

Westfälische  
Wilhelms-Universität  
Münster  
Fachgebiet: Psychologie

REASONING ABILITY:  
RULE-BASED TEST CONSTRUCTION OF A FIGURAL  
ANALOGY TEST

Inaugural-Dissertation  
zur Erlangung des Doktorgrades  
der  
Philosophischen Fakultät  
der  
Westfälischen Wilhelms-Universität  
zu  
Münster (Westf.)

vorgelegt von  
**BARBARA MARIA ESTHER BECKMANN**  
aus Münster

Juli 2008

Tag der mündlichen Prüfung: 26.09.2008

Dekan: Prof. Dr. Bromme

Referent: Prof. Dr. Holling

Korreferent: PD Dr. Gediga



## Vorwort

Im Rahmen der vorliegenden Arbeit wird das Ziel der regelgeleiteten Testentwicklung zur Erfassung der fluiden Intelligenz mittels figuralen Analogieaufgaben verfolgt. Schlussfolgerndes Denken wurde als Testgegenstand gewählt, da es als zentrale Komponente kognitiver Fähigkeiten und analytischer Intelligenz gilt. Der Einfluss der im Test angewandten Regeln und Elemente auf die Aufgabenschwierigkeit wird mittels Linear Logistischer Testmodelle untersucht.

Testinstruktionen und Beispielaufgaben des im Rahmen dieser Dissertation entwickelten figuralen Analogietests werden im Anhang aufgeführt. Da die Arbeit in englischer Sprache geschrieben ist, befindet sich eine deutsche Zusammenfassung am Ende der Arbeit.

Mein herzlicher Dank gilt Herrn Prof. Dr. Holling für die Betreuung meiner Arbeit, seine Unterstützung und Anregungen. Herrn PD Dr. Gediga danke ich für seine Bereitschaft, die Aufgabe des zweiten Gutachters zu übernehmen. Meinen Kollegen des Lehrstuhls gebührt ebenfalls großer Dank für ihre Unterstützung und die beständige und gute Zusammenarbeit.

Münster, im Juli 2008

Barbara Beckmann

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Reasoning Ability: Theoretical and Empirical Background</b>	<b>7</b>
2.1	Reasoning Ability: Assignment and Definition . . . . .	7
2.2	Reasoning Ability in Influential Intelligence Models . . . . .	9
2.3	Information-Processing Theories of Analogy Tasks . . . . .	16
2.4	Empirical Findings between Cognitive Abilities and Academic Achievement . . . . .	35
2.5	Effects of Training and Instruction on Test Performance . . . . .	47
2.5.1	Training Effects . . . . .	47
2.5.2	Effects of Instruction on Test Performance . . . . .	51
2.5.3	Concluding Remarks on the Effects of Training and Instruction . . . . .	56
<b>3</b>	<b>Analysis of Cognitive Structure</b>	<b>58</b>
3.1	Rule-based Test Construction . . . . .	59
3.2	Test Theory . . . . .	61

## *Contents*

---

3.2.1	Classical Test Theory and Item Response Theory . . .	62
3.2.2	Probabilistic Test Models . . . . .	64
3.3	Methods of Validating the Cognitive Structure . . . . .	67
3.3.1	Linear Logistic Test Model . . . . .	68
3.3.2	Linear Logistic Test Model with Random Item Effects	71
3.3.3	Optimal Design . . . . .	73
3.4	Empirical Findings on Information Structure and Components of Item Difficulty . . . . .	74
3.5	Summary . . . . .	82
<b>4</b>	<b>Research Questions and Pretests</b>	<b>83</b>
4.1	Research Questions . . . . .	84
4.2	Pretests . . . . .	88
4.2.1	Pretest 1 . . . . .	91
4.2.2	Pretest 2 . . . . .	95
4.2.3	Pretest 3 . . . . .	99
4.2.4	Summary and Conclusion on Pretests . . . . .	104
<b>5</b>	<b>Main Examination</b>	<b>106</b>
5.1	Material . . . . .	107
5.1.1	Test and Items . . . . .	107
5.1.2	Distractors . . . . .	109
5.2	Instruction, Procedure, and Measures . . . . .	112
5.3	Sample . . . . .	113
5.4	Results . . . . .	115
5.4.1	Item Statistics According to Classical Test Theory . . .	115

*Contents*

---

5.4.2	Confirmatory Factor Analysis . . . . .	123
5.4.3	Estimating Parameters According to IRT-Models . . .	124
5.4.4	Goodness of Fit . . . . .	132
5.4.5	Item Construction in Focus . . . . .	140
5.4.6	Impact of Elements on Item Difficulty . . . . .	153
<b>6</b>	<b>Summary and Discussion</b>	<b>160</b>
6.1	Future Prospects . . . . .	171
	<b>References</b>	<b>173</b>
<b>A</b>	<b>Tables and Figures</b>	<b>184</b>
<b>B</b>	<b>Test Materials</b>	<b>188</b>

# List of Tables

2.1	Analogy Example According to Sternberg (1977a) . . . . .	21
4.1	Descriptive Statistics of Pretest 1 . . . . .	94
4.2	Parameter Estimates of Pretest 1 . . . . .	95
4.3	Descriptive Statistics of Pretest 2 . . . . .	97
4.4	Parameter Estimates of Pretest 2 . . . . .	98
4.5	Descriptive Statistics of Pretest 3 . . . . .	102
4.6	Parameter Estimates of Pretest 3 (Subtests 1, 4, 5) . . . . .	103
4.7	Parameter Estimates of Pretest 3 (Subtests 2, 3) . . . . .	103
5.1	Cognitive Operations . . . . .	108
5.2	Optimal Design Matrix (Subtests 1 & 2) . . . . .	110
5.3	Optimal Design Matrix (Subtests 3 & 4) . . . . .	111
5.4	T-Test Means of the Instruction Group and the Non-Instruction Group . . . . .	118
5.5	Descriptive Statistics: Total Sample . . . . .	119
5.6	Descriptive Statistics: Instruction Group . . . . .	120



*List of Tables*

---

5.7	Descriptive Statistics: Non-Instruction Group . . . . .	121
5.8	Item Parameters 1PL Model (BILOG) . . . . .	129
5.9	Item Parameters 2PL Model (BILOG) . . . . .	130
5.10	Goodness of Fit According to Likelihood and Information- theoretic Criteria . . . . .	133
5.11	Model Test According to Martin-Löf ( $k = 31, N = 484$ ) . . . . .	135
5.12	Model Test According to Martin-Löf ( $k = 44, N = 484$ ) . . . . .	136
5.13	Model Test According to Andersen ( $k = 31, N = 484$ ) . . . . .	138
5.14	Model Test According to Andersen ( $k = 44, N = 484$ ) . . . . .	139
5.15	LLTM vs. Rasch Parameter: Total Sample (LPCM-Win) . . . . .	142
5.16	LLTM: Basic Parameter Estimates, Total Sample (LPCM-Win)	143
5.17	LLTM: Basic Parameter Estimates, Instruction Group (LPCM- Win) . . . . .	144
5.18	LLTM vs. Rasch Parameter: Instruction Group (LPCM-Win) .	145
5.19	LLTM: Basic Parameter Estimates, Non-Instruction Group (LPCM-Win) . . . . .	146
5.20	LLTM vs. Rasch Parameter: Non-Instruction Group (LPCM- Win) . . . . .	147
5.21	Comparison of Parameter Estimates: Instruction Group vs. Non-Instruction Group . . . . .	148
5.22	LLTM with Random Item Effects: Basic Parameter Estimates, Total Sample (SAS) . . . . .	150
5.23	LLTM: Total Sample (SAS) . . . . .	151
5.24	LLTM with Random Item Effects for Rasch Conform Items ( $k = 32, N = 484$ ) (SAS) . . . . .	152

*List of Tables*

---

5.25 LLTM with Random Item Effects: Parsimonious Model for Rasch Conform Items ( $k = 32, N = 484$ ) (SAS) . . . . .	153
5.26 LLTM: Combination of Letters and Digits (LPCM-Win) . . . . .	154
5.27 LLTM: Orientation of A-Elements (LPCM-Win) . . . . .	155
5.28 LLTM: C-Elements (LPCM-Win) . . . . .	157
5.29 LLTM: Elements A & C (LPCM-Win) . . . . .	157
5.30 LLTM: Elements and Transformations (LPCM-Win) . . . . .	159
A.1 LLTM with Random Item Effects: Instruction Group (SAS) . . . . .	185
A.2 LLTM: Instruction Group (SAS) . . . . .	185
A.3 LLTM with Random Item Effects: Non-Instruction Group (SAS) . . . . .	186
A.4 LLTM: Non-Instruction Group (SAS) . . . . .	186

# List of Figures

2.1	Component Processing Times According to Sternberg (1977b)	25
2.2	Model of the Geometric Analogy Solution Process (Mulhol- land et al., 1980) . . . . .	31
2.3	Intelligence-Factor Model According to Rindermann & Neubauer (2004) . . . . .	44
4.1	Cognitive Operations . . . . .	90
5.1	Distribution of the Sum Scores of the Figural Analogy Test . .	116
5.2	Q-Q Diagram: Testing the Sum Scores of the Figural Analogy Test for Normal Distribution . . . . .	117
5.3	ICC Item 1.3 . . . . .	128
5.4	ICC Item 4.9 . . . . .	128
5.5	Test Information for the 1PL Model . . . . .	131
5.6	Test Information for the 2PL Model . . . . .	132
A.1	Item Characteristic Curves of the Figural Analogy Test Items	187

# 1

## **Introduction**

The present study deals with reasoning ability and intelligence. Intelligence is a human ability vitally important at school, work, and day-to-day life as it provides a basis for all cognitive abilities and skills. Reasoning ability constitutes a central construct in numerous theories on human intelligence. It is further is a fundamental part of fluid intelligence and of particular importance to new situations: The ability to reason is indispensable when problem solving skills are required, thus in situations in which experienced operations and algorithms for problem solution are not avail-

able or cannot be retrieved. Without reasoning, already acquired knowledge and experiences could not be applied to new situations.

The topic of the dissertation refers to the construction of a rule-based figural analogy test. This study is part of an extensive research project of the chair of statistics and quantitative methods of the psychological institute of the University of Münster, Germany. This research project engages in rule-based test construction and development of computer-generated adaptive tests capturing various cognitive abilities, in particular reasoning. In the course of this project, rule-based test construction and computerized generation algorithms had been successfully implemented for figural matrix items (Freund, Hofer, & Holling, 2008). In line with the addressed project, this study issues another type of reasoning task, the figural analogy, and accomplishes exploratory research to the rule-based test construction with the superior, long-term aim of an adaptive computer-generated test version. Besides the potential of identifying item components and their difficulty parameters, rule-based tests enable targeted practicing of the rules or transformations. Practicing enables equal a priori test conditions for subjects and hence differences in test scores can be almost exclusively ascribed to differences in ability and not to differences in test familiarity. This study chooses and applies an adequate experimental design to elucidate whether the knowledge of rules has impact on test scores.

The following objectives are pursued with the present research. This study aims to derive and provide psychometric and content fundamentals for the development of a test measuring reasoning ability. Analogy tasks of figural content are chosen for subsequent account: In literature reasoning

tasks hold a status as valid indicator of general intelligence. Contentwise, geometric or figural analogies constitute culture-fair tasks because they enable measurement of intellectual abilities regardless of environmental influences and cultural background of the subject tested. Thus, with applying a reasoning test, measurement of this important component of analytic intelligence is pursued as well as further access to fundamental human thought-processes and its structure.

However, the new test must account for complex exigencies and criteria. Scaling according to probabilistic test theories must be accomplished since the fit with the unidimensional model of this theory displays a condition precedent for item component analysis. Item response theory is chosen as a basis for characterizing these psychometric properties since it illustrates the relationship between the item response of the examinee and the ability score, thus the extent of the latent trait to be captured. Means of identifying item components provide an important tool to predict item difficulties. This is necessary in order to construct test items of particular difficulty to precisely measure different ability levels. With regard to the aptitude level of the subject, only test items of adequate difficulty and thus a small measurement error can provide sufficient information. Knowing the item components and their impact on item difficulty is inalienable, for example, in the instance of adaptive or tailored testing.

The study provides the following structure: Subsequently to this introduction the theoretical and empirical background of reasoning in general and of analogical reasoning in particular is presented in chapter 2. The chapter

is composed of different sections, starting with an assignment and definition of reasoning in section 1. Section 2 illustrates why, considering the wide range of intelligence research, engaging in reasoning ability in particular. Scientific debates on the property and measurement of reasoning are motivated by factor-analytic findings and numerous theories about the relationship between this ability and what is generally understood as intelligence. Factor-analytic approaches proved that reasoning ability is a central constituent of intelligence. How different theories integrate reasoning into models of intelligence is analyzed and presented in this section. Section 3 of the second chapter outlines information-processing theories of analogical reasoning. Different theories on the components involved in the solution process of analogy tasks are reviewed to reveal what constitutes reasoning. Section 4 of the second chapter refers to the findings between reasoning ability and academic achievement. Reasoning tests, with their property as indicator of general intelligence, have become popular among entrance and recruiting tests, thus empirical evidence concerning their predictive validity for academic achievement seems appropriate and is outlined. The last section of chapter 2 presents effects of training and instruction on test performance.

Chapter 3 analyzes the cognitive structures in analogical reasoning tasks. An introduction to rule-based test construction and analysis of its advantages and practical implications is presented at the beginning of this chapter. Subsequently, before methods of validating the cognitive structure are outlined, a short review on modern probabilistic test theory as opposed to classical test theory is given in section 2. Understanding of item response

theory in general, and of the 1-parameter logistic model in particular, is important for the linear logistic test model as means of analyzing and validating cognitive structures presented in section 3. Besides the linear logistic test model, the linear logistic test model including random item effects is introduced as means to examine the parameters that constitute item difficulty. The linear logistic test model with random effects accounts for additional item variation and refers to models with both random person and random item effects. This model presents a very realistic approach to validate the cognitive structure. Finally, empirical findings on information structure and components of analogy tasks are presented in the fourth section of chapter 3 before findings are summarized.

Chapter 4 addresses the research topics of the study and presents the three pretests that were conducted before the final test version for the main examination was decided on and composed.

Subsequently, chapter 5 presents the main examination of the dissertation. Before the test material is presented in section 2, the hypotheses of the study are specified in the beginning of the chapter (section 1). Then the test instruction, the testing procedure (section 3), and the sample (section 4) of the main examination are described. Section 5 investigates the results in the following order: First, data is analyzed according to classical test theory, referring to item statistics, test reliability and test validity. Confirmatory factor analysis is then conducted and provided. Further, parameters are estimated according to item response theory; tests to estimate the goodness of model fit are presented. By focusing on item construction a main topic of the study is addressed: The linear logistic test model and the lin-



ear logistic test model with random item effects are applied as methods to test predefined hypotheses on parameters influencing item difficulty. The impact of the transformations as well as the impact of the elements on the difficulty of the figural analogy tasks are further explored.

The sixth chapter summarizes and discusses preceding results and prepares future prospects for further examinations and test applications.

Summarized, the present study provides the design of new rule-based reasoning tasks, aiming to meet the demand for methodologically sound tasks. It is beyond doubt that reasoning is a central construct in human intelligence and necessary to accomplish easy and complex real life cognitive tasks. It is also assumed to also hold predictive validity for academic and educational performances. However, former or classic reasoning tasks are mostly scalable according to classical test theory only, with its - to some extent deficient - assumptions of ability measurement. Therefore, scalability to item response theory is aimed for.

# 2

## **Reasoning Ability: Theoretical and Empirical Background**

### **2.1 Reasoning Ability: Assignment and Definition**

Reasoning is essential in everyday life. When we have to make a decision in a surrounding that is new to us or when the decision refers to content that is unknown, we tend to relate to similar past experiences to find an answer. To simply illustrate what constitutes everyday reasoning, one might refer

to Sternberg (1977a):

We reason analogically whenever we make a decision about something new in our experience by drawing a parallel to something old in our experience. When we buy a new pet hamster because we liked our old one or when we listen to a friend's advice because it was correct once before, we are reasoning analogically. (p. 99)

Different modes of reasoning can be distinguished. Carroll (1989) defined three kinds of reasoning: deductive, inductive, and quantitative (inductive and deductive mathematical reasoning) reasoning. In logic, deduction means inferring or deducing specific statements from premises that have general character. A certain statement is therefore inferred from one or several preceding statements, namely from the general to the individual case. If the premises are concerned with true statements, the conclusion inferred from these statements must also be true if the principle of deduction had correctly been applied. It is therefore a matter of syllogisms which means deductive argumentation that draws a logical conclusion from two premises. Each premise shares a term with the other premise and one with the conclusion. Deduction is fulfilled via rules of reasoning. Three kinds of deductive reasoning can be distinguished: implications from a more general statement to a less general statement, from one generality to the same generality, and the implication from the specific to the particular. Induction refers to a procedure of inferring from the particular to generality and recognizing principles. It is about concluding from a basic set to theories. Induction, therefore, plays an important role in testing hypotheses and co-

herence. Induction thus constitutes an established procedure in science as various scientific disciplines attempt to conclude, for example, from samples to the total population. However, in contrast to deductive reasoning, the premises in inductive reasoning do not inevitably lead to the right conclusion since hypotheses are only tested. Thus, deductive and inductive reasoning differ in the certainty of a right conclusion. In deductive reasoning a true premise leads to a true conclusion, but in inductive reasoning true conclusions cannot necessarily be derived from the premise.

## **2.2 Reasoning Ability in Influential Intelligence**

### **Models**

Models of intelligence vary in the structural concept they assume for the construct of intelligence. By means of factor analytic approaches researchers developed different models with different implications in terms of the constitution of intelligence. In the meantime some of these models have evolved to classic and well-known intelligence models that are different in quantity, content, generality or specificity of factors. Furthermore, these models differ in the acceptance or non-acceptance of whether a general factor of intelligence. It is further distinguished between hierarchical models and non-hierarchical models. Hierarchical models are models in which factors represent a certain arrangement and ranking and thus differ in their extent of generality. Non-hierarchical models assume the same generality for all factors. Below, some classical models of intelligence are elucidated in terms of the reasoning component implied in the model. The follow-

ing paragraphs thus do not provide an exhaustive description of classical intelligence models but aim to illustrate the role of reasoning in selected models.

### **Spearman's General Factor Theory**

Spearman's *g*-factor in his *General Factor Theory* or the *Theory of two factors* (Spearman, 1904) is connected with reasoning ability: Since most tests that proved loadings on *g* implied abstract reasoning, he concluded that *g* was an essential part of human intelligence. Spearman documented empirical support for his thesis (Spearman, 1927) by presenting a correlation of .84 of Otis (1918) between tests containing analogies and *g*.

Raven (1936) constructed his first figural matrices test, the Standard Progressive Matrices, according to Spearman's theory, designing the tasks in order to capture the factors of intelligence defined by Spearman.

### **Thurstone's Primary Factor Model**

Thurstone (1938), in his *Primary Factor Model*, interpreted one of his seven primary abilities as reasoning and detecting rules. He defined induction as rule finding. In Thurstone's model, according to Wilhelm (2005), "the reasoning factor is marked mostly by inductive tasks" (p. 377). Thurstone reported factor loadings of .39 for pattern analogies on the inductive reasoning factor and of .60 for verbal analogies on a verbal factor.

### **Cattell's Fluid and Crystallized Intelligence**

Cattell (1971) assumed that intelligence is composed of two factors, fluid and crystallized intelligence. Fluid intelligence refers to the ability of developing problem solving approaches and strategies for situations that do not provide ad hoc solving techniques and procedures. Fluid intelligence generally implies the ability to detect relations, and is therefore a highly relevant process for reasoning. Referring to Horn and Cattell (1966), Sternberg (1986, p. 282) stated that "the Horn-Cattell (1966) theory of fluid and crystallized abilities also suggests that performances on induction tests can be understood in terms of a single factor (namely, that of fluid ability)". Crystallized intelligence is the ability of applying afore acquired knowledge. Further, crystallized intelligence is environmental because knowledge and its acquirement are not free of cultural impact. Regarding the relationship between fluid and crystallized intelligence it can be assumed that fluid intelligence is a requirement for crystallized intelligence. The status or extent of crystallized intelligence is in turn the result of fluid intelligence. Cattell (1971) also considered reasoning indispensable in forming crystallized intelligence, thus learning and acquisition of knowledge.

### **Jäger's Berlin Model of Intelligence Structure**

In Jäger's Berlin model of intelligence structure (BIS; Jäger, 1982) the operation facet processing capacity corresponds to reasoning ability as it refers to processing of complex information, detecting relations, and logical thinking in situations or tasks when no experienced solving algorithms are available. In this model every intelligent performance is defined by one of the

four operational facets and one content facet referring to the stimulus material. This can be figural-spatial, verbal or numerical, thus reasoning tasks can be further distinguished depending on the kind of material involved.

### **Carroll's Three Stratum Theory**

In the *Three-Stratum Theory* of Carroll (1993), fluid intelligence is - to a certain degree - defined by reasoning ability that in turn consists of three factors from the lowest stratum: sequential reasoning, induction, and quantitative reasoning. This structure can also be found in the model of Horn and Cattell (1966) apart from that their model does not provide a level of higher generality and thus no *g*. Reasoning ability (*Gf*) is the fundamental component for the more specific factors that can be measured by several tasks. Factor analytic research revealed high loadings of figural reasoning on fluid intelligence. It is therefore assumed that the relationship between figural reasoning and *Gf* can be judged as almost perfect (Süß & Beauducel, 2005). Reasoning ability is often a synonym for the general intelligence factor *g* (Carroll, 1993). Carroll, who developed a hierarchical model of intelligence, considered the ability to reason the crucial point of intelligence (Carroll, 1989).

### **Gustafsson: Documentation of Intelligence-Reasoning Coherence**

To conclude the discussion on intelligence and reasoning, findings of Gustafsson (1984) concerning this issue are illuminated. Thus, not a model but empirical evidence for the relation of general intelligence and fluid intelligence is presented.

Gustafsson (1988) described that general intelligence is commonly equated with fluid intelligence and furthermore with inductive reasoning. Reasoning ability seems to be rather similar to fluid intelligence and thus constitutes a central factor in numerous cognitive tasks and learning aptitude tests. Gustafsson believed that the extent of benefit from instruction is determined by reasoning ability. Given that, the performance especially in novel tasks could only be successful if the subject transferred instruction onto the task to be processed via reasoning. However, it seems that reasoning ability and fluid intelligence constitute at least similar if not same constructs and reasoning ability can thus be assumed to be an important component of intelligence. Sternberg (1986) states as follows:

An interesting finding that emerges from the literature attempting to relate cognitive task performance to psychometrically measured intelligence is that the correlations of task performance and IQ seems to be a direct function of the amount of reasoning involved in a given task, independent of the paradigm or label given to the paradigm ... Thus, reasoning ability appears to be central to intelligence. (pp. 309-310)

Gustafsson (1984) applied a battery of 16 tests and examined a sample of more than 1000 subjects. Conducted factor analysis resulted in a model with factors on three levels. According to the author the factors on the first level can be interpreted following Thurstone or Guilford. The model further contained three factors of second order which are fluid intelligence (*Gf*), crystallized intelligence (*Gc*) and visuo-spatial ability (*Gv*). These three second-order factors in turn loaded on one single factor of third order



which is interpreted as general intelligence ( $g$ ).  $Gv$  showed factor loadings of .80,  $Gc$  of .76 and  $Gf$  of 1.00 on  $g$ . According to this model general intelligence and fluid intelligence constitute identical constructs. Further inspection of the factors on second and first level revealed that the  $Gf$ -factor of the second and therefore more general level has a very strong, almost perfect relationship with one first order factor construed as induction (the tasks number series and letter grouping loaded on this factor). Expressing this relation in terms of a numerical coefficient, a correlation coefficient of .99 was obtained. To prove this nearly perfect correlation by more than one empirical reference, Gustafsson alluded to another study (Gustafsson, Lindström, & Björck-Åkesson, 1981) that illustrated loadings of the fluid factor on the third-level factor  $g$  of 1.00.

