{\mathcal{S}}\ SC_N$). For all pairs $\mathcal{T}_i\ \dot{\mathcal{S}}\ \mathcal{T}_j$, it is expected that some item class in $\mathcal{T}_j$ has a prerequisite in $\mathcal{T}_i$ (see also Section 5.1.5, prediction Ia).

A detailed description of the item classes, their respective items, and the postulated prerequisite relationships is given in the next section.

5.4.2 Method

5.4.2.1 Participants

I collected data of 122 participants for Investigation III. The sample included 80 high–school students and 42 first year university students, who were enrolled in an introductory psychology course (Seminar on General Psychology I). High–school students stem from two Styrian schools, the “Bundesgymnasium Rein” (secondary academic school) and the “Höhere Bundeslehranstalt für wirtschaftliche Berufe Schrödingerstraße” (secondary vocational school) with 46 and 34 participants respectively. Ninety–nine participants are female and 23 participants are male. The participants’ average age was 19.18 years ranging between 17 and 42 years, with a standard deviation of 3.7. Thirty–five participants indicated to be familiar with this type of tests. The investigation was conducted in the class–rooms of the schools and a lecture room at the University of Graz in January 2001. All participants attended the study voluntarily.

The data of one male participant had to be excluded from further analyses, because his response pattern was incomplete (four unprocessed items). Thus, the final number of participants equals 121.

5.4.2.2 Material

Materials consist of 20 inductive reasoning problems, which can be grouped into four problem types (tests) with five items each. The four problem types include verbal analogies (AN_V), geometric analogies (AN_G), numerical series completion problems (SC_N), and geometric-figural matrices (MT_G). Items of the first three problem types are taken from the BIS test by Jäger et al. (1997, see Section 2.4.2.2), geometric matrices from the WMT by Formann and Piswanger (1979, see Section 2.4.2.4). Except for the analogy items, answer alternatives have been added (series completion items) or removed (matrix items) to achieve a common answer format of five answer alternatives per item. Table 5.19 gives an overview of the used items, including the original item numbers and abbreviations for the subtests as reference. Figure 5.13 shows an example for each problem type. Except for the verbal analogy item, the depicted items are self–constructed, but show are similar to the presented ones (copyrights). The verbal analogy item is taken from Bejar et al. (1991). The selection of items and answer alternatives was conducted under consideration of the required solution strategies, i. e. the attribute combinations inherent in each item. As far as possible only items with different attribute combinations were chosen. Exceptions are two verbal analogies and two matrix problems with the same attribute combinations each[11] (see Table 5.20). The exact assignments of the items and the modified answer alternatives used in Investigation III are presented in Appendix B.2, Table B.4.

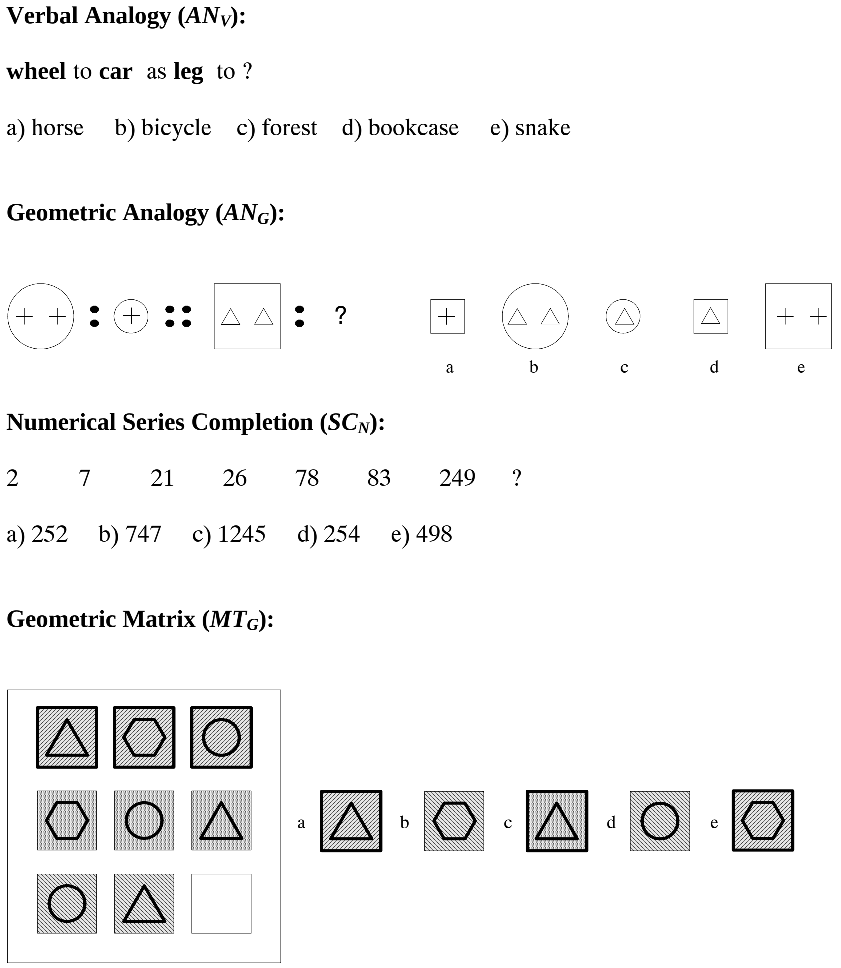
Table 5.19: Selected items and tests for Investigation III

Problem type	Test	Subtest	Instructions	Item numbers
Verbal analogies	BIS	WA	Exp.2	1,3,4,6,7
Geometric analogies	BIS	AN	Exp.1	1,3,4,6,7
Numerical series completion	BIS	ZN	Exp.1	1,4,5,7,9
Geometric matrices	WMT	–	Exp.A	C,2,9,22,23

Note. Subtest abbreviations, example (Exp.) and item numbers refer to the original tests (BIS and WMT).

Table 5.20 summarizes the item descriptions for each of the 20 items, including the attribute combinations (item classes), item numbers, operation types, and downsets.

[11] The verbal analogy test did not contain any items with other attribute combinations; for the geometric matrix test, deleting the component task ambiguity (see Section 7, classification scheme) lead to a reduction in the number of attribute combinations.



Verbal Analogy (AN_V):

wheel to **car** as **leg** to ?

a) horse b) bicycle c) forest d) bookcase e) snake

Geometric Analogy (AN_G):

Numerical Series Completion (SC_N):

2 7 21 26 78 83 249 ?

a) 252 b) 747 c) 1245 d) 254 e) 498

Geometric Matrix (MT_G):

Figure 5.13: Examples for each problem type presented in Investigation III

The five verbal analogy items (in German) are all of the type $A : B = C : ?$. Below the stem terms, five answer alternatives (labeled a through e) are presented. Participants had to choose the alternative that completed the analogy best. Stem terms as well as alternatives are nouns (3 items), verbs (1 item), or mixed with nouns for terms A and B and verbs for term C and the five answer alternatives (1 item). The types of operations (component A) needed to solve the problems include the semantic relations cause–purpose (CP), part–whole (PW), and the more **D**ifficult relation similar/comparative (SI). A description of the semantic relations is given in Section 2.2.1.1, Table 2.1. Three item classes require two operations, one item class three operations (component B). The number of operations refers to the number of elements in the analogies' rationales (see Section 2.2.1.1, Box 2.3). In Table 5.20 only the semantic relations between the terms are listed (fourth column), while the complete rationales are given in Appendix B.1, Table B.3. **H**igh demands on constraint (component C) are given in three item classes, **L**ow demands in one item class. Components D and E are identical for all item classes, viz. **V**erbal material and **5** answer alternatives. Figure 5.13 gives an example of a verbal analogy item for the item class (O,1,L,V,5). A detailed description is given in Section 5.2.2.2.

The geometric analogies are of the same type as the verbal analogies, except that stem terms as well as the answer alternatives are geometric figures, such as triangles, squares, or circles. On component *A* (operation difficulty) the items require transformations in shape (SH) and shading (SD) as well as the **D**ifficult operations in number (NO) and space (SP). The number of operations (component *B*) ranges between one and three per item. **H**igh demands on constraint (component *C*) are given in two item classes, **L**ow demands in three item classes. Components *D* and *E* are identical for all item classes, viz. **G**eometric material and **5** answer alternatives. The example in Figure 5.13 requires one transformation in size and one in number, which are both low in demand. Thus, the corresponding item class is (D,2,L,G,5) and the correct answer is 'd'.

The third test contains five numerical series completion problems with five (one item) and seven (four items) numbers per series. The last number is followed by a question mark, for which the participants had to choose one out of five alternatives (labeled a through e) to complete the series correctly. The arithmetic operations needed to solve the problems include additions (AD), subtractions (SU), multiplications (MU), and divisions (DI). The hierarchical sequences for the two **D**ifficult series completion problems (component *A*) are $n-1$ for item 14 and $n+1$ for item 15. The number of operations (component *B*) ranges between one and three per item. **H**igh demands on constraint (component *C*) are given in two item classes, **L**ow demands in three item classes. Components *D* and *E* are identical for all item classes, viz. **N**umerical material and **5** answer alternatives. The example in Figure 5.13 requires one addition (+5) and one multiplication (x3) with no hierarchical sequences and low demand. The respective item class is therefore (O,2,L,N,5) and the correct answer is 'd'.

Finally, the five geometric matrix items contained in the fourth test, are all constructed as 3 by 3 matrices. The items contain different types of geometric forms (squares, circles, lines, etc.) colored in black and a blank square on the lower right. Five answer alternatives (labeled a through e) are presented on the right of each matrix. Participants had to decide which alternative belongs in the blank square to complete the matrix correctly. The types of operations included in the five matrix problems are constant in a row (CR), constant in a column (CC), pairwise progression (PP), figure addition (FA), distribution of three values (D3), and as **D**ifficult operations (component *A*) the Boolean AND operator (BA) and the exclusive-OR rule (XO). A description of the rules is given in Section 2.2.3, Table 2.2. By analyzing the geometric forms row by row and/or column by column between one and four relational rules can be induced (component *B*). **H**igh demands on constraint (component *C*) are given in three item classes, **L**ow demands in one item class. Components *D* and *E* are identical for all item classes, viz. **G**eometric material and **5** answer alternatives. The example in Figure 5.13 requires the operations D3 (on the geometric form) and CR (on the background shading), which are both low in demand. Thus, the corresponding item class is (O,2,L,G,D) and the correct answer is 'b'.

Regarding the prerequisite relationships postulated in Section 5.1, Table 5.20 depicts the resulting item classes (2nd column) and their respective downsets (5th column). Component *E* (number of answer alternatives) is not listed, because it is identical (five alternatives) for all items and does therefore not contribute to the order on items and tests. The downsets for each item class only include the prerequisite item classes

Table 5.20: Item descriptions and downsets for the four tests in Investigation III

Problem types	Item classes[a]	Item numbers	Operation types[b]	Downsets for item classes (within tests)[c]	Downsets for items (within tests)
AN	(O,2,H,V)	1	CP	(O,2,H,V)	1
	(O,3,H,V)	4	PW	(O,2,H,V),(O,3,H,V)	1,4,5
	– ‖ –	5	PW		
	(D,2,L,V)	2	SI	(D,2,L,V)	2
	(D,2,H,V)	3	SI	(O,2,H,V),(D,2,L,V), (D,2,H,V)	1,2,3
AN	(O,1,L,G)	8	SH	(O,1,L,G)	8
	(D,1,H,G)	7	SP	(O,1,L,G),(D,1,H,G)	7,8
	(D,2,L,G)	6	NO,NO	(O,1,L,G),(D,2,L,G)	6,8
	(D,2,H,G)	9	NO,SD	(O,1,L,G),(D,1,H,G), (D,2,L,G),(D,2,H,G)	6,7,8,9
	(D,3,L,G)	10	NO,SP,SH	(O,1,L,G),(D,2,L,G), (D,3,L,G)	6,8,10
SC	(O,2,L,N)	11	SU	(O,2,L,N)	11
	(O,3,L,N)	13	DI,AD,MU	(O,2,L,N),(O,3,L,N)	11,13
	(O,3,H,N)	12	DI,SU,MU	(O,2,L,N),(O,3,L,N), (O,3,H,N)	11,12,13
	(D,2,L,N)	14	SU,MU	(O,2,L,N),(D,2,L,N)	11,14
	(D,2,H,N)	15	MU,AD	(O,2,L,N),(D,2,L,N), (D,2,H,N)	11,14,15
MT	(O,1,H,G)	16	FA	(O,1,H,G)	16
	(O,2,L,G)	17	CR,CC	(O,2,L,G)	17,18
	– ‖ –	18	PP,D3		
	(D,2,H,G)	19	XO,BA	(O,1,H,G),(O,2,L,G), (D,2,H,G)	16,17,18,19
	(D,4,H,G)	20	XO,XO,XO,XO	(O,1,H,G),(O,2,L,G), (D,2,H,G),(D,4,H,G)	16,17,18,19,20

Note. [a] Components are ordered alphabetically, i. e. A through D, component E is identical for all item classes, viz. 5 answer alternatives; [b] CP = cause–purpose, PW = part–whole, SI = similar/comparative, SH = shape, SP = space, NO = number, SD = shading, SU = subtraction, DI = division, AD = addition, MU = multiplication, FA = figure addition, CR = constant in a row, CC = constant in a column, PP = pairwise progression, D3 = distribution of three values, XO = exclusive-OR, BA = Boolean AND. [c] See Table 5.1 for the derivation of downsets.

of the same test. The prerequisite relationships across tests are given in Table 5.21. Table 5.20 shows that each item class has between one (which is due to the property of reflexivity) and four prerequisite item classes within its test. Regarding the three major components A, B, and C, the attribute combinations range from the easiest item class (O,1,L) in the classification scheme (see Table 5.1 and Figure 5.3) to the most difficult item class (D,4,H).

A combined analysis of the item classes of all four tests results in the prerequisite relationships across tests shown in Table 5.21. By analyzing the prerequisite relationships between each pair of tests, the hypotheses on surmise relations between tests listed in Section 5.4.1 have been inferred. Summarized, the general and the left–covering surmise relation contain five test pairs each and the right–covering surmise relation contains three pairs. Furthermore, there is one total–covering surmise relation (which is contained in the left– as well as right–covering relation) from the geometric matrix test to the verbal analogy test.

Figure 5.14 illustrates the postulated surmise relations within and between the four tests. For reasons of more clarity, the relationships across tests are not depicted in this Figure. However, they can easily be transferred from Table 5.21. The relation files for the $SRbI$, $SRxT$, and the four $SRwT$ are listed in Appendix E.1, the corresponding base files for the $KSbI$, $KSxT$, and the four $KSwT$ in Appendix E.2.

5.4.2.3 Procedure

The investigation was conducted in group sessions with 17 to 24 (high–schools) and 42 (university) participants. Students' teachers were present during instructions and testing periods. At the beginning of each session the general aim of the study and its theoretical background was outlined. Participants were told that the study is about the structure of knowledge and that it belongs to the field of cognitive psychology. As remote aim I mentioned the development of an adaptive testing system. I pointed out that the presented problems can be found in many intelligence tests and that they are sometimes used for the recruitment of personnel. Furthermore, I explained that the problems were analyzed and ordered with respect to their difficulty and that participants' data are important for the evaluation of our assumptions. Participant's were told, that the test is not about measuring their intelligence but about which problems can be solved. Finally, participants were requested to try to solve every single item and not to cheat.

Subsequently, test–papers were handed out and written instructions as well as example problems were reviewed with the participants. Instructions and examples were also presented on overhead–transparencies. Participants were instructed to process the items in the presented order (see below) and to try to solve every single item (the complete written instructions are provided in Appendix C.1).

After the instructions, participants specified a personal code and filled out questions concerning demographic data (type of school, level of education, age, sex). Then, a detailed explanation of one example for each problem type followed (see Appendix C.1). The main purpose was to familiarize participants with the relevant problem solving concepts. After all participants had signaled their understanding of the problems, they started with the test. Beforehand, they were once more reminded to process all of the problems and to follow the given order.

The presentation of items was randomized by four different sequences generated with

Table 5.21: Prerequisite relationships across the tests in Investigation III

		AN_V						AN_V						AN_G				
		(O,2,H)	(O,3,H)	(D,2,L)	(D,2,H)			(O,2,H)	(O,3,H)	(D,2,L)	(D,2,H)			(O,1,L)	(D,1,H)	(D,2,L)	(D,2,H)	(D,3,L)
AN_G	(O,1,L)	◇	◇	◇	◇	SC_N	(O,2,L)	◇	◇	◇	◇	MT_G	(O,1,H)	x	◇		◇	
	(D,1,H)				◇		(O,3,L)		◇				(O,2,L)	x		◇	◇	◇
	(D,2,L)			x	◇		(O,3,H)	x					(D,2,H)	x	x	x		
	(D,2,H)	x		x	x		(D,2,L)				◇		(D,4,H)	x	x	x	x	x
	(D,3,L)			x			(D,2,H)	x		x								
		MT_G						MT_G						SC_N				
		(O,1,H)	(O,2,L)	(D,2,H)	(D,4,H)			(O,1,H)	(O,2,L)	(D,2,H)	(D,4,H)			(O,2,L)	(O,3,L)	(O,3,H)	(D,2,L)	(D,2,H)
AN_V	(O,2,H)	x	x	◇	◇	SC_N	(O,2,L)		◇	◇	◇	AN_G	(O,1,L)	◇	◇	◇	◇	◇
	(O,3,H)	x	x		◇		(O,3,L)		x		◇		(D,1,H)					◇
	(D,2,L)		x	◇	◇		(O,3,H)	x	x		◇		(D,2,L)	x			x	◇
	(D,2,H)	x	x	◇	◇		(D,2,L)		x	◇	◇		(D,2,H)	x			x	x
							(D,2,H)	x	x	◇	◇		(D,3,L)	x	x		x	

Note. An x indicates that the item class in column i is prerequisite for the item class in row j (iSj); a ◇ indicates that the item class in row j is prerequisite for the item class in column i (jSi). AN_V = verbal analogies, AN_G = geometric analogies, SC_N = number series completions, and MT_G = geometric matrices.

the program `permute` (see Appendix C.2 for the generated sequences). The test–papers were handed out by alternating the four sequences.

After processing all of the problems, participants were prompted to check once more, whether they had answered all items. If yes, participants had to indicate their familiarity with inductive reasoning problems in general (yes or no) and to rate each problem type with respect to difficulty (very easy, rather easy, rather difficult, or very difficult).

The entire session lasted about 50 minutes, the time for solving the problems up to 40 minutes. All but one participant processed all problems in the given time limit. Participants who finished the test early were asked to occupy themselves quietly and to stay in class until the end of the session.

At the end of each session, participants received information on the type of feedback provided after the evaluation of the results. The evaluation followed between two and three weeks after the single sessions. The feedback included the relative solution frequencies for each participant, the quartiles for the respective groups (school or university classes) and for the entire sample of 122 particpants, as well as an explanation of the given results.

5.4.3 Results

As for Investigations I and II, the validation procedures are divided into two parts, namely the results derived via the surmise relation and the results derived via the knowledge space. Comparisons with simulated data sets and statistical analyses are also included. The data set for this investigation contains 2420 responses from 121 participants who processed all of the 20 items.

5.4.3.1 Validation of hypotheses via the surmise relation

Percentage of correct solutions All in all, the participants gave 1556 correct responses, i. e. 64.3% of all answers. The minimal number of correctly solved items amounts to four items (one response pattern), the maximal number to 18 items (6 response patterns). The frequencies for each single item are provided in Appendix F.1, Table F.3. Figure 5.14 shows the postulated surmise relation within and between tests together with the relative solution frequencies for each item class. The relative solution frequencies for the two item classes containing two items (see Table 5.20) represent the averaged relative solution frequencies of both items in the respective class. The difference in the percentage of correct solutions for the two items within one class each amounts to 20.66% for the analogy item class (O,3,H,V) and 12.4% for the matrix item class (O,2,L,G). The remaining item classes contain only one item each. The percentages of correct solutions for the item classes range between 16.53% for the matrix class (D,4,H,G) and 93.80% for the matrix class (O,2,L,G). For reasons of more clarity, the relationships across tests are not depicted in Figure 5.14, but can easily be inferred from Table 5.21.

The results for the four surmise relations within tests ($SRwAN_V$, $SRwAN_G$, $SRwSC_N$, $SRwMT_G$) show that the solution frequencies confirm all but one of the 21 postulated

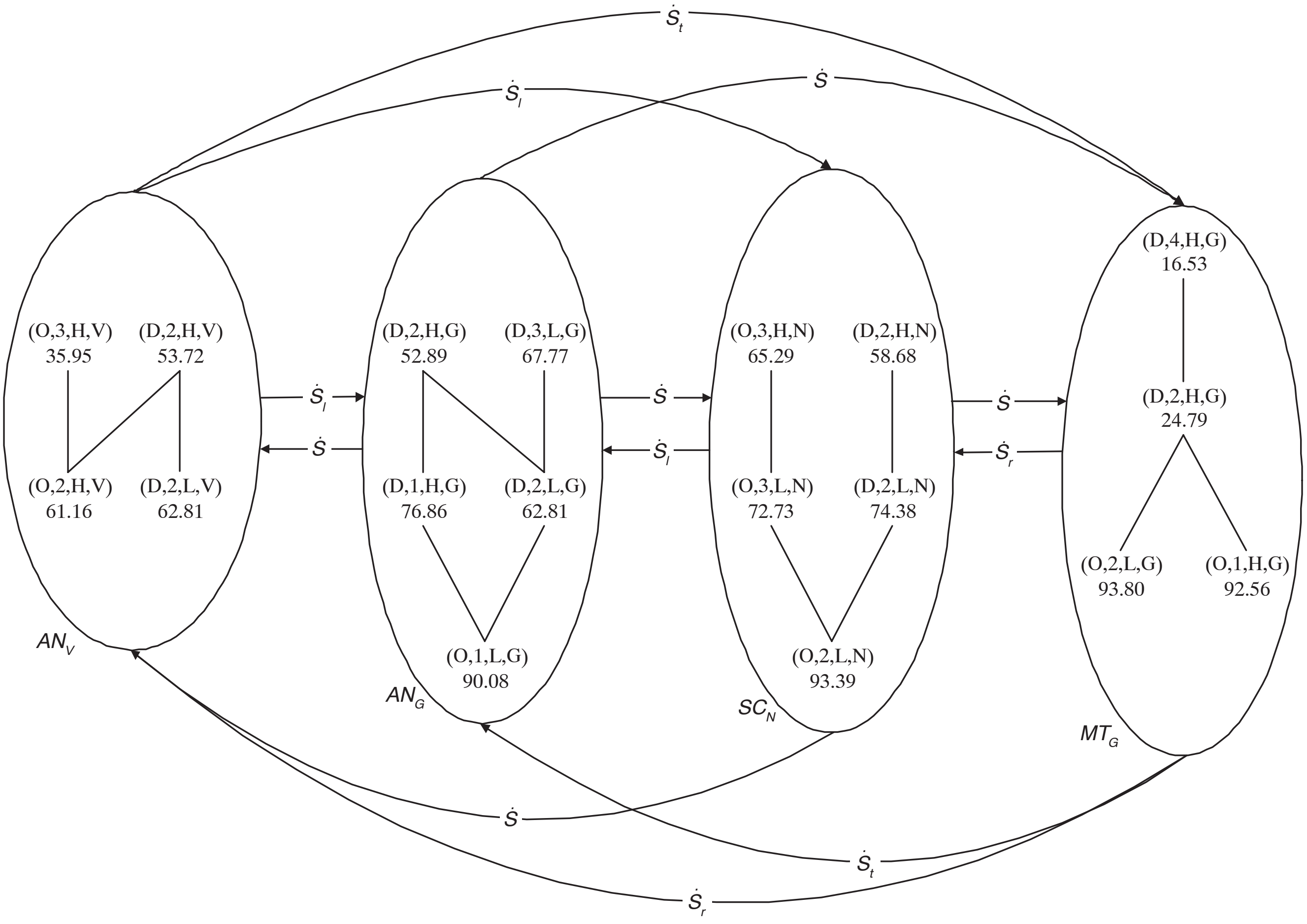


Figure 5.14: Hypothesis and relative solution frequencies for the surmise relation within and between tests in Investigation III

pairs of item classes (not counting reflexive pairs). Within the geometric analogy test ($SRwAN_G$; 2nd ellipse from the left), the percentage of correct solutions for item class (D,3,L,G) is higher than that of its postulated prerequisite (D,2,L,G), but the difference is statistically not significant ($\chi^2(1, N = 158) = 0.23$, p $<$ .05).