However, the hypothesis of coherence between  $Gf$  and general intelligence is contentious and depends on methodological procedures. Gustafsson (1984) referred to the rotation procedures in factor analysis and argued that “when Multiple Factor analysis is used with orthogonal rotation, the general factor is ‘roted away’ ” (p. 200), resulting in only weak loadings of general intelligence on all existing factors. Oblique rotation, however, displays the general factor as correlations between the other factors. According to Gustafsson though, oblique rotation underestimates the extent of correlations among factors. Since in oblique rotation the rotation angle is arbitrary, “ ‘objective’ empirical information on the amount of actual correlation between factors” cannot be obtained (Gustafsson, 1984, p. 200). Because of this criticism Gustafsson emphasized the advantages of confirmatory factor analysis and structure equation modeling (LISREL), which

he also applied onto his data.

Undheim and Gustafsson (1987) considered the LISREL-method an adequate method of testing hypotheses in hierarchical models. They again applied it to examine possible equivalence of  $Gf$  and  $g$ . Structural equation modeling was based on data of 144 eleven-year old pupils that were tested with a test battery embodying 12 subtests (Undheim, 1976). The resulting hierarchical model contained ten first-order factors, four second-order factors and  $g$  as the only factor of third order. Undheim and Gustafsson (1987) evaluated their main hypothesis that  $Gf$  and  $G$  are uniform constructs as confirmed although factor loadings of  $Gf$  on  $g$  were with 1.15 larger than 1 and therefore the Heywood case occurred. The Heywood case refers to negative eigenvalues or communalities above one, which however could only vary as sum of squared correlations between 0 and 1. Since the loadings of  $Gf$  on  $g$  were above one the factor variance was subsequently negative. Undheim and Gustafsson however argued that  $t$ -values for negative residual variance did not prove to be significant and their hypothesis on the equivalence of  $Gf$  and  $g$  would therefore be supported. Modification of the model resulted in one in which the previous four second-order factors now turned into first-order factors and the previous latent variable of third order now represents a second-order factor. Due to their numerical relationship of 1.01 Undheim and Gustafsson considered  $Gf$  and  $g$  as unitary constructs.

In a second study Undheim and Gustafsson (1987) reanalyzed data of 149 thirteen-year-olds. However, the model proved inadequate fit with the data and the hypothesis of a perfect relationship between  $g$  and  $Gf$  could

therefore not be approved. A modified model with five first-order factors and  $g$  as the only second-order factor,  $Gf$  proved factor loadings of .95 on  $g$  which was judged non-significantly deviating from 1.00.

In a third study, testing 148 fifteen-year-old subjects (Undheim & Gustafsson, 1987), the hypothesis of uniform constructs could again be supported: In a first LISREL-model,  $G$  and  $Gf$  showed correlations of .97. In a simplified one-factor model,  $Gf$  proved with .94 the highest loading on  $g$ .

## 2.3 Information-Processing Theories of Analogy

### Tasks

Looking at the precedent models of intelligence that have coined and developed an idea on the structure and components of human intelligence, one might conclude that reasoning ability constitutes a central construct in human intelligence. These models ascribe different significance to reasoning ability within each model, but all models consider it a core and fundamental ability of human intelligence, and reasoning ability is represented in each model. However, the models do not describe components and processes, neither of reasoning in general nor of analogical reasoning in particular. But for a solid understanding of reasoning, the processes involved have to be analyzed. The following section therefore presents several selected information-processing theories of analogical reasoning.

### **Spearman's Information Processing Theory**

Besides his theory of a general factor, Spearman also proposed an information-processing theory of reasoning (Spearman, 1923). He put forth three principles of cognition. In the first principle *Apprehension of Experience*, Spearman stated that "any lived experience tends to evoke immediately a knowing of its characters and experienter" (p. 48). The second principle, called *Eduction of Relations*, means that "the mentally presenting of any two or more characters (simple or complex) tends to evoke immediately a knowing of relation between them" (p. 63). Spearman defines relation as "any attribute which mediates between two or more fundamentals" (p. 66). He distinguishes between real and ideal relations: Real relations are according to Spearman (1923) *attribution* (relating something to its fundament), *identity* (refers to the preservation of an object), *time*, *space*, *cause*, *objectivity*, and *constitution*. Ideal relations are *likeness* (refers to resemblance as well as dissimilarity), *evidence* (in terms of reasoning), *conjunction*, and *intermixture* (intermixture of any relations stated before). The third principle *Eduction of Correlates* represents "the presenting of any character together with any relation tends to evoke immediately a knowing of the correlative character" (p. 91).

To illustrate his principles, Spearman (1923) provided several examples from different domains. For the domain of perception he gave following example of a quint of musical harmony: When two tones are given, the subject tested might recognize that these tones are related by a musical fifth (second principle). When a third tone is presented and the subject mentally represents the concept of the musical fifth, he might be able to

generate or mentally represent a tone that corresponds to the afore heard tone, thus to educe a correlate (third principle) in the relation of a musical fifth. Spearman also demonstrated examples for analogies. In the analogy “WHITE is to BLACK as GOOD is to ...” (p. 100) it may be recognized that the first two terms constitute opposites (second principle) and by finding the opposite of the third term C, and thus generating the same relation that connects A and B, the third principle of educing correlates is fulfilled.

### **Sternberg’s Componential Model of Analogical Reasoning**

In literature several models of analogical reasoning have been proposed. Among these, Sternberg’s componential model of analogical reasoning is a popular model (Sternberg 1977a). Sternberg referred to four processing models of analogical reasoning that differ in the sequence of processes applied. In all models, processes were the same and serial execution of processes is given. Models however, differed in the sequence and frequencies of processes. Sternberg presented the models with steady regard to one example (Washington : 1 :: Lincoln : (a.10, b.5)) presented in Table 2.1. In Model 1 the first step is the encoding of term A (Washington) followed by the encoding of term B (1). Attributes for each term are recalled from long-term memory and stored in working memory. Then attributes of the A term (president, portrait on currency, revolutionary war hero) and the B term (counting number, ordinal position, amount) are related to each other, representing the third step of inferring a relation between A and B. The attribute president is related to the attribute ordinal position (first president), portrait on currency is related to the attribute amount (portrait on one dol-

lar note), and between revolutionary and counting number no relation is inferred and marked with  $\emptyset$  in Table 2.1. In the next step, called mapping, the C term (Lincoln) is encoded and related to the A term (Washington). The result of relating the two terms to each other might be the discovery that both were presidents, are portrayed on currencies and were heroes in wars. Then the response options (a.10 , b.5) are encoded and applied onto the C term in the last step: The A term Lincoln is related to the answer options in terms of the attributes (president, currency, war hero). Relating Lincoln to answer option a (10) fails (Lincoln was not the 10th president and not portrayed on a 10 dollar bill), but answer option b (5) can be related to the attribute currency (portrayed on 5-dollar note). Thus, A and C are related by the attribute currency.

Summarized, Sternberg's first model presents a sequence of seven steps (encode A  $\longrightarrow$  encode B  $\longrightarrow$  infer A to B  $\longrightarrow$  encode C  $\longrightarrow$  map A to C  $\longrightarrow$  encode D  $\longrightarrow$  apply C to D) to solve the verbal analogy. The model can be divided into two basic components: Attribute identification (encoding) and attribute comparison. Attribute comparison refers to three obligatory processes: inferring, mapping, and application. These represent the steps of finding a rule relating A to B (inferring), finding a rule relating A to C (mapping), and apply the rule from C to D (application). Sternberg (1977a) stated that the additional and optional attribute-comparison component of justification is necessary in forced choice answering formats as a validation process when none of the response choices exactly presents the answer generated by the subject. A control component was also implied in the model, and necessary in order to guide the process of solving an item and

responding to the item.

Whenever attribute lists are needed in the model, working memory capacity is required since attributes have to be stored and retrieved for conducting the steps of inference, mapping, and application.

Models II-IV referred to the same information-processing components but components were differently organized. Sternberg (1977a) summarized that in Model I, all attribute comparison components (inferring, mapping, applying) are “exhaustive processes” (p. 139) because each component refers to all attributes. In Model II, only the processes of inference and mapping are exhaustive since they compare all attributes, “but application is self-terminating” (p. 140) because not all attributes have to be applied. If the subject compares the first attribute and finds a match with the answer option, the process of applying is terminated. In Model III, inference is exhaustive. The subject might then find a correct solution after mapping the first attribute and the processes of mapping and applying are thus self-terminating. In Model IV, all three processes can be self-terminating whenever not all attributes have to be inferred to solve the analogy.

Empirical research for model validation was carried out by Sternberg conducting three experiments using analogies of different content: People-piece analogy, verbal analogy, and geometric analogy (Sternberg 1977a). Impact of the components was quantified in latencies, i.e. time for each operation and errors (difficulty of operation).

By creating different experimental groups, Sternberg enabled the measurement of the component latencies and thus parameter estimation. The experimental groups differed in the numbers of cues presented before the

Table 2.1: Analogy Example According to Sternberg (1977a)

Process	Analogy Term/Relation	Relevant Attributes & Values
Encoding	Washington	president (first), portrait on currency (dollar), war hero (Revolutionary)
	1	counting number (one), ordinal position (first), amount (one unit)
	Lincoln	president (sixteenth), portrait on currency (five dollars), war hero (Civil)
	10	counting number (ten), ordinal position (tenth), amount (ten units)
	5	counting number (five), ordinal position (fifth), amount (five units)
Inference Mapping	Washington $\longrightarrow$ 1	president (ordinal position (first)), portrait on currency (amount (\$)), $\emptyset$
	Washington $\longrightarrow$ Lincoln	presidents (first, sixteenth), portraits on currency (dollar, five dollars), war heroes (Revolutionary, Civil)
Application	Lincoln $\longrightarrow$ 10	$\emptyset, \emptyset, \emptyset$
	Lincoln $\longrightarrow$ 5	$\emptyset$ , portrait on currency (amount (five dollars)), $\emptyset$

*Note.*  $\emptyset$  indicates that no relation between terms is inferred.



total analogy was presented. The precuing condition was defined by the number of cues presented: In the 0-cue condition, none of the analogy terms was presented before the whole analogy was shown. In the 1-cue condition, only the A term was presented before the full term, in the 2-cue condition the A and B terms were presented and in the 3-term condition the A, B and C terms were shown. After precuing, subjects had to indicate whenever they were ready to proceed. After having seen the full analogy, subjects were asked to judge if the analogy was true or false (people-piece experiment and verbal-analogy experiment) or to choose between two response choices offered (geometric analogy).

Thus a regression model for the solution time could be generated with the latencies of encoding, inference, mapping, and application as regressors. The equations for the cuing conditions differed in terms of the regressors involved (e.g. the 2-cue condition had no inference parameter since the relation between the A term and the B term was inferred in the precuing phase; the 3-cue condition had no inference and no mapping parameter).

Across all conditions Model III accounted for the most variance in the people-piece experiment (92%) and the verbal analogies (86 %), whereas in the experiment of geometric analogies Models II, III and IV all explained 80% of variance each, and Model I 74 % only. Looking at the regression models with all components as parameters, analogies from the 0-cue condition were analyzed only. In the people-piece and verbal analogy experiments Model III yielded the best fit (89% in people-piece experiment and 62% in verbal experiment) and Model II and III each explained 80% of variance in the geometric analogy experiment. Sternberg (1977b) concluded

that Model III was the predominantly best model and Model I with its fully exhaustive operations yielded the worst statistical explaining power.

Referring to response errors as the variable to be explained, Models III and IV yielded the best fit for all three experiments. Discriminating between exhaustive and self-terminating operations, Sternberg (1977b) inferred that only the self-terminating parameters were useful to predict errors. Thus, besides contributing to shorter solution times, these components convey error making.

Sternberg (1977a) also provided two strategies for applying a corresponding analogous rule from the C term to the D term. These two strategies are called *sequential option scanning* and *altering option scanning*. In the first one, all response options are evaluated one after another by applying all attributes onto the first option. Then all attributes are applied onto the next answer option until all options have been scanned. The strategy of checking all options is used to prevent incorrect responses. The second strategy refers to the alternate checking of the answer options. Single attributes are applied to each option before the application process continued to apply the next attribute onto the options. The application of strategies is only useful when a forced-choice response format is used. Sternberg (1977b) therefore analyzed model fit only for the geometric analogy experiment to compare alternating and sequential scanning procedures. The alternating scanning method was judged as superior since alternating scanning models accounted for more variance than the sequential scanning models: The alternating Model III explained 80% of variance as opposed to the sequential Model III explaining only 68%.

Another interesting topic emerging from Sternberg's analysis (1977b) refers to the processing time of the components in the process of the analogy solution (according to Model III). Figure 2.1 illustrates the distribution of processing times for the people-piece analogy, the verbal analogy task, and the geometric analogy task. Compared to the people-piece analogy (1435 msec) and the verbal analogy (2408 msec), geometric analogies (5.7 sec) have longer processing times. In terms of percentaged portions of processing times for components, encoding claims 54% of processing time in the verbal analogy vs. 36% in the geometric analogy. In the geometric analogy the processes of inferring, mapping, applying, and justifying require the most time (57%) whereas only about 30% of processing time is spent on attribute comparison processes in the verbal analogy. Sternberg (1977b) states "that encoding of words is much more time-consuming than encoding of simple schematic figures" (p. 372) and that the processes of attribute comparisons in geometric analogies take longer because they are of a higher level of difficulty than the ones of the verbal analogy or the people-piece analogy.

Sternberg's information-processing models can be applied to forced-choice formats with different numbers of answering options as well as to true-false answer formats. Regarding content, Sternberg (1977a) stated that the models provide information-processing flows not only for verbal analogies but also for geometric analogies.

Sternberg (1977a) evaluated and judged his own componential theory of information-processing as complete (process of solving an item is provided from the beginning to the end), specific (describes details of attribute-

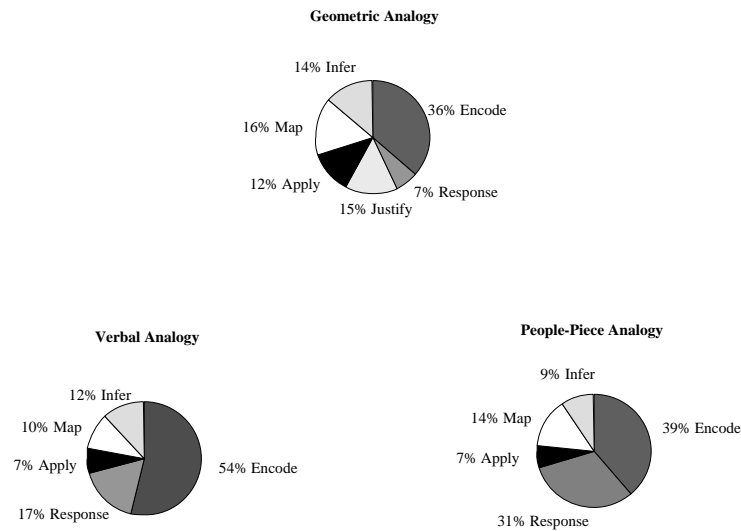


Figure 2.1: Component Processing Times According to Sternberg (1977b)

comparisons components), general (analogies with a variety of contents can be applied), parsimonious (only five mandatory and one optional process), and plausible (in contrast to other models, Sternberg's model contains the process of mapping, and provides experimental data to prove that the component of mapping also accounts for predicting solution latencies, error rates, and individual differences in experiments).

In his unified theory of human reasoning (Sternberg, 1986), Sternberg proposed "that reasoning can be understood in terms of the mediated, controlled application of selective encoding, selective comparison, and selective combination to inferential rules" (p. 310). Thus, according to Sternberg, these elementary information processes define reasoning. Selective encoding refers to selecting relevant problem-solving information. Selective comparison refers to comparing information - selected in the first step and stored in working memory - with declarative and procedural knowl-

edge from long-term memory relevant to the problem. The third process, selective combination, combines the selectively encoded or compared information. However, Sternberg (1986) stated that task and test taker interact and influence the degree of automatization since the degree of task novelty to the test taker determines to which extent the processes are controlled. Sternberg explained that the processes of encoding and comparison are rather inductive and the selective combination is rather deductive. Thus, analogy tasks with their inductive nature (“analogies ... are considered almost prototypical of inductive reasoning tasks” (Sternberg, 1986, p. 294)) do not involve the process of combination as elements or terms never have to be combined to new units.

Selective combination is therefore not required in analogical reasoning but in deductive tasks such as those involving syllogisms. However, within the analogy task one must distinguish between verbal and geometric/figural analogies in terms of the degree of selective encoding and selective comparison involved. In verbal analogies, comparison of the encoded task information with declarative knowledge from long-term memory is more essential than in geometric analogies where selective encoding of information is an important process. According to Sternberg, because of the processes of encoding and comparison involved (in contrast to combination), analogies are representative of inductive reasoning tasks.

### **Induction and Deduction as Components of Analogical Reasoning**

Johnson (1962) identified two core processes of reasoning: induction and deduction. He investigated verbal analogies by means of serial exposure.

He divided the problem solving process of analogies into two stages. The first phase called *preparation period*, in which the relation between the A term and the B term has to be deduced, is referred to as induction. Deduction is executed when the inferred relation is applied onto the C term to generate the D term (*second period*). According to Johnson, item difficulty can be led back to either induction or deduction, depending on the degree of familiarity of the stimuli. He assumed that in items, in which the words applied in the A term and the B term are familiar and the C term is unfamiliar (deduction problem), more time will be spent on the deduction compared to time spent on the deduction part in induction problems (A and B are unfamiliar, C is familiar). In induction problems, time spent on the induction part or preparation period is supposed to be longer than time spent on deduction. Johnson analyzed three task formats regarding these hypotheses. The first task format required the subject to produce the D term, i.e. write it down. The second task format was a multiple choice question format with five response options. The third format was like a multiple choice test format but offered initial letters only. A mixture of 25 deduction tasks and 25 induction tasks was constructed for each format and processed by 20 subjects in each condition.

His results proved, that for all three test formats, significantly more time was needed for the induction phase in induction problems than in deduction problems and that, vice versa, time for deduction was significantly longer in deduction problems than in induction problems (not in the multiple choice format). Total processing time between induction and deduction only significantly differed in the multiple choice format with less time

needed for the deduction part. In terms of error rates, no significant differences between induction and deduction were reported for all test formats. As expected, deduction time was significantly longer in production problems (due to writing) than in initial and multiple choice formats.

### **Evans (1968)**

Evans (1968) designed a computer program called ANALOGY that was constructed to solve geometric analogy items. The - here extremely simplified - basic solving process implemented in the computer algorithm to solve the item involves the finding of the transformation that changes A to B, and C to the answer options presented. The algorithm provides two phases in the solution process. In the first phase, figures are decomposed and the relations and properties of the resulting subfigures are analyzed. Similarity between figures is calculated and, together with the information that originates from the decomposition and relation process, forwarded to the second part of the program dealing with rule construction. Rules, according to which objects of figure A are transformed into figure B, are constructed by considering changes in properties, relations, and sub-components. Potential rules are then gathered and generalized before deciding on the most adequate one.

### **Semantic Relationships of Concepts**

The model of Rumelhart and Abrahamson (1973) cannot be described as a process theory of analogical reasoning but provides a model of semantic relationships of concepts. They differ in reasoning and remembering, with

the former referring to relationships and the latter referring to the retrieval of specific information from memory. Following Henley (1969), Rumelhart and Abrahamson assumed that the memory structure of certain concepts can be organized in an Euclidean space. Each element can be represented as a point in multidimensional space. Judging similarities and dissimilarities of concepts constituted reasoning and the closeness of concepts in memory served as indicator for the degree of similarity. Considering the analogy format  $A:B::C:D$  in a multidimensional representation, the vector distance between A and B represents the similarity of concepts between these two terms. The term C should be equally similar to a chosen concept (D) as A is to B. The terms C and D should thus have the same vector distance as the concepts of A and B. Animal terms were chosen for the empirical examination of compliance between predicted responses by the model and empirical responses. Subjects were presented 30 analogy problems with four response options each. They were instructed to rank the options according to their analogical degree. Results showed that almost 71% of the rank 1 answers were given to the closest concept (closest concept to the ideal answer). Further, a correlation of  $r = .93$  was computed between the predicted and observed number of subjects.

### **Components of Geometric Analogy Solution**

In their study *Components of geometric analogy solution*, Mulholland, Pellegrino, and Glaser (1980) suggested a combined model for the solution process of geometric analogies. Their model was based on the models of Sternberg (1977a) and Evans (1968) and is illustrated in Figure 2.2.



The process model relates to true or false analogies and presents solution processes and emerging products and latencies. The first step in the model is the comparison and decomposition of the A term and the B term resulting in element lists. Mulholland et al. assumed that the reaction time (RT) needed for such processes is a function of the number of elements (E) involved in the A-B term. The second step refers to analyzing the transformations that change A into B and generating rules. The resulting list of objects and transformations is thought to be stored in working memory. Mulholland et al. (1980) set up the equation so that the reaction time for this process is a function of the number of transformations (T) of an item, since "previous research on the processing of spatial transformations of geometric stimuli (e.g., Bundesen & Larsen, 1975; Shepard, 1975) has shown that separate rotations, reflections, and size changes appear to be individually and additively executed" (p. 257). These processes are again executed when the C term and the D term are analyzed and reaction times are again functions of the number of elements, respectively transformations. Before responding, the rule comparison process is conducted to evaluate the degree of correspondence between rules. The total processing time of an item is thought to be predicted by the latencies of the different steps ( $RT = xE + yT + k$ ).

### **Summary**

Before making a statement on which theory is the most adequate one and adopted as underlying information-processing model for the present study, theories outlined beforehand are evaluated and summarized where neces-

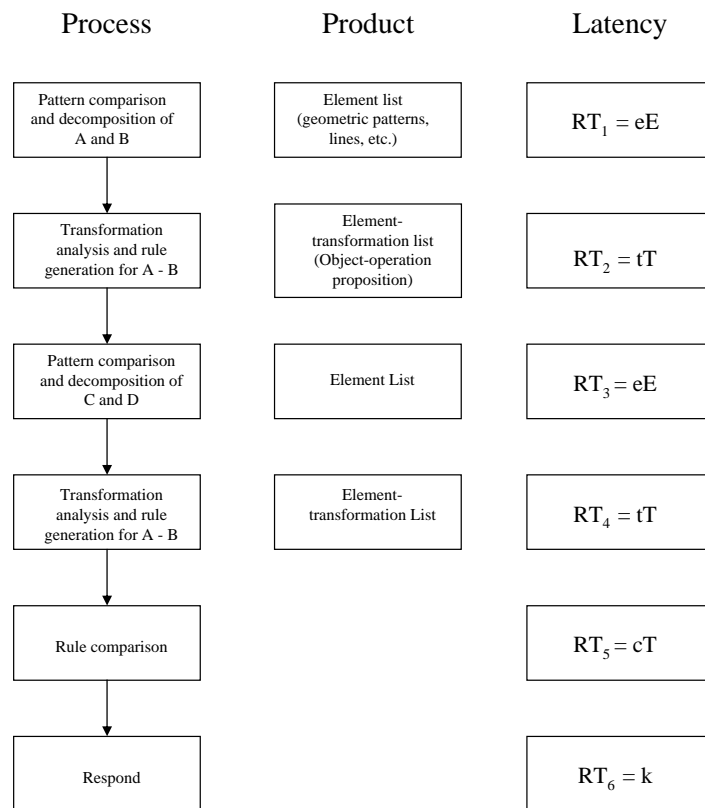


Figure 2.2: Model of the Geometric Analogy Solution Process (Mulholland et al., 1980)

sary. Here, guidelines of evaluation are some criteria according to Sternberg (1977a), such as empirical support provided by the authors, generality of the theory, completeness, and parsimony.

Spearman's theory (Spearman, 1923) provides three principles: apprehension of experience, eduction of relations, and eduction of correlates. Yet, the actual information processes are not really explained, thus sufficient understanding of the solution process is not allowed for. However, the theory is very general and probably applicable to a wide range of analogy tasks. This, nevertheless, remains an assumption since no experimental

data is reported to reinforce his theory.

Johnson identified induction and deduction as features of analogical reasoning and components of item difficulty in his experiment. However, the theory does not offer detailed insight into the cognitive processes involved in solving an analogy task.

Evans' computer program ANALOGY (Evans, 1968) consisted of two parts. The first referred to decomposing of the figures and the second to generating a rule according to which A was transformed into B and accordingly C into D. However, the theory might not be judged very parsimonious due to the complex problem solving algorithm implemented in the computer program. It is further not very general since it was applied to geometric stimuli only. Taking over this theory for the present study was however not considered since the solving process constituted a computer algorithm that was not available and adoptable for the purposes of this research.