Regarding the surmise relation across tests ($SRxT$), 73 out of 79 pairs confirm the assumed relationships across the various tests (see also Table 5.21). A more differentiated inspection of the relation's subsets shows that for the surmise relation from the verbal analogy test (AN_V) to the other three tests all of the 19 pairs in the hypothesized model are confirmed, while for the geometric analogy tests (AN_G) 17 out of 18 pairs, for the numerical series completion test (SC_N) 14 out of 16 pairs, and for the geometric matrix test (MT_G) 23 out of 26 pairs are in accordance with the hypothesis. For all six of the reversed pairs, one–dimensional χ^2 tests yield no significant difference on an α–level of .05 ($.001 \leq \chi^2 \leq .23$; see Appendix F.2, Table F.6 for the exact values). With respect to the hypothesis that the percentages of prerequisite item classes are higher than or equal to the classes they are surmised from (prediction IIa in Section 5.1.5), there is no statistically significant contradiction of the postulated model.

The results for the postulated general, left–, right–, and total–covering surmise relations between tests (see Figure 5.14) show that 10 out of the 12 test pairs are confirmed by the relative solution frequencies (see predictions Ia – Id in Section 5.1.5). These 10 pairs of tests include three pairs in the left–covering surmise relation ($AN_G\ \dot{\mathcal{S}}_l\ AN_V, SC_N\ \dot{\mathcal{S}}_l\ AN_V$, and $MT_G\ \dot{\mathcal{S}}_l\ AN_V$), two pairs in the right–covering surmise relation ($AN_V\ \dot{\mathcal{S}}_r\ MT_G$ and $SC_N\ \dot{\mathcal{S}}_r\ MT_G$), and the five pairs in the general surmise relation ($AN_V\ \dot{\mathcal{S}}\ AN_G$, $SC_N\ \dot{\mathcal{S}}\ AN_G$, $MT_G\ \dot{\mathcal{S}}\ AN_G$, $AN_V\ \dot{\mathcal{S}}\ SC_N$, $MT_G\ \dot{\mathcal{S}}\ SC_N$). Deviations are found for the postulated left–covering surmise relation from the series completion test to the geometric analogy test and the total–covering surmise relation from the geometric matrix test to the geometric analogy test. In the first case ($AN_G\ \dot{\mathcal{S}}_l\ SC_N$), item class (O,2,L,N) does not have a prerequisite in the geometric analogy test. Regarding the total–covering surmise relation, each of the geometric analogy items is prerequisite for a geometric matrix item, hence, there is a right–covering surmise relation from the matrix tests to the analogy test ($AN_G\ \dot{\mathcal{S}}_r\ MT_G$). However, the left–covering surmise relation between the same pair of tests ($AN_G\ \dot{\mathcal{S}}_l\ MT_G$) is not confirmed, since the two matrix item classes (O,1,H,G) and (O,2,L,G) both have higher solution frequencies than their postulated prerequisite (O,1,L,G) in the analogy test. However, since the performed χ^2 tests (see above) show no significant differences for the reversed pairs of item classes, the postulated left– and total–covering surmise relations between the tests can also be accepted.

Summarized, the surmise relation between items is confirmed by 93% of the 100 postulated pairs, the surmise relation across tests by 92.4%, and the surmise relations within the four tests by 95.24%. Considering only the statistically significant differences, the *SRbI* and its subrelations are confirmed in 100% of all postulated pairs.

Indices for the fit of a surmise relation As for Investigations I and II, the fit of the postulated model on an item–level was validated by calculating the violational coefficient (VC) and the gamma–index (γ). Table 5.22 shows the results for the *SRbI* and its subsets, viz. the *SRxT* and the relations within the four single tests ($SRwAN_V$, $SRwAN_G$, $SRwSC_N$, and $SRwMT_G$). The values refer to relationships between single items.

The values of the indices VC and global γ_G both signify that the surmise relation on the geometric matrix test fits the set of data best and that the surmise relation on the verbal analogy test deviates most from the data. The percentages of item pairs with positive γ–indices amount to 85.71% for the $SRwAN_G$ and to 100% for the three remaining surmise relations within tests. This last result corresponds to the results derived from the solution frequencies. Thus, also on an item level, the postulated surmise relation on the set of tests shows a good fit to the empirical data.

The McNemar χ^2 values for the *SRbI* and its subrelations show that the number of concordant pairs is always significantly higher than the number of discordant pairs, which is in accordance with prediction IIb (Section 5.1.5).

5.4.3.2 Validation of hypotheses via the knowledge space

Symmetric distances and distance agreement coefficient The mean symmetric distances (*ddat*) for Investigation III range between 0.14 ($KSwMT_G$) and 0.47

Table 5.22: VC and γ for the surmise relation between items and its subsets (N = 121)

	No. of items	No. of pairs[a]	VC	γ_G	$\gamma > 0$		χ^2
$SRbI$	20	123	0.07	0.69	115	(93.50%)	3225.37
$SRxT$	20	99	0.07	0.70	92	(92.93%)	2593.28
$SRwAN_V$	5	4	0.13	0.40	4	(100%)	32.32
$SRwAN_G$	5	7	0.11	0.47	6	(85.71%)	72.86
$SRwSC_N$	5	6	0.06	0.63	6	(100%)	96.27
$SRwMT_G$	5	7	0.03	0.92	7	(100%)	497.33

Note. $SRbI$ = surmise relation between the items of the set of tests, $SRxT$ = surmise relation across the tests, $AN_V/AN_G/SC_N/MT_G$ = surmise relation within the verbal analogy / geometric analogy / series completion / matrix test. [a]Reflexive pairs are not counted.

($KSwAN_G$) for the knowledge spaces of the four single tests and reach a value of 2.31 for the knowledge space between items ($KSbI$; see Table 5.23). A comparison with the theoretical maxima ($dmax$ = 2 and 10) and the powersets' distances ($0.66 \leq dpot \leq 5.66$) show significant differences in favor of the hypothesis for all (sub)structures (see prediction IIIa in Section 5.1.5). χ^2 values range between 28.75 $(2, N = 242)$ and 1945.75 $(6, N = 242)$, p < .001 (see Appendix F.5 for the remaining values). Therefore, it can be concluded that the test knowledge space as well as its substructures fit the empirical data better than unstructured response vectors. The distance distributions for the $KSbI$, its substructures, and the corresponding powersets are provided in Appendix F.3, Tables F.11 and F.12. Regarding the proportion of careless errors and lucky guesses that contribute to the mean empirical distance, an analysis of the items' invalidity (see Appendix F.4, Table F.15) yielded about one and a half times as many careless errors than lucky guesses (on the average, the mean distance is composed of .070 careless errors and .045 lucky guesses per item and person)[12]. This means that the contradicting response patterns mainly arise from incorrect solutions to items that are assumed to be prerequisite for some other correctly solved item(s).

A comparison of the various structures' fit by means of the distance agreement coefficient (DA) yields results that are comparable to those of the indices VC and γ_G (see Section 5.4.3.1). The lowest (best) DA value results for the geometric matrix test (DA = 0.17), the highest value for the verbal analogy test (DA = 0.53).

Simulations In Table 5.24 the mean symmetric distances and the DA values for the empirical data ($ddat$) are compared to the same values for random ($dsim_r$) and probability ($dsim_p$) simulations. The values for both types of simulations are the averaged mean distances and standard deviations from 1000 data sets each. The number

[12] See footnote on page 104.

Table 5.23: Symmetric distances for the test knowledge space, its substructures, and their powersets (N = 121)

	m	$\|\mathcal{K}\|$	$dmax$	$ddat$ (SD)	Mdn	$dpot$ (SD)	Mdn	DA
$KSbI$	20	255	10	2.31 (1.41)	2	5.66 (1.54)	6	0.41
$KSxT$	20	484	10	2.03 (1.30)	2	5.15 (1.44)	5	0.39
$KSwAN_V$	5	14	2	0.35 (0.51)	0	0.66 (0.64)	1	0.53
$KSwAN_G$	5	9	2	0.47 (0.60)	0	0.94 (0.70)	1	0.50
$KSwSC_N$	5	10	2	0.25 (0.45)	0	0.86 (0.70)	1	0.29
$KSwMT_G$	5	10	2	0.14 (0.35)	0	0.81 (0.63)	1	0.17

Note. m denotes the number of items, $|\mathcal{K}|$ the number of knowledge states; $KSbI$ = knowledge space between the items of the set of tests, $KSxT$ = knowledge space across the tests, $KSwAN_V/KSwAN_G/KSwSC_N/KSwMT_G$ = knowledge space within the verbal analogy / geometric analogy / series completion / matrix test.

Table 5.24: Symmetric distances for the test knowledge space, its substructures, and simulated data sets (N = 121)

			random simulations			probability simulations		
	$ddat$	DA	$dsim_r$ (SD)	Mdn	DA	$dsim_p$ (SD)	Mdn	DA
$KSbI$	2.31	1.41	5.67 (1.54)	6	1	1.86 (1.26)	2	0.33
$KSxT$	2.03	1.30	5.14 (1.44)	5	1	1.78 (1.21)	2	0.35
$KSwAN_V$	0.35	0.51	0.66 (0.64)	1	1	0.22 (0.43)	0	0.33
$KSwAN_G$	0.47	0.60	0.94 (0.70)	1	1	0.33 (0.53)	0	0.35
$KSwSC_N$	0.25	0.45	0.87 (0.69)	1	1	0.31 (0.51)	0	0.36
$KSwMT_G$	0.14	0.35	0.81 (0.63)	1	1	0.34 (0.50)	0	0.42

Note. For the simulated data sets SD refers to the mean standard deviations; $KSbI$ = knowledge space between the items of the set of tests, $KSxT$ = knowledge space across the tests, $KSwAN_V/KSwAN_G/KSwSC_N/KSwMT_G$ = knowledge space within the verbal analogy / geometric analogy / series completion / matrix test.

of response patterns per data set corresponds to the number in the empirical data set, i. e. 121 patterns. For the probability simulation, the probability for lucky guesses corresponds to the number of answer alternatives, namely $\eta = 0.2$ for five alternatives. The probability for careless errors was varied in 10 steps with $0.05 < \beta \leq 0.15$. For each probability level 100 sets of data were simulated. The averaged distance distributions obtained from each of the 1000 simulated data sets are provided in Appendix F.3, Tables F.11 and F.12.

The results for random simulations ($dsim_r$ in Table 5.24) are comparable to those found for the powersets of the various structures. With $0.66 \leq dsim_r \leq 5.67$, the postulated $KSbI$ and its substructures fit the empirical data clearly better than random data sets.

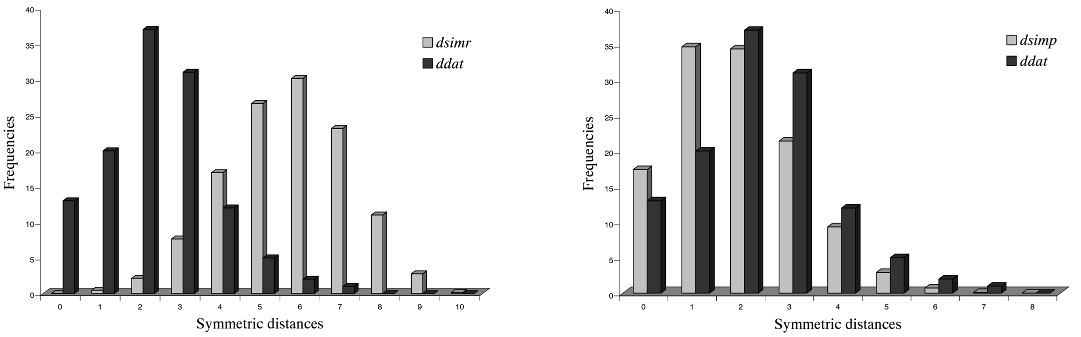


Figure 5.15: Distance distribution of the empirical data set (*ddat*) compared to the distributions of random ($dsim_r$) and probability simulations ($dsim_p$)

Regarding the results for the simulations on the postulated knowledge spaces ($dsim_p$ in Table 5.24), the mean symmetric distances as well as the DA values for the $KSbI$, the $KSxT$, and the knowledge spaces within the two analogy tests ($KSwAN_V$ and $KSwAN_G$) are below, i. e. better than those for the empirical data set. For the series completion ($KSwSC_N$) and the matrix ($KSwMT_G$) test, the respective knowledge spaces show a better fit to the empirical data than to the data derived from the probability simulations.

With respect to the $KSbI$, Figure 5.15 compares the distribution of the empirical distances to the averaged distance distributions derived from random and probability simulations. As already found in Investigations I and II, the distribution for the random data sets (left figure) is located further to the right on the distance scale than the empirical distribution and there is only a small overlap of the two distributions. The probability distribution (right figure), on the other hand, shows higher peaks on the left of the distance scale and a more positive skew than the empirical data. Thus, the number of response patterns with low distances is higher than for the empirical data. To test the significance of the found differences, the following statistical analyses have been performed.

Statistical analyses The standardization of the mean empirical distances to the distribution of random simulations ($dsim_r$) reveals significant differences for the $KSbI$ and its substructures. The obtained z–scores (see Table 5.25) range between -23.96 for the $KSbI$ and -5.2 for the $KSwAN_V$. Thus, the results of the random simulations support the hypothesis according to prediction IIIb (Section 5.1.5).

At an α–level of 0.01, the z–scores derived for the probability simulations ($dsim_p$) show a significant difference in favor of the hypothesis (prediction IIIc in Section 5.1.5) for the matrix test ($z = -4.09$), and a significant difference against the postulated knowledge space for the verbal analogy test ($z = 2.67$). The values for the remaining (sub)structures are not significant[13]. The same is true for the results of the χ^2 statistics. Regarding the central tendency of the distributions, the results of the U test show,

[13] For $\alpha = .05$, the data for the $KSwAN_G$ also shows a significant difference ($z = 2.15$), whereas the data for the remaining knowledge spaces differ still not significantly from the simulated data sets.

Table 5.25: Comparison of the empirical and simulated symmetric distances for the test knowledge space and its substructures (N = 121)

	random simulations			probability simulations				
	$dsim_r$	SD	z	$dsim_p$	SD	z	$U(z)$	χ^2 (df)
$SRbI$	5.67	0.14	-23.96	1.86	0.27	1.67	2.47	13.41 (5)
$SRxT$	5.14	0.13	-23.31	1.78	0.26	0.97	1.55	6.37 (4)
$SRwAN_V$	0.66	0.06	-5.20	0.22	0.05	2.67	2.17	11.34 (1)
$SRwAN_G$	0.94	0.06	-7.25	0.33	0.06	2.15	1.78	6.58 (1)
$SRwSC_N$	0.87	0.07	-9.42	0.31	0.06	-1.01	-19.10	1.02 (1)
$SRwMT_G$	0.81	0.06	-11.75	0.34	0.05	-4.09	-8.60	18.30 (1)

Note. SD refers to the distributions' standard deviations; SRbI = surmise relation between the items of the set of tests, SRxT = surmise relation across the two tests, $AN_V/AN_G/SC_N/MT_G$ = surmise relation within the verbal analogy / geometric analogy / series completion / matrix test; $U(z)$ denotes the standardized U–score for large samples with tied ranks.

that the empirical means for the series completion and the matrix test are significantly below the means for the simulated data sets. The $U(z)$ values for the tests knowledge space ($KSbI$), and the remaining substructures reveal no significant difference between the empirical and the simulated data sets ($1.55 \leq U(z) \leq 2.47, \alpha = 0.01$). A more differentiated analysis of the distributions obtained by the probability simulations for the $KSbI$ shows that the simulated data sets yield significantly lower mean distances ($z \geq 3.72$, $\alpha = 0.01$) at β levels $\leq .11$. For $.12 \leq \beta \leq .15$, the differences are not significant ($z \leq 2.54$, $\alpha = 0.01$), which indicates that the probability for careless errors amounts to about 12%.

The reported results indicate that, except for a substructure in the verbal analogy test, the postulated test knowledge space reliably explains the empirical data.

In order to estimate the influence of the predicted knowlege states, frequency simulations were computed. Table 5.26 depicts the the mean symmetric distances ($dsim_f$), the DA values, and the results of the statistical analyses for all but one substructures. For the verbal analogy test, frequency simulations were not computed, because the postulated model already showed significant deviations regarding the probability simulation. The results for all substructures yield no significant differences between the simulated and the empirical response patterns. Thus, it has to be concluded that the marginal frequencies of the matrix predict the solution behavior of the participants equally well as the postulated knowledge states.

5.4.4 Discussion

In this last investigation four types of inductive reasoning tests have been related according to the componentwise ordering principle. As already mentioned above, the

Table 5.26: Results of the frequency simulations for the test knowledge space and its substructures (N = 121)

	frequency simulations					
	$dsim_f$ (SD)	Mdn	DA	z	$U(z)$	χ^2 (df)
$KSbI$	2.29 (1.32)	2	0.41	0.39	0.05	2.86 (5)
$KSxT$	2.03 (1.23)	2	0.40	-0.05	-0.06	1.29 (4)
$KSwAN_G$	0.46 (1.59)	0	0.49	0.68	0.17	0.18 (1)
$KSwSC_N$	0.27 (1.49)	0	0.31	-1.25	-0.15	0.04 (1)
$KSwMT_G$	0.15 (0.36)	0	0.19	-0.82	-0.18	0.07 (1)

Note. $KSbI$ = knowledge space between items of both tests, $KSxT$ = knowledge space across the two tests, $KSwMT$ = knowledge space within the matrix test, $KSwAN$ = knowledge space within the analogy test; $U(z)$ denotes the standardized U–score for large samples with tied ranks.

results obtained in this study confirm the assumed prerequisite relationships among the set of tests. For the single tests, only the deviations within the verbal analogy test cannot be explained by the assumed noise variables in the data. Locating most contradictions of the postulated model in this test corresponds to the findings of Investigations I and II. This issue will be discussed in more detail in Chapter 7 (classification scheme).

As the postulated surmise relations and knowledge spaces between items, across tests, and within tests proved, for the main part, appropriate to explain the data set, I will concentrate on the discussion of the surmise relation between tests.

The presentation of four different types of inductive reasoning problems constitutes a first demonstration of the applied methods for a larger set of tests. Although the analysis of items by solely common components simplifies the description of item classes, the results show that a surmise relation between tests can be established by the componentwise ordering principle. The reported results (percentage of correct solutions in Section 5.4.3.1) clearly confirm 10 out of the 12 test pairs in the postulated subrelations between tests. The right–covering surmise relation is confirmed by both test pairs, the left–covering surmise relation by three test pairs, and the general surmise relation by all five of the contained test pairs. The left–covering surmise relation from the numerical series completion test to the geometric analogy test and the total–covering surmise relation from the geometric matrix test to the geometric analogy test contain pairs of item classes with reversed solution frequencies. However, the differences in the solution frequencies of each of the reversed pairs are not statistically significant. Thus, the hypothesized surmise relation between the set of all four tests can be accepted.

With respect to the confirmed relationships between the four tests, several inferences can be drawn. From $AN_V \ \dot{\mathcal{S}}_r \ MT_G$, $AN_G \ \dot{\mathcal{S}}_r \ MT_G$, and $SC_N \ \dot{\mathcal{S}}_r \ MT_G$ we can surmise that a person who solves all geometric matrix problems will also be able to solve all of the items in the remaining three tests. From the left–covering surmise relation, we can infer that a person won't solve any item in the verbal analogy test, if he or she doesn't solve any item in either one of the remaining three tests ($AN_G\ \dot{\mathcal{S}}_l\ AN_V$, $SC_N\ \dot{\mathcal{S}}_l\ AN_V$, and $MT_G\ \dot{\mathcal{S}}_l\ AN_V$). Similar conclusions can be drawn for the test pairs $AN_G\ \dot{\mathcal{S}}_l\ SC_N$ and $AN_G\ \dot{\mathcal{S}}_l\ MT_G$ (the left–covering part of $AN_G\ \dot{\mathcal{S}}_t\ MT_G$). A person, who doesn't solve any item in the geometric analogy test, will also fail in solving the items in the series completion and the matrix test. Regarding the general surmise relation, which contains the remaining test pairs, it can be concluded that at least some of the item classes are related, i. e. that the demands on the set of tests are not independent of each other.

Reconsidering the methodological problems brought up in Section 5.3.4 (results obtained via the surmise relation vs. the knowledge space), the results show the same trend as those for Investigation II. The results derived via the surmise relation (percentage of correct solutions, VC, and γ) are better fitting than those for Investigation I, while the values derived via the knowledge space (DA and simulations) are less fitting. A comparison of Investigations II and III, on the other hand, yields the same results for the methods derived via the surmise relation as for the knowledge space. Irrespective of the applied validation method, the model of Investigation III fits the empirical data always better than the model of Investigation II. The only exception is the verbal analogy test, for which the model in Investigation II yields better results with all of the applied validation methods. Since the data in both Investigation II and III show a wider range in solution frequency than the data in Investigation I, the very well fitting results derived via the knowledge space in Investigation I are obviously an artifact of the high percentage of correct solutions found in Investigation I (see also Sections 5.2.4 and 5.3.4 for a discussion of this issue). For the interpretation of results derived via the two different methods, it should therefore be considered that ceiling effects (and most likely also floor effects) have a positive influence on the results for the knowledge space but not for the surmise relation.

To answer the question, whether the results of the probability simulations in Investigation II are primarily due to deficiencies in the model or to a greater amount of noise in the data (see Section 5.3.4), the results for this last investigation have to be taken into account. The basic hypothesis on the relationships between various types of inductive reasoning problems is the same for all three investigations (see Section 5.1). Differences among the investigations occur only with respect to the item classes realized in each set of tests. Contrary to Investigations I and II, this study was conducted according to the standards of empirical knowledge space research. The response patterns in Investigation I stem from corporal and officer candidates of the Austrian military, the pattern in Investigation II from draftees of the German military, while the participants in Investigation III processed the test items under my supervision and were asked to give there best and to process all of the items. Thus, supposing that the validity of the model does not depend on the sampling, this investigation is most informative with respect to the model's validity. Hence, the assumption that the noise specified in the probability simulation of Investigation II was too low for the respective sample (draftees do not necessarily give their best), will also be maintained.

6 Adaptive Testing

As already mentioned several times (cf. Chapter 4), one of the main goals of my research is the development of test structures, which can be used as a basis for adaptive assessment procedures. Adaptive testing has a long tradition. In 1904 Binet began with the development of the first adaptive intelligence test for children, which started with questions that matched the children's age and stopped after a few subsequent questions could not be answered correctly. Today, adaptive tests are often computerized, which has the advantage that the selection of questions and the estimation of a person's ability level are computed automatically.

In spite of the long tradition of adaptive tests (see e. g., Wainer, 1990, for an introduction and historical overview), most psychometric aptitude and intelligence tests are constructed for a presentation in a standardized fashion, i. e. the testee has to answer all items contained in the given test. The advantages of standardized presentations are that the administration is easy (e. g., in a paper–pencil format), the requirements on technology are low, and the results represent an exact estimation of the testee's performance at that point in time. However, for large sets of items or evaluations of the testee's knowledge in various domains, a standardized presentation becomes very time–consuming. Furthermore, the item pool has to cover a broad range of difficulty, and therefore, many of the asked questions are often not suitable for the testee's level of aptitude. For all testees, there are usually questions that are either too easy or too difficult for them. Answering easy questions correctly and answering difficult questions incorrectly doesn't provide a lot of information about the testee's ability level.

Adaptive tests, on the other hand, tailor themselves to the testee's ability level by taking into account, whether the testee answered previous questions correctly or incorrectly. Thus, the sets of items presented to low–ability testees and to high–ability testees differ. The low–ability testee has to answer relatively easy questions, while the high–ability testee has to answer more difficult questions. The ability level of each person is iteratively estimated during the testing procedure and the selection of items is based on the preliminary ability estimate after each response. By this, the testee receives questions that maximize the information about his or her ability level. Items that are too easy or too difficult for the testee are not informative and are therefore not presented. The information of an item is higher, the more we can learn about the testee's ability level, i. e. the better the item discriminates among a set of plausible ability levels (Foster, 1998).

The main advantages of adaptive tests over the standardized presentation of items are that they are more efficient by reducing the number of presented items and that participants are usually more motivated, because the selected items are for the main part personally challenging without being too easy or too difficult. Moreover, the personalized selection of items leads to a higher degree of information gained from each given response (Foster, 1998; Wainer, 1990).

6.1 Approaches to adaptive knowledge assessment

Traditionally (Wainer, 1990; Wainer and Mislevy, 1990), adaptive tests are based on linear orders derived from probabilistic models, such as item response theory (IRT). The basic assumption in IRT is a single underlying trait or ability dimension (e. g. verbal or mathematical proficiency), on which the test items occupy different positions according to their difficulty. A testee's observable response to an item is related to the unobservable ability level of the testee. Thus, in the statistical framework of IRT, the probability to answer an item correctly is related to the characteristics of the item and the ability of the person. If the person's ability is much higher than the item's difficulty, the probability for a correct response will be large. Or, vice versa, if the difficulty of an item is much higher than the person's ability, the probability for a correct response will be small. In both cases only little information is gained from the response. The item characteristic curve depicts this relationship between item difficulty parameters and person ability parameters, and thereby specifies the proportion of information each item is contributing to the determination of the ability parameter, i. e. the latent variable. Maximal information on the person's ability is gained when the probability of a correct answer equals 0.5. The selection of items is based on the available estimates of the person's ability at each point in time. After each response, the ability of the testee is re–estimated and as next item that one is selected which provides most information with respect to the current ability estimate. Criteria for deciding when to stop the process are often based on the desired accuracy level of the ability estimation, measured by a precision indicator, such as the standard error (Fischer and Molenaar, 1995; Foster, 1998; Kubinger, 2000). Representative models of the IRT are, for example, the family of Rasch models or the two and three parameter models by Birnbaum (see e. g. Kubinger, 1992, for an overview).

The advantages of IRT models are that they allow a precise and efficient assessment of person abilities and that they have to fulfill the criterion of specific objectivity. Thus, the ability ratio for two persons is independent of the presented items and at the same time, the difficulty ratio for two items is always the same, irrespective of the sample taking the test (Kubinger, 2000). Adaptive testing procedures that are based on IRT require large sets of homogeneous items (which assess only one ability dimension) with known item characteristics. Since most of the available instruments do not fulfill these requirements, it is usually necessary to construct new item pools. Examples for adaptive Rasch homogeneous aptitude tests are the paper–pencil test AID ('Adaptives Intelligenz Diagnostikum') by Kubinger and Wurst (1985) or the two computerized adaptive tests BBT ('Begriffsbildungstest') by Kubinger, Fischer, and Schuhfried (1993) and AMT ('Adaptiver Matrizen Test') by Hornke, Rettig, and Etzel (1999; see also Hornke et al., 2000).

Still, the construction of unidimensional tests is often problematic. Even when applying strictly defined construction principles, a subset of the resulting items does often not fulfill the requirement of homogeneity (see e. g., Section 2.4.2.4 in which the construction of the item set for the WMT is outlined).

A more general approach is the use of non–linear structures. In connection with the theory of knowledge spaces (see Chapter 3), adaptive tests can be developed which are based on partial orders, and therefore overcome the restrictions of the traditionally used linear orders. Using a partial order, it is no longer necessary that the test items are unidimensional (i. e. that they assess only a single ability dimension). Several procedures for the deterministic (Degreef, Doignon, Ducamp, and Falmagne, 1986; Dowling and Hockemeyer, 2001; Hockemeyer, 2002) and non–deterministic (Doignon, 1994b; Falmagne and Doignon, 1988a,b) adaptive assessment of knowledge have been developed. Based on the set of knowledge states specified within a knowledge space, the testing procedure can adaptively be reduced to the presentation of a relatively small amount of problems. A central requirement for an accurate and not only efficient assessment procedure is the existence of a valid knowledge space. Otherwise, the algorithms will still reduce the number of posed questions, but the diagnosed knowledge states will not reflect the testee's true knowledge state. Considering item descriptions that are derived by a theory–driven generation of hypotheses (see Section 3.3.1), a further advantage of testing procedures within the knowledge space framework is that they provide detailed information on the problem demands a person is (un)able to fulfill.

The general principles of the IRT and the knowledge space based adaptive testing procedures are similar. Both approaches are based on a difficulty order (or prerequisite relation) on the set of items, which is either derived from data or from a theoretical analysis of the items. The selection of items within an assessment follows the principle of maximal information gain and participants' previous answers (correct or incorrect) are used for the selection of items. Differences between the traditional and the knowledge space approaches are, on the one hand, the type of underlying order (linear vs. partial) which results in different requirements on the dimensionality (uni– vs. multidimensional), and on the other hand, the diagnostic result (test score or ability parameter vs. knowledge state).

6.2 Adaptive assessment algorithms based on knowledge space theory

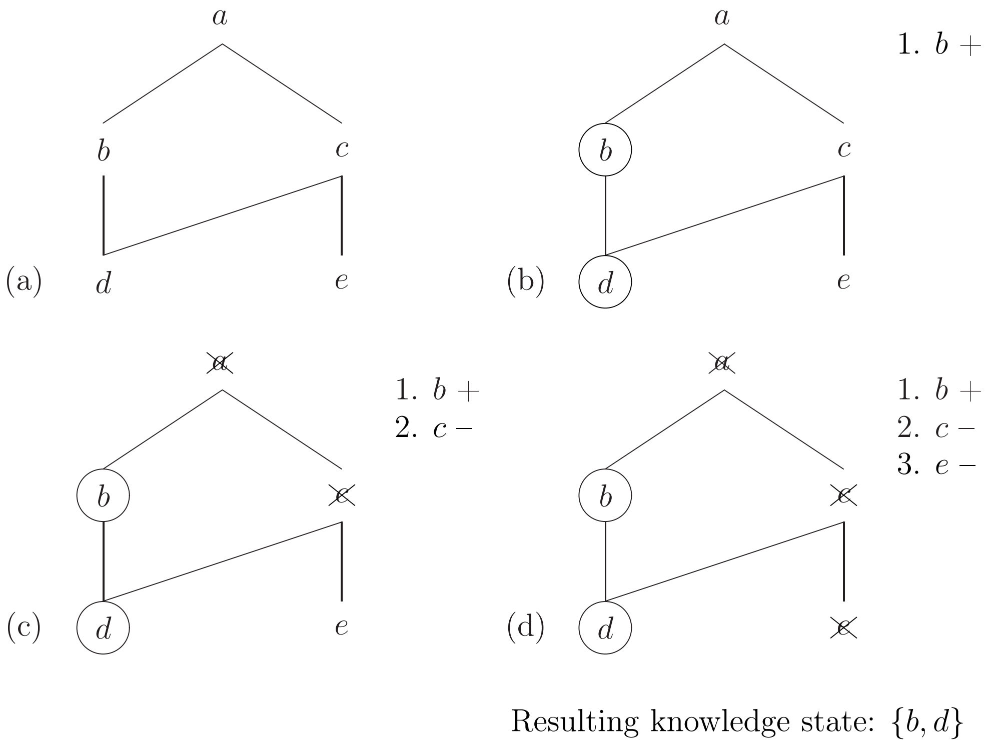
For my research, I applied two algorithms for the adaptive assessment of knowledge that are both based on knowledge space theory. The first algorithm is a deterministic procedure, which acts on the assumption that the testees' responses reflect their true knowledge states. The second algorithm is probabilistic and takes into account the probabilities for careless errors and lucky guesses. Both algorithms use the postulated knowledge states for the selection of items, i. e. the corresponding surmise relation and its properties (e. g. the partial order) are only considered indirectly.

6.2.1 Deterministic assessment algorithm

Based on the work by Degreef et al. (1986), Dowling and Hockemeyer (2001; Hockemeyer, 2002) developed an algorithm for the adaptive assessment of knowledge, which is based on prerequisite relationships (see Section 3.1). Assuming a valid knowledge structure, the assessment procedure selects each item according to the information obtained from previous answers. The first question is always the same for all persons (i. e. the algorithm does not use potential pre–information on the participants). Working with a binary decision tree, the algorithm chooses an item that, in the optimal case, is able to split the set of knowledge states in half (*half–split rule*). Depending on the testee's answer, the procedure will now either disregard items that are prerequisite for the first question (if the answer was correct) or disregard the items the first question is prerequisite for (if the answer was incorrect). Then the algorithm selects a question from the remaining items which will again split the set of remaining knowledge states in half. This procedure continues until the testee's knowledge state is determined.

Figure 6.1 gives an example for a set of five items. Figure 6.1a shows the surmise relation on the set of items, which corresponds to a quasi ordinal knowledge space with nine knowledge states. Item b occurs in four of the states and can therefore split the knowledge space in two parts of approximately the same size (four states including item b and five states not including item b). Figure 6.1b illustrates a correct response to item b (b +). From the given surmise relation, it can be inferred that the testee will also master item d and exclude the five knowledge states that do not contain items b and d. In the remaining four states item c occurs twice. Thus, a question testing the mastery of item c will again split the remaining set of states in half. Figure 6.1c illustrates an incorrect response to item c (c –) and the inferred incorrect response to item a. Finally, in Figure 6.1d an item testing the mastery of item e is presented. The response is incorrect (e –) and the testee's knowledge state is determined as $K = \{b, d\}$. The maximal number of five questions in this example is reduced by only two items, but for larger sets of items considerable reductions can be obtained.

Percevic and Wesiak (2001) applied the procedure to sets of data from over 4000 psychotherapeutic patients who answered various questionnaires on psychological stress (90 items), social functioning (64 items), and psychosomatic complains (24 items). Using a data–driven method for the generation of hypotheses (item tree analysis by van Leeuwe, 1974, see Section 3.3), we derived knowledge spaces with 21,323 states, 659 states, and 76 states respectively. The application of the described adaptive assessment algorithm to the data reduced the number of questions from the 90, 64, and 24 questions to an average of 14.15 ($SD = 0.6$), 9.28 ($SD = 0.63$), and 6.17 ($SD = 0.44$) questions respectively. The mean differences between the empirical and the estimated answer patterns mount up to 8.3 ($SD = 9.8$), 6.36 ($SD = 5.18$), and 3.25 ($SD = 2.86$) items. Thus, the adaptive strategy leads to large reductions in the number of posed questions. The mentioned differences between the empirical and the estimated answer patterns can be attributed to noise variables in the data and invalidities in the hypotheses, which often occur when the generation process allows a certain amount of noise in the data. The latter means, that the data–driven method accepts all prerequisite relationships between items, which are not contradicted by more than a specified percentage of empirical responses (in the described study this level was set to 15% of



$$Q = \{a, b, c, d, e\}, \mathcal{K} = \{\emptyset, \{d\}, \{e\}, \{d, e\}, \{b, d\}, \{b, d, e\}, \{c, d, e\}, \{b, c, d, e\}, Q\}$$

Figure 6.1: Illustration of the adaptive assessment procedure by Dowling and Hockemeyer (2001; adapted from Dowling et al., 1996)

allowed contradictions).

In general, it needs to be noted that deterministic assessment procedures assume that a person's responses are completely determined by that person's knowledge, i. e. they do not account for noise variables, such as careless errors or lucky guesses. The advantage of deterministic procedures is that new information is gained with each question, which leads to a high efficiency of these procedures. On the other hand, a careless error or a lucky guess might lead to further invalid inferences and therefore, deterministic procedures are often less accurate than non–deterministic ones.

6.2.2 Non–deterministic assessment algorithms

There are two main non–deterministic approaches to the adaptive assessment of knowledge, both developed by Falmagne and Doignon (1988a,b; Doignon, 1994b) and based on the deterministic model by Degreef et al. (1986). As opposed to the deterministic algorithm, the procedures do not assume that each answer reflects the true knowledge state of the testee. One (discrete) algorithm determines a testee's knowledge state by asking a few additional questions towards the end of the assessment procedure, while the other (continous) algorithm uses a probability distribution on the knowledge space.

The discrete procedure (Doignon, 1994b; Falmagne and Doignon, 1988b) starts with the deterministic algorithm described above (Section 6.2.1) and assumes that the resulting knowledge state is close to the testee's true knowledge state. This means, that potential errors in the diagnosis should only occur with respect to the neighboring knowledge states, i. e. those states which differ by exactly one item (Falmagne et al., 1990). Thus, the algorithm presents some additional items (mostly questions that have already been asked), which distinguish the deterministically diagnosed state from its neighbors and, if necessary, updates the assessment. As stopping criterion, a number of loops is specified, which indicates for how many questions the assessed state should remain unchanged before a final diagnosis is reached.

The second procedure applies a continuous algorithm (Falmagne and Doignon, 1988a), which works with a probability distribution on the knowledge space. This means, that for each state in the knowledge space, the algorithm estimates the probability that the respective state is the testee's true knowledge state. At the beginning of the assessment procedure each knowledge state has the same probability, viz. $\frac{1}{|K|}$, with $|K|$ denoting the number of knowledge states. After each response, the algorithm updates the probability distribution on the knowledge space and the probabilities to answer an item correctly. The probability to answer an item q correctly corresponds to the probability that the testee is in one of the states containing item q and the noise probabilities for careless errors and lucky guesses. The noise probabilities have to be specified for each item. The selection of items is based on the half–split rule described in Section 6.2.1. Thereby, the procedure selects those items, which have a probability of about 0.5 to be solved correctly. As mentioned above (Section 6.1), items with a solution probability of 0.5 provide most information on the testee's ability level. The assessment procedure stops, whenever a specified probability estimate of the testee's knowledge state is reached, i. e. the probability mass on a single state has to exceed a given threshold. The specification of this threshold depends on the particular needs of the assessor. Higher thresholds usually lead to more accurate but less efficient assessments.

Hockemeyer (2002) compared the discrete and the continuous procedure with respect to the accuracy and the efficiency of the algorithms. Therefore, he simulated data sets (1000 patterns each) with varying probabilities for lucky guesses and careless errors (0%, 5%, and 10%) that are based on three knowledge spaces of varying size ($\mathcal{K}$ = 2261, 14,569, and 41,395 states). The item set contained 28 items from the domain 'usage of AutoCAD', which were structured by means of expert queries (see Section 3.3). The stopping criteria for the discrete procedure were specified with one, three, and five loops, the criteria for the continuous procedure with probability estimates of 51%, 70%, and 90%. With regard to accuracy (measured in assessment errors, i. e. the number of wrongly assessed items per answer pattern), the simulation study showed that for both procedures accuracy increases with lower noise rates and stricter stopping criteria. Furthermore, the continuous procedure yielded more accurate results than the discrete procedure, especially for higher noise rates (0.90 vs. 1.00 wrongly assessed item at 10% lucky guesses and careless errors) and stricter stopping criteria (0.17 vs. 0.40 wrongly assessed items at probability estimates of 90% or 5 loops). The efficiency (measured in the number of posed questions) of both procedures decreased with an increasing number of knowledge states and with stricter stopping criteria. Furthermore, Hockemeyer found that the continuous algorithm was always more efficient than the discrete procedure, irrespective of the knowledge spaces' sizes and the applied stopping criteria.

6.3 Adaptive assessment for Investigations I–III

After validating the three sets of tests presented in Investigations I through III, it is possible to apply the obtained knowledge spaces for a first evaluation of the models in an adaptive testing system. Since the adaptive algorithms are based on item as opposed to test structures, the knowledge spaces between items ($KSbI$) have been used as underlying models. The postulated test knowledge spaces are therefore only considered indirectly via the subset of pairs contained in the knowledge space across tests. Considering the deterministic approach for establishing the hypothesis on the structure of inductive reasoning problems, I first selected the deterministic procedure for the adaptive assessment of knowledge developed by Dowling and Hockemeyer (2001; Hockemeyer, 2002; see Section 6.2.1). However, taking into account that the empirical data are noisy, I also applied a probabilistic procedure, namely the continuous assessment algorithm by Falmagne and Doignon (1988a), which proved to be more accurate and more efficient than the discrete algorithm in earlier investigations (see Section 6.2.2).

6.3.1 Adaptive testing procedure

For each of the three test knowledge spaces, the corresponding answer patterns have been used to simulate persons who take an adaptive version of the respective tests presented in the three investigations. For example, imagine an existing response vector $\langle 0, 0, 1, 1, 1 \rangle$ for the items a, b, c, d, and e and the knowledge space used for the illustration of the deterministic assessment algorithm (see Figure 6.1 for the corresponding surmise relation).

Both algorithms start with item b and select the corresponding response from the given vector which is an incorrect answer ('0'). Based on this response, the deterministic algorithm infers the incorrect answer to item a and eliminates all knowledge states containing items a and b. The non–deterministic procedure considers the possibility of a careless error and will therefore not eliminate all knowledge states containing item b (and a) right away. It just decreases the probability that the testee's knowledge state is one of the states containing items a and b and at the same time increases the likelihood of the states not containing items a and b. Then, the next items are presented and the corresponding answers are selected from the response vector until the final knowledge state can be determined.

For the non–deterministic procedure, it is necessary to specify the probabilities for careless errors and lucky guesses, as well as the stopping criterion. The guessing probabilities correspond to the number of answer alternatives in each of the three investigations (with 8 and 5 alternatives, $\eta = .16$ for Investigations I and II, and $\eta = .2$ for Investigation III with 5 alternatives). Since the probabilities for careless errors are not known, I simulated several assessments with $\beta = 5\%$, 10%, or 15%. For the stopping criterion (i. e. the probability that the assessed knowledge state reflects the true knowledge state of the participant), the threshold (th) was set to $th = .7$ for a moderate criterion and to $th = .9$ for a strict criterion.

Discrepancies between the empirical and the estimated response patterns occur, whenever the distance of an empirical pattern to the nearest knowledge state is greater than zero. For example, the response vector $\langle 0, 0, 1, 0, 1\rangle$ contains either a lucky guess for item c or a careless error for item d (regarding the surmise relation in Figure 6.1). Since the algorithms' assessments are based on the postulated knowledge space, the best estimates are the knowledge states $\{e\}$ or $\{c, d, e\}$ with the corresponding response vectors $\langle 0, 0, 0, 0, 1\rangle$ and $\langle 0, 0, 1, 1, 1\rangle$ respectively. More generally said, the minimal symmetric distance between an empirical response pattern and the knowledge space will always represent the smallest possible distance between the empirical and the estimated pattern.

Thus, considering the deviations of the empirical data sets from the postulated models (see Section 5), the accuracy of the assessment procedure has to be evaluated by simultaneously considering the results of the models' empirical validation (see Sections 5.2.3.2, 5.3.3.2, and 5.4.3.2). Additionally, with larger knowledge spaces, as they are given in Investigations I through III, this distance can grow with the number of inferences drawn by the algorithms.

6.3.2 Adaptive testing results and discussion

In order to estimate the algorithms' accuracy, I computed for each of the three $KSbI$ the mean symmetric distances (di) between the estimated and the original response vectors. The minimal possible distance di between the two response vectors equals the distance between the empirical response patterns and the postulated knowledge spaces ($ddat$), i. e. di is composed of $ddat$ and the discrepencies arising from the adaptive assessment. In order to determine how much the adaptive assessments contribute to the distance between the empirical and the estimated response vectors, I determined the difference between the two averaged symmetric distances [(diff(di, $ddat$)]. With regard to the algorithms' efficiency, I calculated the average number of questions posed and the percentage of saved questions. Table 6.1 shows the obtained results for the three investigations.