The model of Rumelhart and Abrahamson (1973) referred to the relationship of semantic concepts and thus did not come into question as basic principle for a figural analogy test due to its content and specificity. Apart from that the model does not constitute an information-processing model as it does not state processes involved in the solution process of an analogy item. It was introduced as an alternative model for analogy tasks of verbal content.

Based on the processing theory of Sternberg and the work of Evans, the component model of Mulholland et al. (1980) is very specific regarding the solution processes involved in the course of solving analogy tasks. It

additionally takes into account reaction times of the components and defines the total time needed to solve an item as a function of the number of elements and transformations involved. The authors proved empirical support for the model, although only for geometric stimuli, thus generality might be restricted.

Sternberg's componential model of analogical reasoning (Sternberg 1977a, 1977b) can be evaluated as a very complex and specific model. Despite differentiating between different models within his componential model, Sternberg defined six component processes: encoding, inference, mapping, application, justification, and control. He rightly judged his own theory as complete and specific since it describes the full process and cognitive components involved when solving an analogy task (Sternberg 1977a). He concluded that "the theory thus manages to be complete while at the same time retaining parsimony" (p. 146). Generality is clearly given by the support of empirical data concerning different analogy contents: Besides geometric analogies and verbal analogies the theory could be applied to people-piece analogies and animal analogies. Since Sternberg's componential model of analogical reasoning fulfills the evaluation criteria the best and reflects and explains the cognitive processes involved in solving an analogy item in most detail, it is consulted as underlying information-processing theory in the present study.

Summarized, the different models of information processing can also be classified and assigned to the subsequent three processing theories, following Sternberg (1977b). Since researchers agree on the processes of encoding

and responding, only the processes of inference, mapping, and application guide the assignment of the antecedent information-processing models to the following three theories.

*Process theory with inference and application, but no mapping:* The theories of Spearman (1923) and Johnson (1962) can be assigned to the process theory with inference and application without mapping. Their introduced terms of inductive and deductive operation (Johnson, 1962) and education of relations and education of correlates (Spearman, 1923) corresponds to Sternberg's concept of encoding, inference, and application.

*Process theory with inference and mapping, but no application:* According to Sternberg (1977a), theories based on Evans (1968) can be subsumed under the process theory with inference and mapping but no application. After terms are encoded and relations between A and B and between C and the response alternatives are inferred, mapping occurs between the relation of A and C, and C and the answer options until the correct answer is found. Thus mapping induces the response and the process of application is not explicitly performed.

*Process theory with inference, mapping, and application:* The information-processing theory of Sternberg (1977a, 1977b) takes into account inference, mapping, and application. The different models within the theory only differ in whether the components are exhaustive or self-terminating.

## 2.4 Empirical Findings between Cognitive Abilities and Academic Achievement

In the following paragraphs empirical findings between intelligence and academic achievement are outlined. The coherence of reasoning and academic achievement is of particular interest because reasoning tests are frequently applied among selection procedures. Proving its predictive validity is therefore of particular importance.

However, specific findings on the relationship between analogical reasoning and academic achievement are seldomly examined or reported. Thus analysis of this relation had to be extended to findings regarding academic achievement and intelligence in general presented in the first part of this section. Hence, the repertory of studies is accounted for. First, findings on the coherence of intelligence and academic or scholastic performance are reported, considering different intelligence dimensions and measures as well as different operationalizations for scholastic achievement. Further, results of the study of Rindermann and Neubauer (2004) are presented, testing hypotheses of the relation between processing speed, intelligence, creativity, and school performance by means of structure equation models, to elucidate the causal relationship among variables. The second part of this section presents findings regarding the Miller Analogy Test, representing evidence of the predictive validity of analogy tests.

Thus, with regard to the predictive validity (concerning scholastic achievement) of the analogy test constructed in the present study, the aim of this section is to analyze and exhibit evidence for the relationship between in-

telligence and academic performance, and to provide benchmarks for the evaluation of the predictive validity of the figural analogy test.

### **Intelligence and Scholastic Achievement**

In the United Kingdom reasoning tests have a long history of assessing and predicting academic performances. The Cognitive Ability Test (CAT) is the most commonly applied test assessing academic achievement of about one million pupils each year (Deary, Strand, Smith, & Fernandes, 2007). Deary et al. applied the subsequent version of the CAT, CAT2E (Thorndike, Hagen, & France, 1986) to measure verbal, quantitative and nonverbal reasoning abilities. The General Certificate of Secondary Education (GCSE) referred to is a national public examination school students take in year 11 at the age of 15 or 16 to prove and certify their educational performances of secondary school. A correlation of  $r = .69$  between the GCSE overall score and the  $g$  factor was obtained. Further, Deary et al. report following results: Among science subjects, mathematics provided the highest correlation of  $.77$  with the  $g$  factor measured by the CAT. Subjects such as physics ( $.50$ ), chemistry ( $.46$ ) and biology ( $.51$ ) showed lower correlations probably due to restricted range because of high ability students. Among arts and humanities the subject English provided the highest correlations ( $.67$ ) and the subject drama ( $.47$ ) the lowest. Looking at the subjects representing social sciences, geography correlated  $.65$  and information technology only  $.47$  with the CAT  $g$  score. The correlations with the practical subjects such as art and design, music and physical education ranged from  $.43$  to  $.55$ .

Deary et al. (2007) applied structural equation modeling to analyze the

structure of CAT-measures and GCSE-measures and to quantify the relationship between the latent trait of reasoning ability and educational achievement. A correlation of .81 between the latent mental ability trait measured by the CAT2E (at the age of 11) and educational achievement obtained as scores from the GCSE is reported.

According to Strand (2006) reasoning tests such as the Cognitive Ability Test are applied by about two-thirds of secondary schools in England. Strand refers to Thomas and Mortimore (1996) and reports correlations ranging from .67 to .74 between CAT scores and GCSE results with the cognitive abilities measured six years before the GCSE was taken. These results support the assumption that reasoning abilities as measured by the CAT can validly predict educational performances even over a long period of time. Practical implications, apart from predicting future academic performances, involve diagnosing strengths and weaknesses of students (Strand, 2006).

Süß (2001) reviewed different studies referring to the predictive validity of intelligence tests concerning scholastic achievement. Validity of school grades was obtained by meta-analyses that showed that grades of the school subject mathematics had high validity for *Studienerfolg* [academic success] (Baron-Boldt, Schuler, & Funke, 1988) and *Ausbildungserfolg* [educational success] (Baron-Boldt, Funke, & Schuler, 1989), followed by the grades obtained in physics. Meta-analytic studies yielded different correlations of intelligence and school grades: Süß reported correlations of  $r = .34$



(Steinkamp & Maehr, 1983),  $r = .43$  (Fleming & Malone, 1983),  $r = .48$  (Boulanger, 1981), and  $r = .51$  (Hattie & Hansford, 1982). According to Jensen (1980), correlations vary with the type of school: Elementary schools show highest correlations of between .60 and .70, high schools of .50 to .60 and graduate schools of .30 to .40. Correlations thus decrease with increasing school level (Jensen, 1998). Amongst other reasons for decreasing correlations, Süß assumed that higher education is accompanied by restricted variance in ability.

Süß provided results of regression analyses referring to the predictive validity of *The Berlin intelligence structure model* (Jäger, 1982). Single school subjects were aggregated to two separate factors, a science factor consisting of grades of mathematics, physics and chemistry, and a factor referring to language consisting of German and foreign languages. Among all model components, the operative component reasoning capacity had the highest predictive validity for the science factor and the verbal ability was most predictive for the language factor. Together, the three content-related components (numerical, figural, verbal) explained more variance of the language factor ( $R^2 = .281$ ) than of the science factor ( $R^2 = .134$ ). Contrary, the components of the operative abilities (reasoning capacity, speed, memory, creativity) explained more variance of the science factor ( $R^2 = .238$ ) than of the language factor ( $R^2 = .000$ ).

Freund, Holling, and Preckel (2007) also analyzed the relationship between cognitive abilities and scholastic achievement. They generated composite scores (natural sciences & mathematics, language, and social science) of the

grades reported by the subjects. Following cognitive abilities according to the BIS (Jäger, 1982) were assessed: reasoning capacity, speed, creativity, and memory. *The Berlin Structure of Intelligence Test for Youth: Assessment of Talent and Giftedness* (BIS-HB; Jäger et al., 2006) was applied to measure these cognitive abilities and to examine their predictive validity for scholastic achievement. Multivariate analysis with the factors reasoning capacity, creativity, memory, and speed as independent predictors of the dependent variable school achievement, operationalized as composite scores of maths and natural sciences, languages, and social sciences, showed that reasoning capacity and creativity had highly significant influence ( $p < .001$ ) on all three subjects. However, reasoning capacity had more impact on maths and natural sciences (.472) and languages (.268) than creativity had on maths and natural sciences (.104) and on languages (.100). Influence on social sciences was greater for creativity (.228) than for reasoning capacity (.163). Memory had no significant influence on maths & natural sciences ( $p > .05$ ), but significant impact on languages (.109,  $p < .001$ ) and social sciences (.065,  $p < .05$ ). Speed was nonsignificant for languages and social sciences and of minor influence for maths and natural science (.069,  $p < .05$ ). General intelligence had the strongest impact on maths & natural sciences (.542), followed by languages (.443) and social sciences (.408).

The findings of Freund et al. (2007) thus indicate that among the cognitive abilities captured by the test based on the BIS, especially reasoning ability influences scholastic performance in mathematics and sciences, whereas creativity best accounts for performances in social sciences.

Holling, Preckel, and Vock (2004) referred to Amelang and Bartussek (1997) and Jensen (1998) and stated that intelligence and scholastic achievement correlate at about  $r = .50$  in meta-analytic studies. They report findings that specific cognitive operations measured by the BIS scales correlated with single subjects: Dimensions such as *Verbale Denkfähigkeit* [verbal cogitation] correlated higher with language-related subjects ( $r = .52$ ), than with maths and sciences ( $r = .38$ ). The negative values were due to the scoring of the grades, with low values indicating better achievement.

In their study, Luo, Thompson, and Detterman (2003) analyzed the source of the relation between intelligence and scholastic achievement. Since “it is not clear whether the basic cognitive processes involved in scholastic performance are *qualitatively* equivalent to those involved in psychometric *g*” (p. 69), they conducted complex analyses to examine the role of cognitive abilities as mediator of the correlation between intelligence and scholastic performance. Psychometric *g* was measured by subtests of the Wechsler Intelligence Scale for Children-Revised (WISC-R) and scholastic performance was assessed by the Metropolitan Achievement Test (MAT). The Cognitive Ability Test (CAT) was adopted to capture the cognitive processes. These tests were thus applied to test the assumption that the relationship between *g* and scholastic achievement (operationalized by applying the WISC-R, respectively MAT) is causally determined by the cognitive abilities as measured by the CAT. Results show that variability between psychometric intelligence and scholastic performance can be well explained by the CAT. Zero-order correlations, thus no control of variables, of .53 between

the general factor of the WISC-R and the MAT were obtained, indicating shared variance between the two variables of about 30%. Controlling the general factor of the cognitive ability variable, semipartial correlations of .248 were yielded and thus only 6% of shared variance between psychometric  $g$  and scholastic performance. Further, the measures of the general factor of the CAT explained about 60% of the variance in the general factor of the WISC-R and about 25% of the variance in scholastic performance.

Rindermann and Neubauer (2004) analyzed the relationship between processing speed, intelligence, creativity and school performance. They assumed that school performance is used “as probably the most valid external criterion for intelligence” (p. 575). However, they further aimed at testing how processing speed, intelligence and creativity are related to school performance. By means of structure equation modeling they tested the following models. The first model (three-factor model) assumes that school performance is influenced by the factors processing speed, intelligence and creativity. The impact of these three factors is supposed to be of direct and independent nature. The second model (speed-factor models) incorporates two versions but generally assumes that processing speed has direct impact on intelligence and creativity, but direct and indirect impact on school performance. The third model (intelligence-factor model) refers to direct influence of intelligence on processing speed and creativity, but direct and indirect influence on school performance. The last model, model four (creativity-factor model), ascribes creativity direct impact on processing speed and intelligence, and direct and indirect impact on school

performance.

Intelligence was operationalized by applying Raven's Advanced Progressive Matrices (Heller, Kratzmeier, & Lengfelder, 1998; Raven, 1958) and the *Kognitiver Fähigkeits-Test* [cognitive ability test] (KFT; Heller, Gaedike, & Weinläder, 1985), measuring verbal, numerical and figural competency. Processing speed was assessed using tests (ZVT; Oswald & Roth, 1978 and KDT; Lindley, Smith, & Thomas, 1988) that required solving tasks within a given limited period of time. Creativity was measured applying two creativity tests (VKT; Schoppe, 1975 and VWT; Facaoaru, 1985). Grades were obtained by the school reports of the testees and assigned to the categories of languages, mathematics & physics, natural sciences, and humanities to operationalize school performance. Regarding the relation of school performance with the other variables, correlations of the means of the four variables (processing speed, intelligence, creativity, school performance) showed that intelligence and school performance correlated the highest ( $r = .52$ ) followed by processing speed ( $r = .35$ ) and creativity ( $r = .25$ ). Processing speed correlated .31 with intelligence and .33 with creativity and intelligence and creativity showed a correlation of only .14. Stepwise regression proved that all three regressors had significant impact on school performance ( $R^2 = .33$ ). Intelligence ( $\beta = .45$ ) had the highest explaining power as regressor on school performance, followed by processing speed ( $\beta = .17$ ) and creativity ( $\beta = .13$ ).

Testing the models, Rindermann and Neubauer (2004) concluded that the speed-factor model provided the best fit with the data, indicating that intelligence and creativity are influenced by processing speed and further

influence school performance. According to this model, processing speed provided a total effect on school performance of  $\beta = .39$  with a direct effect of  $\beta = .08$  and an indirect effect of  $\beta = .31$  (indirect effect greater via intelligence ( $\beta = .21$ ) than via creativity ( $\beta = .10$ )). In the three-factor model, intelligence had the highest direct effect on school performance ( $\beta = .53$ ), followed by creativity ( $\beta = .19$ ) and processing speed ( $\beta = .09$ ). In the intelligence-factor model, intelligence yielded a total effect of  $\beta = .63$  on school performance, with a direct effect of  $\beta = .54$  and thus an indirect effect, mediated by creativity and processing speed, of  $\beta = .09$  (cf. Figure 2.3). According to the creativity-factor model, creativity obtained a direct effect of  $\beta = .26$  on school performance. Figure 2.3 shows the intelligence-factor model according to the structure equation modeling of Rindermann and Neubauer (2004). Although it is not the model with the best model-fit, this model is illustrated due to the focus on the influence of intelligence on school performance.

Across all models, explained variance of school performance by the variables (intelligence, processing speed and creativity) constitutes at  $R^2 = .43$  and is thus higher than in stepwise regression due to assumptions of error-free measurement in structure equation modeling.

### **Analogical Reasoning and Scholastic Achievement**

In the following paragraph the predictive validity of analogical reasoning tests is analyzed by presenting findings of the Miller Analogy Test (MAT; Miller, 1960).

The Miller Analogy Test has a long history as an admission test to grad-

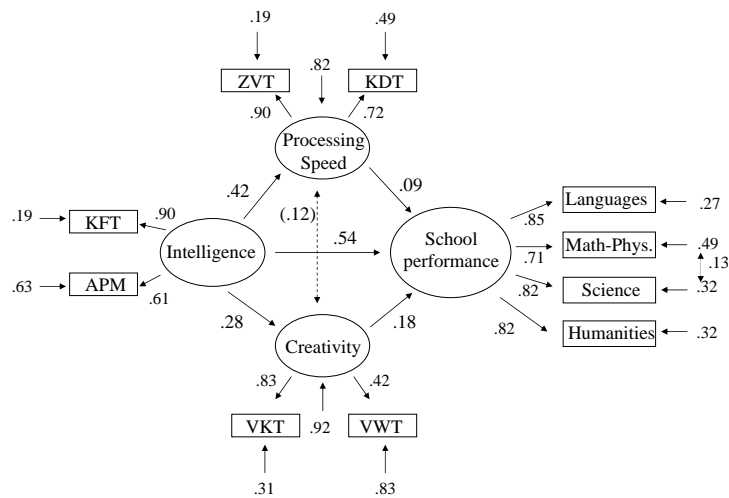


Figure 2.3: Intelligence-Factor Model According to Rindermann & Neubauer (2004)

uate schools in the United States. The MAT is composed of analogy items from different domains, such as humanities, language, mathematics, natural sciences, and social sciences. Relationships refer to word meanings, logical and mathematical objectives, classification and association. Items have the format of A:B::C:D and one of the terms is always missing and the correct term has to be chosen from answer options presented. Kuncel, Hezlett, and Ones (2004) conducted a meta-analysis to examine the measurement of cognitive abilities by the MAT and to estimate the predictive validity of cognitive abilities measured by the MAT for different academic and work criteria. Regarding correlations to other ability measures, besides the Graduate Record Examination-Verbal test (GRE-V; Briel, O'Neill, & Scheuneman, 1993) and the GRE-Quantitative test (GRE-Q), tests of the studies included in the meta-analysis were assigned to the category of either verbal ability tests, mathematical ability tests or general cognitive ability and rea-

soning tests. Results showed that the MAT proved correlations with other measures of cognitive abilities (Kuncel et al. 2004). The MAT provided high estimated true score correlations of  $\rho = .88$  ( $k = 15$ ,  $N = 8,328$ ) to the GRE-V and  $\rho = .88$  ( $k = 23$ ,  $N = 3,614$ ) to other verbal ability tests. Further correlations of  $\rho = .57$  ( $k = 15$ ,  $N = 7,055$ ) to the GRE-Q were obtained and of  $\rho = .68$  ( $k = 18$ ,  $N = 2,874$ ) to other mathematical ability tests. High true score correlations of  $\rho = .75$  ( $k = 15$ ,  $N = 1,753$ ) were reported with tests measuring general cognitive ability and reasoning tests. The validity of the MAT was analyzed in terms of three criteria categories: academic criteria, school-to-work transition criteria, and job performance criteria. Following eight academic criteria were generated with the estimated true score validity in parentheses: graduate grade point average (.39), first-year graduate grade point average (.41), faculty ratings (.37), comprehensive examination scores (.58), research productivity (.19), degree attainment (.21), time to finish degree (.35), and number of courses/credits completed (-.06). Apart from one criteria, the MAT could validly predict academic criteria with an average coefficient of  $\rho = .32$ . Six transitional school-to-work criteria were composed, among these internship ratings as well as potential ratings and creativity ratings. The MAT proved to be a valid predictor of all criteria, apart from the criteria student-teaching performance ratings, with an average coefficient of  $\rho = .29$ . To analyze the criterion-related validity for work settings, four criteria were examined: job performance ( $\rho = .41$ ,  $k = 7$ ,  $N = 598$ ), counseling performance ( $\rho = .51$ ,  $k = 2$ ,  $N = 92$ ) educational administration ( $\rho = .27$ ,  $k = 10$ ,  $N = 225$ ) and membership in a professional organization ( $\rho = .27$ ,  $k = 3$ ,  $N = 278$ ). Thus, all work related criteria could



be validly predicted by the MAT with an average coefficient of  $\rho = .37$ .

Summarized, the meta-analysis proved that an analogy test such as the MAT, which was originally constructed as admission test to educational settings, is not only capable of predicting academic-related performances, but also provides a tool to validly predict job performance.

### **Summary**

The preceding sections examined the potential of cognitive ability tests to predict scholastic achievement. The validity of a variety of such measures could be confirmed for different scholastic achievement criteria. For the Miller Analogy Test, as an example for analogical reasoning tests, validity was not only proven for several educational and academic criteria but further for several job-related performance criteria in work settings.

Besides expounding the empirical findings and status quo for the coherence of intelligence and reasoning measures with scholastic achievement, the section is insofar of high relevance for the figural analogy test of the present study, as it provides informative basis for the hypotheses regarding its relationship with academic achievement. Further, benchmarks for critical evaluation of the empirical found coherence of the figural analogy test with academic performance criteria are thus provided.

## 2.5 Effects of Training and Instruction on Test

### Performance

The following section presents several empirical findings on the effects of training and instruction types on the performance in cognitive ability tests and in analogy tasks. Training effects and effects of different instructions imply important practical implications. Different benefit from different instructions can, for example, influence test performance. Awareness of such effects is therefore crucial to correctly attribute and interpret inter-individual differences in test performance to such testing conditions and maybe not to real differences in performance. Effects must also be known in order to choose the most adequate instruction type or to create equal conditions for all subjects via practicing.

The first part of this section reviews and reports findings on training effects of analogical reasoning tasks and meta-analytic findings of training effects of cognitive abilities before effects of instruction on test performance are outlined. The last part summarizes findings of the effects of training and instruction on test performance and features implications for the research questions regarding the figural analogy test constructed for the present study.

#### 2.5.1 Training Effects

In their study referring to training of analogical reasoning processes, White and Caropreso (1989) analyzed the effect of instruction on test performance of low socioeconomic status preschool children. In a 2 x 3 test design, they

tested three groups in pre- and posttests to assess the effect of training on test performance. Besides the training group, one control group and one play group were established. They applied geometric analogies of the format A:B::C:D. The training group was instructed in three sessions. Training referred to the explanation of the solution process of analogies, following the model of Sternberg (1977a). Concrete and abstract toy objects were applied in the first, respectively second training session and practice of the solution process of geometric analogies was topic of the third training session. The three groups showed no significant inter-group differences in vocabulary knowledge (however, all three groups were below the standard score mean) and in an analogy pretest. Results proved that training significantly influenced test performance as test score means were higher in the posttest for the trained group. White and Caropreso concluded that training had significant impact on test performance in an geometric analogy test for the trained group by posttest scores ( $F = 3.72$ ,  $df = 2,23$ ,  $p = .038$ ).

The impact of training was already examined by White and Alexander (1986) in a precedent study, again testing preschool children without a focus on the socioeconomic status. The performance of two groups (trained vs. non-trained) was investigated by applying geometric analogy tasks. Performance was measured two weeks before the training sessions (pretest), within one week after the training (immediate posttest) and a month after the training (delayed posttest). The results of White and Alexander indicated a significant impact of training on test performance when comparing

the scores of trained and untrained. At both points in time (immediate and delayed), significant differences between the groups could be observed ( $F_{(1,27)} = 41.47, p < .0001$ ). Time itself did not significantly influence test scores of the trained and untrained as comparisons between the immediate test and the delayed test showed ( $F_{(1,28)} = 2.31, p > .05$ ). Inter-group-wise, test scores significantly increased for the trained group from pretest to the immediate posttest ( $t = 8.48, df = 1,38, p < .0001$ ) and from pretest to delayed posttest ( $t = 9.88, df = 1,38, p < .0001$ ). The untrained group did not show increased test scores compared to the pretest, neither at the immediate posttest ( $t = 1.16, df = 1,18, p > .05$ ), nor at the delayed posttest ( $t = .41, df = 1,18, p > .05$ ).

Effects of training could also be proven for performance on verbal analogies (Alexander, Haensly, Crimmins-Jeanes, & White, 1986). The effect of age and training as well as the effect of ability and training was analyzed. The *Woodcock Reading Mastery Test*, Forms A and B (Woodcock, 1973) were applied measuring verbal analogy ability. Task format was again A:B::C:D. Operationalizing the age variable, subjects from different grades (4th grade, 8th grade, 10th grade) were recruited. The eighth graders were judged as gifted as they were enrolled in gifted language art programs according to different test criteria. Training was conducted as in-class training sessions of the analogy components according to Sternberg (1977a). Half of the subjects from each age level took part in the componential analogy training and half of the subjects were controls. For the fourth graders (whose results were separately analyzed since they received a different test version),

subjects having received the training scored significantly higher on immediate ( $F = 17.82$ ,  $df = 1,33$ ,  $p < .001$ ) and delayed ( $F = 11.54$ ,  $df = 1,33$ ,  $p < .002$ ) posttests. On the test performance of the 8th and 10th graders, training also had significant impact ( $F = 27.33$ ,  $df = 1,121$ ,  $p < .0001$ ). The trained 8th graders showed higher posttest scores than the trained 10th graders indicating significant impact of ability ( $F = 10.29$ ,  $df = 1,121$ ,  $p < .002$ ). Higher gifted students thus seemed to gain more benefit from training than students with average ability.