Accuracy of the algorithms With respect to the accuracy of the two algorithms, the distances di (2nd column in Table 6.1) between the empirical and the estimated response patterns yield the best results for Investigation III ($2.61 \leq di \leq 2.64$), followed by Investigations I ($3.70 \leq di \leq 3.87$) and II ($6.68 \leq di \leq 6.83$). This order corresponds to the number of items presented in each of the Investigations (30, 40, and 20 items in Investigations I, II, and III respectively) and to the obtained distances $ddat$ between the empirical patterns and the postulated knowledge spaces ($ddat = 2.99$, 6.08, and 2.31 for Investigations I through III). As mentioned above, the smallest possible value for di corresponds to $ddat$. Thus the obtained order of accuracy is simply caused by the differences in the fit of the three knowledge spaces (see also Sections 5.2.3.2,

Table 6.1: Results for the adaptive testing procedures

Investigation I (N = 572, 2 tests, 30 items)
$|KSbI| = 293$, $ddat$ (SD) = 2.99 (2.17), prob for η =.16

Algorithm	di (SD)	diff($di, ddat$)	NQ (SD)	% saved
deterministic	3.85 (3.04)	0.86	8.33 (0.69)	72.23
prob ($\beta = .05, th = .7$)	3.84 (3.04)	0.85	8.42 (0.66)	71.93
prob ($\beta = .10, th = .7$)	3.84 (2.99)	0.85	8.48 (0.61)	71.73
prob ($\beta = .15, th = .7$)	3.72 (2.99)	0.73	9.17 (0.81)	69.43
prob ($\beta = .05, th = .9$)	3.87 (3.05)	0.88	8.79 (0.77)	70.70
prob ($\beta = .10, th = .9$)	3.83 (3.02)	0.84	9.40 (1.25)	68.67
prob ($\beta = .15, th = .9$)	3.70 (2.92)	0.71	10.22 (1.20)	65.93

Investigation II (N = 2628, 2 tests, 40 items)
$|KSbI| = 7633$, $ddat$ (SD) = 6.08 (2.26), prob for η =.16

Algorithm	di (SD)	diff($di, ddat$)	NQ (SD)	% saved
deterministic	6.68 (2.62)	0.60	12.96 (0.90)	67.60
prob ($\beta = .05, th = .7$)	6.83 (2.91)	0.75	13.37 (1.21)	66.58
prob ($\beta = .10, th = .7$)	6.83 (2.93)	0.75	15.82 (1.98)	60.45
prob ($\beta = .15, th = .7$)	6.83 (2.92)	0.75	17.45 (2.34)	56.38
prob ($\beta = .05, th = .9$)	6.83 (2.93)	0.75	17.26 (2.35)	56.85
prob ($\beta = .10, th = .9$)	6.82 (2.92)	0.74	18.68 (2.14)	53.30
prob ($\beta = .15, th = .9$)	6.83 (2.92)	0.75	20.81 (3.88)	47.98

Investigation III (N = 121, 4 tests, 20 items)
$|KSbI| = 255$, $ddat$ (SD) = 2.31 (1.41), prob for η =.2

Algorithm	di (SD)	diff($di, ddat$)	NQ (SD)	% saved
deterministic	2.62 (1.57)	0.31	8.01 (0.65)	59.95
prob ($\beta = .05, th = .7$)	2.64 (1.56)	0.33	8.15 (0.65)	59.25
prob ($\beta = .10, th = .7$)	2.61 (1.53)	0.30	8.57 (1.05)	57.15
prob ($\beta = .15, th = .7$)	2.63 (1.57)	0.32	9.42 (1.31)	52.90
prob ($\beta = .05, th = .9$)	2.64 (1.56)	0.33	9.27 (1.38)	53.65
prob ($\beta = .10, th = .9$)	2.62 (1.55)	0.31	10.88 (1.50)	45.60
prob ($\beta = .15, th = .9$)	2.63 (1.57)	0.32	11.21 (1.67)	43.95

Note. $KSbI$ = knowledge space between items, $ddat$ = distance between the empirical patterns and the $KSbI$, β/η = error/guessing probabilities, th = threshold for stopping the assessment, di = distance between the empirical and estimated response patterns, diff($di, ddat$) = difference between di and $ddat$, NQ = number of posed questions, % saved = percentage of saved questions.

5.3.3.2, and 5.4.3.2). In order to estimate the assessment errors of the algorithms, the differences between *di* and *ddat* have to be taken into account (3rd column in Table 6.1). Both the deterministic and the non–deterministic procedure provide the most accurate estimates for Investigation III ($.30 \leq \text{diff}(di, ddat) \leq .33$) and the least accurate estimates for Investigation I ($.71 \leq \text{diff}(di, ddat) \leq .88$). This result should be viewed in consideration of the structure defined on each of the three sets of items. The surmise relations between items (*SRbI*) in Investigations I, II, and III contain 42.33% (381 out of 900), 36.44% (583 out of 1600), and 35.75% (143 out of 400) of all possible pairs respectively . Thus, the order of accuracy corresponds to the percentage of pairs contained in each of the relations. More generally, with an increasing number of pairs in the surmise relation (or, correspondingly, with a decreasing number of knowledge states) the number of assessment errors increases. This result is plausible, since each additional pair in a surmise relation leads to an increase in the number of possible inferences.

Deterministic versus non–deterministic accuracy results A comparison of the deterministic and the non–deterministic algorithm with respect to accuracy yields ambiguous results[1]. The deterministic procedure is more accurate for Investigation II, irrespective of the assumed noise probabilities and the chosen stopping criterion in the non–deterministic procedure ($\text{diff}(di, ddat) = .60$ for the deterministic and $\text{diff}(di, ddat) \geq .74$ for the non–deterministic procedure, see Table 6.1, third column). For Investigation III the deterministic procedure is equally or more accurate than the non–deterministic procedure for all but one probability specification. With probabilities of $\eta = .2$, $\beta = .10$, and $th = .7$ the non–deterministic procedure shows slightly better results ($\text{diff}(di, ddat) = .30$ versus .31 for the deterministic procedure). Thus, regarding Investigations II and III, the deterministic procedure should be preferred over the non–deterministic one[2]. However, for Investigation I the non–deterministic procedure proved to be more accurate in five out of the six probability specifications. Compared to a $\text{diff}(di, ddat)$ value of .86 for the deterministic procedure, the five more accurate non–deterministic assessments yielded values between .71 and .85. Thus, for Investigation I, the non–deterministic procedure should be preferred over the deterministic one, at least with respect to accuracy.

As one reason for the found ambiguity, C. Hockemeyer (personal communication, February 25, 2003) suggests that the distribution of lucky guesses and careless errors in the empirical data sets influence the accuracy of the non–deterministic algorithm's estimates. Regarding the symmetric distances between the data sets and the respective knowledge spaces in the three investigations, the found proportions of careless errors to lucky guesses (as far as they are contributing to the obtained distances) are as follows: In Investigation I the relative frequency of careless errors per item and person amounts to about five times the frequency of lucky guesses ($\beta = .084$, $\eta = .016$). In Investiga-

[1] Because of the ambiguity of the results derived by the continous procedure, the discrete non–deterministic algorithm was also tested. The results are similar to those found by Hockemeyer (2002), namely that the discrete procedure is less accurate and less efficient than the continous procedure, irrespective of the specified stopping criteria and noise level.

[2] Even with a specification of higher error probabilities, the accuracy of the non–deterministic procedure does not improve, whereas the number of questions steadily increases.

tion II the relative frequency of careless errors equals that of lucky guesses ($\beta = .076$, $\eta = .075$) and in Investigation III there are one and a half times as many careless errors than lucky guesses ($\beta = .070$, $\eta = .045$)[3]. Thus, careless errors and lucky guesses are distributed differently in the three investigation, whereas the specified values for the non–deterministic algorithm are identical for all three investigations (except for η in Investigation III). The influence of the high number of careless errors in Investigation I is reflected by the rapid increase of the non–deterministic algorithm's accuracy with increasing β–values. For the other two investigations, where the differences between the two noise variables are much smaller, the larger error probabilities do not enhance the accuracy of the estimates. Nevertheless, probability specifications on the basis of the values found for the distances between the empirical answer patterns and the knowledge spaces (see above) yield the same results as the assumed probabilities (with thresholds of $th = .7/.9$, $\text{diff}(di, ddat) = .89/.85$, $.75/.74$, and $.32/.32$ for Investigations I, II, and III respectively). The results concerning the better performance of the deterministic algorithm are still not resolved but under discussion.

Efficiency of the algorithms Looking at the efficiency of the two algorithms, the deterministic procedure is always the more efficient one, meaning that it asked fewer questions (4th column in Table 6.1) and therefore lead to greater savings in the number of necessary questions (last column in Table 6.1). This result was expected, since the deterministic procedure does not account for noise in the data, but interpretes each given response as the true knowledge of the testee. With the non–deterministic assessment the number of posed questions continuously increases (and the savings decrease) with higher probability specifications for the noise variables or the stopping criterion. Overall, the savings range from 65.93% to 72.23% for Investigation I, from 47.98% to 67.6% for Investigation II, and from 43.95% to 59.95% for Investigation III. The order of the savings corresponds to the percentage of possible pairs in the postulated relations (see above, accuracy of the algorithms). Overall, the application of both algorithms leads to substantial reductions in the number of presented items.

One noteable point is the unexpected result that the specification of a stricter stopping criterion ($th = .9$) does not necessarily lead to more accurate assessments, although the number of questions increases. For example, in Investigation I the probability specifications $\eta = .16$, $\beta = .05$, and $th = .7$ lead to an assessment with an average of 8.42 questions and a resulting distance di of 3.84 ($\text{diff}(di, ddat) = .85$). Keeping the error and guessing probabilities, but using the stricter stopping criterion of $th = .9$ leads to an assessment with 8.79 questions and a resulting distance di of 3.87 ($\text{diff}(di, ddat)$ $= .88$). Thus, with 1.23% less savings the accuracy of the algorithm decreases by .03. Although this difference is very small, one would expect that a higher number of questions should at least result in equally accurate assessments. For Investigations II and III, the higher threshold yields almost always the same accuracy but less savings.

Summarizing, the results of the adaptive assessments yield a slight preference for the deterministic procedure, especially with regard to the trade–off between accuracy and efficiency. With an always lower amount of presented items, the deterministic algorithm yields results that are either more accurate (Investigations II and III) or only slightly

[3] See footnote on page 104.

below (Investigations I) the accuracy level reached with the non–deterministic procedure. The overall estimation errors, by which the assessment algorithms contribute to the distances between the empirical and the estimated response patterns ($.30 \leq \text{diff}(di, ddat) \leq .88$), amount to less than one wronlgy assessed item per response pattern). With savings up to 72.23% of all possible questions, also the efficiency of the algorithms could be demonstrated.

7 General Discussion

The purpose of this study was to develop a model for inductive reasoning tests, which integrates various problem types and can be implemented into an adaptive testing system. In order to reach these goals, the theory of knowledge spaces (see Chapter 3) was selected as the methodological framework throughout this study. The theory proved to be suitable for several reasons.

First of all, by interpreting relationships among problems or tests as surmise or prerequisite relations, a set of problems can be structured by means of a partial order, i. e. independencies between certain items are allowed. The use of partial orders becomes especially important when problems from different types of tests are integrated into a common problem structure. The advantage of prerequisite relations is that the correct or incorrect solution to a given set of problems can be inferred from previously obtained answers. Most of the previous research on the establishment of surmise relations covered exactly one domain of information or one test. In this study, problem types that are usually presented in different tests or subtests have been related. Thus, the approach of surmise relations between tests (*SRbT*) was applied (see Section 3.2). This generalization of a surmise relation between items permits assumptions about the relationships among sets of items, i. e. tests, and yields therefore more general predictions. This means that it is possible to draw inferences from the solution behavior in one test to the solution behavior in on or more other tests.

In all three of the reported investigations, the derived *SRbT* allowed predictions from the obtained response patterns for one of the tests to the response patterns for the other test(s). Regarding the results for the postulated *SRbT* in Investigations I through III, the derived solution frequencies confirmed all of the postulated relationships between tests (χ^2 statistics showed that the differences for the respective reversed solution frequencies were statistically not significant). Reconsidering the first scientific problem (i) presented in Chapter 4, the obtained hypotheses and results show that it is possible to define meaningful relationships between sets of tests and that the derived models accurately predict participants' solution behavior across tests. Thus, it can be concluded that the establishment of surmise relations between tests is a suitable approach to structure the domain of inductive reasoning.

A second reason for the selection of the applied methodology is that the knowledge space theory provides approaches for the generation of theoretically founded item and test structures (see Section 3.3). In order to determine not only the number of items that are solved by the participants, but also which problem demands are met, the componentwise ordering principle was chosen for the establishment of the test knowledge spaces (see Section 3.3.1). The specification of common components and their attributes followed the psychological findings outlined in Chapter 2 (esp. Section 2.2) and Section 3.5. By forming the Cartesian product of the components and by defining difficulty orders on the attributes of each component, a surmise relation between items and tests was derived for the domain of inductive reasoning (see Section 5.1). The thereby established hypothesis incorporates all possible attribute combinations (item classes) and thus, represents the core of the three investigations. Demand analyses of the presented problems show that each item can be assigned to exactly one of the postulated item classes. Thus, for the first part of the scientific question (ii) presented in Chapter 4, it can be concluded that various types of inductive reasoning problems can be described by a set of common components and attributes. The derived test knowledge spaces provide a set of empirically testable hypotheses and allow precise predictions of individual response patterns. For the validation of the postulated surmise relations and knowledge spaces, the knowledge space theory provides several methods for each approach (see Section 3.4). The second part of question (ii) inquires, whether the specified components, their attributes, and the postulated order on the components and attributes are a valid representation of participants' solution behavior. To answer this question, the pairs contained in the postulated surmise relations and the knowledge states contained in the postulated knowledge spaces have been compared to the empirical response patterns of three investigations. Investigation I and III clearly confirm the postulated model, whereas the results for Investigation II are ambiguous. The methods via the surmise relation (percentage of correct solutions, VC, and γ–index, see Section 5.3.3.1) are in accordance with the hypothesis, but the simulation studies on the postulated knowledge space indicate that either the model assumptions are incorrect or that the noise rate in the data was underestimated (i. e. $\beta > 0.16$ and/or $\eta > 0.15$). Considering that the basic model is the same for all three investigations (see general hypothesis in Section 5.1) and that the participants in Investigation II were draftees and therefore probably not doing their best while processing the items, a higher percentage of noise seems to be a reasonable explanation. Thus, with respect to the second part of question (ii) (see above), it can be concluded that the component based model accuratly represents the solution behavior of participants.

The final reason for the choice of the methodology is that there already exist algorithms for adaptive assessments, which are based on knowledge space theory and therefore account for the theoretical model assumptions. The availability of adaptive assessment programs facilitated a first implementation and evaluation of the postulated models. The results (cf. Chapter 6) show that the derived model provides a good basis for the applied adaptive assessment procedures. For both the deterministic and the probabilistic assessment algorithm, the mean symmetric distances arising from the adaptive assessments ranged between .30 and .89 items per response pattern, while the savings ranged between 43.96% and 72.23% of the questions (see Table 6.1). Thus, the algorithms constitute an efficient way to the assessment of knowledge with only very little costs in the estimations' accuracy, which answers question (iii) in Chapter 4.

Summarized, the general framework of knowledge space theory served in a variety of areas, including the generation of hypotheses by establishing surmise relations between inductive reasoning tests, the prediction of observable behavior, the validation of the models, and the efficient diagnosis of knowledge within adaptive testing procedures. In the remaining part of this chapter, I will take up some issues concerning the selected test model, the established classification scheme, the used validation methods, and the applied adaptive assessment algorithms. A short outlook on further research will conclude this report.

The test model There are two main factors distinguishing the non–numerical knowledge space approach from other approaches to the investigation and assessment of inductive reasoning problems and abilities (see Sections 2.2 and 2.4.2). Firstly, knowledge space theory provides a framework, in which theoretically founded item and test structures can be represented without the requirement of homogeneous item sets. In classical test theory, there are no explicit assumptions made on the factors that determine item difficulty and the selection of items as well as the estimation of item parameters are mostly based on an a posteriori analysis. Probabilistic test theory requires a set of homogeneous items and it is therefore usually necessary to exclude some of the items from the intended item pool. Moreover, a combination of different problem types seems not appropriate, because the unidimensionality of items cannot be expected. Secondly, knowledge space theory directly links the problem descriptions to the resulting interindividual differences in performance. Both classical and probabilistic test theory describe participants' performance in the form of test scores or ability parameters, i. e. the approaches are purly quantitative. The knowledge space approach directly links the theoretically specified item demands (components or skills) to the observable solution behavior of participants. That is, it provides, on the one hand, a formal description of the components that influence task difficulty and, one the other hand, a formal description of the processes in which individuals differ (i. e. the ability to deal with the specified problem components for each item). This connection between item descriptions and empirical response patterns permits an evaluation of the assumed requirements of inductive reasoning problems by providing falsifiable predictions about the postulated response patterns. This means that a cognitive theory on the problem components, which are required for the solution of an item, can be evaluated directly by observing the compatibility of the theoretical knowledge states and the empirical response patterns. Furthermore, the mentioned link renders not only quantitative but also qualitative descriptions of participants' performance on the test(s). The advantage arising from the description of individuals as being in a certain performance state instead of reaching a certain test score, is that the obtained diagnostic information is detailed enough to be used for personalized adaptive testing as well as training. With respect to personalized training, each diagnosed knoweldge state contains information on the problem demands the testee was able to meet and those he or she is lacking and therefore needs further training on.

The classification scheme Coming back to the main objective of this study, namely the construction of a valid test knowledge space for inductive reasoning tests, some remarks on the specified components are necessary. In order to develop a classification scheme which integrates various types of inductive reasoning problems, the findings of earlier research in this field (cf. Sections 2.2 and 3.5) have been considered to define common components and attributes. Although the description of item classes was reduced to those aspects which are inherent in all problem types, the results showed that the established difficulty order is able to explain the empirical data for the most part.

In earlier analyses (Albert and Wesiak, 2002; Wesiak and Albert, 2001) of the tests, we also differentiated between the constraint coming from the salience of the operations (i. e. the easiness to detect the relevant relations) and the constraint coming from the set of answer alternatives (referred to as task ambiguity, i. e. the difficulty or similarity of the alternatives). However, further investigations of the sets of tests and data showed that the exclusion of the component task ambiguity renders a more informative structure with a higher number of pairs in the relation (and fewer states in the knowledge space) by simultaneously maintaining the goodness of fit for the derived test structure. More exactly, the earlier model with six components resulted in a surmise relation between items with 231 pairs and a corresponding knowledge space with 553 states (as compared to 351 pairs and 293 states contained in the present model). The differences in the structure are even more evident when considering that the earlier model included only 27 instead of 30 items[1]. With respect to models' fit, the validation via the surmise relation yielded a global γ–index of .28 for the earlier model and .36 for the present model, the validation via the knowledge space resulted in DA coefficients of .29 for the earlier and .30 for the present model. Thus the more informative model with only five components shows a better fit with respect to the surmise relation and only slight differences with respect to the knowledge space.

Coming back to classification scheme presented in this report, the results of the various substructures in each of the investigations have to be taken into account in order to decide which parts of the classification scheme contribute most to the found deviations. Considering the obtained results via the surmise relation as well as via the knowledge space for the investigations' $SRbI$, $SRxT$, and $SRwT$, refinements of the hypothesis appear necessary for problems of the type verbal analogy. The results obtained in all three investigations yield most contradictions for this test. It is necessary to distinguish between two aspects, namely (a) the comparability of verbal analogies with other types of problems and (b) the difficulty orders within the analogy tests. With regard to (a), the component number of operations (component B in the classification scheme) is probably not as adequate to compare verbal material with other contents. For numerical and geometric material, the number of operations is based on the varying number of independent operations that need to be performed to solve the problems. The terms of verbal analogy items are usually connected by only one semantic relation. Variations in the number of operations are based on the rationale complexity, i. e. the number of relevant concepts or elements in the relation (see Section 2.2.1). A better correspondence for relevant concepts in a semantic relation might be found in the number of constituent elements in geometric material (number of lines, forms, etc.), while there are no corresponding features for numerical contents. A thinkable solution to this discrepancy is to go back to a more general model which does not account for

[1] Due to the different model assumptions the preediting of data and tests (see Section 5.2.3) resulted in the elimination of 18 items, which lead to a final set of 12 items in the matrix test and 15 items in the analogy test (as compared to 14 and 16 items in the present investigation). The simultaneous elimination of incomplete response patterns resulted in a set of 809 patterns (as compared to the present 572).

the number of operations, but only for operation difficulty. In this study, the more specific variant was chosen, because several earlier studies (cf. Section 2.2) indicate that the number of operations is an important factor contributing to item difficulty. Another reason for the assumed comparability of the two kinds of relational complexity (number of independent operations vs. elements in the semantic rationale) is that both constitute problem features that influence working memory.

The second issue (b) concerns the item descriptions within the verbal analogy tests. In Section 2.2.1 several factors that contribute to item difficulty but are not included in the classification scheme have been reported. Included are word frequency, semantic distance between the terms, and abstract versus concrete material. In order to find out whether the components specified in the applied classification scheme need to be supplemented by components that are specific to certain problem types, several alternative models have been developed. The component abstract versus concrete material was investigated for the problems presented in Investigations I and II. Word frequency was incorporated into a model for Investigation III by specifying analogies with at least one term of low frequency (less than 20 entries) as more difficult to solve. The results show that neither the inclusion of abstract versus concrete material nor the component word frequency improved the model. Of course, the adequacy of the only available word statistic for German vocabulary (Meier, 1964) has to be questioned for the language use of today's participants. In addition, word frequency as contributing factor to item difficulty has lost its importance in newer research (R. Muhr, personal communication, June 22, 2001). A more prominent factor is the influence of semantic distances and collocations. Unfortunately, there is no statistic available yet for German collocations and a investigation of this component was therefore not possible. However, these factors are important aspects for future research in this field.

Validation methods The results for the three investigations were obtained by applying two different methods, namely the validation via the surmise relation and via the knowledge space. This has the advantage that not only the validity of the various (sub)structures but also the validity of the applied methods is evaluated. In knowledge space theory there exists a one–to–one correspondence between a knowledge space and its respective surmise relation. Hence, the application of methods based on the knowledge space should yield the same results as methods based on the surmise relation. In addition, the empirical validation of surmise relations between tests requires a separated analysis of the test knowledge structure and its substructures. Otherwise it is not possible to determine in which way different parts of the structure contribute to the results. Therefore, the validation procedures always included an evaluation of the surmise relation between items, across, and within tests. Thus, for the two methods of validation, it is expected that the results for the various structures are equivalent. As already mentioned in Sections 5.2.4 and 5.3.4, all of the applied validation methods lead to the same results with respect to the relative fit of the four structures in Investigations I and II. The same is true for the last investigation, except that the analysis of relative solution frequencies yields better results for the verbal than for the geometric analogies. This result is most likely due to the fact, that the analysis of solution frequencies is the only method, which does not work on an individual level and gives therefore only a rough estimation of a model's fit.

In Section 3.4, I pointed out that floor or ceiling effects can lead to an overestimation of a model's fit, because trivial response patterns as well as item pairs where either both or none of the items are solved correctly are always in accordance with the hypothesis. The results of the three investigations indicate that overestimations of the goodness of fit occur mainly for the validation methods via the knowledge space. As already discussed in Sections 5.3.4 and 5.4.4, the high solution frequencies in Investigation I had a positive influence on the symmetric distances (cf. Section 5.2.3.2) but not on the percentage of correct solutions, VC, or the γ–index (cf. Section 5.2.3.1). As a consequence, the model in Investigation I shows the best fitting knowledge space but the least fitting surmise relation (including the substructures between items, across, and within tests). Comparing the results of Investigations II and III, which are not effected by a floor or ceiling effect, the various validation methods correspond to each other. All of the measures via the surmise relation (Sections 5.3.3.1 and 5.4.3.1) and via the knowledge space (Sections 5.3.3.2 and 5.4.3.2) indicate that the models between items, across tests, and within the matrix test show a better fit in Investigation III, while the model for the verbal analogy test shows always a better fit in Investigation II. Thus, it can be concluded, that the various validation methods applied for this research lead to the same results as long as the respective data set does not contain any floor or ceiling effects (in this study only a ceiling effect occured, but it is plausible that floor effects influence the methods via the knowledge space in the same way as ceiling effects do).

With regard to the surmise relation between tests, one problem of the approach is that the only existing method to validate a postulated $SRbT$ is by means of the solution frequencies, which constitutes the weakest method to validate knowledge space hypotheses. All of the remaining validation methods can only be applied to the derived subrelations or substructures. They are, however, valuable measures to evaluate the influence of these structures on the overall hypothesis. Thus, there is definitely more research needed on the development of further validation methods for the $SRbT$. This research should also include the development of methods for the statistical validation of the obtained indices. As for now, only the γ–index permits a statistical evaluation be means of a χ^2 test, but even this measure is rather weak because it only judges whether the number of concordant pairs is significantly higher than the number of discordant pairs. For the knowledge space, which renders a frequency distribution of the symmetric distances, the obtained results have been constrasted with the results obtained via the powersets and simulated data sets. Comparisons with the respective knowledge spaces' powersets and random data sets also constitute relatively weak measures, because the only implication is that the model explains the investigated domain better than random knowledge structures. The simulations on the hypothesis (probability simulations) are a much stronger test of the model and therefore a better indicator for the model's fit. However, the noise probabilities are unknown and can therefore only be roughly estimated. The strictest test of the postulated knowledge spaces are the frequency simulations which consider the solution frequencies of the items and participants. In this study only the surmise relation between items of Investigation I performed better than the simulated data matrices. This basically means, that for all other knowledge spaces the frequencies are an equally well predictor for testee's solution behavior as the postulated knowledge states. However, this conclusion might be premature, because the postulated model still needs to be refined with respect to part of its substructures (see above).

A last point concerning the validation methods is the fact that the application of deterministic models usually leads to difficulties with regard to the interpretation of deviating response patterns. The problem is that it is not possible to decide whether contradicting response patterns are a consequence of false model predictions or a consequence of noisy data, i. e. of lucky guesses or careless errors. There are two possible solutions to this problem. One is the application of probabilistic models, which is up to now difficult for larger sets of items, because the number of model parameters quickly increases with the number of states and items (Doignon and Falmagne, 1999). Wickelmaier (2002), for example, applied Doignon and Falmagne's (1999, Chapter 7) simple learning model for a probabilistic validation of the set of data and tests presented in Investigation III. Based on a data–driven knowledge space with 438 states (derived via item tree analyis by van Leeuwe, 1974, see Section 3.3) and the investigations' 121 response patterns, Wickelmaier tried to estimate the probabilities for careless erros, lucky guesses, and a learning parameter for each item, in order to obtain a probabilistic knowledge space. The validation of the derived model (with 2^{20} possible answer vectors and 31 estimated parameters) was only possible by the deterministic method of symmetric distances, whereas the intended probabilistic validation via a χ^2 statistic (distributed as a random variable) was not computable. Thus, with larger sets of items, the application of probabilistic models is still problematic. Another approach to the solution of this problem is the presentation of several items per attribute combination (item class) combined with the definition of a threshold specifying the percentage of items that has to be solved per item class. Thereby, it should be possible to determine whether or not a person has the ability to master the given problem requirements. In this study, the material was not self–constructed in order to make sure that the items already proved to be aedequate representatives of the assessed ability domain. Therefore, the number of items per class varied (between one and five items per item class) and the described scoring method could not be applied.

Adaptive knowledge assessment In Chapter 6, I outlined several approaches to the adaptive assessment of knowledge, focusing on two adaptive algorithms that are based on knowledge space theory. The test knowledge spaces[2] established for the three investigations were implemented into a deterministic and a non–deterministic adaptive assessment algorithm (see Section 6.2). The results presented in Section 6.3.2 show that the deterministic procedure surpasses the non–deterministic one in efficiency as well as accuracy. Since this unexpected result is still unresolved, I will in the following refer to both algorithms. In this research, the adaptive assessment was simulated by using the already obtained response patterns. Thereby, an evaluation of the models when applied to adaptive testing procedures was possible by comparing the estimated patterns to the original ones. The estimation errors arising from the two algorithms amounted to less than one item per response pattern, while the number of presented problems was reduced by up to 72.23% (see Table 6.1). Thus, the administration of

[2] Note that the postulated surmise relation between tests is only considered indirectly by using the relation across tests for inferences between the tests.

adaptive tests instead of standard tests with fixed sets of items renders an efficient approach to the assessment of inductive reasoning abilities.

Using adaptive procedures that are based on knowledge space theory, furthermore, has the advantage that the assessed knowledge state exactly specifies, which requirements or problem demands a participant is able to meet. Knowledge of the specific attribute combinations somebody is able to master allows for an eventual training that can concentrate directly on the set of lacking skills. The combination of an efficient computer–aided diagnosis of a person's ability level and a precise feedback of the already acquired and the still lacking problem demands or skills, has the advantage that personalized training can be delivered on individual demand and without the need of a human tutor. With respect to the knowledge domain investigated in this research, the importance of inductive reasoning abilities was also emphasized by Klauer (2001), who pointed out that the training of inductive reasoning skills improves the development of fluid intelligence as well as it facilitates learning in academic settings.

Outlook Before implementing the derived test knowledge space into a real diagnostic system, the problems concerning the order on verbal analogies (see above for the discussion of the classification scheme) have to be resolved. Furthermore, the development of a system that is able to diagnose the exact knowledge space of a testee, requires an expansion of the set of items to an item pool, that covers all possible attribute combinations.

Starting with the components and attributes used in this investigation, the classification system could be generalized to include further materials (such as letters or pictorial material) and answer formats. Figure 7.1 illustrates such a classification scheme in form of a mapping sentence, which corresponds to building the Cartesian product of the componenents (see also footnote on page 28). The mapping sentence is based on Klauer's model of inductive reasoning (see Section 2.3.1, Figure 2.3), but additionally specifies the differentiating requirements for items belonging to the same problem types.

Based on the possible attribute combinations that are specified in Figure 7.1, the method of systematical item construction (Albert and Held, 1994, 1999; Held, 1999) could be applied to obtain a complete set of item classes with several representatives per class. The presentation of more than one item per class has the advantage that lucky guesses and careless errors can be accounted for by coding only those item classes as mastered for which a minimal percentage of items has been answered correctly (e. g., two out of three). Such a coding system would of course decrease the efficiency of the adaptive algorithm. A thinkable solution is to present several items belonging to the same class only towards the end of the assessment in order to decide which of the remaining knowledge states represents the true state of the testee. This procedure is similar to the non–deterministic algorithms described in Section 6.2.2, but instead of presenting the same item(s) once more, a new item of the same item class would be presented. Furthermore, the diagnostic system should provide information on the knowledge state of the person and specify the demands, for which further training is needed. In addition, the assessment system could be connected to a tutorial system, which provides immediate learning opportunities. It has already been shown that such

The solution of inductive reasoning problems requires the extraction of

$$\begin{array}{ccccc} A & B & & C & \\ \left\{\begin{array}{ll} a_1 & 1 \\ a_2 & 2 \\ \vdots & \vdots \\ a_n & n \end{array}\right\} & \left\{\begin{array}{l} b_1 \text{ difficult} \\ b_2 \text{ other} \end{array}\right\} & \text{operations of} & \left\{\begin{array}{l} c_1 \text{ low} \\ c_2 \text{ high} \end{array}\right\} & \text{constraint,} \end{array}$$

$$\begin{array}{ccc} & D & \\ \text{their application to} & \left\{\begin{array}{l} d_1 \text{ geometric–figural} \\ d_2 \text{ numerical} \\ d_3 \text{ verbal} \\ d_4 \text{ letter} \\ d_5 \text{ pictorial} \\ d_6 \text{ other} \end{array}\right\} & \text{materials, and} \end{array}$$

$$\begin{array}{ccc} & E & \\ \text{an overt response in} & \left\{\begin{array}{l} e_1 \text{ forced choice} \\ e_2 \text{ open} \end{array}\right\} & \text{answer format.} \end{array}$$

Figure 7.1: Mapping sentence for the solution requirements of inductive reasoning problems

a tutorial system can be implemented on the basis of knowledge space theory (Albert and Hockemeyer, 1997; Doignon and Falmagne, 1999; Dowling et al., 1996; Hockemeyer, Held, and Albert, 1998).

As mentioned above, further developments of the classification scheme, the corresponding surmise relation between tests, and the set of items per test are necessary to implement the psychological findings into a comprehensive diagnostic and tutorial system. In the reported study, the basic work for this research has been carried out. In order to obtain more detailed information on the underlying processes participants use to solve a problem, eye tracking data and verbal reports should be collected. Regarding the results of the various surmise relations within tests, the model for the matrix test was best fitting in all three of the investigations. The task analysis of the matrix items followed for a great part the rules found by Carpenter et al. (1990), who also used eye tracking data and verbal reports for the specification of the rules. Thus, similar approaches for the remaining tests should also render more precise information on the relevant components and their attributes. In the case of several tests, such an investigation should of course include different problem types in order to compare the found processes with each other.

Concluding this report, I believe that the established surmise relation between inductive reasoning tests and the first implementations of the three models into adaptive assessment algorithms provide a promising foundation for further research in this area.

References

Airasian, P. W. & Bart, W. M. (1973). Ordering theory: A new and useful measurement model. *Educational Technology, May*, 56–60.

Albert, D. (1995). *Surmise relations between tests.* Talk at the 28th Annual Meeting of the Society for Mathematical Psychology, University of California, Irvine, August.

Albert, D., Brandt, S., Hockemeyer, C., Ünlü, A., & Schappacher, W. (2003). *Properties of surmise relations between tests.* Manuscript in preparation.

Albert, D. & Held, T. (1994). Establishing knowledge spaces by systematical problem construction. In D. Albert (Ed.), *Knowledge Structures* (pp. 78–112). New York: Springer Verlag.

Albert, D. & Held, T. (1999). Component based knowledge spaces in problem solving and inductive reasoning. In D. Albert & J. Lukas (Eds.), *Knowledge Spaces: Theories, Empirical Research, Applications* (pp. 15–40). Mahwah, NJ: Lawrence Erlbaum Associates.

Albert, D. & Hockemeyer, C. (1997). Adaptive and dynamic hypertext tutoring systems based on knowledge space theory. In B. du Boulay & R. Mizoguchi (Eds.), *Artificial Intelligence in Education: Knowledge and Media in Learning Systems* (pp. 553–555). Amsterdam: IOS Press.

Albert, D. & Hockemeyer, C. (1999). Developing curricula for tutoring systems based on prerequisite relationships. In G. Cumming, T. Okamoto, & L. Gomez (Eds.), *Advanced Research in Computers and Communications in Education: New Human Abilities for the Networked Society* (Vol. 2, pp. 325–328). Amsterdam: IOS Press. Proceedings of the 7th International Conference on Computers in Education (ICCE), Chiba, Japan.

Albert, D. & Lukas, J. (Eds.). (1999). *Knowledge spaces: Theories, empirical research, applications.* Mahwah, NJ: Lawrence Erlbaum Associates.

Albert, D., Schrepp, M., & Held, T. (1994). Construction of knowledge spaces for problem solving in chess. In G. H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (pp. 123–135). New York: Springer–Verlag.

Albert, D. & Wesiak, G. (2002). *How to generate and validate hypotheses on surmise relations among tests. exemplified for inductive reasoning tests.* Manuscript in preparation.

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

Amthauer, R. (1973). *Intelligenz-Struktur-Test (IST 70).* Göttingen: Hogrefe.

Bart, W. M. & Krus, D. J. (1973). An ordering-theoretic method to determine hierachies among items. *Educational and Psychological Measurement*, *33*, 291–300.

Baumunk, K. & Dowling, C. E. (1997). Validity of spaces for assessing knowledge about fractions. *Journal of Mathematical Psychology*, *41*, 99–105.

Beauducel, A. & Brocke, B. (1993). Intelligence and speed of information processing: Further results and questions on hick's paradigm and beyond. *Personality and Individual Differences*, *15*, 627–636.

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving.* New York: Springer.

Birkhoff, G. (1937). Rings of sets. *Duke Mathematical Journal*, *3*, 443–454.

Bisanz, J., Bisanz, G. L., & Korpan, C. A. (1994). Inductive reasoning. In R. J. Sternberg (Ed.), *Thinking and Problem Solving* (pp. 179–213). San Diego: Academic Press.

Brähler, E., Holling, H., Leutner, D., & Petermann, F. (Eds.). (2002). *Brickenkamp Handbuch psychologischer und pädagogischer Tests. Band 1.* (3rd ed.). Göttingen: Hogrefe.

Brandt, S., Albert, D., & Hockemeyer, C. (1999). Surmise relations between tests - preliminary results of the mathematical modelling. *Electronic Notes in Discrete Mathematics*, *2*.

Brandt, S., Albert, D., & Hockemeyer, C. (2003). Surmise relations between tests - mathematical considerations. *Discrete Applied Mathematics*, *127*(2), 221–239.

Breuer, F. (1977). *Einführung in die Wissenschaftstheorie für Psychologen* (2nd ed.). Münster: Aschendorf.

Brocke, B. & Beauducel, A. (2001). Intelligenz als Konstrukt. In E. Stern & J. Guthke (Eds.), *Perspektiven der Intelligenzforschung* (pp. 13–42). Lengerich: Pabst Science.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego: Academic Press.

Büchel, F. P. & Scharnhorst, U. (1993). Training des induktiven Denkens bei Lern- und Geistigbehinderten. In K. J. Klauer (Ed.), *Kognitives Training* (pp. 95–123). Göttingen: Hogrefe.

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. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carroll, J. B. (1994). Cognitive abilities: Constructing a theory from data. In D. Detterman (Ed.), *Current Topics in Human Intelligence* (pp. 43–63). Norwood, N.J.: Ablex Publishing Corporation.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22.

Chaffin, R. & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory and Cognition*, *12*(2), 134–141.

Cosyn, E. & Thiéry, N. (2000). A practical procedure to build a knowledge structure. *Journal of Mathematical Psychology*, *44*, 383–407.

Daniels, J. C. (1993). *Figure Reasoning Test (FRT)* (12th ed.). Nottingham: D.Daniels.

Davey, B. A. & Priestley, H. A. (1990). *Introduction to lattices and order.* Cambridge Mathematical Textbooks. Cambridge, UK: Cambridge University Press.

Degreef, E., Doignon, J.-P., Ducamp, A., & Falmagne, J.-C. (1986). Languages for the assessment of knowledge. *Journal of Mathematical Psychology*, *30*, 243–256.

Doignon, J.-P. (1994a). Knowledge spaces and skill assignments. In G. H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (pp. 111–121). New York: Springer–Verlag.

Doignon, J.-P. (1994b). Probabilistic assessment of knowledge. In D. Albert (Ed.), *Knowledge Structures* (pp. 1–56). New York: Springer Verlag.

Doignon, J.-P. & Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, *23*, 175–196.

Doignon, J.-P. & Falmagne, J.-C. (1999). *Knowledge spaces.* Berlin: Springer–Verlag.

Dowling, C. E. (1993a). Applying the basis of a knowledge space for controlling the questioning of an expert. *Journal of Mathematical Psychology*, *37*, 21–48.

Dowling, C. E. (1993b). On the irredundant construction of knowledge spaces. *Journal of Mathematical Psychology*, *37*, 49–62.

Dowling, C. E. & Hockemeyer, C. (2001). Automata for the assessment of knowledge. *IEEE Transactions on Knowledge and Data Engineering*, *13*(3), 451–461.

Dowling, C. E., Hockemeyer, C., & Ludwig, A. H. (1996). Adaptive assessment and training using the neighbourhood of knowledge states. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *Intelligent Tutoring Systems* (pp. 578–586). Berlin: Springer Verlag.

Dowling, C. E., Koch, U., & Quante, K. A. (1996). A new interface for querying experts on prerequisite relationships. In J. Grundy & M. Apperley (Eds.), *Proceedings of the Sixth Auatralian Conference on Computer–Human Interaction (OzCHI 96)* (pp. 320–321). Los Alamitos, California: IEEE Computer Society Press.

Dörner, D. (1976). *Problemlösen als Informationsverarbeitung.* Stuttgart: Kohlhammer.

Düntsch, I. & Gediga, G. (1995). Skills and knowledge structures. *British Journal of Mathematical and Statistical Psychology*, *48*, 9–27.

Egan, D. E. & Greeno, J. G. (1974). Theory of rule induction: Knowledge acquired in concept learning, serial pattern learning, and problem solving. In L. W. Gregg (Ed.), *Knowledge and Cognition* (pp. 43–103). Potomac, Md.: Erlbaum.

Ernst, G. W. & Newell, A. (Eds.). (1969). *GPS: A case study in generality and problem solving.* New York: Academic Press.

Evans, J. S., Newstead, S. E., & Byrne, R. M. (Eds.). (1993). *Human reasoning. The psychology of deduction.* Hillsdale, NJ: Laurence Erlbaum Associates.

Falmagne, J.-C. (1989a). A latent trait theory via stochastic learning theory for a knowledge space. *Psychometrika*, *53*, 283–303.

Falmagne, J.-C. (1989b). Probabilistic knowledge spaces: A review. In F. S. Roberts (Ed.), *Applications of Combinatorics and Graph Theory to the Biological and Social Sciences* (pp. 283–303). New York: Springer Verlag.

Falmagne, J.-C. (1994). Finite markov learning models for knowledge structures. In G. H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (pp. 75–89). New York: Springer–Verlag.

Falmagne, J.-C. & Doignon, J.-P. (1988a). A class of stochastic procedures for the assessment of knowledge. *British Journal of Mathematical and Statistical Psychology*, *41*, 1–23.

Falmagne, J.-C. & Doignon, J.-P. (1988b). A Markovian procedure for assessing the state of a system. *Journal of Mathematical Psychology*, *32*, 232–258.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test and search them. *Psychological Review*, *97*, 201–224.

Fay, E., Trost, G., & Gittler, G. (1998). *Intelligenz-Struktur-Analyse ISA. Ein Test zur Messung der Intelligenz.* Frankfurt: Swets Test Services.

Fischer, G. H. & Molenaar, I. W. (Eds.). (1995). *Rasch Models. Foundations, recent developments, and applications.* New York: Springer-Verlag.

Formann, A. K. (1973). *Die Konstruktion eines neuen Matrizentests und die Untersuchung des Lösungsverhaltens mit Hilfe des linearen logistischen Testmodells.* Dissertation, Universität Heidelberg, Germany.

Formann, A. K. & Piswanger, K. (1979). *Wiener Matrizen-Test (WMT).* Weinheim: Beltz.

Foster, D. (1998). Adaptive testing [on-line]. Available: `http://www.galton.com/research`.

Garnier, R. & Taylor, J. (1992). *Discrete mathematics for new technology.* Bristol: Institute of Physics Publishing.

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

Gentner, D. & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277–300.

Gick, M. L. & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Gick, M. L. & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38.

Goertzel, B. (1993). *The structure of intelligence. A new mathematical model of mind.* Berlin: Springer.

Goodman, L. A. & Kruskal, W. H. (1972). Measures of association for cross classifications. *Journal of the American Statistical Association*, *67*, 415–421.

Greeno, J. G. (1978). Natures of problem-solving abilities. In W. K. Estes (Ed.), *Handbook of Learning and Cognitive Processes* (Vol. 5, pp. 239–270). Hillsdale, N.J.: Erlbaum.

Greeno, J. G. & Simon, H. A. (1988). Problem solving and reasoning. In R. C. Atkinson, R. J. Herrenstein, G. Lindzey, & R. D. Luce (Eds.), *Steven's Handbook of Experimental Psychology, Vol.2: Learning and Cognition* (pp. 589–672). New York: Wiley.

Guilford, J. P. (1965). The structure of intellect. *Psychological Bulletin*, *53*(4), 267–293.

Guilford, J. P. (1981). Higher-order structure-of-intellect abilities. *Multivariate Behavioral Research*, *16*, 411–435.

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179–203.

Haygood, R. C. & Bourne, L. E. (1965). Attribute- and rule-learning aspects of conceptual behavior. *Psychological Review*, *72*, 175–195.

Held, T. (1993). *Establishment and empirical validation of problem structures based on domain specific skills and textual properties.* Dissertation, Universität Heidelberg, Germany.

Held, T. (1999). An integrated approach for constructing, coding, and structuring a body of word problems. In D. Albert & J. Lukas (Eds.), *Knowledge Spaces: Theories, Empirical Research, Applications* (pp. 67–102). Mahwah, NJ: Lawrence Erlbaum Associates.

Held, T. & Korossy, K. (1998). Data analysis as a heuristic for establishing theoretically founded item structures. *Zeitschrift für Psychologie*, *206*, 169–188.

Held, T., Schrepp, M., & Fries, S. (1995). Methoden zur Bestimmung von Wissensstrukturen — Eine Vergleichsstudie. *Zeitschrift für Experimentelle Psychologie*, *XLII*(2), 205–236.