To summarize training effects of cognitive abilities in general, meta-analytic reviews should be considered. The effect of training in inductive reasoning was examined in a meta-analysis by Klauer (2001) who showed that training inductive thinking has mediocre effect ( $d = .60$ , 61 studies). Score increases can also be obtained due to retesting without any interventions between test administrations. These gains in test scores are usually referred to as practice effects (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Freund, 2008). In their meta-analysis of practice effects for cognitive ability tests, Hausknecht et al. (2007) showed that overall practice effects of  $d = .24$  from the first to the second test and of  $d = .18$  from the second to the third test and of  $d = .51$  from the first to the third test were yielded. Besides mere repetition, Hausknecht et al. related different moderators to increases in test scores. For samples that received test coaching ( $k = 23$ ) effect sizes of  $d = .64$  were obtained as opposed to samples that did not receive test coaching ( $k = 75$ ,  $d = .21$ ). Effect sizes for the moderator variable study context did not significantly differ and were quite similar for

operational contexts ( $d = .27$ ) and research contexts ( $d = .22$ ). Hausknecht et al. further supported their hypothesis that the effect size for identical test forms ( $d = .40$ ) is larger than the effect size for alternate forms ( $d = .22$ ). With regard to the present study, a relevant hypothesis of the meta-analysis referred to effect size estimates of different cognitive ability dimensions. Practice effects for tests referring to verbal ability (such as analogies, vocabulary, reading comprehension tasks), qualitative ability (eg. arithmetic computation, mathematical reasoning) and analytical ability (use of logic, inductive and deductive reasoning items) were analyzed. Hausknecht et al. assumed that larger effect sizes for quantitative and analytical ability tests than for verbal ability tests, since score gains of verbal ability tests are tied to the acquisition of new information as opposed to quantitative and analytical ability tests, where performance improvement might be led back to increased test familiarity when retested and the use of skills involving general problem solving. However, the meta-analytic results proved larger effect sizes for analytical tests ( $d = .29$ ) and quantitative tests ( $d = .27$ ) than for verbal tests ( $d = .17$ ) but the differences between the effect size estimates were not significant.

### **2.5.2 Effects of Instruction on Test Performance**

Bisanz, Bisanz, and LeVevre (1984) analyzed the effect of instructions on individual differences in analogical reasoning. They referred to incomplete instruction and assumed that the degree of benefit from incomplete instruction indicates aspects of intelligence. Individual differences were thought to stem from selecting the right task-relevant strategies from in-

complete instruction. According to Bisanz et al. instruction frequently applied in intelligence tests, is often not very explicit. In analogy testing, usually instruction of the test format is given such as finding the relation between A and B and then picking an adequate D that is related to C. The authors judged such instructions as confusing, especially for younger children as task-appropriate strategies have to be derived from such instruction and a few examples. Bisanz et al. assessed subjects of different age (9, 11, 13 and 19 years) to analyze the different profit of incomplete instruction. They applied dot analogies that were true to different criteria (true by magnitude and direction, true in direction but false in magnitude, true in magnitude but false in direction; for a more detailed explanation see Bisanz et al., 1984). Direction referred to the calculating operation of plus or minus, and magnitude referred to the value added or subtracted. They distinguished between low-demand (values 1-7 in the C & D terms) and high-demand (values 7-12) tasks. Subjects worked through eight examples before processing the actual test. Results showed that performance varied with age with older subjects yielding higher scores. The authors related the improvement of performance with age to a more frequent use of the magnitude and direction strategy. Significant interaction of age and demand was found. However, contrary to expectance, high-demand problems had impact on the performance of adults and not of younger subjects.

In their study entitled *Do written examples need instruction?*, LeFevre and Dixon (1986) examined the impact of information gained by more specific example instruction as opposed to general, written instruction on the

performance in inductive reasoning tasks (series completion and classification). They compared the use of instruction information versus example information by conducting several experiments using a 15-item test version in experiments 1-3 and a revised version in experiments 4-6.

In their first experiment the influence of conflicting information, i.e. instruction and example suggested different procedures (eg. one suggests series procedure and the other classification procedure), was determined. Conflicting information was tested against non-conflicting information to determine the source of information subjects referred to. Thus, two non-conflicting and two conflicting conditions were generated. Results showed that in nonconflicting items 96% of the consistent (consistently follow one procedure vs. ambiguous or inconsistent behavior) subjects applied the correct procedure, and in conflicting items 92% of the subjects were geared to the example. LeFevre and Dixon attempted to explain the strong preference of following the example and applying its suggested procedure by the similarity of the examples and actual items, although written abstract instruction might represent cognitive procedures used in the task more distinctly. They conducted a series of experiments to elucidate this responding behavior, testing two sets of hypotheses. One referred to the good composition of the examples and stressed the nature of the example (*strong example hypothesis*) and the other referred to the insufficient character of the instructions (*weak instruction hypotheses*) and thus emphasized the nature of the instruction.

Experiment 2 tested a weak instruction hypothesis comparing the application of instruction procedures and example procedures after instructions



or examples were presented to the subjects. The weak instruction hypothesis could not be supported since no significant difference was found and procedures were followed equally often.

Experiment 3 tested if the example effect corresponded to a recency effect due to a higher degree of salience of the example appearing after the instruction. A strong example hypothesis was thus tested and orders of instruction and example were reversed. Results however showed that 93% of the consistent subjects followed the example and only 7% the instruction. The favor of the example did therefore not depend on the order of presentation.

In Experiment 4, LeFevre et al. analyzed if the information regarding incorrect answers, provided by the examples, had impact on the choice to adopt the example procedure. By presenting the correct as well as incorrect answer alternatives, examples indicated inappropriate procedures, referred to as *disconfirming information* which was thought to be responsible for the example effect. Thus another strong example hypothesis was tested. Disconfirming information was shifted from the example to the instruction. Again results showed the favor of the example indicating that disconfirming information did not affect the choice of procedure. Testing a weak instruction hypothesis, Experiment 5 analyzed if the disregard of the instruction information was to be led back to the short and not very detailed nature of the instruction. Instructions were therefore made long and redundant. However, results showed that subjects did not assign such importance to the length of instruction as they still preferred the example and no effect of length of instruction was obtained. Questioning the

subjects on their responding behavior, a higher usefulness and importance was ascribed to the examples. Significant differences between the percentage of subjects preferring the examples (76%) and the instructions (24%) were yielded. Referring to this type of task, 50% thought that examples were better in general and the authors concluded that the results proved a weak instruction hypothesis and that disregard of instruction was not due to its insufficient nature but to the assumption that "instructions are weak by virtue of subjects' belief about the usefulness or importance of instructions" (p. 24). In their last experiment, LeFevre and Dixon (1986) therefore tested the belief hypothesis assuming that subjects followed the example because they assigned the instruction no importance. Importance of instructions was manipulated by not only making them visually more salient but also by telling the subjects that the instructions were very important. The use of instruction procedures increased but still only a third of the consistent subjects followed the instructions in the conflicting condition.

In general, these experiments showed a robust example effect insofar as examples were generally preferred compared to instructions regardless of order, disconfirming information, length of instruction, or emphasis on instruction. However, presenting instruction or examples individually, no preference or superior performance could be observed. Possible practical implications inferred are, that when information is meant to be transferred via instruction information, additional examples should not be presented as well since the instruction would be disregarded. In general, however, application of examples should also depend on their accurateness, appropriateness and distinctiveness.

### 2.5.3 Concluding Remarks on the Effects of Training and Instruction

According to Freund (2008) “concluding from the empirical evidence on practice effects in intelligence testing, it seems that the extent of a general practice effect is rather small” (p. 24). However, specific practice effects of different extent result and depend on test, item, and context features and cognitive dimensions as shown in the moderator analysis by Hausknecht et al. (2007). In terms of analogical reasoning, training effects, as results of in between test interventions, could be exemplarily demonstrated on the basis of three studies showing score gains in geometric, respectively verbal analogy tests when subjects were retested.

With regard to the present study, implications of these findings are important when regarding the construction of this study’s analogy test in terms of pure research for higher ranking objectives. Adaptive testing and computerized item generation could be considered as such as objectives as they represent means to construct numerous alternate and parallel test forms to enable retesting and experimental manipulation to examine training effects for the figural analogy test.

Empirical findings regarding test instruction on test performance appeared less frequent when literature was reviewed. Findings referring to effects of instruction on individual differences in analogical reasoning were reported in the preceding section. Further the impact of instruction information vs. example information on an inductive reasoning test was outlined. Evidence for the influential power of instruction on performance in reasoning tests was thus obtained and aroused interest regarding the rule-

based figural analogy test of this study. Since one research question should address the effect of instruction - in fact the effect of explaining the cognitive operations involved in the solution process - on test performance, the findings reported above guided the expectancies of that research question (see chapter 4).

# 3

## **Analysis of Cognitive Structure**

The topic of this chapter refers to the analysis of cognitive structures. The intention of this chapter is to emphasize on the importance and practical implications of rule-based test construction and to provide a basis for understanding the procedure of analyzing cognitive task structures. The issue and advantages of rule-based test construction are elucidated in the first section of this chapter. Then test theories are shortly outlined before methods of validating the cognitive structure are introduced in the third part of this chapter. Subsequently, empirical findings on information struc-

ture and components of item difficulty are presented. This will conclude this chapter of the analysis of the cognitive structure.

### **3.1 Rule-based Test Construction**

As already stated in the introduction of this dissertation, the present study is part of a complex research project concerning the rule-based test construction of various cognitive ability tests and the implementation of computerized generation algorithms. Along with this research project, Freund, Hofer & Holling (2008) previously illustrated that “matrix items can be decomposed into different basic components. This composite character makes the task format a good choice for rule-based item construction” (p. 195). Thus, the ability to rationally construct items of that task type enabled Freund et al. (2008) to implement algorithms and generate figural matrix items by the computer.

Since rule-based test development and implementing computerized algorithms for item generation already proved to be successful for figural matrix items (Freund et al., 2008), the same process should be accomplished for another task type. The present study was therefore designed to provide basics for the superior objective of implementation of computerized generation algorithms for another task type, namely figural analogies. Thus specification of task parameters, fit with the Rasch model, and analyses within the linear logistic test models (which are introduced later on in this chapter) are prominent topics of this study.

Specification of task parameters refers to identifying the parameters in-

volved in solving the items, such as the cognitive operations that have to be applied to correctly answer the items. Like figural matrix items, figural analogy items feature a composite character. Cognitive operations, defined by the relation of elements (eg. relation between the A and B term of an item), can be pre-specified and constitute construction rules according to which items are generated. Besides applying only one cognitive operation per item, different cognitive operations can be applied in single items. As opposed to arbitrarily construct items, a method of systematically and rationally composing items is possible. Besides the advantage of knowing the contributions of the task parameters to item difficulty, a cognitive theory of the item solving process can be developed. If the cognitive theory can then be validated, statements on the construct validity of the items can be made.

The ability to sufficiently explain item difficulty by the task parameters, entails several important implications. Knowing the contribution of each parameter to item difficulty enables one to create items of designated difficulties. This is a tremendous advantage, for example, regarding adaptive testing: By applying items of adequate difficulty, regarding the ability level of the subject tested, psychometric advantages such as minimizing the standard errors of the ability estimates result. Usually in traditional paper-pencil testing all subjects process the same items without considering that certain items might be too difficult or too easy for some subjects and in turn provide only little, if any, information. This is at the expense of motivation and test economy because a large number of items has to be applied for a reasonable precise ability estimate.

According to Freund et al. (2008) “the main advantages of computerized item generation are increased economy, avoidance of construction errors, and higher comparability of items due to more stringent construction algorithms” (p. 201). Further issues of computerized item generation (for a more detailed outline see Freund et al., 2008) are not elucidated at this point. Computerized item generation, however, still remains a long-term objective for the figural analogy test, but its premises have to be analyzed and accounted for first and are thus the object of the present study. Therefore, the rule-based construction of a figural analogy test according to an underlying rationale and the evaluation of the cognitive structure, not only by means of the linear logistic test model but also by means of the linear logistic test model with random item effects, constitute the central topic of the present research.

## **3.2 Test Theory**

In general, test theory refers to assumptions on the relationship between traits and empirically obtained test scores. Objectives of test application are reliable statements on the characteristic of the measured trait. In this section, besides stating the characteristics of the classical test theory (CTT), important features of the probabilistic test theory are outlined and its different models are presented.



### 3.2.1 Classical Test Theory and Item Response Theory

The item response theory (IRT) is based on the fundamental assumption that the response to a certain task is determined by the underlying ability. Since the latent trait is not observable but assumed to be continuous, measurable characteristics influenced by the latent trait are operated. Item responses are used as indicators for this underlying basic disposition and the probability of making a certain response is a function of the latent trait. This theory, making use of the probability, is therefore also referred to as probabilistic test theory which describes the relationship between a subject's ability level and the likelihood to solve a task. IRT assumes that the probability of solving a task varies as a function of the feature characteristic. High-ability subjects are more likely to solve tasks than less competent subjects. The relationship between ability and solution is therefore described in terms of likelihoods: The probability of certain discrete responses to an item is described as the function of the latent trait or person parameter. It is assumed that the resulting function that represents the relation between ability score and solving probability is monotonically increasing. Higher ability therefore corresponds to a larger likelihood in solving the item.

Amongst other differences, theories differ in assumptions concerning test reliability. In classical test theory reliability is defined by the error of measurement which equals the discrepancy between the individual's true score and observed score. The extent of measurement precision is usually expressed by an index describing the test's average reliability. Unlike classical test theory, the item response theory enables one to calculate test

reliabilities for different ability scores as it presumes that reliability is not the same for different test scores. Statements on measurement accuracy for different ability levels can therefore be made. Models based on the IRT, such as the 1-parameter logistic model (1PL model), also assume that estimated parameters do not depend on items applied or samples tested. On the contrary, in CTT the ability estimate does not only depend on the test taker but also on the difficulty and content of the tasks applied.

Task parameters of classical test theory are therefore population and sample dependent, which means that according to the sample tested and its characteristics, different values for test parameters can be obtained. The one-parameter logistic model therefore suggests specific objectivity which refers to the idea that the estimation of item parameters is independent of the estimation of person parameters and vice versa. Thus, due to advantages offered by the probabilistic test theory, the present study aims that items are scalable according to the IRT to examine the disposition measured (ability score, latent construct) and item features such as difficulty and item discrimination.

The item response theory offers great opportunities to further examine tasks components. The claim of rule-based item construction can be reassessed by means of the linear logistic test model (LLTM) by Fischer (1972). The characteristics of the LLTM are described in the subsequent part of this chapter after different probabilistic test models have been presented.

### 3.2.2 Probabilistic Test Models

Probabilistic test models differ in their number of parameters assumed but share some joint features. Common constituents of probabilistic test theory are unidimensionality and local stochastic independence. Unidimensionality refers to the assumption that all items applied in the test measure the same latent construct in all examinees. Unidimensionality is therefore a “property of the items” (Lord, 1980, p. 20). In the 1PL model unidimensionality implies that all items measure the same construct. In multidimensional models however, thus in tests measuring more than one latent trait, the requirement of unidimensionality refers to groups of items among which unidimensionality should be found.

For test items that are dichotomously scored, IRT distinguishes three different models that commonly assume one-dimensional underlying latent traits but differ in the number of parameters specified in each model. These models are presented below and are generally referred to as the one-parameter logistic (1PL) model, also known as the Rasch model (Rasch, 1960), the two-parameter logistic (2PL) model, and the three-parameter logistic (3PL) model (Birnbaum, 1968). Person parameters and item parameters are located on the same scale and are therefore directly comparable.

*1PL model.* Within item response theory the Rasch model (Rasch, 1960), also referred to as the 1PL model, constitutes a well-known model and is fundamental to all item response models. The response format is dichotomous and the 1PL model contains only one parameter, the item difficulty parameter ( $\sigma$ ). It therefore postulates that all items have the same item

discrimination and can thus be displayed as parallel functions on a continuous ability axis. Location on this axis represents item difficulty. In the 1PL model (equation 3.1) the relation between the latent trait, item difficulty and the likelihood to solve the item is defined as follows:

$$P(x_{vi} = 1) | \theta_v, \sigma_i = \frac{e^{(\theta_v - \sigma_i)}}{1 + e^{(\theta_v - \sigma_i)}} \quad (3.1)$$

In equation 3.1,  $P(x_{vi} = 1)$  denotes the probability that examinee  $v$  solves item  $i$ ,  $\theta_v$  denotes the person parameter of person  $v$ , and  $\sigma_i$  the item difficulty parameter of item  $i$ .

Item characteristic curves for items of the 1PL model are on the same latent scale and parallel positions of the items indicate identical item discrimination parameters and only different difficulty parameters.

*2PL model.* The 2PL model goes back to Birnbaum (1968) and contains two parameters: item difficulty ( $\sigma$ ) and item discrimination ( $\beta$ ). Contrary to the Rasch model, items may vary in their discrimination, thus in their ability to differentiate between test takers. Highly discriminating items effectively distinguish between good and bad test takers. The function describing the relationship between item difficulty, item discrimination, and likelihood to solve the task can be expressed as follows:

$$P(x_{vi} = 1) | \theta_v, \sigma_i, \beta_i = \frac{e^{\beta_i(\theta_v - \sigma_i)}}{1 + e^{\beta_i(\theta_v - \sigma_i)}} \quad (3.2)$$

In equation 3.2,  $P(x_{vi} = 1)$  denotes the probability that examinee  $v$  solves item  $i$ ,  $\theta_v$  denotes the person parameter of person  $v$ ,  $\sigma_i$  the item difficulty parameter of item  $i$ , and  $\beta_i$  indicates the item discrimination parameter of item  $i$ .

*3PL model.* The 3PL model is not only composed of the parameters item difficulty and item discrimination, but also contains a guessing parameter ( $\gamma$ ). The parameter refers to the probability with which examinees can solve an item through guessing. For items with a multiple choice response format this probability can be deducted from the number of response alternatives. The function describing the 3PL model is displayed below:

$$P(x_{vi} = 1) | \theta_v, \sigma_i, \beta_i, \gamma_i = \gamma_i + (1 - \gamma_i) \frac{e^{\beta_i(\theta_v - \sigma_i)}}{1 + e^{\beta_i(\theta_v - \sigma_i)}} \quad (3.3)$$

In equation 3.3,  $P(x_{vi} = 1)$  again denotes the probability that examinee  $v$  solves item  $i$ ,  $\theta_v$  describes the person parameter of person  $v$ ,  $\sigma_i$  refers to the item difficulty parameter of item  $i$ ,  $\beta_i$  is the item discrimination parameter of item  $i$ , and  $\gamma_i$  indicates the guessing parameter of item  $i$ .

Since the 2PL model and the 3PL model have item discrimination parameters, specific objectivity, an important property of the 1PL model, cannot be sustained in these models. Rasch used the term specific objectivity to describe that comparisons between persons and therefore the ranking of examinees according to their ability remain the same even if different items

are used. Reciprocally, specific objectivity implies that estimates of item parameters do not depend on the test takers.

Psychometric information on items is provided by item characteristic curves. Item response functions, also referred to as item characteristic curves, display the characteristics of an item. The item parameter difficulty determines the item's location on the ability scale. The item discrimination parameter determines the slope and the guessing parameter determines the intercept on the y-axis. Item characteristic curves can be transformed into item information functions which offer enormous information about the property of the item. They provide information on how informative an item is (i.e information on how accurate an item measures the latent construct). This information is important when choosing items to measure a certain ability level. Item information serves as a parameter introduced by IRT, similar to the standard error of measurement and reliability. The item information is important for adaptive testing as it enables item selection so that items adequate in difficulty can be selected for each examinee.

### **3.3 Methods of Validating the Cognitive Structure**

In order to predict item difficulty, to construct or choose items with an adequate difficulty regarding the ability of the test taker, sufficient knowledge on how each item component contributes to item difficulty is required. The item difficulty can be specified using the cognitive operations involved in that item by relating the psychometric property (item difficulty) to item components (cognitive operations). Performance on items can thus be de-

scribed as a function of cognitive operations involved in that item.

Gaining insight into this relationship enables one to design items of a desired difficulty and selection of items appropriate to the examinee's ability level. This, in turn, offers tremendous opportunities in assessing abilities since procedures such as adaptive testing may be applied. However, mere knowledge of the cognitive operation involved is not sufficient. The impact of the operation needs to be quantified and a validation of the cognitive structure is required.

Two of such methods are described in the following: The linear logistic test model (LLTM) and its extension the linear logistic test model with random item effects are introduced as means to validate cognitive structures. Subsequently a short outline of optimal designs as means to provide balanced test designs is presented.

#### **3.3.1 Linear Logistic Test Model**

Among the models of the item response theory, the 1PL model enables the application of the linear logistic test model (Fischer, 1972) to examine if rule-based item construction fulfills its demand. The LLTM can only be applied to 1PL models as other parameters than item difficulty such as item discrimination are not accounted for in the model. In an analogy test each item has certain structure characteristics such as elements, transformation rules or cognitive operations. The resulting item structure can be related to item difficulty. The impact of these features on item difficulty can be analyzed and the claim of rule-based item construction can thus be validated. The LLTM therefore constitutes a means that requires pre-experimental hy-

potheses on item structure characteristics which are then to be tested. To analyze the factors that constitute item difficulty, procedures decomposing the items into its components can be applied. Embretson (1984) described the LLTM as “a unidimensional model in which components are identified from item scores on complexity factors that are postulated to determine item difficulty” (p. 176).

In the LLTM the item parameters are a function of basic component parameters representing the cognitive components involved in solving an item. The components, assumed to influence the solving process, can be specified in a design matrix in which every item has a certain predefined value on each component. The difficulty ( $\sigma$ ) of each item is additively composed of the basic components ( $\eta_j$ ) required to solve the item  $i$  and a normalizing constant ( $c$ ) (equation 3.4) (Fischer, 1983). Equation 3.5 again illustrates the constitution of the item difficulty.

$$\sigma_i = \sum_j q_{ij}\eta_j + c \quad (3.4)$$

$$\sigma_i = q_{i1}\eta_1 + q_{i2}\eta_2 + q_{i3}\eta_3 + \dots + q_{ij}\eta_j \quad (3.5)$$

The basic parameters ( $\eta_j$ ) represent the difficulty of each cognitive operation involved in the item and  $Q$  is a matrix of weights ( $q_{ij}$ ). The matrix therefore represents single weights indicating if, and if applicable how often, the cognitive component or operation is needed.  $Q$  can thus adopt integers such as 0 (operation/cognitive component not required), or 1 (op-



eration/cognitive component required), or  $> 1$  if that operation needs to be applied more than once to solve the item. However, the weights ( $q_{ij}$ ) are not estimated but represent predetermined hypotheses of the component structure of the item.

The LLTM, as a unidimensional model, constitutes an extension of the Rasch model (equation 3.1), taking into account the cognitive operations involved in solving an item. Equation (3.6) expresses the linear logistic test model according to Fischer (1972):

$$P(x_{vi} = 1) = \frac{\exp\left(\theta_v - \sum_{j=1}^h q_{ij}\eta_j - c\right)}{1 + \exp\left(\theta_v - \sum_{j=1}^h q_{ij}\eta_j - c\right)} \quad (3.6)$$

Further, the goodness of fit of models with different numbers of complexity factors can be compared (Whitley & Schneider, 1981) and the difficulty of items in terms of operations etc. can be explored and explained. Whitley and Schneider analyzed analogical reasoning tasks and reported that different transformations necessary to compose the analogy have different impact on the difficulty of the item. These findings are presented in detail in the section of empirical findings on information structure and components of item difficulty.

Generally, the LLTM as method of explaining the cognitive components that constitute item difficulty has important practical implications as the knowledge of the factors composing the item difficulty helps predicting the difficulty of new items. Concerning adaptive testing, the necessity of large

item pools for examinees with different proficiency levels is facilitated.

### 3.3.2 Linear Logistic Test Model with Random Item Effects

Since in item response theory correct responses to an item depend on characteristics of the person tested and on item properties, the probability of solving the item can be described as a function of person and item effects. “Person effects are usually defined as a random sample from a population distribution” (Van den Noortgate, De Boeck, & Meulders, 2003, p. 369). In the traditional LLTM item effects are fixed and person effects are random. However, when additional item variation is assumed, “models with *crossed random effects* are obtained” (Janssen, Schepers, & Peres, 2004, p. 189). According to Janssen et al. random item effects are, for example, referred to when applying “domain-referenced testing” as items are randomly chosen from a predefined item pool of that domain. However, Janssen et al. distinguish between random item variation due to item properties and item variation due to belonging to item groups. Thus item property or item group can be predictors of item difficulty.

Janssen et al. introduce a relaxation of the LLTM in the following regression model where the Rasch item difficulties  $\beta_i$  are explained by the item predictors  $X_{ik}$  and  $\varepsilon_i$  describing the item-specific error term with  $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ :

$$\beta_i = \sum_{k=0}^K \beta_k X_{ik} + \varepsilon_i \quad (3.7)$$

Thus by applying the traditional linear logistic test model accurate pre-

dictions of item difficulties by item properties can hardly be yielded due to the strict assumption of the LLTM. The LLTM including random item effects can, however, be described as an LLTM with relaxed assumptions: According to Rijmen and De Boeck (2002, p. 277) the LLTM is extended and relaxed “by adding an item-specific error term to the decomposition of the item difficulty in terms of the basic parameters”. Thus by adding an item-specific error and obtaining a model with random person effects and random item effects, a model with diminished assumption for decomposing item difficulty is created. Referring to the error term, Rijmen and De Boeck (2002) explain that normal distribution is presumed and “hence, the item difficulties are considered to be random effects” (p. 277). They further state that the incorporation of the error term enables to condemn the requirement “that items with the same combination of values on the predictors ... are of equal difficulty” (p. 277). Rijmen and De Boeck ascribe variability in item difficulty to item characteristics such as item content. They further accomplish that “in contrast with the RWLLTM [random weights LLTM] in which the effects are random across persons, the effects in the random-effects LLTM are random across items belonging to the same item group” (p. 277).

The linear logistic test model with the assumption of additional item variation therefore represents a model with less strict assumptions and thus represents a more realistic approach. Both models, the LLTM and the LLTM with random item effects, are applied to the data of the present study. Accordingly, results are presented in chapter five.

### 3.3.3 Optimal Design

Not only methods of evaluating the cognitive structure, such as the LLTM, are important among the process of examining the rule-based test construction approach, but also means to combine task parameters within the test design present the first step of this process. The aims of the present study to construct a rule-based figural analogy test and to analyze the difficulty of the items by decomposing the items into its components shall be pursued referring to efficient designs as basis of the test. In a full-factorial design all factor levels (cognitive operations) are combined, and thus all possible combinations are presented in the experiment. However, depending on the number of factor levels such an experiment could be very time consuming and might have an impact on motivation and test performance and, regarding practical aspects, such an experiment could not efficiently be conducted.

The efficiency of designs captures the degree of balance and orthogonality of an design and can be quantified. "The D-optimal design minimizes the generalized variance of the parameter estimators. This is done by minimizing the determinant of the variance-covariance matrix of the parameter estimators or, equivalently, by maximizing the determinant of the information matrix" (Goos, 2002, p. 14). Since the determinant of the variance-covariance matrix specifies the standard errors of estimating the parameters, D-efficient designs yield small standard errors and the aim of well-estimated parameters can be pursued. In conclusion optimal designs can be regarded as an approach to efficiently select factor level combinations and reciprocally optimal efficiency is obtained when designs are

orthogonal and balanced. The analogy test of the present study was constructed according to efficient designs and the test design and test material of the main examination are presented in chapter 5.1.1.

### **3.4 Empirical Findings on Information Structure and Components of Item Difficulty**

Different studies regarding the information structure and components of item difficulty of figural or geometric analogy tests were reviewed and are presented in the following paragraphs. The empirical findings presented serve to examine the impact of transformation rules on item difficulty regarding the choice of operations for the test construction. Findings are of particular importance for the research questions stated in chapter 4 and when evaluating and discussing the results of this study's figural analogy test.

#### **Whitely & Schneider (1981)**

Whitely and Schneider (1981) explored the information structure for geometric analogies and judged that information structures contain important practical implications as tests or items of certain difficulty can be constructed "by controlling information structure" (p. 396). Information processes influence performance and can be examined by cognitive component analysis. Whitely and Schneider analyzed how information processes are related to item difficulty. They applied 30 analogy items of the Cognitive Ability Test (Thorndike & Hagen, 1974) and used the linear logistic test

model to investigate how transformations and elements contributed to item difficulty. Different transformations occurred in the test: number (adding, removing, dividing elements), shade, size, shape, rotation, reflection, and spatial exchanges. In one model, these transformations were assigned to two groups of transformations: spatial displacement and spatial distortion. The transformations referring to disorientation of elements from A to B such as rotation, reflection and exchanges were assigned to the group of displacement. The transformations size, shade, shape, and number constituted the group of spatial distortion. Three models with different levels of parsimony were proposed. Besides transformations, each model also took the elements of the A stimulus into account. Whitely and Schneider compared the different models in terms of how different transformations related to item difficulty. Different transformations yielded different impact on item difficulty.

The first model did not distinguish between the different transformations but referred to the number of transformations only and was judged as the most parsimonious. However, results showed that the number of transformations did not have a significant impact on item difficulty. The second model categorized the number of transformations into two classes referring to spatial displacements and distortions. This model obtained better model fit and proved that the two types of transformations significantly influenced item difficulty. With the number of displacement-transformations, the prediction of items with high difficulty was possible, whereas the number of distortion-transformations enabled the prediction of items with low difficulty. The number of elements did not have a significant impact on

item difficulty in this model. The third and least parsimonious model took into account each transformation separately. This model yielded the best model fit but led only to an increase of accounted variance of 7%, which is, according to the authors, disproportionate to the complexity of the model due to five more parameters. Results proved that transformations had different impact on item difficulty. The transformations number and shading did not significantly influence item difficulty. All other transformations were highly significant in impact with again opposite effects on item difficulty for different types of transformations: The transformations shape and size, thus referring to distortion, led to lower item difficulty and transformations referring to displacement (reflection, spatial exchanges, rotation) led to higher item difficulties.

Summarizing the findings of Whitely and Schneider (1981), transformations referring to spatial displacements led to an increase in item difficulty whereas transformations referring to distortion caused an opposite effect (the “number of spatial displacement transformations was positively related to item difficulty, while number of spatial distortions was negatively related”, p. 395). Important practical implications therefore concern test construction and item design as the item parameter difficulty can systematically be influenced by the type and number of transformations chosen.

#### **Mulholland, Pellegrino & Glaser (1980)**

Besides proposing a solution process model of geometric analogies (cf. chapter 2), Mulholland, Pellegrino and Glaser (1980) also analyzed the nature and impact of transformations applied in analogy tasks by exam-

ining the performance on true and false analogy tasks. They stated that “item difficulty appears to be related to increases in both the number of elements and transformations in an item” (p. 258). They applied six types of elements (line, triangle, circle, cross, rectangle, pentagon) and six transformations (identity, increase in size, 45 degree rotation to the right, reflection about the x-axis, addition of an element, removal of one-half of an element). Analyzing the factors contributing to item difficulty, the number of transformations yielded a significant effect on error rates in true analogy tasks. Number of transformations and number of elements interacted and significantly effected the amount of errors made. However, the number of elements alone did not influence item difficulty in terms of verification errors but an increase of elements led to an increase in solution times. The processing of transformations was more time consuming than the processing of elements. Analyses of the performance on false analogies showed that the detection of the first false partition of an item led to termination of processing and judging the item as false. The information on the false partition could stem from an element or transformation, but it seemed that elements were analyzed before transformations were analyzed. Processing time increased with an increase in the number of elements and transformations and an increase in transformations conveyed error rates.

Mulholland et al. concluded that “the largest single source of errors was multiple transformations of single elements” (p. 282), thus an increase in the number of transformation to be performed. This can be referred to as transformational complexity which promotes verification errors and therefore item difficulty. Working memory load is assumed to account for



an increase in error rate due to an increase in transformational complexity and thus an increased amount of information needs to be stored and processed in working memory.

Regarding the results of the precedent study, Whitely et al. (1981) stated that the findings of Mulholland et al. (1980) did not comply with their findings in which types of transformations differ in their impact on item difficulty.

#### **Novick & Tversky (1987)**

Novick and Tversky (1987) analyzed the order of cognitive operations in geometric analogy tasks. They proposed that subjects processing geometric analogy tasks would perform cognitive operations in a specific order of precedence and that this sequence would be applied to reduce working memory load. They assumed that the difficulty of each transformation would be predictor of transformation order and that operations would be performed in order of decreasing difficulty to diminish working memory load. In the first experiment of the study of Novick and Tversky (1987) two groups of subjects worked on 21 geometric analogy items with two transformations each. One group of 48 subjects was instructed to identify the operations that had to be applied to transform the A term into B. Subsequently they had to mentally construct the D term by applying the identified operations onto the C term. They then had to sequence the operations by numbering them according to the order of application. With an average Spearman-rank order correlation of .44, the authors concluded on concordance ( $\omega = .45$ ,  $\chi^2(6, N = 48) = 129.12$ ,  $p < .001$ ) between sub-

jects on the sequence of cognitive operations (move, rotate/reflect, remove, size, add, shading) . The difficulty of the eight different types of transformations (rotate, reflect, move, size, add half, add part, remove, shading) applied was examined to analyze the assumption that cognitive operations were serially performed in order of decreasing difficulty to reduce working memory load. The transformation rotate was the most difficult one (15.8 % errors) followed by size (13.1%), reflect (11.0%), shading (10.6%), add half (10.4%), move (7.7%), add part (6.3%) and remove (4.2)%. Transformations were not applied in decreasing order of difficulty and the working-memory hypothesis could not be confirmed.

This hierarchy of transformation difficulty is almost consistent with the research of Whitely and Schneider (1981) who confirmed higher item difficulty for tasks containing spatial displacement in contrast to tasks implying distortions. Therefore - neglecting the operation size being the second most difficult operation - in order to construct tests of higher difficulty level, tasks should obtain more spatial displacements. To create easier tasks, transformations involving distortions should be applied.

To examine if the order of transformations was influenced by item difficulty, Novick and Tversky (1987) constructed a test of 12 analogy items with three transformations each and tested a group of 59 students (Experiment 3). Pairwise comparison across 51 subjects was conducted and a Spearman correlation of .19 was obtained. A Spearman correlation referring to rank-order of .91 between the transformation sequence of this experiment and Experiment 1 was yielded. The transformations move and rotate/reflect changed position and move became the second transforma-

tion in the sequence instead of the first (Experiment 1). Contrary to Experiment 1, where high and low ability subjects did not differ in preferred transformation order or in the degree of consistency concerning the ordering, in Experiment 3 ability groups differed: High- and low-ability groups showed diverse performance. Low-ability subjects did not process items with a consistent pattern of transformation ordering and for middle-ability subjects the application of a certain order was not as consistent as for high-ability subjects, but transformations were applied in a likewise sequence. Regarding the very similar transformation sequence of Experiment 1 and Experiment 3, Novick and Tversky inferred that "...there must be cognitive constraints other than difficulty operating to determine transformation ordering" (p. 62).

#### **Bethell-Fox, Lohman, & Snow (1984)**

Among their complex study on componential and eye movement analysis of geometric analogy performance, Bethell-Fox, Lohman, and Snow (1984) also analyzed the impact of figural and spatial transformations, respectively the cognitive component processes accompanying these transformations. They suggested that not all processed items involved the same processing components but differed in the cognitive components activated depending on item characteristics. They tested spatial and figural transformations, assuming that spatial transformations evoked additional processes. Spatial transformations referred to mental rotation and reflection and figural transformations referred to halving, size-change and doubling of elements. Results revealed that spatial transformations were positively

related to latencies and errors. They also yielded higher correlations with the fluid-visualization ability composed of scores of tests capturing the constructs fluid-analytic ability and visualization.

Bethell-Fox et al. assumed that these findings were due to additional cognitive components involved in items containing spatial transformations. They suggested spatial inference and spatial application as such components. Spatial inference was thought to be the process of inferring the spatial transformation between the terms A and B, and spatial application was thought to be necessary to apply this inferred transformation onto the C term to find the appropriate D term. Analyzing the fit of different models, Bethell-Fox et al. were able to validate their suggested components. The components of spatial inference and spatial application were only relevant for items involving spatial transformations and it seemed that these processing components could be activated or deactivated depending on the item. An additional finding referred to the justification component of Sternberg (1977a): Justification was only important when ambiguous items had to be processed and not when unambiguous items had to be processed. In ambiguous items the surmised correct answer was not presented among the response choices whereas in non-ambiguous items it was. Thus the justification component could also be activated depending on requirements of the tasks.

### 3.5 Summary

Chapter 3 was designed to make the origination of the research questions presented in chapter four, and the statistical analyses conducted and presented in chapter five, comprehensible, by consecutively introducing the three parts of the chapter. First, the issues and amenities of rule-based test construction as an approach to rational and objective item generation were outlined. Then, in order to understand the procedures applied in this study to evaluate, if rule-based test construction was successful and to judge if the pre-defined task parameters could reasonably well explain item difficulty, part two referring to test theory and part three presenting methods of validating the cognitive structure were included. Classical test theory as opposed to item response theory was presented and three item response theory models were shortly introduced. As methods of validating the cognitive structure two approaches were presented differing in their strictness of implied assumptions. Both approaches aim at explaining item difficulty by relating the psychometric property (item difficulty) to item components. Contrary to the LLTM, the linear logistic test model with random item effects assumes an additional error and represents a more realistic way to evaluate the contribution of different task parameters to item difficulty. Finally, empirical evidence concerning the information structure and components of difficulty of geometric analogy items was featured.

# 4

## Research Questions and Pretests

The first section of this chapter addresses research questions of the present study. Research questions resulted from two sources. The first source refers to the general superior objective of this study: rule-based test construction of a figural analogy test. The first question addresses the requirements of different analyses. The second source relates to the theoretical assumptions and empirical findings of chapter 2 and 3. Based on these findings, the second and third research questions evolved. The second part of this chapter presents the pretests that were conducted before deciding on a final test

version and procedure. Administration of these pretests resulted from the first research question regarding the rule-based test construction of the figural analogy test. Pretests were required to provide basic research and first results.

## 4.1 Research Questions

Looking at the role of reasoning in more ancient as well as recent theories on human intelligence, reasoning ability or fluid intelligence constitutes a central construct among the theories. Thus, since it seems that fluid intelligence is such a distinct indicator of general intelligence, it is beyond question why construction and examination of tasks measuring reasoning ability, and therefore fluid intelligence, shall be conducted and why the application of such tasks to assess general intelligence is subject-matter of this dissertation.

Three research topics can be defined: The first research question arose within the topic of rule-based test construction of a figural analogy test. It refers to the rational construction of test items and thus the task parameters as fundamental components of the items. Because of the composite character of figural analogy items, items can in turn be decomposed into its components such as transformation rules and elements. To predict item difficulty, decomposing items into its components and analyzing the impact of the single components on item difficulty is necessary. Practical relevance of the prediction of item difficulty - amongst others - lies in the

application of adaptive testing and accurate estimation of ability. The first research question therefore is:

1. What constitutes item difficulty?

In the context of this question the impact of the transformation rules shall eminently be analyzed. According to empirical findings among the analysis of the cognitive structure (eg. Whitely & Schneider, 1981) the following specific hypotheses are formulated: Transformations referring to spatial displacements (eg. reflect, rotate) increase item difficulty and transformations referring to spatial distortions (eg. size) decrease item difficulty. Further, the increase in the number of transformations per item is assumed to enhance item difficulty, too.

The linear logistic test model as means to estimate the contribution of the task parameters to item difficulty is applied to analyze the data of the main examination. However, contrary to previous studies that only examined components of item difficulty by means of the linear logistic test model, the present study additionally applies the LLTM with random item effects. The advantages of the LLTM with random item effects (cf. chapter 3) account for a substantial progress of this research. Further, besides analyzing the impact of the transformation rules, the impact of elements as constituents of item difficulty is also addressed. However, a directional hypothesis concerning the impact of elements on item difficulty is not presented, but their influence is nevertheless evaluated.



The second research question addresses the effect of different instruction modes on test performance. It is assumed that the performance of subjects introduced to the transformation rules differs from the test performance of subjects that did not receive an introduction to the transformation rules. Thus the second research question is formulated:

2. Does the introduction to the rules that determine the relation between A and B influence test performance?

To explore this issue the research design provided randomly assigned subjects to two groups (one with instruction of rules and one without instruction of rules). On basis of findings of effects of training and instruction it is assumed that the instructed group gains benefit from the knowledge of the transformation rules and thus outperforms the non-instructed group. The specific hypothesis thus reads as follows: The instructed group is expected to perform significantly better than the non-instructed group.

Further, subgroup analyses are supposed to examine if the impact of the basic constituents on item difficulty differ between the instruction group and the non-instruction group. Thus, besides analyzing the difference in performance of the two groups with the assumption that the instruction group outperforms the non-instruction group it shall be analyzed if difference in performance goes along with the different impact of the transformation rules on item difficulty. The specific hypotheses for this issue reads as: The impact of the transformation rules as basic components of the item does not differ for the groups in terms of order of difficulty and absolute

value. It is assumed that, although the subjects of the instruction group are expected to perform significantly better on the analogy test due to benefiting from the knowledge of the transformation rules, the degree of influence of the single transformation rules on item difficulty are the same for both groups as they are not expected to differ in their cognitive structure and processes.

The third and last research question concerns scholastic achievement as external validation criteria. Literature reviews in chapter two proved that cognitive abilities can validly predict academic performances. The criterion-related validity regarding academic achievement is analyzed for the rule-based analogy test of this study. The corresponding research question is:

3. Can performance on the analogy test be related to academic performance?

Referring to the analysis of the relation between intelligence and academic performance following assumptions due to the empirical findings reported in the theoretical background are made: Intelligence as measured by the figural analogy test is expected to moderately correlate with average school grades (still a high predictive validity compared to other criteria). Among the school grades, mathematics and sciences are supposed to show higher correlations with the analogy test score than subjects referring to languages, social sciences or music and art. The corresponding hypothesis is: Compared to other school grades, maths & natural sciences correlate higher

with the analogy test score.

## 4.2 Pretests

Three different pretests were performed prior to the main examination. The pretests pursued different purposes. Firstly, by accomplishing the pretests, the rational item construction approach applied in this study should be scrutinized for the first time. Thus, apart from specific hypotheses and research questions expressed for the main examination, examination of the rule-based construction of the test served to evaluate whether the transformation rules applied in the test significantly contributed to item difficulty. The direction and magnitude of the impact of the transformations were also a matter of particular interest. The psychometric properties of the items had to be analyzed before deciding on a final test version. Psychometric properties involve parameters such as item difficulties and item discrimination and these parameters were examined by applying the preliminary test versions. Further the reliability and validity of the test should be examined. Before presenting the pretests and their results, the test format underlying the pretests is briefly outlined.

A general note concerning the results of all pretests refers to the statistical analyses conducted. Since the number of participants of each pretest was rather small, analyses were restricted to classical test theory and analyses according to probabilistic test theory could therefore not be applied in the pretests but in the main examination. Thus, in the pretests, regression analyses were conducted instead of LLTM analyses.

The test format  $A : B = C : ?$  was chosen as the basic format of this figural analogy test. Figural analogy studies were reviewed in order to decide on cognitive transformations (cf. chapter 3). Transformations such as adding, removing, or dividing elements were not applied as transformation rules since they could only be applied to composite elements and not to holistic elements such as letters and digits that were chosen for the present test. Further, transformations such as shade and shape were excluded as they were expected to make items too easy. Operations such as rotation of elements, reflection of elements on its x-axis or y-axis and change in size are frequently applied transformations in analogy research (eg. Whitely and Schneider, 1981; Mulholland et al., 1980) and were thus selected as operations for the figural analogy test. Although the operation size was expected to be related to the easiness of an item (Whitely and Schneider, 1981) it was nevertheless included to ensure variability. Further, the direction operation applied to the dot analogies in the study of Bisanz, Bisanz, and LeVevre (1984) (cf. chapter 2) that referred to the calculating operation of plus or minus gave reason to include such an operation in the present study, namely sequence. Thus, categories of cognitive transformations included in the analogy test were size, rotation, reflection, and sequence. Referring to the category of size, two transformations were allowed: Elements could be increased in size, *size plus* (sp) or decreased in size, *size minus* (sm). In Figure 4.1 the cognitive operations chosen for the analogy test are illustrated. Among the rotation category, three different types of rotation were distinguished: 180° rotation (r180), 90° rotation right (rr), and 90° rotation left (rl). The category reflection contained two types of

transformations: *reflection about the x-axis* (rfx) and *reflection about the y-axis* (rfy). Sequence referred to *sequence plus* (sqp) and *sequence minus* (sqm).

As stimuli, alphabetical letters and single digits were chosen. The rules sequence plus and sequence minus required the participant to go forward or backward in the alphabet when letters were presented and to add or subtract when digits were presented. Elements of the A term and C term had to be different from each other but could vary in their initial position, i.e they could already present transformed elements in the A term or C term (for example a reflected or rotated element etc.). Letters in the A & B terms could be paired with letters or digits in the C & D terms and vice versa. However, A and B, and C and D had to be of the same element category. Five distractors per item were presented plus the alternative that the right solution was not among the response choices.










Transformations of the category rotation	Transformations of the category size
R :  = J : ?    Rotation 90° right	R :  = J : ?    Size plus
R :  = J : ?    Rotation 90° left	R :  = J : ?    Size minus
R :  = J : ?    Rotation 180° right	
Transformations of the category reflection	Transformations of the category sequence
R :  = J : ?    Reflection x-axis	R :  = J : ?    Sequence plus
R :  = J : ?    Reflection y-axis	R :  = J : ?    Sequence minus

Figure 4.1: Cognitive Operations

### 4.2.1 Pretest 1

#### Sample and Procedure

Forty-eight males and 78 females, thus a total of 126 subjects represented the sample of the first pretest. All subjects tested were German year 11 and year 12 students from a local high school (German Gymnasium). The sample tested had a mean age of 18.49 years ( $SD = .74$ ) with a minimum age of 17 years and a maximum age of 20 years. A figural analogy test of 20 items, containing the transformations outlined above, was constructed and applied to provide information on item difficulties, item discrimination and testing procedure. Further, the impact of difficulty of the different cognitive operations chosen should be evaluated. To estimate the extent to which this new analogy test corresponds to other tests measuring reasoning ability, a total of 106 students were also tested with the revised German version of Cattell's Culture Fair Test of Intelligence (CFT-20 R; Weiß, 2006) for validity reasons.

The students were first familiarized with the figural analogy test format and examples were then presented to illustrate the cognitive operations. All testees were told that testing time was limited and that they should follow the test instructor's orders. The analogy test of the first pretest consisted of 20 items, distributed to two subtests for time management reasons. For each subtest the start and halt sign was given by the test instructor and five minutes testing time was allowed for each subtest. Subjects were also instructed that taking any kind of notes was not allowed.

## Results

Item difficulty, defined as the percentage of correct answers per item, ranged from .10 (item 2.7) to .92 (item 1.6) with a mean difficulty of .40 ( $SD = .22$ ). According to Lienert and Raatz (1998) difficulties should vary between .15 and .85. and only one item exceeded the upper bound of .85 and three items the lower bound of .15 and can be regarded as too easy, respectively too difficult, and might therefore not provide sufficient information. Thus, in terms of item difficulty the descriptive statistical analysis of these tasks provided satisfactory results.

Item discrimination measured by corrected item-total correlation ( $r_{it}$ ) ranged from -.07 (item 2.2) to .27 (item 1.1) with a mean of .08 ( $SD = .09$ ). Items therefore did not meet the demand of mediocre discrimination ( $.30 < r < .50$ ) according to Lienert and Raatz (1998). However, looking at the three items with negative discrimination no salient indication was obtained for these insufficient parameters.

Reliability measured by Cronbach's  $\alpha$  for the 20 items applied was  $\alpha = .37$  and hence too low. Statistics for each item are presented in Table 4.1.

To estimate validity, the figural analogy sum scores of 106 students that additionally took the intelligence test CFT-20 R (Weiß, 2006), were correlated with the sum scores of the CFT-20 R. The two sum scores did not significantly correlate ( $r = .08, p = .42$ ).