Heller, J. (2001). *Statistischer Test der empirischen Validität einer Wissenstruktur.* Internal manuscript, Karl–Franzens–Universität Graz, Austria.

Hockemeyer, C. (2001). *Tools and utilities for knowledge spaces* (2nd ed.). Unpublished technical report, Institut für Psychologie, Karl–Franzens–Universität Graz, Austria.

Hockemeyer, C. (2002). A comparison of non–deterministic procedures for the adaptive assessment of knowledge. *Psychologische Beiträge*, *44*, 495–503.

Hockemeyer, C., Albert, D., & Brandt, S. (1998). Surmise relations between courses. *Journal of Mathematical Psychology*, *42*, 508. Abstract of a talk presented at the 29th EMPG meeting, Keele, UK, September 1998.

Hockemeyer, C., Held, T., & Albert, D. (1998). RATH — a relational adaptive tutoring hypertext WWW–environment based on knowledge space theory. In C. Alvegård (Ed.), *CALISCE'98: Proceedings of the Fourth International Conference on Computer Aided Learning in Science and Engineering* (pp. 417–423). Göteborg, Sweden: Chalmers University of Technology.

Hockemeyer, C. & Pötzi, S. (2001). *Documentation of the libsrbi library.* Unpublished technical report, Institut für Psychologie, Karl–Franzens–Universität Graz, Austria.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction.* Cambridge, MA: MIT Press.

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1982). Cognitive dimensions of numerical rule induction. *Journal of Educational Psychology*, *74*(3), 360–373.

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*(4), 603–618.

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, *57*, 253–270.

Hornke, L. F. & Habon, M. (1984). Erfahrungen zur rationalen Konstruktion von Testaufgaben. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *5*, 203–212.

Hornke, L. F., Küppers, A., & Etzel, S. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. *Diagnostica*, *46*(4), 182–188.

Hornke, L. F., Rettig, K., & Etzel, S. (1999). *Adaptiver Matrizentest (AMT).* Mödling: Schuhfried.

Hunt, E. (1974). Quote the Raven? Nevermore! In L. W. Gregg (Ed.), *Knowledge and Cognition* (pp. 129–157). Potomac, Md.: Erlbaum.

Hunt, E. (1978). Mechanics of verbal ability. *Psychological Review*, *85*, 109–130.

Hunt, E. (1985). Verbal ability. In R. Sternberg (Ed.), *Human abilities. An information-processing approach* (pp. 31–58). New York: W. H. Freeman and Co.

Jacobs, P. I. & Vandeventer, M. (1972). Evaluating the teaching of intelligence. *Educational and Psychological Measurement*, *32*, 235–248.

Jäger, A. O., Süß, H. M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test (BIS-Test), Form 4.* Göttingen: Hogrefe.

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, *28*(3), 195–225.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, *35*(1), 21–35.

Jäger, A. O. & Tesch-Römer, C. (1988). Replikation des Berliner Intelligenzstrukturmodells (BIS) in den "Kit of Reference Test for Cognitive Factors" nach French, Ekstrom & Price (1963). *Zeitschrift für Differentielle und Diagnostische Psychologie*, *9*(2), 77–96.

Johnson-Laird, P. N. & Byrne, R. M. J. (1993). Precis of deduction. *Behavioral and Brain Sciences*, *16*, 323–380.

Johnson-Laird, P. N., Byrne, R. M. J., & Shaeken, W. (1992). Propositional reasoning by model. *Psychological Review*, *99*, 418–439.

Kail, R. V. & Pellegrino, J. W. (1988). *Menschliche Intelligenz.* Heidelberg: Spektrum.

Kambouri, M., Koppen, M., Villano, M., & Falmagne, J.-C. (1994). Knowledge assessment: Tapping human expertise by the QUERY routine. *International Journal of Human–Computer–Studies*, *40*, 119–151.

Kinder, A. & Lachnit, H. (1994). Erwerb und Anwendung logischer Relationen bei mehrdimensionalen Reizen. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *41*, 173–183.

Klahr, D. & Wallace, J. G. (1970). The development of serial completion strategies: An information processing analysis. *British Journal of Psychology*, *61*, 243–257.

Klauer, K. J. (1996). Begünstigt induktives Denken das Lösen komplexer Probleme? *Zeitschrift für Experimentelle Psychologie*, *43*(1), 85–113.

Klauer, K. J. (1997). Induktives Denken: Definition, Theorie und Training. *Zeitschrift für Experimentelle Psychologie, 44*(2), 213–219.

Klauer, K. J. (2001). Training des induktiven Denkens. In K. J. Klauer (Ed.), *Handbuch Kognitives Training* (2nd revised ed.) (pp. 165–209). Göttingen: Hogrefe.

Klauer, K. J. & Phye, G. D. (1994). *Cognitive training for children.* Seattle: Hogrefe and Huber.

Klix, F. (1978). Analoges Schließen: Kognitive Analyse einer Intelligenzleistung. In H. Ueckert & D. Rhenius (Eds.), *Komplexe menschliche Informationsverarbeitung. Beiträge zur Tagung "Kognitive Psychologie" in Hamburg 1978* (pp. 162–174). Bern: Hans Huber.

Klix, F. (1992). *Die Natur des Verstandes.* Göttingen: Hogrefe.

Koppen, M. (1993). Extracting human expertise for constructing knowledge spaces: An algorithm. *Journal of Mathematical Psychology, 37*, 1–20.

Koppen, M. (1994). The construction of knowledge spaces by querying experts. In G. H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology* (pp. 137–147). New York: Springer–Verlag.

Koppen, M. & Doignon, J.-P. (1990). How to build a knowledge space by querying an expert. *Journal of Mathematical Psychology, 34*, 311–331.

Körner, C. (1998). *Comprehension of visualized ordered sets — an empirical approach based on the theory of knowledge spaces.* Inauguraldissertation, Karl–Franzens–Universität Graz, Graz, Austria.

Körner, C. (2000). *An index for testing surmise hypotheses.* Talk at the 31st European Mathematical Psychology Group (EMPG) Meeting, University of Graz, Austria.

Korossy, K. (1997). Extending the theory of knowledge spaces: A competence–performance approach. *Zeitschrift für Psychologie, 205*, 53–82.

Korossy, K. (1998). Solvability and uniqueness of linear-recursive number sequence tasks. *Methods of Psychological Research Online, 3*(1).

Kotovsky, K. & Simon, H. A. (1973). Empirical test of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology, 4*, 399–424.

Krause, B. (1985). Zum Erkennen rekursiver Regularitäten. *Zeitschrift für Psychologie, 193*, 71–86.

Kubinger, K. D. (1992). Testtheorie: Probabilistische Modelle. In R. S. Jäger & F. Petermann (Eds.), *Psychologische Diagnostik* (2nd ed.) (pp. 322–334). Weinheim: Psychologie Verlags Union.

Kubinger, K. D. (2000). Replik auf Jürgen Rost “Was ist aus dem Rasch-Modell geworden?”: Und für die Psychologische Diagnostik hat es doch revolutionäre Bedeutung. *Psychologische Rundschau*, *51*, 33–34. [On-line longversion. Available: `http://www.univie.ac.at/Psychologie/diagnostik/forsch/`].

Kubinger, K. D., Fischer, D., & Schuhfried, G. (1993). *Begriffs-Bildungs-Test (BBT).* Mödling: Wiener Testsystem/Schuhfried.

Kubinger, K. D. & Wurst, E. (1985). *Adaptives Intelligenzdiagnostikum (AID).* Weinheim: Beltz.

Lakshminarayan, K. & Gilson, F. (1998). An application of a stochastic knowledge structure model. In C. E. Dowling, F. S. Roberts, & P. Theuns (Eds.), *Recent Progress in Mathematical Psychology* (pp. 155–172). Hillsdale, USA: Lawrence Erlbaum Associates Ltd.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist*, *43*, 431–442.

Lehman, D. R. & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology*, *26*, 952–960.

Lukas, J. (1997). Modellierung von Fehlkonzepten in Wissensstrukturen. *Kognitionswissenschaft*, *6*(4), 196–204.

Lukas, J. & Albert, D. (1993). Knowledge assessment based on skill assignment and psychological task analysis. In G. Strube & K. F. Wender (Eds.), *The Cognitive Psychology of Knowledge* (pp. 139–160). Amsterdam: North–Holland.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*, 107–128.

Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: W.H. Freeman and Company.

Meier, H. (1964). *Deutsche Sprachstatistik.* Hildesheim: Georg Olms.

Michalski, R. S. (1983). A theory and methodology of inductive learning. In R. S. Michalski, J. C. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning. An Artificial Intelligence Approach* (pp. 83–134). Palo Alto: Tioga Publishing Company.

Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*, 252–284.

Musch, J. & Albert, D. (2003). *Knowledge space modelling of inductive reasoning in an intelligence test.* Unpublished manuscript.

Nährer, W. (1980). Zur Analyse von Matrizenaufgaben mit dem linearen logistischen Testmodell. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *27*, 553–564.

Neubauer, A. (1995). *Intelligenz und Geschwindigkeit der Informationsverarbeitung.* Wien: Springer.

Paul, S. M. (1986). The Advanced Raven's Progressive Matrices: Normative data for an American university population and an examination of the relationship with Spearman's *g*. *Journal of Experimental Education*, *54*, 95–100.

Pellegrino, J. W. (1985). Inductive reasoning ability. In R. J. Sternberg (Ed.), *Human abilities. An information-processing approach* (pp. 195–225). New York: W. H. Freeman and Co.

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. Federico, & W. E. Montague (Eds.), *Aptitude, Learning, and Instruction: Cognitive Process Analyses of Aptitude* (Vol. 1, pp. 177–217). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pellegrino, J. W. & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 269–245). Hillsdale, NJ: Erlbaum.

Percevic, R. & Wesiak, G. (2001). *Adaptive-sequential testing in psychotherapy outcome monitoring.* Talk at the 32nd European Mathematical Psychology Group Meeting, Lisbon, Portugal.

Ponocny, I. & Waldherr, K. (2002). *A nonparametric goodness-of-fit procedure for unidimensional polytomous Rasch models.* Talk at the 33rd European Mathematical Psychology Group Meeting, Bremen, Germany.

Popper, K. R. (1972). *The logic of scientific discovery* (3rd ed.). London: Hutchinson.

Posner, M. I. & Mitchell, R. F. (1967). Chronometric analysis of classification. *Psychological Review*, *74*, 392–409.

Pötzi, S. (2001). *Documentation of the libsrbt library.* Unpublished technical report, Institut für Psychologie, Karl–Franzens–Universität Graz, Austria.

Pötzi, S. & Wesiak, G. (2001). *SRbT tools user manual.* Unpublished technical report, Institut für Psychologie, Karl–Franzens–Universität Graz, Austria.

Ptucha, J. (1994). Kognitive Operationen beim Fortsetzen von Zahlenfolgen: Eine experimentelle Untersuchung zur Theorie der Wissensräume. *Zeitschrift für Psychologie*, *202*, 253–274.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1–48.

Raven, J. C. (1948). The comparative assessment of intellectual ability. *British Journal of Psychology*, *39*, 12–19.

Raven, J. C. (1958). *Standard Progressive Matrices, Sets A, B, C, D and E.* Cambridge: University Press.

Raven, J. C. (1965). *Advanced Progressive Matrices, Sets I and II.* London: H.K. Lewis.

Raven, J. C. (1976). *The Coloured Progressive Matrices.* London: Lewis.

Riegler, K. (1999). Konstruktion von Wissensräumen im Symptombereich der Panikattacke. Diplomarbeit, Karl–Franzens–Universität Graz, Austria.

Rips, L. J. (1990). Reasoning. *Annual Review of Psychology*, *41*, 321–353.

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

Schaefer, R. E. (1985). *Denken: Informationsverarbeitung, mathematische Modelle und Computersimulation.* Berlin: Springer.

Scharroo, J. & Leeuwenberg, E. (2000). Representation versus process in simplicity of serial pattern completion. *Cognitive Psychology*, *40*, 39–86.

Scheiblechner, H. (1997). Corrections of theorems in Scheiblechner's treatment of ISOP models and comments on Junker's remarks. *Psychometrika*, *63*(1), 87–91.

Schmidt, J. (1984). Simultane Überprüfung der Zweimodalität im Berliner Intelligenzstrukturmodell. *Diagnostica*, *30*, 93–103.

Schrepp, M. (1993). *Über die Beziehung zwischen kognitiven Prozessen und Wissensräumen beim Problemlösen.* Dissertation, Universität Heidelberg, Germany.

Schrepp, M. (1995). Modeling interindividual differences in solving letter series completion problems. *Zeitschrift für Psychologie*, *203*, 173–188.

Schrepp, M. (1999). An empirical test of a process model for letter series completion problems. In D. Albert & J. Lukas (Eds.), *Knowledge Spaces: Theories, Empirical Research, Applications* (pp. 133–154). Mahwah, NJ: Lawrence Erlbaum Associates.

Schrepp, M., Held, T., & Albert, D. (1999). Component–based construction of surmise relations for chess problems. In D. Albert & J. Lukas (Eds.), *Knowledge Spaces: Theories, Empirical Research, Applications* (pp. 41–66). Mahwah, NJ: Lawrence Erlbaum Associates.

Schweizer, K. (1995). *Kognitive Korrelate der Intelligenz.* Lehr- und Forschungstexte Psychologie. Göttingen: Hogrefe.

Shye, S. (1988). Inductive and deductive reasoning: A structural analysis of ability tests. *Journal of Applied Psychology*, *73*, 308–311.

Simon, H. A. & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *70*, 534–546.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 47–103). Hillsdale, NJ: Erlbaum.

Sternberg, R. J. (1977a). Component processes in analogical reasoning. *Psychological Review*, *84*(4), 353–378.

Sternberg, R. J. (1977b). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Sternberg, R. J. (Ed.). (1999). *Cognitive psychology* (2nd ed.). Fort Worth: Harcourt Brace.

Sternberg, R. J. & Gardner, M. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, *112*(1), 80–116.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*, 406–427.

Tziner, A. & Rimmer, A. (1984). Examination of an extension of Guttman's model of ability tests. *Applied Psychological Measurement*, *8*, 59–69.

Undheim, J. O. & Gustafsson, J.-E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, *22*, 149–171.

van Buggenhaut, J. & Degreef, E. (1987). On dichotomisation methods in boolean analysis of questionaires. In E. E. Roskam & R. Suck (Eds.), *Progress in Mathematical Psychology* (pp. 447–543). Amsterdam: Elsevier.

van Leeuwe, J. F. J. (1974). Item Tree Analysis. *Tijdschrift voor de Psychologie*, *29*, 475–484.

Verguts, T. & De Boeck, P. (1999). *The induction of solution rules in complex analogy problems.* Research Report 99-1, Katholieke Universiteit Leuven, Belgium.

Verguts, T., De Boeck, P., & Maris, E. (1999). Generation speed in Raven's Progressive Matrices Test. *Intelligence*, *27*(4), 329–345.

Verguts, T., Van Nijlen, D., & De Boeck, P. (1999). *Variation and retention in two intelligence tests.* Research Report 99-2, Katholieke Universiteit Leuven, Belgium.

Villano, M. (1991). *Computerized knowledge assessment: Building the knowledge structure and calibrating the assessment routine.* Doctoral Dissertation, New York University, New York.

Vodegel Matzen, L. B., van der Molen, M. W., & Dudink, A. C. (1994). Error analysis of Raven test performance. *Personality and Individual Differences*, *16*(3), 433–445.

Wainer, H. (1990). Introduction and history. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (pp. 1–22). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer* (pp. 65–102). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Wason, P. C. & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.

Wesiak, G. (1998). Construction of attribute spaces for personality disorders by querying experts. Diplomarbeit, Karl–Franzens–Universität Graz, Austria.

Wesiak, G. & Albert, D. (2001). Knowledge spaces for inductive reasoning tests. In K. Kallus, N. Posthumus, & P. Jimenez (Eds.), *Current psychological research in Austria. Proceedings of the 4th scientific conference of the Austrian Psychological Society (ÖGP)* (pp. 157–160). Graz: Akademische Druck- u. Verlagsanstalt.

Whitely, S. E. (1977). Relationships 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). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement*, *18*, 67–84.

Whitely, S. E. & Barnes, G. M. (1979). The implications of processing event sequences for theories of analogical reasoning. *Memory and Cognition*, *7*(4), 323–331.

Wickelmaier, F. (2002). Empirische Untersuchung zur Erfassung von Wissen durch deterministische und probabilistische Wissensstrukturen. Diplomarbeit, Universität Regensburg, Germany.

Wriessnegger, S. (2000). Cognitive processes and knowledge spaces: An experimental approach of eye movements in solving letter series problems. Diplomarbeit, Karl–Franzens–Universität Graz, Austria.

Wriessnegger, S., Janzen, G., & Albert, D. (2002). Eye movements in solving letter series completion problems. *Psychologische Beiträge*, *44*, 512–520.

Appendix A: Mathematical Basics

List of Symbols

$\wedge$	logic and
$\vee$	logic or (inclusive)
$\neg$	negation
$\Rightarrow$	implication
$\Leftrightarrow$	equivalence
$\forall$	for all
$\exists$	exists
$\in$	is element of
$\notin$	is not element of
$\cup$, $\bigcup$	union of subsets
$\cap$, $\bigcap$	intersection of subsets
$\subset$	binary subset relation
$\subseteq$	reflexive binary subset relation
$\triangle$	set difference
$\lfloor x \rfloor$	floor of x, denotes the greatest integer less than or equal to x
$\mathcal{K}_q$	family of all knowledge states containing item q
B_q	family of all knowledge states containing item q and some item $b \in B$ $(B_q := \cap \bigcap \mathcal{K}_q)$

Relations and their properties

Relations are subsets of a Cartesian product, i. e. they are relating elements of sets to one another. If an ordered pair (x, y) is part of the subset, we write $(x, y) \in R$ or xRy, while a binary relation P is defined as $P \subset AxB$.

Properties of relations

Reflexivity $\forall\, x \in A\ [(x, x) \in R\ ]$
A relation is reflexive, iff any of its elements bear a relation to themselves.

Irreflexivity $\forall\, x \in A\ [(x, x) \notin R\ ]$
A relation is irreflexive, iff none it its elements bear a relation to themselves.

Symmetry $\forall\, x, y \in A\ [(x, y) \in R \Rightarrow (y, x) \in R\ ]$
A relation is symmetric, iff xRy implies yRx for all $x, y \in A$.

Asymmetry $\forall\, x, y \in A\ [(x, y) \in R \Rightarrow (y, x) \notin R\ ]$
A relation is asymmetric, iff xRy implies $\neg yRx$ for all $x, y \in A$.

Antisymmetry $\forall\, x, y \in A\ [(x, y) \in R \wedge (y, x) \in R \Rightarrow x = y\ ]$
A relation is antisymmetric, iff xRy and yRx imply that $x = y$.

Transitivity $\forall\, x, y, z \in A\ [(x, y) \in R \wedge (y, z) \in R \Rightarrow (x, z) \in R\ ]$
A relation is transitive, iff xRy and yRz together imply xRz.

Connectedness $\forall\, x, y \in A\ [(x, y) \in R \vee (y, x) \in R\ ]$
A relation is connected, if all elements are related and therefore comparable.

Order relations

An order R is defined on a set Q of problems, such that the order forms the pair (Q, R). Order relations are defined by one or more of the properties presented above. However, orders are always transitive. In the following, a few examples for order relations are presented together with their descriptive properties.

- **Linear orders, chains, or complete orders**
  are transitive, reflexive, antisymmetric, and connected.
- **Partial orders**
  are transitive, reflexive, and antisymmetric.
- **Antichains**
  are transitive, reflexive, symmetric, and antisymmetric.
- **Quasi orders**
  are transitive and reflexive.

Appendix B: Material

B.1 Semantic rationales

Table B.1: Semantic rationales for the analogy items in Investigation I

Item class[a]	Items	Operation types[b]	Rationales
O1LV4	21	PW	B is part of A
	26	CP	A causes B
O1LV5	15	CO	A is the opposite of B
	17	CO	A is the opposite of B
	25	PW	A is part of B
	30	PW	A is part of C
O1HV5	27	AT	C is an attribute of A
	29	AT	C is an attribute of A
O2LV5	18	ST	B is a device to protect A
O2HV5	22	ST	B is an area located in A
	24	ST	B is a process of regulation in A
O3HV5	20	CP	A is a prerequisite action to undertake B
	28	CP	B is art produced by a set of different As
D1LV5	16	SI	B is a comparative of A
D1HV5	19	SI	B is a comparative of A
D2LV5	23	SI	B is a pleasant variant of A

Note. [a]The components for the description of the item classes (*A* through *E*) are ordered alphabetically. [b]PW = part–whole, CP = cause purpose, CO = contrast, AT = attribute, ST = space–time, SI = similar/comparative.

Table B.2: Semantic rationales for the analogy items in Investigation II

Item class[a]	Items	Operation types[b]	Rationales
O1LV5	21	ST	B takes place in A
	23	CO	A is the opposite of B
O1HV5	30	AT	B is an action of A
O2LV5	36	PW	A is a part of B that covers B
O2HV5	35	PW	B is a marked off part of A
O3HV5	39	PW	A is a part of B that connects other parts of B
D1LV5	22	CI	A is a member of B
	24	SI	A is a coordinate of B
	27	SI	A is a coordinate of B
D1HV5	25	SI	A is a coordinate of B
	28	SI	A is a comparative of B
	32	SI	A is a coordinate of B
D2LV5	26	SI	B is a coordinate of A that follows A
	29	SI	A is a two–dimensional coordinate of B
	34	SI	A is a larger coordinate of B
	37	SI	A is a coordinate of B living in the same natural environment
D2HV5	31	SI	A is a form of B, which converts into B
	33	SI	B is a harmonic comparative of A
	38	SI	A is a more valuable coordinate of B
	40	SI	A is a consumable coordinate of B

Note. [a]The components for the description of the item classes (*A* through *E*) are ordered alphabetically. [b]PW = part–whole, CO = contrast, AT = attribute, ST = space–time, CI = class inclusion, SI = similar/comparative.

Table B.3: Semantic rationales for the verbal analogy items in Investigation III

Class[a]	Items	Operation types[b]	Rationales
O2HV5	1	CP	A's purpose is to function as B
O3HV5	4	PW	B is a part of A that adds zest to A
	5	PW	B is a period in A occurring in the beginnings of A
D2LV5	2	SI	B is a musical coordinate of A
D2HV5	3	SI	B is an exaggerated coordinate of A

Note. [a]The components for the description of the item classes (*A* through *E*) are ordered alphabetically. [b]PW = part–whole, CP = cause purpose, SI = similar/comparative.