A regression analysis was conducted to estimate the influence of the cognitive operations on item difficulty. Item difficulty was used as dependent variable and the cognitive operations specified in the design matrix of the test were used as regressors. In some items the operations rotation ( $90^\circ$

left or 90° right) and reflection (about the x-axis or about the y-axis) were used simultaneously (items 1.10, 2.1, 2.2, 2.6, 2.7, 2.9, 2.10). In the design matrix, whenever these operations appeared isochronally, they were not coded as both reflection and rotation but a new joined operation (rot & ref) was created and coded with 1 for these items. Of the 20 items applied, one test item implied only one cognitive transformation (item 7), whereas 12 items had two transformations (items 1-6, 8-11, 13, 14), five items had three transformations (items 15-19) and two items (items 12 and 20) had four transformations. In the description above, the transformation rot & ref was always counted as two transformations.

Table 4.2 presents the regression weights of the parameters. Negative estimates indicate an increase in difficulty, thus a decrease in the probability of solving an item, when applying the respective transformation. The transformation rotation 90° right seemed to be the most difficult one as it reduced the probability of solving an item the most, followed by sequence minus, size plus, rot & ref, reflection x-axis, size minus, and sequence plus. Contrary, the transformations reflection y, rotation 180°, and rotation 90° left made tasks easier. Summarizing the regression analytic evaluation of this pretest one may find that no coherent statement on the impact of categories of cognitive operations may be found. In contrast to the findings of Whitely and Schneider (1981) who concluded that spatial distortions such as size make items easier, the operations size plus and size minus of this pretest decreased the probability of solving an item. According to Whitely and Schneider spatial displacements such as the category of rotation (90° right, 90° left, 180°) and reflection (x and y) should decrease the probability



of solving an item. The results of the first pretest are, however, not consistent with these findings since operations of the same category had opposed impact: The transformations 90° rotation right and reflection about the x-axis increased item difficulty as expected, and others such as rotation left and rotation 180° decreased item difficulty, i.e. made items easier. None of these parameters yielded significant influence ( $p > .05$ ) on item difficulty. This is not surprising since the number of items and subjects tested was rather small. It can be concluded that the parameter estimates are more informative and important as they provide information on the impact of the transformation on item difficulty.

Table 4.1: Descriptive Statistics of Pretest 1

Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>	Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>
1.1	.25	.44	.27	2.1	.32	.47	.23
1.2	.28	.45	.11	2.2	.12	.33	-.07
1.3	.63	.49	.08	2.3	.45	.50	.19
1.4	.79	.41	.00	2.4	.13	.33	.20
1.5	.50	.50	-.02	2.5	.33	.47	.15
1.6	.92	.27	.11	2.6	.53	.50	.01
1.7	.60	.49	.19	2.7	.10	.31	.04
1.8	.25	.44	.09	2.8	.53	.50	.13
1.9	.36	.48	.15	2.9	.29	.46	.14
1.10	.33	.47	-.01	2.10	.18	.39	.02

### Summarized Findings Pretest 1

Evaluation of the first pretest leads to the following conclusions. Item difficulties proved satisfactory: values scattered over almost the whole ability range and showed, with an average difficulty of .40, a medium test difficulty. Item discrimination parameters fell far short and need enhancement.

Table 4.2: Parameter Estimates of Pretest 1

Variables	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Constant	.572	.255	2.248	.051
sp	-.208	.154	-1.351	.210
sm	-.033	.117	-.283	.784
rr	-.288	.307	-.936	.374
rl	.381	.312	1.219	.254
r180	.113	.261	.433	.675
rot & ref	-.199	.224	-.887	.398
rfx	-.126	.251	-.501	.628
rfy	.057	.273	.209	.839
sqm	-.277	.167	-1.659	.131
sqp	-.005	.129	-.037	.971

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rot & ref = rotation and reflection, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

Reliability measured by Cronbach's  $\alpha$  and external validity were both dissatisfying. Regression analytic evaluation revealed no taxonomy of impact of the transformations on item difficulty. Within transformational categories, impact on item difficulty was sometimes contrary. Transformations of mental rotation were positively as well as negatively related to item difficulty, just like reflection. Both size transformations increased item difficulty as did both sequence transformations.

Shortcomings of pretest 1 thus provided the job definition of pretest 2.

#### 4.2.2 Pretest 2

To find out if testing time might have influenced item discrimination and to again analyze the cognitive operations as regressors on item difficulty, a second pretest with extended testing time was conducted. For this test, the

test format, design matrix and cognitive operations and elements remained the same. Again two subtests of 10 items each were applied. In subtest 1 none of the items were changed, but in subtest 2 some items were modified. In item 2.4 the element applied as distractors was exchanged: The digit 7 was used instead of the digit 3 to facilitate visual discrimination. The orientation of each distractor remained the same. In items 2.8 and 2.10 the elements applied in the A, B and C terms were altered and thus the elements used as distractors were altered, too.

Testing procedure and instructions corresponded to the ones of pretest 1 and are therefore not presented again.

### **Sample**

Fifty-eight year 11 to year 13 students were tested. Their age ranged from 16 to 19 years with an average age of 17.53 years ( $SD = .68$ ). Of the 58 subjects, 22 were male and 36 female.

### **Results**

Descriptive statistics are presented in Table 4.3. For this pretest item difficulty ranged from .07 (item 1.10) to .72 (item 1.6) with a average item difficulty of .37 ( $SD = .20$ ). Item discrimination parameters varied from -.10 (item 1.2) to .41 (item 1.5) with a mean of .14 ( $SD = .16$ ).

For the three items amended, descriptive statistics changed from pretest 1 to pretest 2. In item 2.4 the digit 7 was used instead of the digit 3 as distractors, since difficulties in discrimination between the differently orientated digit 3 as distractor might have caused a low percentage of correctly

Table 4.3: Descriptive Statistics of Pretest 2

Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>	Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>
1.1	.36	.48	-.08	2.1	.36	.48	.21
1.2	.26	.44	-.10	2.2	.17	.38	-.02
1.3	.50	.50	.11	2.3	.64	.48	.36
1.4	.69	.47	.07	2.4	.36	.48	.14
1.5	.52	.50	.41	2.5	.57	.50	.26
1.6	.72	.45	.38	2.6	.47	.50	.29
1.7	.38	.49	.31	2.7	.21	.41	.18
1.8	.09	.28	.00	2.8	.38	.49	-.05
1.9	.10	.31	.06	2.9	.24	.43	.17
1.10	.07	.26	.10	2.10	.22	.42	.08

solved items. Applying the digit 7 in item 2.4 of pretest 2 led to a difficulty index of .36. Exchanging the elements in item 2.8 did not lead to a large difference in item difficulty but worsened item discrimination. For item 2.10 correct response probability increased from .18 in pretest 1 to .22 in pretest 2.

Evaluating the leverage of the cognitive operations by means of regression analysis (Table 4.4), the cognitive operations proved different influence. Only the basic parameters sequence minus and size plus were positively related to item difficulty (i.e. increased item difficulty) but all other basic parameters decreased item difficulty.

Reliability measured by Cronbach's  $\alpha$  resulted in a coefficient of  $\alpha = .49$ , and to estimate validity the figural analogy sum score was correlated with the CFT-20 R sum score ( $N = 58$ ). A highly significant correlation of  $r = .45$  ( $p < .001$ ) was yielded.

Table 4.4: Parameter Estimates of Pretest 2

Variables	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Constant	.078	.236	.332	.748
sp	-.071	.142	-.496	.632
sm	.089	.108	.825	.431
rr	.085	.284	.298	.773
rl	.557	.289	1.926	.086
r180	.421	.242	1.743	.115
rot & ref	.139	.208	.670	.519
rfx	.269	.233	1.156	.277
rfy	.387	.252	1.533	.160
sqm	-.111	.155	-.718	.491
sqp	.096	.119	.804	.442

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rot & ref = rotation and reflection, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

### Summarized Findings Pretest 2

In this second pretest item difficulty was again satisfactory. With the extended testing time the objective of improved item discrimination parameters was pursued. The average discrimination parameters consequently improved but still did not attain adequate magnitude. Further, reliability of the 20 item test increased to  $\alpha = .49$ . To estimate external test validity, the CFT-20 R was again referred to. A correlation of  $r = .45$  between the figural analogy test score and the CFT-20 R test score was obtained and proved a coherence of an expected and acceptable extent. Regression parameters were estimated but most parameters had positive algebraic signs indicating their property of only facilitating tasks.

### 4.2.3 Pretest 3

To give consideration to the findings of pretest 2 and the involved requirements and criteria for sound tasks, the third pretest had following characteristics for several reasons. In this third pretest, five subtests were generated and applied. Two of these (subtest 2 and 3) were taken from pretest 1 and partly contained items including rotation transformations as well as reflection transformations, since for these subtests no restrictions concerning the combination of transformations were made. The other three subtests were newly compiled in order to meet the demand of task with distinct cognitive requirements. Subtest 1 contained nine items with only one transformation each and six distractors, including the response option “no solution right”, were presented per item. For subtests 4 and 5 the restriction was implemented that the cognitive operations of the category reflection (x and y) would never be combined with the cognitive operations of the category rotation (90° right, 90° left and 180°). Subtests 4 and 5 included 12 items each, with two cognitive operations per item for subtest 4 and three cognitive operations per item for subtest 5. Instead of displaying six distractors a new format with ten distractors, amongst the “no solution right” - option, was introduced.

These different types of subtests were chosen in order to compare subtest characteristics before deciding on a final test version for the main examination.

## Sample

The resulting figural analogy test was applied testing a sample of 52 subjects of which 35 were year 12 pupils and 17 were first year psychology students. The mean age of the test takers ranged from 17 to 23 years with an average age of 19.08 ( $SD = 1.30$ ). The sample consisted of 24 females and 28 males.

## Results

First, descriptive analyses (Table 4.5) were conducted to statistically evaluate the items applied. For subtest 1, difficulty (percentage of correctly solved items) varied between .12 and .96 with a mean difficulty of .56 ( $SD = .28$ ). Item discrimination averaged at .12 ( $SD = .21$ ). For subtest 2 difficulty ranged from .12 to .81 with a mean difficulty of .50 ( $SD = .21$ ). A mean item discrimination of .21 ( $SD = .26$ ) was obtained. Subtest 3 contained items with difficulties between .19 and .63 ( $M = .39$ ,  $SD = .14$ ) and an average item discrimination of .07 ( $SD = .18$ ). Subtest 4 ranged from .19 to .75 in item difficulty with a mean of .47 ( $SD = .18$ ) and items had a mean discrimination of .29 ( $SD = .21$ ). In subtest 5, item difficulty varied between .06 and .67 ( $M = .37$ ,  $SD = .21$ ) and the mean item discrimination was .21 ( $SD = .20$ ).

Reliability was estimated calculating Cronbach's alpha. Across all subtests (53 items) a coefficient of  $\alpha = .73$  was obtained: For the 20 items of the subtests 2 and 3 a reliability coefficient of  $\alpha = .44$  was yielded. For the items of subtests 1, 4 and 5 (33 items)  $\alpha$  reached .71.

For criterion related validity estimates, a subsample of 29 students addi-

tionally took the CFT-20 R. Sum scores of the analogy test and the CFT-20 R were correlated and a significant correlation of .43 ( $p < .05$ ) resulted.

Two separate regression analyses were conducted. The first regression analysis (Table 4.6) refers to the subtests for which the constraint concerning the combination of the transformations rotation and reflection was implemented (subtests 1, 4, and 5). The second analysis (Table 4.7) was applied to the data resulting from subtests 2 and 3: Here no restriction regarding the combination of the cognitive operations reflection and rotation was made. Looking at the regression equations of subtests 1, 4 and 5 (Table 4.6), significant ( $p < .05$ ) estimations of the basic parameters rotation right, rotation left, reflection x, reflection y, sequence plus and sequence minus were obtained. The impact of size plus, size minus and rotation 180° was nonsignificant. Among the significant basic parameters, the operation reflection about the x-axis increased difficulty the most, followed by rotation left, reflection about the y-axis, rotation right, sequence plus and sequence minus.

### **Summarized Findings Pretest 3**

Regarding the results of pretest 3 following findings can be summarized. Looking at item difficulty, again no fault could be found with the parameters, both across all test items and within single subtests. Average item discrimination parameters were still too low, but tended to be higher for the subtests that included items in which the transformations rotation and reflection were not combined.

Reliability for all items was  $\alpha = .73$  and  $\alpha = .71$  for the 33 items of the



Table 4.5: Descriptive Statistics of Pretest 3

Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>	Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>
1.1	.37	.49	.06	4.1	.52	.51	.40
1.2	.38	.49	-.34	4.2	.75	.44	.30
1.3	.96	.19	.34	4.3	.19	.40	.19
1.4	.90	.30	.22	4.4	.56	.50	.41
1.5	.12	.32	.10	4.5	.25	.44	-.10
1.6	.54	.50	.11	4.6	.56	.50	.45
1.7	.77	.43	.18	4.7	.50	.51	.43
1.8	.35	.48	.38	4.8	.42	.50	-.11
1.9	.65	.48	.05	4.9	.27	.45	.28
				4.10	.31	.47	.31
				4.11	.67	.47	.40
				4.12	.63	.49	.57
2.1	.35	.48	.22	5.1	.46	.50	.49
2.2	.38	.49	.47	5.2	.06	.24	-.17
2.3	.71	.46	.34	5.3	.15	.36	.07
2.4	.71	.46	-.01	5.4	.62	.49	.09
2.5	.52	.51	.48	5.5	.50	.51	.38
2.6	.81	.40	.29	5.6	.65	.48	.56
2.7	.52	.51	.39	5.7	.67	.47	.23
2.8	.12	.32	-.23	5.8	.31	.47	.26
2.9	.52	.51	.29	5.9	.25	.44	.25
2.10	.37	.49	-.17	5.10	.25	.44	.08
				5.11	.37	.49	.15
				5.12	.15	.36	.15
3.1	.31	.47	.06				
3.2	.29	.46	.14				
3.3	.63	.49	-.09				
3.4	.19	.40	.43				
3.5	.60	.50	.29				
3.6	.46	.50	.08				
3.7	.33	.47	.06				
3.8	.42	.50	-.20				
3.9	.37	.49	-.05				
3.10	.27	.45	.01				

Table 4.6: Parameter Estimates of Pretest 3 (Subtests 1, 4, 5)

Variables	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Constant	.748	.083	9.018	.000
sp	.036	.069	.519	.609
sm	.082	.069	1.184	.249
rr	-.236	.103	-2.286	.032
rl	-.284	.090	-3.153	.004
r180	-.027	.093	-.286	.777
rfx	-.435	.097	-4.505	.000
rly	-.279	.098	-2.858	.009
sqp	-.214	.069	-3.104	.005
sqm	-.158	.070	-2.246	.035

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rly = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

Table 4.7: Parameter Estimates of Pretest 3 (Subtests 2, 3)

Variables	<i>B</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Constant	.608	.235	2.587	.029
sp	-.167	.142	-1.178	.269
sm	.010	.108	.088	.932
rr	-.307	.284	-1.083	.307
rl	.192	.288	.666	.522
r180	.085	.241	.354	.731
rot & ref	-.243	-.653	-1.175	.270
rfx	-.213	.232	-.915	.384
rly	-.059	.252	-.233	.821
sqm	-.146	.154	-.945	.369
sqp	.079	.119	.663	.524

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180=rotation 180°, rot & ref = rotation and reflection, rfx = reflection about x-axis, rly = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

subtests 1, 4 & 5 and .43 for the 20 items of the subtests 2 & 3. Validity was again estimated by the correlations with the CFT-20 R test score and a mediocre coherence was obtained.

To evaluate the results from the regression analyses one should distinguish between the subtests. For subtests 1, 4 & 5 six out of nine transformations yielded significant basic parameter estimates. Apart from the size minus operation, all other transformations were positively related to item difficulty, i.e. impeded items and thus reduced the probability to solve an item. As in the preceding pretests, the transformations of subtests with items involving rotation as well as reflection transformations (subtests 2 & 3) did not have a significant impact on item difficulty.

#### **4.2.4 Summary and Conclusion on Pretests**

Pretests were designed and conducted to provide an informative basis for the rule-based construction of the figural analogy test applied in the main examination. Since the cognitive operations chosen and their application to the elements involved in the test had not been empirically analyzed before, exploratory research concerning several issues had to be accomplished. For each pretest, data was analyzed according to classical test theory as the number of subjects tested in the pretests was too little to apply test procedures according to item response theory. For each test descriptive statistics, i.e. item difficulty and item discrimination, were inspected for each item. Compared to pretest 1, time permitted to process the items was varied in Pretest 2. This was done in order to examine if item discrimination parameters would improve with extended testing time. Difficulty, defined

as the percentage of subjects that correctly solved the item, was calculated to evaluate if difficulty was generally of an adequate level and variability. Regression analyses provided information on the magnitude and direction of the leverage that the transformations had on item difficulty. A first impression concerning the construction approach was thus obtained. The findings of the three pretests provided a basis and implied consequences for the rule-based test construction of the figural analogy test material for the main examination.

The most important implications from the pretests concerned the rational and rule-based item construction: Of all pretests, parameter estimates of the three subtests (1, 4 & 5) of pretest 3 showed that, apart from the size operations, the application of other rules contributed to item difficulty, i.e. increased item difficulty. Due to their property of incrementing item difficulty these rules should be adopted in the final version of the figural analogy test. Although the application of the size rules led to an increase in the probability of solving an item they were still chosen as task parameters to retain the variation of item difficulty and to further analyze their property as contributors to item easiness. Therefore, the rationale for the construction of the analogy test for the main examination included the constraint regarding the combination of the rules rotation and reflection that should be implemented in the design for the test of the main examination.

# 5

## **Main Examination**

The following chapter presents the main examination of this study. The first section refers to the test material. Then instruction, testing procedure, research design and the different measures collected are pointed out. In the third section, the sample is presented and sample comparisons due to the research design are made. The last and most extensive section of this chapter presents the results of this study. The presentation of the results is arranged according to different test theories and according to the objectives of the respective analysis.

## **5.1 Material**

The several pretests conducted proved that subtests with no restriction regarding the combination of the rules rotation and reflection always yielded nonsignificant basic parameter estimates. Further, to lead item difficulties back to its components, the cognitive operations involved in solving the item have to be clearly defined. Thus only the application of the predefined and distinct operations should induce the correct response to an item. The cognitive operations of each item therefore have to be well-defined in order to apply statistical evaluation methods such as the linear logistic test model. Subtests 1, 4 and 5 of the third pretest, for example, only contained items in which the transformations of the category rotation and reflection were never applied at the same time. These subtests also proved that almost all parameter estimates had significant impact on item difficulty in the regression analytic evaluation. Thus information on the contribution of each cognitive operation to item difficulty could be provided. Deeper insight into task structure and basic assumptions on contributors to item difficulty were obtained by the different pretests and pioneered the test construction for the main examination.

### **5.1.1 Test and Items**

Having conducted the pretests, the final version of the figural analogy tasks consisted of four subtests with a total of 45 items. The division of the test into different subtests was chosen for time management reasons and to arrange groups of items with the same number of rules per item.

The cognitive operations already introduced in the pretests are again presented in Table 5.1. The category size with its facets size plus and size minus was implemented as well as the category rotation of elements that referred to 180° rotation, 90° left rotation, and 90° right rotation. The transformation group reflection included reflection of the elements on its x-axis or y-axis and the transformation sequence differed in whether the subject had to increase or decrease the number or alphabet sequence.

Table 5.1: Cognitive Operations

Operation	level 1	level 2	level 3
Size	size plus (sp)	size minus (sm)	
Rotation	180° (r180)	90° right (rr)	90° left (rl)
Reflection	x-axis (rfx)	y-axis (rfy)	
Sequence	sequence minus (sqm)	sequence plus (sqp)	

*Note.* Level 1, 2 and 3 represent the facets of the operations.

The construction of each subtest was based on an efficient design provided by the software SAS (version 9.1) with the implemented restriction that the cognitive operations of the category reflection (x and y) would never be combined with the cognitive operations of the category rotation (90° right, 90° left, and 180°). Considering the constraints, designs with a maximum of efficiency were thus obtained. Four subtests were constructed that differed in the number of cognitive operations involved per item. Each subtest was constructed according to its efficient design that was generated in order to provide balanced and orthogonal fractional factorial designs. The software SAS was also chosen to implement the restriction concerning the combination of operations. Subtest 1 contained nine items with one rule each. Subtest 2 had 12 items with two rules each and subtests 3 and 4 had

12 items each with three rules applied per item. Subtest 4 functioned as a parallel version of subtest 3, thus these two subtests had identical design matrices but different elements.

Tables 5.2 and 5.3 present the efficient design matrices generated by SAS for the subtests. In the first column the item is indicated. Subsequent columns then indicate which transformation rule was applied in the item and the last column of each table indicates the accumulated number of cognitive rules per item.

### 5.1.2 Distractors

The design of the distractors generally requires high precision. In terms of the figural analogy test it should be assured that subjects solve the items by applying the cognitive operations required to solve the particular item. Salient features among the distractors or features that occur often across distractors should be minimized to prevent subjects from developing their own solving strategies. Thus, constructing the distractors, it was aimed to equally distribute the occurrence of element features among the answer options.

In subtest 1 six answer options (a - f) per item were displayed. Distractors a - e presented elements (letters and digits) and distractor f always presented the answering option "no correct solution". Subtests 2 - 4 had ten distractors per item and of these, distractors a-i again presented stimuli related to the item and distractor j offered the choice "no correct solution". Subtest 1 had less distractors than the other subtests because only one transformation rule per item was applied. Since all other subtests had



Table 5.2: Optimal Design Matrix (Subtests 1 & 2)

Item	sp	sm	rr	rl	r180	rfx	rfy	sqp	sqm	$\Sigma$ rules
1.1	0	0	0	0	0	1	0	0	0	1
1.2	0	0	0	1	0	0	0	0	0	1
1.3	1	0	0	0	0	0	0	0	0	1
1.4	0	1	0	0	0	0	0	0	0	1
1.5	0	0	1	0	0	0	0	0	0	1
1.6	0	0	0	0	0	0	0	1	0	1
1.7	0	0	0	0	1	0	0	0	0	1
1.8	0	0	0	0	0	0	0	0	1	1
1.9	0	0	0	0	0	0	1	0	0	1
2.1	0	0	0	0	1	0	0	1	0	2
2.2	0	1	0	0	1	0	0	0	0	2
2.3	0	0	0	0	0	1	0	1	0	2
2.4	0	0	1	0	0	0	0	0	1	2
2.5	0	0	0	0	0	0	1	1	0	2
2.6	1	0	0	0	0	0	0	0	1	2
2.7	1	0	0	1	0	0	0	0	0	2
2.8	0	0	0	0	0	0	1	0	1	2
2.9	0	0	0	1	0	0	0	0	1	2
2.10	0	1	0	0	0	1	0	0	0	2
2.11	0	1	0	0	0	0	0	1	0	2
2.12	1	0	1	0	0	0	0	0	0	2

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqp = sequence plus, sqm = sequence minus,  $\Sigma$ rules = number of rules applied in the item

Table 5.3: Optimal Design Matrix (Subtests 3 &amp; 4)

Item	sp	sm	rr	rl	r180	rfx	rfy	sqp	sqm	$\Sigma$ rules
3.1	0	1	0	1	0	0	0	0	1	3
3.2	0	1	0	0	0	1	0	0	1	3
3.3	1	0	0	0	0	0	1	0	1	3
3.4	1	0	0	0	1	0	0	0	1	3
3.5	0	1	0	0	1	0	0	1	0	3
3.6	0	1	0	0	1	0	0	1	0	3
3.7	0	1	1	0	0	0	0	0	1	3
3.8	1	0	0	1	0	0	0	1	0	3
3.9	1	0	0	0	0	1	0	1	0	3
3.10	0	1	0	0	0	0	1	1	0	3
3.11	1	0	0	1	0	0	0	0	1	3
3.12	1	0	1	0	0	0	0	1	0	3
4.1	0	1	0	1	0	0	0	0	1	3
4.2	0	1	0	0	0	1	0	0	1	3
4.3	1	0	0	0	0	0	1	0	1	3
4.4	1	0	0	0	1	0	0	0	1	3
4.5	0	1	0	0	1	0	0	1	0	3
4.6	0	1	0	0	1	0	0	1	0	3
4.7	0	1	1	0	0	0	0	0	1	3
4.8	1	0	0	1	0	0	0	1	0	3
4.9	1	0	0	0	0	1	0	1	0	3
4.10	0	1	0	0	0	0	1	1	0	3
4.11	1	0	0	1	0	0	0	0	1	3
4.12	1	0	1	0	0	0	0	1	0	3

*Note.* sp = size plus, sm = size minus, rr=rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqp = sequence plus, sqm = sequence minus,  $\Sigma$ rules = number of rules applied in the item

two or three transformation rules per item, more distractors had to be displayed in order to account for the different possible combinations evolving from the increase in transformations.

## **5.2 Instruction, Procedure, and Measures**

Two test versions (A and B) were designed to administer the figural analogy test. Items were the same in both tests but the instruction of the two versions differed. Both test versions started with a general instruction to the test. Test version A had this general instruction only and one sample item, and in test version B, besides the general instruction, participants were introduced to each of the nine transformation rules prior to the test, illustrating the rules with two examples each (cf. Appendix B). This design was chosen to test the hypothesis that participants having received the rule explanations (test version B) performed better and achieved a significant higher overall test result compared to those subjects that only received the general instruction (test version A). Subjects were randomly assigned to two groups: the non-instruction group (receiving test version A) and the instruction group (receiving test version B). All participants were instructed to follow the examiner's start and stop signals. Items of each subtest were always displayed on two pages and two minutes and 30 seconds processing time were allowed for subtest 1, five minutes and 30 seconds for subtest 2, and seven minutes each for subtests 3 and 4.

Before taking the figural analogy test, all subjects were asked to complete a questionnaire of demographical items such as age and sex and to

self-report recent grades for a variety of school subjects. After the figural analogy test was conducted, most subjects also took the revised German version of Cattell's Culture Fair Test of Intelligence (CFT-20 R; Weiß, 2006). In this test, subjects had to work on 56 items measuring general intelligence.

Additionally, most subjects also completed the German version of the NEO-FFI (Costa & McCrae, 1992; Borkenau & Ostendorf, 1993). The questionnaire was applied to assess the five central dimensions of personality (neuroticism, extraversion, openness, agreeableness, and conscientiousness). Further, a general interest test (AIST; Bergmann & Eder, 1999) measuring scholastic and professional interests was applied in most testing sessions. The AIST consists of 60 items and interests are measured on six scales: realistic, investigative, artistic, social, enterprising, and conventional interests. However, the NEO-FFI and AIST were applied for motivational purposes and results are not presented as these tests were not administered for research purposes but for feedback reasons only.

Subjects were tested in groups and classrooms served as testing area. General instructions were given at the beginning of the testing session and task-specific instructions preceded each test.

### **5.3 Sample**

The sample taking the Figural Analogy Test included 212 male and 272 female students and thus a total of 484 subjects. All participants were Germans and recruited from various German high-schools. Age ranged

from 15 to 20 years with a mean of 17.8 years ( $SD = .84$ ).

Students were randomly assigned to the experimental group: 249 subjects were members of the instruction group and 235 subjects were members of the non-instruction group. Of the 484 subjects, 312 additionally took the CFT-20 R and 480 self-reported their school grades. Besides serving as criteria for validity estimates of the test, the data of the CFT-20 R and the school grades were applied for sample comparisons concerning the instruction versus non-instruction group. Results of the sample comparison are subsequently reported before results of the figural analogy test are presented.

#### **Sample Comparison: CFT-20 R**

Of all 312 subjects that took the CFT-20 R, 128 belonged to the instruction group and 184 belonged to the non-instruction group. The instruction group obtained mean scores of 43.25 ( $SD = 5.85$ ) and the non-instruction group yielded mean scores of 43.55 ( $SD = 4.79$ ). Inferential statistical analysis revealed that the two groups did not significantly differ in general intelligence measured by this test ( $t = -.504$ ,  $df = 310$ ,  $p = .615$ ).

#### **Sample Comparison: Scholastic Achievement**

Subjects were asked to fill out a questionnaire on their recent grades. The common German point format was used to assess academic achievement (15 points = best grade versus 0 points = worst grade). Looking at the average self-reported grades, groups did not significantly differ ( $t = -.836$ ,  $df = 478$ ,  $p = .404$ ). For the instruction group a mean of 8.75 points ( $SD = 1.84$ )

and for the non-instruction group a mean of 8.89 ( $SD = 1.61$ ) was calculated from the data. To evaluate if the groups differed in different subjects, an analysis looking beyond the average points obtained was conducted. Composite scores of different subjects were computed to provide a well arranged presentation of results. A composite score (social) referring to the subjects history, geography, politics, religious studies, education science was generated as well as a composed language score (language) consisting of the subjects German, Latin, English, and French. A combined score for art & music was computed and the subjects mathematics, physics, chemistry and biology were integrated to a composite score (maths & sciences). The instruction group did not significantly differ from the non-instruction group in any of these composite scores.

## 5.4 Results

In the next sections the results of the data, analyzed according to different test theories and procedures, are presented. With regard to the figural analogy test, item 24 (3.3) was excluded in all following analyses due to an construction error. The number of relevant test items was therefore decreased to a total of 44 items.

### 5.4.1 Item Statistics According to Classical Test Theory

To analyze the data, scoring at item-level was applied: subjects always obtained one point per correctly solved item. The average sum score of the figural analogy test was 20.67 ( $SD = 6.12$ ) and a minimum score of 3 and a

maximum score of 37 were obtained. The distribution of the sum scores of the figural analogy test seemed to be normally distributed as illustrated in Figure 5.1 showing the histogram of the sum scores, and Figure 5.2 showing the Q-Q diagram plotting the observed values against the expected normal values. The statistical analysis by means of the Kolmogorov-Smirnov test ( $Z = 1.076, p > .05$ ) confirmed the assumption that the distribution did not significantly differ from the normal distribution ( $p = .197, df = 484$ ).

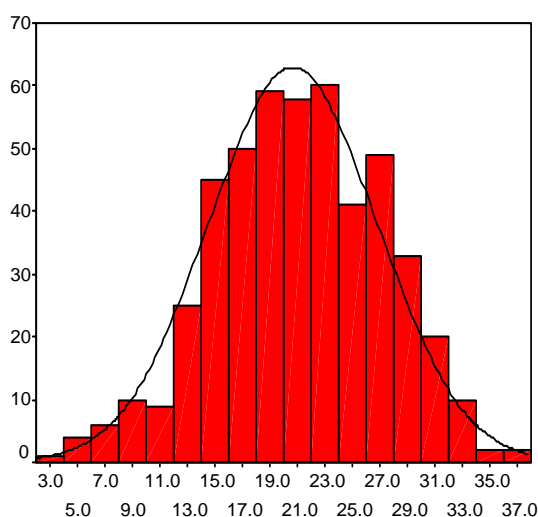


Figure 5.1: Distribution of the Sum Scores of the Figural Analogy Test

The instruction group, having received an explanation of the operations involved, significantly differed from the non-instruction group in the mean of correctly solved items (Table 5.4). Subjects that had been introduced to the transformation rules (instruction group) averagely solved 21.74 items and subjects that had not been introduced to the transformation rules (non-instruction group) averagely solved 19.54 items out 44. This difference was significant ( $t = 4.024, df = 482, p < .001$ ) and an effect size of  $d = .36$  was yielded. An introduction to the cognitive operations applied in the test

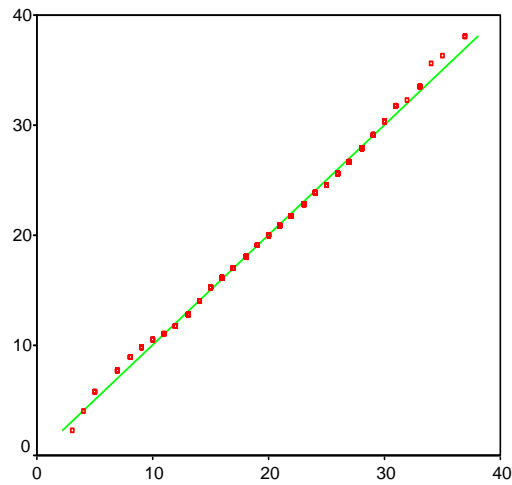


Figure 5.2: Q-Q Diagram: Testing the Sum Scores of the Figural Analogy Test for Normal Distribution

therefore had impact on the average sum score, and participants who were introduced to the transformation rules before taking the test, significantly obtained a higher overall test score on the figural analogy test.

Comparing the sum scores of the figural analogies between male and female no significant difference could be observed ( $t = -.319$ ,  $df = 482$ ,  $p = .750$ ). Taking the group membership in terms of transformation explication received vs. not received into account, again no significant difference between the sum scores of males and females was yielded, neither for the instruction group ( $t = -1.032$ ,  $df = 247$ ,  $p = .303$ ) nor for the non-instruction group ( $t = .504$ ,  $df = 233$ ,  $p = .615$ ).

Table 5.5 shows item statistics according to classical test theory for the total sample: Item difficulty, standard deviation, item discrimination and internal consistency (Cronbach's  $\alpha$  excluding single items) are presented. Item difficulty for each item was defined as the percentage of subjects that



Table 5.4: T-Test Means of the Instruction Group and the Non-Instruction Group

Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>df</i>	<i>t</i>	<i>p</i>
Instruction Group	249	21.74	6.19	482	4.024	.000
Non-Instruction Group	235	19.54	5.85			

correctly solved that item. Item difficulties ranged from .11 (item 3.2) to .93 (item 1.3) with two items of subtest 1 (with one transformation rule each) yielding difficulty coefficients above .90. These two items are items 1.3 and 1.4, applying the transformation rules size plus and size minus respectively. Item 3.2 was the most difficult one with only 11% of participants solving that item, and the only item with  $p < .20$ . This item contained size minus and reflection on the x-axis as cognitive operations. To which degree these transformation rules or the property of the elements or the distractors chosen had impact on the item difficulty is analyzed when the item construction and cognitive structure is focused on. Looking at the total of the 44 items applied, a mean item difficulty of .47 ( $SD = .14$ ) was obtained.

Items yielded a mean item discrimination of .22 ( $SD = .08$ ) with a maximum of .40. Item 3.2 (the most difficult item of the test) had a negative item discrimination of -.08 indicating that more subjects of the lower ability group answered that item correctly than of the higher ability group. Eliminating item 3.2 from calculation an average item discrimination of .23 ( $SD = .09$ ) was obtained. Overall item discriminations were rather low and might be attributed to a shortage of testing time per subtest.

Looking at the average difficulties of the subtests, subtest 1 with one transformation rule per item yielded an average difficulty of .63 ( $SD = .43$ )

5 Main Examination

and was therefore easier than subtests 2 ( $M = .44, SD = .47$ ), 3 ( $M = .40, SD = .46$ ) and 4 ( $M = .45, SD = .48$ ). Thus, as expected, the first subtest was the easiest one, since cognitive requirements were restricted to one transformation rule per item. On average, subtest 2, with two transformation rules per item, obtained almost the same item difficulty as subtest 4, but was, as expected, easier than subtest 3 with three transformation rules per item.

Table 5.5: Descriptive Statistics: Total Sample

Item	$M$	$SD$	$r_{it}$	$\alpha$	Item	$M$	$SD$	$r_{it}$	$\alpha$
1.1	.36	.48	.24	.76	3.1	.48	.50	.24	.76
1.2	.45	.50	.29	.76	3.2	.11	.31	-.08	.77
1.3	.93	.26	.23	.76	3.4	.67	.47	.31	.75
1.4	.92	.27	.19	.76	3.5	.43	.50	.20	.76
1.5	.48	.50	.18	.76	3.6	.57	.50	.31	.75
1.6	.71	.45	.30	.76	3.7	.52	.50	.31	.75
1.7	.71	.46	.12	.76	3.8	.24	.43	.25	.76
1.8	.44	.50	.27	.76	3.9	.37	.48	.20	.76
1.9	.64	.48	.19	.76	3.10	.27	.44	.19	.76
					3.11	.47	.50	.40	.75
					3.12	.25	.43	.10	.76
2.1	.45	.50	.26	.76	4.1	.63	.48	.24	.76
2.2	.63	.48	.21	.76	4.2	.44	.50	.26	.76
2.3	.21	.41	.14	.76	4.3	.33	.47	.21	.76
2.4	.45	.50	.28	.76	4.4	.64	.48	.30	.76
2.5	.21	.41	.01	.77	4.5	.38	.49	.25	.76
2.6	.61	.49	.18	.76	4.6	.63	.48	.34	.75
2.7	.58	.49	.34	.75	4.7	.51	.50	.34	.75
2.8	.39	.49	.13	.76	4.8	.54	.50	.23	.76
2.9	.29	.45	.21	.76	4.9	.20	.40	.04	.76
2.10	.32	.47	.23	.76	4.10	.27	.45	.10	.76
2.11	.60	.49	.22	.76	4.11	.41	.49	.33	.75
2.12	.49	.50	.39	.75	4.12	.44	.50	.23	.76

Note.  $\alpha = \alpha$  without item

Statistics according to classical test theory are also separately presented for the two groups (instruction vs. non-instruction group) in Table 5.6 and

Table 5.7. The items applied in the instruction group had a mean item discrimination of .23 ( $Min = -.14$ ,  $Max = .45$ ) and were thus slightly higher in average discrimination than the items applied in the non-instruction group ( $M = .21$ ,  $Min = -.09$ ,  $Max = .42$ ). Looking at the item difficulties for both groups, items were on average easier ( $M = .49$ ,  $Min = .11$ ,  $Max = .94$ ) in the instruction group than in the non-instruction group ( $M = .44$ ,  $Min = .11$ ,  $Max = .94$ ).

Table 5.6: Descriptive Statistics: Instruction Group

Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>	Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>
1.1	.39	.49	.26	3.1	.49	.50	.28
1.2	.53	.50	.27	3.2	.11	.32	-.08
1.3	.92	.27	.30	3.4	.70	.46	.30
1.4	.94	.25	.16	3.5	.45	.50	.20
1.5	.52	.50	.18	3.6	.64	.48	.33
1.6	.86	.34	.31	3.7	.51	.50	.35
1.7	.69	.46	.14	3.8	.25	.44	.27
1.8	.52	.50	.28	3.9	.39	.49	.19
1.9	.68	.47	.22	3.10	.29	.45	.06
				3.11	.51	.50	.37
				3.12	.23	.42	.14
2.1	.44	.50	.25	4.1	.62	.49	.31
2.2	.64	.48	.27	4.2	.46	.50	.29
2.3	.22	.41	.09	4.3	.34	.48	.29
2.4	.49	.50	.26	4.4	.66	.48	.38
2.5	.20	.40	-.14	4.5	.43	.50	.27
2.6	.63	.48	.20	4.6	.66	.47	.35
2.7	.61	.49	.40	4.7	.56	.50	.40
2.8	.43	.50	.10	4.8	.55	.50	.22
2.9	.31	.47	.15	4.9	.22	.42	-.01
2.10	.37	.48	.23	4.10	.26	.44	.10
2.11	.59	.49	.21	4.11	.41	.49	.39
2.12	.53	.50	.45	4.12	.47	.50	.16

Table 5.7: Descriptive Statistics: Non-Instruction Group

Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>	Item	<i>M</i>	<i>SD</i>	<i>r<sub>it</sub></i>
1.1	.32	.47	.21	3.1	.47	.50	.12
1.2	.36	.48	.28	3.2	.11	.31	-.09
1.3	.94	.24	.18	3.4	.64	.48	.31
1.4	.90	.30	.20	3.5	.41	.49	.20
1.5	.44	.50	.15	3.6	.49	.50	.25
1.6	.54	.50	.24	3.7	.54	.50	.30
1.7	.72	.45	.11	3.8	.23	.42	.23
1.8	.36	.48	.21	3.9	.35	.48	.19
1.9	.59	.49	.14	3.10	.24	.43	.17
				3.11	.42	.49	.42
				3.12	.27	.44	.06
2.1	.46	.50	.29	4.1	.65	.48	.17
2.2	.62	.49	.15	4.2	.42	.49	.22
2.3	.21	.41	.20	4.3	.32	.47	.13
2.4	.40	.49	.27	4.4	.63	.49	.20
2.5	.23	.42	.18	4.5	.34	.48	.21
2.6	.59	.49	.15	4.6	.59	.49	.32
2.7	.55	.50	.26	4.7	.45	.50	.24
2.8	.34	.48	.15	4.8	.53	.50	.24
2.9	.26	.44	.26	4.9	.18	.38	.08
2.10	.27	.45	.20	4.10	.29	.45	.13
2.11	.60	.49	.24	4.11	.41	.49	.28
2.12	.45	.50	.31	4.12	.41	.49	.30

### **Reliability**

Internal consistency according to Cronbach's  $\alpha$  for the total test of 44 items for the total sample constituted to  $\alpha = .76$ . Thus Anastasi's reliability criterion of .80 or .90 (Anastasi, 1982) was almost obtained for this figural analogy test.

### **Validity**

Criterion-related validity was estimated using the CFT-20 R scores. Of all participants 312 subjects were additionally tested using the CFT-20 R. Among these, 188 were female and 123 were male and the average age was 17.91 years ( $SD = .70$ ). Minimum scores of 24 and maximum scores of 54 were obtained, resulting in a mean score of .43 ( $SD = 5.24$ ). A correlation of .36 ( $p = .000$ ) between the sum scores of the figural analogy test and the scores of the CFT-20 R was obtained. However, correlating the figural analogy scores of the instruction-group ( $N=128$ ) with the CFT scores resulted in a coefficient of .54 ( $p = .000$ ). When transformations were not explained in the non-instruction group ( $N = 184$ ) a lower correlation coefficient of .24 ( $p = .001$ ) between the analogy test score and the CFT-20 R was yielded.

Referring to academic achievement as external validation criterion, following observations were made: The figural analogy score correlated significantly ( $p < .01$ ) with the average grade score ( $r = .18$ ). The composite score maths & sciences correlated at  $r = .22$  ( $p < .01$ ) with the analogy score and the art & music score correlated at  $r = .16$  ( $p < .01$ ) with the analogy test score. The composed language score only correlated at  $r = .10$  ( $p < .05$ ) and the score for the social sciences did not correlate significantly with the

analogy score.

For reasons of comparison, the correlation of the CFT-20 R scores with the composed scholastic measures are as follows: The CFT-score significantly correlated at  $r = .19$  with the average scholastic score,  $r = .25$  with the maths & science score and at  $r = .16$  with the art & music score. The composite scores for the social sciences and language did not prove significant correlations with the CFT-20 R score.

### 5.4.2 Confirmatory Factor Analysis

Confirmatory factor analysis was applied as statistical means to test the hypothesis of one underlying latent variable regarding the structure of the data. The software Mplus (version 5.1) was used to estimate the confirmatory model fit. Two confirmatory factor analyses were conducted, the first for the data of all 44 test items, and the second for the data of the 32 Rasch conform items (Rasch conform according to the  $\chi^2$ -statistics; cf. Table 5.8). For the 44 items, the chi-square test of model fit was significant ( $\chi^2 = 516.13$ ,  $df = 273$ ,  $p = .000$ ). Further the comparative fit index (CFI) and Tucker-Lewis index (TLI) were obtained. These indices for model fit, however, did not indicate very good model fit (CFI = .75, TLI = .77). Yet, the RMSEA (root mean square error of approximation) indicated good model fit (RMSEA = .043). For the 32 Rasch items following results were obtained: the chi-square test of model fit was significant ( $\chi^2 = 387.24$ ,  $df = 212$ ,  $p = .000$ ) and the comparative fit index and Tucker-Lewis index again could not confirm very good model fit (CFI = .69, TLI = .71). The RMSEA, however again proved good model fit for the Rasch items (RMSEA = .041). Thus,

in both analyses, the assumption of unidimensionality was only weakly supported.

### 5.4.3 Estimating Parameters According to IRT-Models

BILOG-MG software (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to estimate item parameters according to the item response theory. Expected a posteriori (EAP) estimates of ability were calculated using the software.

For the total sample Table 5.8 presents item statistics for the 1PL model and Table 5.9 for the 2PL model. Besides item difficulty, defined by the latent underlying dimension where the probability to solve the item is 50%, the standard error (*SE*) of the item difficulty is listed. Since the 1PL model assumes equal item discrimination for all items, the discrimination parameters are only given for the 2PL model in Table 5.9. For the 1PL model sigma varied between -4.65 (item 1.3) indicating a very easy item and 3.78 (3.2) indicating a very difficult item. These figures correspond to the difficulty indices of the CTT listed above where item 1.3 represents the easiest item with  $p = .93$  and item 3.2 is the most difficult one with  $p = .11$ . Mean IRT-item difficulty was .21 ( $SD = 1.61$ ). For the 2PL model mean item difficulty was .49 ( $SD = 1.89$ ) and item discrimination parameters dispersed from .17 to .69 with an average discrimination of .40 ( $SD = .13$ ).

In order to determine if single items showed significant deviation from the 1PL or the 2PL model, statistical tests of goodness of fit had to be calculated. The  $\chi^2$ -fit statistics are probably the most widely used in applied

research. Unfortunately, they are often viewed as inconclusive evidence of adequate fit, because of their sensitivity to sample size and their insensitivity to certain forms of misfit. Therefore, additional tests for significance can be applied, given that the results of the  $\chi^2$ -test can only be interpreted with certain restrictions. According to Rost (2004), the  $Q$ -index, here computed by the program Winmira (von Davier, 2001), is a measure of task fit for the 1PL model based on the log-likelihood of observed item-pattern. Tests for significance of  $Q$  therefore allow interpreting underfit ( $p < .05$ ) and overfit ( $p > .95$ ) of tasks.

According to the  $Z$  statistic of the  $Q$ -index, 31 items of the test showed model fit with the 1PL model. Only some items did not prove sufficient model fit with the 1PL model (Table 5.8): Items 2.12, 3.11, 4.6, 4.7, 4.11 showed overfit ( $p > .95$ ) and items 1.7, 2.5, 2.8, 3.2, 3.10, 3.12, 4.9, 4.10 showed underfit ( $p < .05$ ). Referring to the  $\chi^2$ -fit indices, 32 test items fitted the 1PL model. The following items showed deviations from the 1PL model according to the  $\chi^2$ -fit indices ( $p < .05$ ): 1.3, 2.5, 2.7, 2.12, 3.2, 3.6, 3.11, 4.4, 4.6, 4.7, 4.9, 4.11. Thus both model fit indices overlapped in rejecting 8 items due to insufficient model fit with the 1PL model. For the 2PL model the  $\chi^2$ -fit statistic indicated that only items 3.2 and 4.9 significantly deviated ( $p < .05$ ) from the model (Table 5.9).

Besides identifying the items deviating from the 1PL model by means of the fit indices, analyses in terms of deviation from the 1PL model should be extended with regard to item properties. Distinctive features such as item difficulty, item discrimination and the item characteristic curves (Figure A.1 of Appendix A) might explain the deviation of certain items from



the 1PL model. Item 1.3 was very easy: 93 % of the subjects correctly solved that item. The item was thus considerably easier than the average item difficulty of subtest 1. Looking at the item characteristic curve (ICC) for that item (Figure 5.3), the probability of solving the item at medium ability levels was a lot higher than to be expected according to the item function of the 1PL model. This explains the misfit with the 1PL model according to the  $\chi^2$ -index. Misfit according to the  $Q$ -index was also indicated for item 1.7 which was rather easy but not too easy compared to the other items of subtest 1. The ICC however shows that the item provided higher discrimination at lower ability levels and was more informative for these ability levels. Almost no item discrimination power was provided by item 2.5. This item was quite difficult however, and looking at the ICC, a very high ability level was required for a .50 probability to solve the item. The item thus did not correspond to the function of the 1PL model. Item 2.7 and item 2.8 did not correspond to the 1PL model, according to the  $\chi^2$ -index respectively the  $Q$ -index. However, neither the item parameters of the items nor the ICCs seemed to provide explanations for the misfit of the items to the 1PL model. Information provided by the item 2.7 was slightly higher for the lower medium ability level whereas information provided by item 2.8 was slightly higher for the upper medium ability level. Regarding the item difficulty of item 2.12 no distinct feature in terms of difficulty was yielded since about half of all subjects correctly solved the item. The item however, provided rather high discrimination power compared to the other test items and might thus not correspond to the 1PL model that assumes equal item discrimination parameters for all test items. The misfit of that

item according to the  $\chi^2$ -index and the Q-index might thus be warranted. Item 3.2 proved to be a very difficulty item. Only 11% of the subjects correctly solved the item. Looking at the ICC, a medium solving probability could not be obtained at medium ability levels as expected from the 1PL model function. Further, the negative item discrimination parameter explained the misfit of the item to the 1PL model as indicated by the Q-index as well as the  $\chi^2$ -index. With regard to the misfit of item 3.6 to the 1PL model according to the  $\chi^2$ -index, no evident deviation from the 1PL function could be obtained looking at the ICC of that item. The same applied to items 3.10 and 3.11 for which misfit was indicated by the Q-index, respectively by both indices. Item 3.12 was rather difficult and only provided a low discrimination power. Looking at the ICC, only higher ability subjects were able to achieve a medium probability of solving the item. Thus the item function of that item did not correspond to the function of the 1PL model. With 64% of the subjects solving that item, item 4.4 was quite easy compared to the items of its corresponding subtest. However, looking at the ICC, no distinct deviations could be found. Just like item 4.4, item 4.6 appeared to be rather easy with a comparatively high item discrimination power. This might explain the misfit to the 1PL model which assumes equal discrimination parameters for all test items. Like the previous item, item 4.7 might not fit to the 1PL model due to its higher discrimination power compared to the other test items. The ICC of item 4.9 (Figure 5.4) did not comply with the characteristics of the function according to the 1PL model. With only 20% of all subjects solving the item, the item proved to be very difficult. Subjects of medium ability had a clearly minor probability

to solve that item than expected by the 1PL model. Almost the same issue related to item 4.10, in alleviated manner though, since difficulty was not as high and the item discrimination power not as low. Finally the misfit of item 4.11 to the 1PL model was indicated according to the  $Q$ -index and the  $\chi^2$ -index. Looking at the ICC for this item no distinct deviations from the 1PL model function, however, could be found.