B.2 Items presented in Investigation III

Items 1–15 are taken from the "Berliner Intelligenzstruktur–Test" (BIS) by Jäger et al. (1997), items 16–20 from the "Wiener Matrizen–Test" (WMT) by Formann and Piswanger (1979). For Items 11–15 answer alternatives have been added, for items 16–20 three of the eight original answer alternatives have been removed. Table B.4 depicts the numbers of the original items in the respective tests and subtests, the corresponding numbers for the present investigation, and the modified answer formats for items 11–20.

Table B.4: Items and modified answer formats for Investigation III

Problem type	Test (Subtest)	Original item No.	Present item No.	Answer alternatives				
AN_V	BIS (WA)	1	1					
		3	2					
		4	3					
		6	4					
		7	5					
AN_G	BIS (AN)	1	6					
		3	7					
		4	8					
		6	9					
		7	10					
SC_N	BIS (ZN)	1	11	a) 12	b) 17	c) 13	d) 14	e) 11
		4	12	a) 7	b) 4	c) 5	d) 9	e) 36
		5	13	a) 3	b) 16	c) 24	d) 8	e) 48
		7	14	a) 286	b) 584	c) 146	d) 283	e) 1168
		9	15	a) 584	b) 153	c) 292	d) 730	e) 876
MT_G	WMT	C	16	a, b, d, e, g				
		2	17	a, b, c, d, e				
		9	18	a, b, c, d, e				
		22	19	a, b, c, d, e				
		23	20	a, b, c, d, e				

Appendix C: Procedure for Investigation III

C.1 Instructions

Liebe Versuchsteilnehmerin, lieber Versuchsteilnehmer!

Gegenstand dieser Untersuchung ist die Strukturierung Induktiver Denktests. Dazu habe ich einen Test mit 25 verschiedenen Aufgaben zusammengestellt. Je fünf dieser Aufgaben sind nach dem gleichen Prinzip zu lösen. Ich möchte Sie bitten, die folgenden Aufgaben der Reihe nach zu bearbeiten und zu versuchen, jede einzelne Aufgabe zu lösen. Da für uns die Beziehungen oder Zusammenhänge zwischen verschiedenen Aufgaben von Bedeutung sind, ist es wichtig, daß jede einzelne Aufgabe von Ihnen bearbeitet wird. Sie haben insgesamt ca. 30 Minuten Zeit die Aufgaben zu lösen. Da es sich um verschiedene Arten von Aufgaben handelt, finden Sie am Anfang des Tests eine kurze Anleitung zum Lösen der Aufgaben sowie ein Beispiel mit bereits vorgegebener Lösung.

Für statistische Zwecke benötigen wir zuvor noch einige persönliche Angaben sowie einen von Ihnen selbst gewählten Code (z.B. Geburtstag eines Elternteils), durch den Sie dann je nach Wunsch Ihre Ergebnisse richtig zuordnen können. Alle Daten werden selbstverständlich vertraulich behandelt.

Code ____________________

Datum ____________________

Schultyp ____________________

Schulstufe ____________________

Alter ____________________

Geschlecht □ männlich □ weiblich

Herzlichen Dank für Ihre Unterstützung!!!

Beispiel 1:

Links stehen die Gruppen **A** und **B** mit je sechs grafischen Mustern.
Rechts davon stehen **drei Einzel**muster.
Bei jedem Einzelmuster sollen Sie entscheiden, ob es zu Gruppe **A** oder **B** gehört. Finden Sie heraus, wodurch sich die Gruppen **A** und **B** unterscheiden. Es gibt immer eindeutige Unterscheidungsmerkmale.

Streichen Sie dann unter jedem der drei Einzelmuster den Kennbuchstaben der Gruppe (**A** oder **B**) durch, zu der es gehört.

Presentation of the first practice item in the subtest BG of the BIS test.

Beispiel 2:

Welche der Lösungen **a** bis **e** ist anstelle des Fragezeichens einzusetzen, damit zwischen den beiden Wörtern im **ersten** Wortpaar **dieselbe Beziehung** besteht wie zwischen den beiden Wörtern im **zweiten** Wortpaar?

Streichen Sie den Buchstaben vor dem Lösungswort durch.

Presentation of the second practice item in the subtest WA of the BIS test.

Beispiel 3:

Welche der Lösungen **a** bis **e** ist anstelle des Fragezeichens einzusetzen, damit zwischen den beiden Figuren **hinter dem Gleichheitszeichen dieselbe Beziehung** besteht wie zwischen den beiden Figuren **vor dem Gleichheitszeichen**?

Streichen Sie den Buchstaben unter der richtigen Lösung durch.

Presentation of the first practice item in the subtest AN of the BIS test.

Beispiel 4:

Jede der folgenden Zahlenreihen ist nach einer bestimmten Regel aufgebaut.
Welche der Lösungen **a** bis **e** ist anstelle des Fragezeichens einzusetzen, um die Reihe nach dieser Regel fortzusetzen?

Streichen Sie den Buchstaben vor der richtigen Lösung durch.

Presentation of the first practice item in the subtest ZN of the BIS test with following answer alternatives:

a) 19 b) 34 ~~c~~) 20 d) 15 e) 23

Beispiel 5:

Die Figuren links in der Abbildung (eingerahmt) sind nach bestimmten Regeln geordnet. Welche der Lösungen **a** bis **e** ist anstelle des Fragezeichens einzusetzen, damit die Anordnung sinnvoll vervollständigt wird?

Streichen Sie den Buchstaben vor der richtigen Lösung durch.

> Presentation of the practice item A in the WMT without the answer alternatives f, g, and h.

Es folgen nun pro Beispiel 5 Aufgaben, wobei die Reihenfolge der Aufgaben zufällig gewählt wurde. Bitte bearbeiten Sie die Aufgaben **der Reihe nach** und versuchen Sie für **jede** Aufgabe eine Lösung zu finden.

Streichen Sie entweder den Kennbuchstaben der Gruppe (Beispiel 1) oder den Buchstaben vor/unter der richtigen Lösung durch (Beispiele 2–5).

> Presentation of the test items.

Liebe Versuchsteilnehmerin, lieber Versuchsteilnehmer!

Bitte überprüfen Sie nochmals, ob Sie wirklich alle Aufgaben bearbeitet haben. Wenn ja, nehmen Sie sich bitte noch kurz Zeit, um die folgenden Fragen zu beantworten.

Ich habe schon einmal einen Intelligenztest dieser Art durchgeführt. □ JA □ NEIN

Die Aufgaben der Art “Beispiel 1” waren für mich

□ sehr leicht □ eher leicht □ eher schwer □ sehr schwer zu lösen.

Die Aufgaben der Art “Beispiel 2” waren für mich

□ sehr leicht □ eher leicht □ eher schwer □ sehr schwer zu lösen.

Die Aufgaben der Art “Beispiel 3” waren für mich

□ sehr leicht □ eher leicht □ eher schwer □ sehr schwer zu lösen.

Die Aufgaben der Art “Beispiel 4” waren für mich

□ sehr leicht □ eher leicht □ eher schwer □ sehr schwer zu lösen.

Die Aufgaben der Art “Beispiel 5” waren für mich

□ sehr leicht □ eher leicht □ eher schwer □ sehr schwer zu lösen.

Danke für Ihre Mitarbeit!!!

C.2 Randomization

The sequence of the 25 items (5 items per example) was randomized for four different groups. Random orders were computed with the program `permute` by C. Hockemeyer.

Sequences:

A) 19 6 2 5 20 3 21 11 15 8 13 22 12 10 16 23 1 7 25 24 4 18 14 9 17

B) 14 22 23 9 5 2 18 12 11 15 6 19 21 24 10 17 20 3 4 13 8 1 16 25 7

C) 1 19 18 21 7 6 24 23 4 11 2 12 5 13 14 3 16 22 15 9 25 17 20 10 8

D) 3 18 1 8 16 23 11 19 20 7 12 17 4 24 10 13 6 22 15 25 9 21 14 2 5

Appendix D: List of Programs

All programs used for this study are available at the Cognitive Science Section at the University of Graz. For a detailed description the reader is referred to Hockemeyer (2001), Hockemeyer and Pötzi (2001), Pötzi (2001), and Pötzi and Wesiak (2001).

Programs for the generation of hypotheses

- `patt-statistics` counts the number of complete response patterns for various numbers of items
- `delete-not-ans` deletes a specified number of items and the remaining incomplete response patterns (items are deleted in the reversed order of the number of provided responses)
- `bas2srbi` transforms a base into a relation
- `srbi-part2srbt` computes a surmise relation between tests from a surmise relation between items and a partition into tests
- `tests-properties` computes the properties of a surmise relation between tests (left–, right–, and total–coveringness, antisymmetry, and connectedness)

Programs for the validation of hypotheses

- `partitions` computes, among other functions, the relative solution frequencies for each item
- `getca` computes the index VC
- `valid` computes the index γ
- `constr` constructs the knowledge space for a given base
- `distance` computes the frequency distribution plus the mean, standard deviation, and median for the minimal symmetric distances between a set of response patterns and a knowledge space as well as the invalidity of items with respect to careless errors and lucky guesses
- `random-patterns` computes random response patterns

- `simple-sim` simulates response patterns under consideration of the knowledge space and noise variables (careless errors and lucky guesses)
- `polydif*` simulates response patterns under consideration of the solution frequencies for items and persons

Programs for the adaptive assessment algorithms

- `space-assess` deterministic adaptive assessment algorithm, which estimates response patterns on the basis of a knowledge space
- `halfsplit-assess` non–deterministic adaptive assessment algorithm, which estimates response patterns on the basis of a knowledge space under consideration of noise probabilities

Other

- `permute` permutes a number sequence
- `LaTeX` typesetting system for the production of technical and scientific documentation
- `Micrografx` drawing and diagraming software
- `Microsoft Visio` drawing and diagraming software

Appendix E: Hypotheses

E.1 Relation files

Relation files for Investigation I

SRbI	*SRxT*
30	30
110101011111111010111011100 1110	100000000000000101110111001110
010000011010110000010111000100	010000000000000000010111000100
011101011111111010111011100 1110	001000000000000101110111001110
000100011010110000010111000100	000100000000000000010111000100
010111011111111010111011100 1110	000010000000000101110111001110
000001011010110000110101000100	000001000000000000110101000100
010101111111111010111011100 1110	000000100000000101110111001110
000000010010110000010000000100	000000010000000000010000000100
000000001010110000010000000100	000000001000000000010000000100
000000011110110000010111000100	000000000100000000010111000100
000000000010000000000000000000	000000000010000000000000000000
000000011011110000010111000100	000000000001000000010111000100
000000000000100000000000000000	000000000000100000000000000000
000000000000010000000000000000	000000000000010000000000000000
111111111111111011101110 01110	111111111111111000000000000000
000000000000010100100010000000	000000000000010100000000000000
111111111111110111110111001110	111111111111110010000000000000
010100011111110001010111000100	010100011111110001000000000000
000000000000010000100000000000	000000000000010000100000000000
000000000010100000010000000000	000000000010100000010000000000
111111111111111111111111101111	111111111111110000001000000000
000000011010110000010100000100	000000011010110000000100000000
000000000000010000000010000000	000000000000010000000010000000
000000011010110000010001000100	000000011010110000000001000000
111111111111110101110111101110	111111111111110000000000100000
111111111111111111110111111111	111111111111110000000000010000
000001011010110000110101001100	000001011010110000000000001000
000000000010100000000000000100	000000000010100000000000000100
000001011010110000110101000110	000001011010110000000000000010
111111111111110101110111001111	111111111111110000000000000001

$SRwMT$	$SRwAN$
14	16
11010101111111	1101110111001110
01000001101011	0100100010000000
01110101111111	0111110111001110
00010001101011	0001010111000100
01011101111111	0000100000000000
00000101101011	0000010000000000
01010111111111	1111111111101111
00000001001011	0000010100000100
00000000101011	0000000010000000
00000001111011	0000010001000100
00000000001000	0101110111101110
00000001101111	1111110111111111
00000000000010	0000110101001100
00000000000001	0000000000000100
	0000110101000110
	0101110111001111

Relation files for Investigation II

SRbI	*SRxT*
40	40
1011110011111111111101011111111111111111	1000000000000000000001011111111111111111
0111110011111111111101011111111111111111	0100000000000000000001011111111111111111
0010000000000111111100001001001110100111	0010000000000000000000001001001110100111
0001000000000111111100001001001110100111	0001000000000000000000001001001110100111
0000100000001111111100000100101011101111	0000100000000000000000000100101011101111
0000010000000111111100001001001110100111	0000010000000000000000001001001110100111
0011111011111111111101011111111111111111	0000001000000000000001011111111111111111
0011110111111111111101011111111111111111	0000000100000000000001011111111111111111
0000000010001111111100000100101011101111	0000000010000000000000000100101011101111
0000000001001111111100000100101011101111	0000000001000000000000000100101011101111
0000000000101111111100000100101011101111	0000000000100000000000000100101011101111
0000000000011111111100000100101011101111	0000000000010000000000000100101011101111
0000000000001000010100000000001010000101	0000000000001000000000000000001010000101
0000000000000100010100000000001010000111	0000000000000100000000000000001010000111
0000000000000010010100000000001010000111	0000000000000010000000000000001010000111
0000000000000001010100000000001010000111	0000000000000001000000000000001010000111
0000000000000000110100000000001010000111	0000000000000000100000000000001010000111
0000000000000000010100000000000000000000	0000000000000000010000000000000000000000
0000000000000000011100000000001010000111	0000000000000000001000000000001010000111
0000000000000000000100000000000000000000	0000000000000000000100000000000000000000
1111111111111111111111011111111111111111	1111111111111111111110000000000000000000
0000000000001000010101001101101111001101	0000000000001000010101000000000000000000
1111111111111111111101111111111111111111	1111111111111111111100100000000000000000
0000000000001000010100011101101111001101	0000000000001000010100010000000000000000
0000000000000000010100001000001010000101	0000000000000000010100001000000000000000
0000000000001000010100000100001010000101	0000000000001000010100000100000000000000
0000000000001000010100001111101111001101	0000000000001000010100000010000000000000
0000000000000000010100000001001010000101	0000000000000000010100000001000000000000
0000000000001000010100000000101010000101	0000000000001000010100000000100000000000
0011010000000111111100001001011110100111	0011010000000111111100000000010000000000
0000000000000000010100000000001000000000	0000000000000000010100000000001000000000
0000000000000000010100000000001110000101	0000000000000000010100000000000100000000
0000000000000000010100000000000010000000	0000000000000000010100000000000010000000
0000000000001000010100000000001011000101	0000000000001000010100000000000001000000
0000000000000111111100000000001010100111	0000000000000111111100000000000000100000
0000100011111111111100000100101011111111	0000100011111111111100000000000000010000
0000000000001000010100000000001010001101	0000000000001000010100000000000000001000
0000000000000000010100000000000000000100	0000000000000000010100000000000000000100
0000000000000000000100000000000000000010	0000000000000000000100000000000000000010
0000000000000000010100000000000000000001	0000000000000000010100000000000000000001

SRwMT	*SRwAN*
20	20
10111100111111111111	11011111111111111111
01111100111111111111	01001101101111001101
00100000000001111111	01111111111111111111
00010000000001111111	00011101101111001101
00001000000011111111	00001000001010000101
00000100000001111111	00000100001010000101
00111110111111111111	00001111101111001101
00111101111111111111	00000001001010000101
00000000100011111111	00000000101010000101
00000000010011111111	00001001011110100111
00000000001011111111	00000000001000000000
00000000000111111111	00000000001110000101
00000000000010000101	00000000000010000000
00000000000001000101	00000000001011000101
00000000000000100101	00000000001010100111
00000000000000010101	00000100101011111111
00000000000000001101	00000000001010001101
00000000000000000101	00000000000000000100
00000000000000000111	00000000000000000010
00000000000000000001	00000000000000000001

Relation files for Investigation III

SRbI	*SRxT*
20	20
10111000100100100011	10000000100100100011
01100100110000100011	01000100110000100011
00100000100000000011	00100000100000000011
00010000000000000001	00010000000000000001
00001000000000000001	00001000000000000001
00100100110000100011	00100100000000100011
00100010100000100011	00100010000000100011
11111111111111111111	11111001001111111111
00000000100000000001	00000000100000000001
00000000010000000001	00000000010000000001
11111100111111101111	11111100111000001111
00000000000100000001	00000000000100000001
00011000010110000001	00011000010010000001
00100100110001100011	00100100110001000011
00000000100000100011	00000000100000100011
10111010100100110011	10111010100100110000
11111100110111101011	11111100110111101000
11111100110111100111	11111100110111100100
00000000000000000011	00000000000000000010
00000000000000000001	00000000000000000001

$SRwAN_V$	$SRwAN_G$	$SRwSC_N$	$SRwMT_G$
5	5	5	5
10111	10011	11111	10011
01100	01010	01000	01011
00100	11111	01100	00111
00010	00010	00011	00011
00001	00001	00001	00001

E.2 Base files

Base files for Investigation I

$KSbI$	$KSxT$
30	30
30	30
100000000000002020002000220002	100000000000002020002000220002
212020200000002022002000220002	010000000000002022002000220002
001000000000002020002000220002	001000000000002020002000220002
202120200000002022002000220002	000100000000002022002000220002
000010000000002020002000220002	000010000000002020002000220002
202021200000002020002000222022	000001000000002020002000222022
000000100000002020002000220002	000000100000002020002000220002
222222210202002022002202222022	000000010000002022002202222022
222222201202002022002202222022	000000001000002022002202222022
202020200100002022002000220002	000000000100002022002000220002
222222222212002022022202222222	000000000010002022022202222222
202020200001002022002000220002	000000000001002022002000220002
222222222202102022022202222222	000000000000102022022202222222
222222222202012222202222222022	000000000000012222202222222022
000000000000001000002000020000	000000000000001000000000000000
202020200000002120002000220002	202020200000000100000000000000
000000000000000010002000020000	000000000000000010000000000000
202020200000002021002000220002	202020200000000001000000000000
202022200000002220102000222022	202022200000000000100000000000
222222222202002022012202222022	222222222202000000010000000000
000000000000000000001000000000	000000000000000000001000000000
222222200202002022002100222022	222222200202000000000100000000
222220200202002222002010220002	222220200202000000000010000000
222222200202002022002001222022	222222200202000000000001000000
000000000000000000002000120000	000000000000000000000000100000
000000000000000000000000010000	000000000000000000000000010000
202020200000002020002000221002	202020200000000000000000001000
222222222202002022002202222122	222222222202000000000000000100
202020200000002020002000220012	202020200000000000000000000010
000000000000000000002000020001	000000000000000000000000000001

$KSwMT$	$KSwAN$
14	16
14	16
10000000000000	1000002000020000
21202020000000	2120002000220002
00100000000000	0010002000020000
20212020000000	2021002000220002
00001000000000	2220102000222022
20202120000000	2022012202222022
00000010000000	0000001000000000
22222221020200	2022002100222022
22222220120200	2222002010220002
20202020010000	2022002001222022
22222222221200	0000002000120000
20202020000100	0000000000010000
22222222220210	2020002000221002
22222222220201	2022002202222122
	2020002000220012
	0000002000020001

Base files for Investigation II

$KSbI$	$KSxT$
40	40
40	40
1000000000000000000020200000000000000000	1000000000000000000020200000000000000000
0100000000000000000020200000000000000000	0100000000000000000020200000000000000000
2210002200000000000020200000020000000000	0010000000000000000020200000020000000000
2201002200000000000020200000020000000000	0001000000000000000020200000020000000000
2200102200000000000020200000000000020000	0000100000000000000020200000000000020000
2200012200000000000020200000020000000000	0000010000000000000020200000020000000000
0000001000000000000020200000000000000000	0000001000000000000020200000000000000000
0000000100000000000020200000000000000000	0000000100000000000020200000000000000000
2200002210000000000020200000000000020000	0000000010000000000020200000000000020000
2200002201000000000020200000000000020000	0000000001000000000020200000000000020000
2200002200100000000020200000000000020000	0000000000100000000020200000000000020000
2200002200010000000020200000000000020000	0000000000010000000020200000000000020000
2200202222221000000022220220200002022000	0000000000001000000022220220200002022000
2222222222220100000020200000020000220000	0000000000000100000020200000020000220000
2222222222220010000020200000020000220000	0000000000000010000020200000020000220000
2222222222220001000020200000020000220000	0000000000000001000020200000020000220000
2222222222220000100020200000020000220000	0000000000000000100020200000020000220000
2222222222222222212022222222222222222202	0000000000000000010022222222222222222202
2222222222220000001020200000020000220000	0000000000000000001020200000020000220000
2222222222222222222122222222222222222222	0000000000000000000122222222222222222222
0000000000000000000010000000000000000000	0000000000000000000010000000000000000000
2200002200000000000021200000000000000000	2200002200000000000001000000000000000000
0000000000000000000000100000000000000000	0000000000000000000000100000000000000000
2200002200000000000020210000000000000000	2200002200000000000000010000000000000000
2222022200000000000022221020020000000000	2222022200000000000000001000000000000000
2200202222220000000022220120000000020000	2200202222220000000000000100000000000000
2200002200000000000020200010000000000000	2200002200000000000000000010000000000000
2222022200000000000022220021020000000000	2222022200000000000000000001000000000000
2200202222220000000022220020100000020000	2200202222220000000000000000100000000000
2200002200000000000020200000010000000000	2200002200000000000000000000010000000000
2222222222222222202022222222221202222000	2222222222222222202000000000001000000000
2222022200000000000022220020020100000000	2222022200000000000000000000000100000000
2222222222222222202022222222220212222000	2222222222222222202000000000000010000000
2200202222220000000022220020000001020000	2200202222220000000000000000000001000000
2222222222220000000020200000020000120000	2222222222220000000000000000000000100000
2200002200000000000020200000000000010000	2200002200000000000000000000000000010000
2200202222220000000022220020000000021000	2200202222220000000000000000000000001000
2222222222222222202022222222220202222100	2222222222222222202000000000000000000100
2222222222220222202020200000020000220010	2222222222220222202000000000000000000010
2222222222222222202022222222220202222001	2222222222222222202000000000000000000001

$KSwMT$	$KSwAN$
20	20
20	20
10000000000000000000	10000000000000000000
01000000000000000000	21200000000000000000
22100022000000000000	00100000000000000000
22010022000000000000	20210000000000000000
22001022000000000000	22221020020000000000
22000122000000000000	22220120000000020000
00000010000000000000	20200010000000000000
00000001000000000000	22220021020000000000
22000022100000000000	22220020100000020000
22000022010000000000	20200000010000000000
22000022001000000000	22222222221202222000
22000022000100000000	22220020020100000000
22002022222210000000	22222222220212222000
22222222222201000000	22220020000001020000
22222222222200100000	20200000020000120000
22222222222200010000	20200000000000010000
22222222222200001000	22220020000000021000
22222222222222222120	22222222220202222100
22222222222200000010	20200000020000220010
22222222222222222221	22222222220202222001

Base files for Investigation III

$KSbI$	$KSxT$
20	20
20	20
10000002002000022200	10000002002000022200
01000002002000002200	01000002002000002200
22100222002002022200	00100222002002022200
20010002002020022200	00010002002020022200
20001002002020022200	00001002002020022200
02000102002002002200	02000100002002002200
00000012000000020000	00000010000000020000
00000001000000000000	00000001000000000000
22200222102002222200	22200000102002222200
02000202012022002200	02000000012022002200
00000002001000000000	00000002001000000000
20000002002120022200	20000002000100022200
00000002002010002200	00000002000010002200
00000002002001002200	00000002000001002200
22000222002002122200	22000222000000122200
00000002000000010000	00000002000000010000
00000002002000001000	00000002002000001000
00000002002000000100	00000002002000000100
22200222002002222210	22200222002002200010
22222222222222222221	22222222222222200001

$KSwAN_V$	$KSwAN_G$	$KSwSC_N$	$KSwMT_G$
5	5	5	5
5	5	5	5
10000	10200	10000	10000
01000	01200	21200	01000
22100	00100	20100	00100
20010	22210	20010	22210
20001	20201	20021	22221

Appendix F: Results

The data files for Investigations I and II are confidential and therefore only available after consultation with the HPD in Vienna or the PDB in Bonn. The data file for Investigation III is available at the Cognitive Science Section, Department of Psychology, University of Graz.

F.1 Relative solution frequencies

Table F.1: Relative solution frequencies for items in Investigation I (N = 572)

Matrices				Analogies					
1	97.73	8	82.17	15	87.76	22	84.44	29	88.46
2	93.71	9	81.47	16	87.41	23	80.42	30	83.39
3	97.55	10	90.91	17	77.80	24	89.51		
4	89.69	11	62.41	18	89.34	25	89.51		
5	99.48	12	79.02	19	87.06	26	90.91		
6	87.59	13	72.38	20	70.10	27	83.92		
7	95.10	14	59.79	21	89.86	28	70.98		

Table F.2: Relative solution frequencies for items in Investigation II (N = 2628)

Matrices								Analogies							
1	84.21	6	71.50	11	54.15	16	15.83	21	82.84	26	61.68	31	31.85	36	55.25
2	93.38	7	75.61	12	65.03	17	36.04	22	96.35	27	71.39	32	50.68	37	45.09
3	70.66	8	75.53	13	22.83	18	12.60	23	78.35	28	49.35	33	23.63	38	25.19
4	72.64	9	44.56	14	34.93	19	9.13	24	76.29	29	53.92	34	41.36	39	15.41
5	65.98	10	53.69	15	25.57	20	8.98	25	44.60	30	61.26	35	38.70	40	28.77

Table F.3: Relative solution frequencies for items in Investigation III (N = 121)

AN_V		AN_G		SC_N		MT_G	
1	61.16	6	62.81	11	93.39	16	92.56
2	62.81	7	76.86	12	65.29	17	100
3	53.72	8	90.08	13	72.73	18	87.60
4	25.62	9	52.89	14	74.38	19	24.79
5	46.28	10	67.77	15	58.68	20	16.53

F.2 χ^2 Statistics for reversed solution frequencies

With the α–adjustment for item classes involved in multiple χ^2 tests, the α–level should be reduced. However, because the model postulates that there are no significant differences, the α–adjustment would lead to a weaker test of the hypothesis and was therefore disregarded.

Table F.4: χ^2 Statistics for reversed solution frequencies in Investigation I

Item pair	Solution frequencies	χ^2
(O,1,H,V,5)/(O,2,H,V,5)	493/497	0.02, n.s.
(O,1,H,V,5)/(D,1,H,V,5)	493/498	0.03, n.s.
(O,1,L,V,5)/(O,1,H,V,5)	484/493	0.08, n.s.
(O,1,L,V,5)/(O,2,L,V,5)	484/511	0.73, n.s.
(O,1,L,V,5)/(D,1,L,V,5)	484/500	0.26, n.s.
(O,1,L,V,5)/(O,2,H,V,5)	484/497	0.19, n.s.
(O,1,L,V,5)/(D,1,H,V,5)	484/498	0.20, n.s.
(O,1,H,G,8)/(O,1,H,V,5)	501/493	0.06, n.s.
(O,1,H,G,8)/(O,1,L,V,5)	501/484	0.29, n.s.
(O,2,L,G,8)/(O,1,L,V,5)	505/484	0.46, n.s.
(O,1,L,G,8)/(O,1,L,V,5)	558/484	5.19, s.
(O,1,L,G,8)/(O,1,L,V,4)	558/517	1.53, n.s.

Note. For all χ^2 tests, *df* $= 1$; $\chi^2_{(\alpha=.05)} = 3.84$ and $\chi^2_{(\alpha=.01)} = 6.63$

Table F.5: χ^2 Statistics for reversed solution frequencies in Investigation II

Item pair	Solution frequencies	χ^2
(D,1,L,V,5)/(O,1,L,V,5)	2138/2118	0.09, n.s.
(O,1,L,G,8)/(O,1,L,V,5)	2160/2118	0.41, n.s.
(O,1,H,G,8)/(O,1,H,V,5)	1882/1610	21.19, s.s.
(O,2,L,G,8)/(O,2,L,V,5)	1490/1452	0.49, n.s.
(D,2,H,V,5)/(D,2,L,G,8)	719/600	10.74, s.s.
(D,2,H,V,5)/(O,2,H,G,8)	719/639	4.76, s.

Note. For all χ^2 tests, *df* $= 1$; $\chi^2_{(\alpha=.05)} = 3.84$ and $\chi^2_{(\alpha=.01)} = 6.63$

Table F.6: χ^2 Statistics for reversed solution frequencies in Investigation III

Item pair	Solution frequencies	χ^2
AN(D,3,L,G)/AN(D,2,L,G)	82/76	0.23, n.s.
AN(D,3,L,G)/AN(D,2,L,V)	82/76	0.23, n.s.
SC(O,3,H,N)/AN(O,2,H,V)	79/74	0.16, n.s.
SC(O,2,L,N)/AN(O,1,L,G)	113/109	0.07, n.s.
MT(O,2,L,G)/AN(O,1,L,G)	114/109	0.09, n.s.
MT(O,1,H,G)/AN(O,1,L,G)	112/109	0.04, n.s.
MT(O,2,L,G)/SC(O,2,L,N)	114/113	0.01, n.s.

Note. For all χ^2 tests, *df* $= 1$; $\chi^2_{(\alpha=.05)} = 3.84$ and $\chi^2_{(\alpha=.01)} = 6.63$

F.3 Distance distributions

The powersets (*dpot*) for surmise relations with 30 or 40 items are based on 20,000 random response patterns. For the probability simulations ($dsim_p$) the probabilities for lucky guesses correspond to the respective number of answer alternatives (e. g., $\eta = 0.2$ for five alternatives) and the probabilities for careless errors vary between $0.05 < \beta \leq 0.15$.

Table F.7: Distance distributions for the surmise relations between items and across tests in Investigation I

	SRbI					SRxT				
di	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$
0	53	0	0	22.07	55.04	81	1	0.02	82.35	90.39
1	101	0	0	73.57	102.99	169	27	0.74	165.47	159.87
2	117	2	0.03	121.33	108.58	166	169	4.85	161.56	163.26
3	109	7	0.28	132.07	99.28	102	755	21.65	100.34	101.93
4	72	60	1.23	104.74	78.39	42	2118	60.03	44.24	40.94
5	44	152	4.73	65.54	51.75	7	4081	116.94	14.23	12.19
6	33	500	14.01	32.84	33.61	4	5429	154.29	3.28	2.91
7	24	1130	33.82	13.60	20.70	0	4497	129.95	0.47	0.46
8	7	2163	62.47	4.55	11.27	1	2287	65.07	0.06	0.06
9	5	3288	94.68	1.33	5.86	0	585	16.82	0.01	0.01
10	3	4029	113.39	0.29	2.67	0	50	1.62	0	0
11	3	3905	110.10	0.07	1.17	0	1	0.02	0	0
12	1	2868	82.10	0.01	0.35	0	0	0	0	0
13	0	1453	42.42	0	0.05	0	0	0	0	0
14	0	422	12.01	0	0.01	0	0	0	0	0
15	0	21	0.72	0	0	0	0	0	0	0

Note. The values for $dsim_r$, $dsim_p$, and $dsim_f$ denote the averaged distances derived from 1000 data sets each (N = 572 per data set).

Table F.8: Distance distributions for the surmise relations within tests in Investigation I

	Matrices					Analogies				
di	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$
0	260	57	2.06	181.88	253.39	166	71	0.58	112.16	164.83
1	187	506	17.77	222.53	188.97	182	743	6.53	186.00	178.07
2	89	1938	67.73	122.00	97.45	114	3474	30.79	153.50	118.48
3	30	4087	143.02	37.65	27.53	61	9494	82.87	81.68	63.07
4	5	5068	175.89	7.12	4.33	30	16568	144.17	29.79	29.98
5	1	3619	127.00	0.79	0.32	13	18620	162.96	7.60	13.12
6	0	1074	37.48	0.04	0.01	6	12508	108.34	1.16	3.92
7	0	35	1.12	0	0	0	3879	34.19	0.11	0.52
8						0	179	1.58	0	0.02

Note. The values for $dsim_r$, $dsim_p$, and $dsim_f$ denote the averaged distances derived from 1000 data sets each (N = 572 per data set).

Table F.9: Distance distributions for the surmise relations between items and across tests in Investigation II

	SRbI				SRxT			
di	*ddat*	*dpot*	$dsim_r$	$dsim_p$	*ddat*	*dpot*	$dsim_r$	$dsim_p$
0	4	0	0	63.34	7	0	0	247.68
1	23	0	0	237.75	51	0	0.12	563.31
2	95	0	0	441.10	219	14	0.24	667.07
3	196	1	0.08	553.77	436	54	4.24	545.23
4	371	4	0.44	518.34	545	221	30.66	340.04
5	407	13	1.57	383.75	538	788	101.20	169.39
6	450	53	6.82	232.70	427	1929	245.78	67.50
7	396	136	20.44	117.33	250	3423	462.76	21.46
8	298	373	50.44	50.77	108	4659	616.63	5.28
9	207	851	113.53	18.37	40	4585	602.43	0.91
10	104	1534	201.60	5.67	6	2928	381.63	0.10
11	50	2327	305.19	1.58	1	1155	146.58	0.01
12	20	2950	398.10	3.05	0	233	29.37	0
13	4	3402	439.74	0.19	0	8	2.31	0
14	2	3155	404.48	0.09	0	3	0.05	0
15	0	2367	319.49	0.06	0	0	0	0
16	1	1604	206.66	0.04	0	0	0	0
17	0	854	108.27	0.04	0	0	0	0
18	0	284	39.91	0.03	0	0	0	0
19	0	83	10.34	0.03	0	0	0	0
20	0	9	0.91	0.01	0	0	0	0

Note. The values for $dsim_r$ and $dsim_p$ denote the averaged distances derived from 1000 data sets each (N = 2628 per data set).

Table F.10: Distance distributions for the surmise relations within tests in Investigation II

	Matrices				Analogies			
di	*ddat*	*dpot*	$dsim_r$	$dsim_p$	*ddat*	*dpot*	$dsim_r$	$dsim_p$
0	376	343	0.94	572.27	141	484	1.42	435.12
1	809	4188	10.08	925.03	522	5532	14.08	838.35
2	763	23118	58.86	692.87	742	28570	72.80	745.43
3	438	75807	190.52	315.06	678	87831	217.63	405.07
4	170	163443	408.97	96.25	376	178174	447.94	152.49
5	59	242350	607.73	20.94	133	249633	625.18	41.69
6	11	250902	623.11	3.19	31	244653	614.20	8.45
7	2	179511	455.90	0.35	5	164766	411.71	1.25
8	0	84672	211.95	0.03	0	71678	179.24	0.13
9	0	22772	56.66	0.01	0	16529	42.08	0.01
10	0	1470	3.29	0.01	0	726	1.73	0

Note. The values for $dsim_r$ and $dsim_p$ denote the averaged distances derived from 1000 data sets each (N = 2628 per data set).

Table F.11: Distance distributions for the surmise relations between items and across tests in Investigation III

			SRbI					SRxT		
di	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$
0	13	255	0.02	17.40	9.83	15	484	0.04	18.10	
1	20	3222	0.46	33.68	25.33	28	5810	0.74	36.54	11.91
2	37	18555	2.14	34.38	35.26	37	31154	3.63	35.10	31.37
3	31	64135	7.65	21.42	29.06	28	98178	11.44	20.34	37.75
4	12	147361	16.94	9.34	15.26	7	200524	23.20	8.10	25.57
5	5	234974	26.64	3.33	5.01	5	274744	31.38	2.32	10.93
6	2	262215	30.13	1.13	1.12	1	250847	28.77	0.43	2.89
7	1	199331	23.13	0.32	0.13	0	143689	16.87	0.06	0.53
8	0	94605	11.00	0.02	0.01	0	40800	4.69	0.01	0.05
9	0	22685	2.77	0	0.01	0	2346	0.24	0	0.01
10	0	1238	0.14	0	0	0	0	0	0	

Note. The values for $dsim_r$, $dsim_p$, and $dsim_f$ denote the averaged distances derived from 1000 data sets each (N = 121 per data set).

Table F.12: Distance distributions for the surmise relations within tests in Investigation III

			AN_V					AN_G		
di	*ddat*	*dpot*	$dsim_r$	$dsim_p$		*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$
0	81	14	52.92	96.13		71	9	34.15	84.58	71.72
1	38	15	56.72	23.63		43	16	60.40	32.57	42.94
2	2	3	11.54	1.24		7	7	26.46	3.85	6.34
			SC_N					MT_G		
di	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$	*ddat*	*dpot*	$dsim_r$	$dsim_p$	$dsim_f$
0	92	10	37.89	86.91	91.12	104	10	38.06	81.95	0.03
1	28	16	60.55	30.92	27.02	17	18	67.81	37.06	17.91
2	1	6	22.55	3.17	2.86	0	4	15.13	1.99	103.06

Note. The values for $dsim_r$, $dsim_p$, and $dsim_f$ denote the averaged distances derived from 1000 data sets each (N = 121 per data set).

F.4 Invalidity of items

Table F.13: Invalidity of items in Investigation I

Item	Invalidity		Guesses		Errors	
	abs.	rel.	abs.	rel.	abs.	rel.
1	11.50	0.020	0.00	0.000	11.50	0.020
2	30.50	0.053	0.50	0.001	30.00	0.052
3	13.50	0.024	0.50	0.001	13.00	0.023
4	49.50	0.087	0.50	0.001	49.00	0.086
5	2.50	0.004	0.50	0.001	2.00	0.003
6	60.17	0.105	5.00	0.009	55.17	0.096
7	26.50	0.046	0.00	0.000	26.50	0.046
8	78.37	0.137	12.90	0.023	65.47	0.114
9	75.33	0.132	12.83	0.022	62.50	0.109
10	39.80	0.070	0.00	0.000	39.80	0.070
11	47.77	0.084	47.77	0.084	0.00	0.000
12	108.80	0.190	0.00	0.000	108.80	0.190
13	57.58	0.101	57.58	0.101	0.00	0.000
14	56.80	0.099	56.80	0.099	0.00	0.000
15	68.50	0.120	0.00	0.000	68.50	0.120
16	53.17	0.093	0.50	0.001	52.67	0.092
17	125.50	0.219	0.00	0.000	125.50	0.219
18	60.50	0.106	0.50	0.001	60.00	0.105
19	23.83	0.042	13.17	0.023	10.67	0.019
20	94.03	0.164	18.53	0.032	75.50	0.132
21	57.00	0.100	0.00	0.000	57.00	0.100
22	74.87	0.131	9.20	0.016	65.67	0.115
23	25.73	0.045	7.37	0.013	18.37	0.032
24	49.43	0.086	9.33	0.016	40.10	0.070
25	58.50	0.102	0.00	0.000	58.50	0.102
26	50.50	0.088	0.00	0.000	50.50	0.088
27	79.50	0.139	0.50	0.001	79.00	0.138
28	85.82	0.150	18.17	0.032	67.65	0.118
29	53.00	0.093	0.00	0.000	53.00	0.093
30	93.50	0.163	0.00	0.000	93.50	0.163
M	57.067	0.100	9.072	0.016	47.996	0.084
SD	28.705	0.050	16.385	0.029	32.344	0.056

Table F.14: Invalidity of items in Investigation II

Item	Invalidity		Guesses		Errors	
	abs.	rel.	abs.	rel.	abs.	rel.
1	346.58	0.132	2.50	0.001	344.08	0.131
2	142.08	0.054	15.83	0.006	126.25	0.048
3	289.70	0.110	35.64	0.014	254.06	0.097
4	263.07	0.100	42.83	0.016	220.24	0.084
5	285.48	0.109	39.65	0.015	245.83	0.094
6	300.25	0.114	40.46	0.015	259.79	0.099
7	573.55	0.218	5.00	0.005	509.26	0.194
10	474.32	0.180	38.68	0.015	435.63	0.166
11	372.47	0.142	14.95	0.002	568.55	0.216
8	576.63	0.219	3.50	0.001	573.13	0.218
9	522.43	0.199	13.17	0.005	509.26	0.194
10	474.32	0.180	38.68	0.015	435.63	0.166
11	372.47	0.142	14.95	0.006	357.52	0.136
12	290.91	0.111	45.62	0.017	245.30	0.093
13	459.92	0.175	436.09	0.166	23.83	0.009
14	326.07	0.124	307.48	0.117	18.58	0.007
15	186.70	0.071	177.12	0.067	9.58	0.004
16	164.30	0.063	135.13	0.051	29.17	0.011
17	386.46	0.147	365.13	0.139	21.33	0.008
18	316.42	0.120	315.42	0.120	1.00	0.000
19	133.69	0.051	86.86	0.033	46.83	0.018
20	229.58	0.087	229.58	0.087	0.00	0.000
21	411.17	0.156	0.00	0.000	411.17	0.156
22	79.88	0.030	57.62	0.022	22.27	0.008
23	523.50	0.199	0.00	0.000	523.50	0.199
24	300.96	0.115	34.12	0.013	266.85	0.102
25	169.92	0.065	166.59	0.063	3.33	0.001
26	557.22	0.212	542.22	0.206	15.00	0.006
27	226.83	0.086	18.58	0.007	208.25	0.079
28	208.39	0.079	206.89	0.079	1.50	0.001
29	470.71	0.179	442.68	0.168	28.02	0.011
30	872.58	0.332	23.67	0.009	848.91	0.323
31	783.75	0.298	783.17	0.298	0.58	0.000
32	226.70	0.086	222.86	0.085	3.83	0.001
33	568.25	0.216	566.25	0.215	2.00	0.001
34	266.75	0.102	237.42	0.090	29.33	0.011
35	583.49	0.222	220.52	0.084	362.97	0.138
36	975.10	0.371	13.00	0.005	962.10	0.366
37	408.78	0.156	372.67	0.142	36.11	0.014
38	606.50	0.231	604.17	0.230	2.33	0.001
39	360.42	0.137	358.75	0.137	1.67	0.001
40	723.50	0.275	717.58	0.273	5.92	0.002
M	399.125	0.152	198.485	0.075	200.640	0.076
SD	207.062	0.079	218.644	0.083	247.959	0.094

Table F.15: Invalidity of items in Investigation III

Item	Invalidity		Guesses		Errors	
	abs.	rel.	abs.	rel.	abs.	rel.
1	30.25	0.250	1.50	0.012	28.75	0.238
2	27.33	0.226	0.50	0.004	26.83	0.222
3	15.17	0.125	7.92	0.065	7.25	0.060
4	2.33	0.019	1.83	0.015	0.50	0.004
5	3.25	0.027	2.25	0.019	1.00	0.008
6	26.08	0.216	5.33	0.044	20.75	0.171
7	9.58	0.079	0.83	0.007	8.75	0.072
8	12.00	0.099	0.00	0.000	12.00	0.099
9	29.75	0.246	29.75	0.246	0.00	0.000
10	16.92	0.140	16.42	0.136	0.50	0.004
11	7.25	0.060	0.00	0.000	7.25	0.060
12	6.08	0.050	6.08	0.050	0.00	0.000
13	17.17	0.142	0.00	0.000	17.17	0.142
14	15.42	0.127	0.00	0.000	15.42	0.127
15	17.00	0.140	10.25	0.085	6.75	0.056
16	4.67	0.039	0.00	0.000	4.67	0.039
17	0.75	0.006	0.75	0.006	0.00	0.000
18	12.50	0.103	0.00	0.000	12.50	0.103
19	7.50	0.062	7.50	0.062	0.00	0.000
20	18.00	0.149	18.00	0.149	0.00	0.000
M	13.950	0.115	5.446	0.045	8.505	0.070
SD	9.100	0.075	7.905	0.065	9.216	0.076

F.5 χ^2 Statistics for the differences between the empirical and the powersets' distance distributions

Investigation I

$$\begin{aligned}
\chi^2_{KSbI}(9, N = 1144) &= 38,446.09; p < .001\\
\chi^2_{KSxT}(7, N = 1144) &= 30,654.76; p < .001\\
\chi^2_{KSwMT}(5, N = 1144) &= 9,717.07; p < .001\\
\chi^2_{KSwAN}(6, N = 1144) &= 16,954.93; p < .001
\end{aligned}$$

Investigation II

$$\begin{aligned}
\chi^2_{KSbI}(13, N = 5256) &= 264,596.99;\ p < .001\\
\chi^2_{KSxT}(9, N = 5256) &= 68,170.92;\ p < .001\\
\chi^2_{KSwMT}(8, N = 5256) &= 132,162.66;\ p < .001\\
\chi^2_{KSwAN}(8, N = 5256) &= 36,649.51;\ p < .001
\end{aligned}$$

Investigation III

$$\begin{aligned}
\chi^2_{KSbI}(6, N = 242) &= 1945.75;\ p < .001\\
\chi^2_{KSxT}(5, N = 242) &= 627.22;\ p < .001\\
\chi^2_{KSwAN_V}(2, N = 242) &= 28.75;\ p < .001\\
\chi^2_{KSwAN_G}(2, N = 242) &= 59.54;\ p < .001\\
\chi^2_{KSwSC_N}(2, N = 242) &= 115.84;\ p < .001\\
\chi^2_{KSwMT_G}(2, N = 242) &= 169.29;\ p < .001
\end{aligned}$$