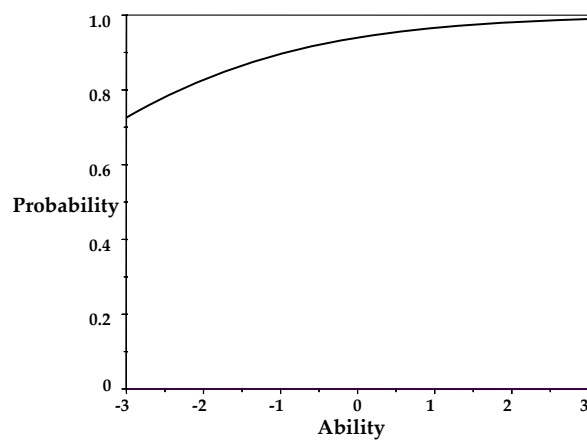


Figure 5.3: ICC Item 1.3

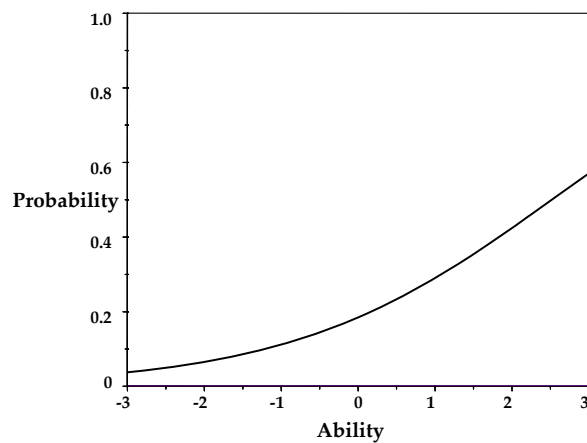


Figure 5.4: ICC Item 4.9

Table 5.8: Item Parameters 1PL Model (BILOG)

Item	$\sigma$	$SE_{\sigma}$	Fit $\chi^2$	Fit $Z_Q$	Item	$\sigma$	$SE_{\sigma}$	Fit $\chi^2$	Fit $Z_Q$
1.1	1.10	.17	.19	.61	3.1	.16	.16	.69	.46
1.2	.39	.17	.35	.85	3.2	3.78	.24	.00	.00
1.3	-4.65	.31	.02	.82	3.4	-1.31	.18	.24	.90
1.4	-4.39	.29	.42	.66	3.5	.52	.16	.67	.28
1.5	.14	.16	.87	.09	3.6	-.50	.17	.02	.88
1.6	-1.63	.18	.35	.86	3.7	-.16	.17	.16	.89
1.7	-1.61	.17	.06	.03	3.8	2.11	.19	.26	.86
1.8	.43	.16	.38	.69	3.9	.98	.17	.61	.24
1.9	-1.03	.17	.92	.20	3.10	1.85	.18	.91	.03
					3.11	.25	.17	.00	.99
					3.12	2.03	.18	.15	.01
2.1	.35	.16	.21	.66	4.1	-1.01	.17	.90	.47
2.2	-.98	.17	.51	.30	4.2	.45	.16	.78	.64
2.3	2.37	.19	.98	.16	4.3	1.28	.17	.97	.41
2.4	.37	.16	.50	.75	4.4	-1.07	.17	.01	.82
2.5	2.39	.19	.00	.00	4.5	.87	.17	.54	.66
2.6	-.83	.16	.76	.13	4.6	-.94	.17	.01	.96
2.7	-.61	.17	.00	.94	4.7	-.04	.17	.01	.96
2.8	.84	.16	.37	.02	4.8	-.29	.16	.76	.38
2.9	1.67	.18	.89	.40	4.9	2.52	.19	.00	.00
2.10	1.38	.17	.61	.53	4.10	1.80	.18	.73	.01
2.11	-.70	.16	.62	.32	4.11	.65	.17	.04	.97
2.12	.07	.17	.00	.99	4.12	.43	.16	.94	.44

Table 5.9: Item Parameters 2PL Model (BILOG)

Item	$\sigma$	$SE_{\sigma}$	$\beta$	$SE_{\beta}$	Fit $\chi^2$	Item	$\sigma$	$SE_{\sigma}$	$\beta$	$SE_{\beta}$	Fit $\chi^2$
1.1	1.03	.22	.37	.06	.55	3.1	.15	.16	.36	.06	.84
1.2	.32	.14	.44	.07	.99	3.2	7.30	2.05	.17	.05	.04
1.3	-2.95	.42	.61	.12	.86	3.4	-.99	.17	.49	.07	.49
1.4	-3.28	.55	.49	.09	.69	3.5	.58	.20	.31	.06	.85
1.5	.16	.19	.29	.05	.99	3.6	-.38	.13	.48	.07	.15
1.6	-1.20	.19	.51	.08	.88	3.7	-.11	.11	.53	.07	.54
1.7	-2.18	.48	.25	.05	.43	3.8	1.57	.24	.50	.08	.77
1.8	.37	.14	.42	.06	.63	3.9	1.13	.27	.30	.06	.70
1.9	-1.17	.28	.30	.06	.77	3.10	2.37	.52	.26	.06	.98
						3.11	.15	.09	.69	.09	.25
						3.12	2.74	.63	.25	.06	.46
2.1	.32	.14	.40	.07	.53	4.1	-.97	.22	.36	.06	.96
2.2	-.99	.23	.34	.06	.47	4.2	.41	.16	.39	.06	.63
2.3	2.59	.53	.31	.07	.82	4.3	1.14	.22	.40	.07	.99
2.4	.29	.13	.46	.07	.87	4.4	-.81	.16	.49	.07	.32
2.5	4.74	1.24	.17	.04	.09	4.5	.73	.17	.43	.07	.70
2.6	-.94	.24	.30	.06	.99	4.6	-.66	.14	.54	.08	.06
2.7	-.42	.12	.56	.07	.62	4.7	-.03	.11	.53	.08	.46
2.8	1.15	.32	.24	.05	.79	4.8	-.29	.17	.34	.06	.52
2.9	1.56	.27	.38	.06	.33	4.9	4.13	1.06	.20	.05	.01
2.10	1.29	.24	.38	.06	.86	4.10	2.50	.57	.24	.05	.79
2.11	-.70	.19	.35	.06	.61	4.11	.44	.11	.58	.08	.26
2.12	.04	.09	.68	.09	.38	4.12	.41	.16	.37	.06	.66

### Test Information

The test information curves below illustrate the measurement accuracy or test information at different ability levels and standard errors of measurement for the 1PL model (Figure 5.5) and 2PL model (Figure 5.6). As to be expected, the information curve for the 1PL model indicates high test information for medium scale scores (thus low standard errors) but larger standard errors for lower and higher scale scores (low test information). For the 2PL model high test information (low standard errors) was also obtained for medium ability levels but ability at low and high levels could only be estimated with higher standard errors as the test provided less information at these ability levels.

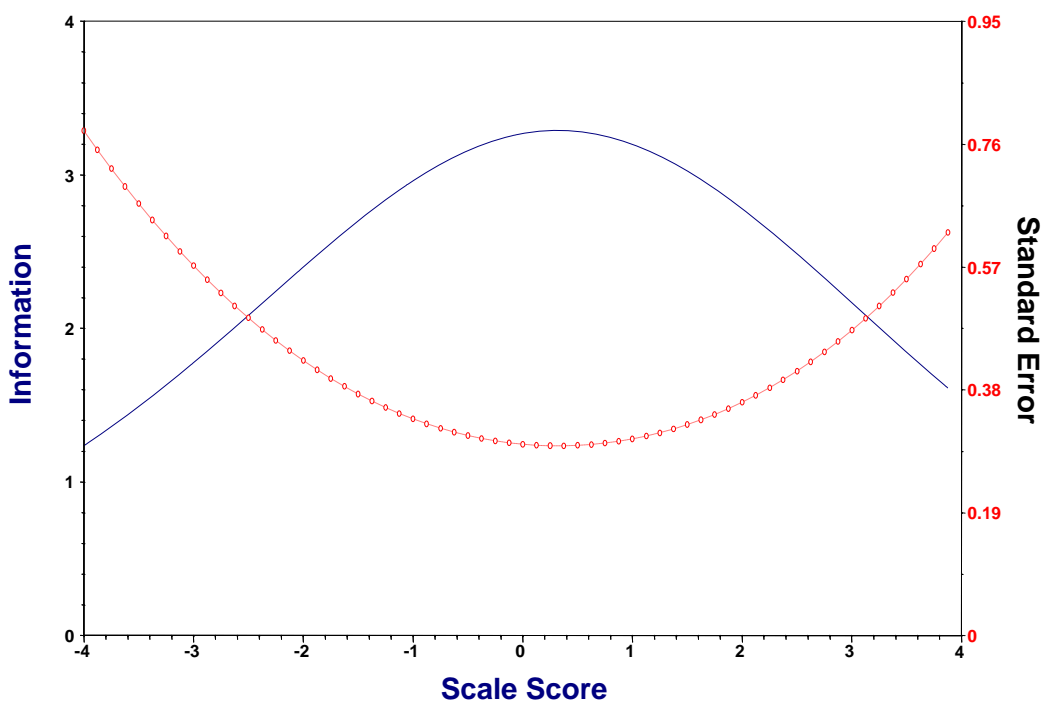


Figure 5.5: Test Information for the 1PL Model

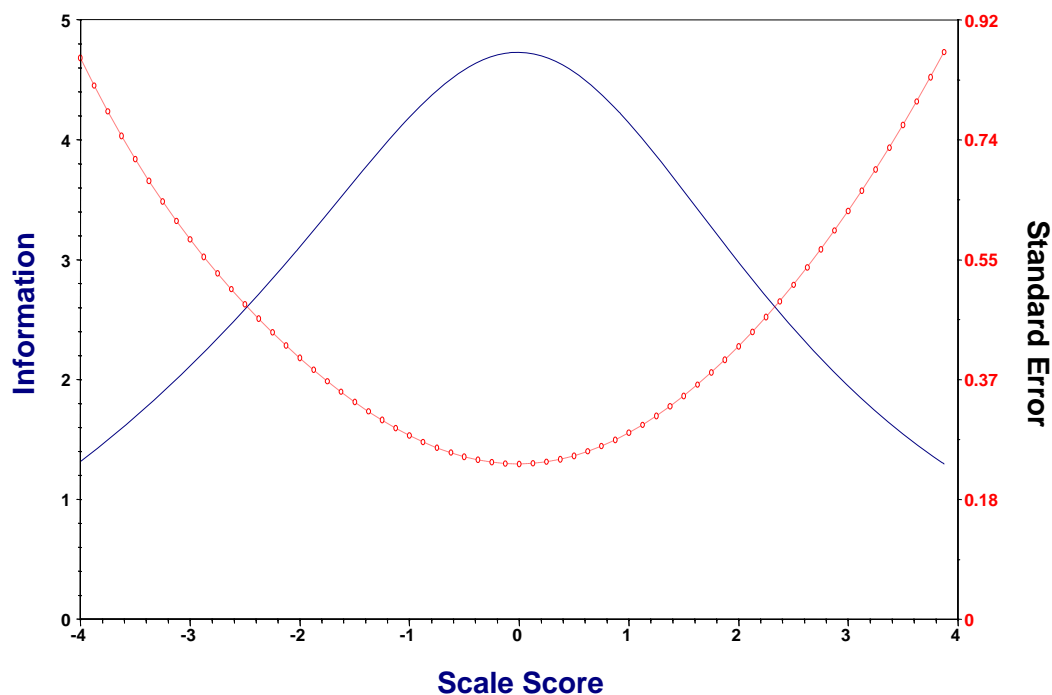


Figure 5.6: Test Information for the 2PL Model

#### 5.4.4 Goodness of Fit

##### Information Theoretic Criteria

Faced with a choice between the fit of competing models to the same data set, one needs to decide which model fits the data best. According to Rost (2004), different information-theoretic selection criteria can be applied as goodness of fit measures. Such measures are the AIC (Akaike information criterion), the BIC (Bayes information criterion), and the CAIC (Consistent AIC). Information criteria penalize models with additional parameters and are therefore based on parsimony. BIC and CAIC do not only penalize for model complexity, but also for sample size. Information criteria have to be compared to determine which model provides the relative best fit

to the data: The lower the AIC, BIC or CAIC values the better the model in comparison to another. As can be seen in Table 5.10 the information-theoretic criteria BIC and CAIC, based on the likelihoods provided by the BILOG-MG software, suggest the relative best fit to the 1PL model while the AIC has a lower index at the 2PL model.

Table 5.10: Goodness of Fit According to Likelihood and Information-theoretic Criteria

Model	Parameter $np$	-2LogL	AIC	BIC	CAIC
1PL	87	25680.45	25854.46	26218.29	26305.29
2PL	131	25545.39	25807.39	26355.23	26486.23

### Specific Model Tests for the 1PL Model: Tests According to Martin-Löf and Andersen

Specific model tests, the Martin-Löf-Test and the Andersen-Test, were applied to test assumptions of the Rasch model in order to judge the goodness of fit. These tests apply conditional likelihood ratio tests to test the Rasch model against a Rasch model with less restrictive assumptions.

#### Martin-Löf-Test

The Martin-Löf-Test estimates item homogeneity according to Martin-Löf (1973). In terms of a significant Martin-Löf-Test the Rasch model's assumption of item homogeneity for the test have to be rejected for less restrictive assumptions. A nonsignificant test would not reject the assumption that different subgroups of items measure the same underlying latent dimension.



For the total sample ( $N = 484$ ) the  $Q$ -index indicated that 13 items did not sufficiently fit the 1PL model and were thus eliminated from the item pool for this analysis. The  $Q$ -index was chosen as reference criterion for item selection for the specific model tests as it is according to Rost and von Davier (1994) sensitive to “item and person misfit” (p. 180). The Martin-Löf-Test (Table 5.11) was then conducted for the 31 Rasch homogeneous test items using the software LPCM-Win 1.0 (Fischer & Ponocny-Seliger, 1998). The item number (even vs. uneven), the number of transformation rules applied per item (1 & 2 rules vs. 3 rules), and the item parameter (defined as the median of the item difficulty parameter  $p$  according to CTT; median = .48) were chosen as item selection criteria. The Martin-Löf statistics were nonsignificant ( $p > .05$ ) for all item selection criteria. Item homogeneity could therefore not be rejected and it can be assumed that subgroups of items measure the same latent dimension.

The Martin-Löf-Test was also applied for all 44 items without pre-selection due to misfit and results are presented in Table 5.12. The same item selection criteria were reapplied and the test statistics were again nonsignificant ( $p > .05$ ). The assumption that items measure the same trait can therefore not be rejected.

### **Andersen-Test**

The Andersen-Test (Andersen, 1973) refers to person homogeneity as core assumption of the Rasch model. A nonsignificant test result in the Andersen-Test implies that the restrictive Rasch model has to be rejected in favor of a less restrictive Rasch model, i.e. the Rasch model is valid in different

Table 5.11: Model Test According to Martin-Löf (k = 31, N = 484)

Selection Criteria	$Gr_1$	$k_1$	$LL_1$	$Gr_2$	$k_2$	$LL_2$	$\chi^2$	df	p
Even-uneven	uneven	16	-3556.66	even	15	-3382.94	154.42	239	1.00
Number of rules	1&2 rules	17	-3656.70	3 rules	14	-3150.58	227.96	237	.65
Item parameter (median = .48)	< median	15	-3478.72	$\geq$ median	16	-3438.12	125.83	239	1.00

*Note.*  $Gr_1$  = Group 1 consisting of all items with an uneven number as item index,  $k_1$  = Number of Rasch homogenous test items of group 1,  $LL_1$  = Log likelihood of the data obtained by group 1,  $Gr_2$  = Group 2 consisting of all items with an even number as item index,  $k_2$  = Number of Rasch homogenous test items of group 2,  $LL_2$  = Log likelihood of the data obtained by group 2

Table 5.12: Model Test According to Martin-Löf ( $k = 44$ ,  $N = 484$ )

Selection Criteria	$Gr_1$	$k_1$	$LL_1$	$Gr_2$	$k_2$	$LL_2$	$\chi^2$	df	$p$
Even-uneven	uneven	22	-5246.37	even	22	-5150.48	207.21	483	1.00
Number of rules	1&2 rules	21	-4766.61	3 rules	23	-5506.76	311.29	482	1.00
Item parameter (median = .45)	< median	21	-4921.61	$\geq$ median	23	-5458.09	248.50	482	1.00

*Note.*  $Gr_1$  = Group 1 consisting of all items with an uneven number as item index,  $k_1$  = Number of test items of group 1,  $LL_1$  = Log likelihood of the data obtained by group 1,  $Gr_2$  = Group 2 consisting of all items with an even number as item index,  $k_2$  = Number of test items of group 2,  $LL_2$  = Log likelihood of the data obtained by group 2

score groups and estimates of item parameters in these score groups can significantly differ. A nonsignificant test result would provide evidence for the assumption of person homogeneity: Item parameters estimated by different subgroups of subjects would not significantly differ.

To conduct the Andersen Test, four person selection criteria were chosen to generate subgroups of subjects: score median, sex, age, and split-half (according to subject number) of the sample. For the 31 Rasch conform items (according to the *Q*-index), the selection criteria age was the only selection criteria that provided subgroups that did not significantly differ in estimates of item parameters by the groups (Table 5.13). It can therefore be assumed that estimates of item parameters are not significantly different in these two subgroups so that all subjects process the items due to the same latent variable (Rost, 2004). However, Andersen tests conducted for all other groups generated by various selection criteria did not support the assumption of person homogeneity.

Table 5.14 presents the results of the Andersen-Test for all 44 items applied in the figural analogy test. Selection criteria to generate subgroups were again median, sex, age, and split-half. The subgroups generated by means of the criteria age and sex did not significantly differ in item parameter estimates ( $p > .05$ ), indicating person homogeneity for these subgroups. The selection criteria median and split-half, however, divided the sample into subgroups for which person homogeneity cannot be claimed as item parameter estimates of the groups significantly differed according to the Andersen-Test.

Table 5.13: Model Test According to Andersen (k = 31, N = 484)

Selection Criteria	$Gr_1$	$N_1$	$LL_1$	$Gr_2$	$N_2$	$LL_2$	$\chi^2$	df	p
Score Median (16)	low score	274	-4386.57	high score	210	-3301.67	43.94	30	.05
Sex	male	212	-3387.86	female	272	4298.06	48.59	30	.02
Age (median)	<18	159	-2495.955	$\geq 18$	325	-5194.40	39.72	30	.11
split half	1st half	242	-3801.00	2nd half	242	-3876.38	65.64	30	.00

*Note.*  $Gr_1$  = Group 1 as obtained by the selection criteria,  $N_1$  = Number of subjects of group 1,  $LL_1$  = Log likelihood of the data obtained by group 1,  $Gr_2$  = Group 2 as obtained by the selection criteria,  $N_2$  = Number of subjects of group 2,  $LL_2$  = Log likelihood of the data obtained by group 2

Table 5.14: Model Test According to Andersen (k = 44, N = 484)

Selection Criteria	Gr <sub>1</sub>	N <sub>1</sub>	LL <sub>1</sub>	Gr <sub>2</sub>	N <sub>2</sub>	LL <sub>2</sub>	$\chi^2$	df	p
Score Median(21)	low score	267	-6156.98	high score	217	-5049.61	136.85	43	.00
Sex	male	212	-4972.98	female	272	-6280.94	43.19	43	.46
Age(median)	<18	159	-3668.51	>=18	325	-7578.40	56.21	43	.09
split half	1st half	242	-5599.25	2nd half	242	-5622.98	105.98	43	.00

Note. Gr<sub>1</sub> = Group 1 as obtained by the selection criteria, N<sub>1</sub> = Number of subjects of group 1, LL<sub>1</sub> = Log likelihood of the data obtained by group 1, Gr<sub>2</sub> = Group 2 as obtained by the selection criteria, N<sub>2</sub> = Number of subjects of group 2, LL<sub>2</sub> = Log likelihood of the data obtained by group 2

## 5.4.5 Item Construction in Focus

### Calculation of a Linear Logistic Test Model

To validate the hypothesized task structure defined by the cognitive operations and to analyze the impact of these components on item difficulty, the linear logistic test model (LLTM) was applied. The goal of this analysis was to explain item difficulty by the underlying construction rationale of the items. Thus item difficulty was assumed to be additively composed of the cognitive operations applied in the items. Further, different cognitive operations were expected to diversely contribute to item difficulty.

The program LPCM-Win 1.0 (Fischer & Ponocny-Seliger, 1998) was used to calculate the linear logistic test model. The LLTM compares its estimated item difficulties to the 1PL model estimates of the item difficulties. By means of the conditional likelihood ratio (CLR) test the maximum of the conditional likelihood of the data under the LLTM was compared to the maximum conditional likelihood of the data under the 1PL model. In a first step all 44 items were included in the calculation of the linear logistic test model in order to estimate the basic parameters. The LLTM was calculated for the data of the total sample of 484 subjects as well as for the instruction and non-instruction groups separately. The following three subsections present the results of the LLTM calculations for the total sample followed by the results referring to the instruction group and the non-instruction group.

### Total Sample

Analyzing the 44 items of the total sample the total likelihood of the LLTM was -11788.45 and the likelihood of the 1PL Rasch model was -11275.02. The statistic of the CLR, the Andersen  $\chi^2$ -test statistic displayed by the program LPCM-Win 1.0, proved that the LLTM provided a significant worse fit to the data than the Rasch model ( $\chi^2 = 1026.86$ ,  $df = 34$ ,  $p < .01$ ). However, the correlation of  $r = .84$  between item difficulties of the LLTM and the 1PL model indicated that the parameters indeed did not show perfect, but all the same high coherence. Thus about 70 % of variance in item difficulty could be explained by rules of the construction rationale. Table 5.15 presents estimates of the item parameters according to the LLTM and the Rasch model calculated by LPCM-Win. Comparing these Rasch parameter estimates with the parameter estimates presented in Table 5.8 as calculated by BILOG, to some extent considerable differences between parameter estimates occur. This is due to the different estimate algorithms applied by the programs: The discrimination parameter of the items presented by BILOG is not 1 but the mean of the item discrimination parameters of the 2PL model.

As can be seen in Table 5.16, all parameters were significant ( $p < .01$ ). Looking at the estimates for the basic parameters one can derive their impact on item difficulty. The cognitive operations referring to size facilitated items and thus increased the probability of solving the items involving such operations. The operation reflection on x-axis contributed the most to item difficulty, i.e. applying this operation in an item decreased the probability of solving an item. Reflection-y had the second highest impact on item



difficulty followed by rotation right, rotation left, sequence plus, sequence minus, and rotation 180°.

Table 5.15: LLTM vs. Rasch Parameter: Total Sample (LPCM-Win)

Item	LLTM (rescaled)	LLTM <sup>1</sup>	RM	Item	LLTM (rescaled)	LLTM <sup>1</sup>	RM
1.1	-.49	-.69	-.52	3.1	-.06	-.08	.03
1.2	.25	.36	-.10	3.2	-.80	-1.13	-2.10
1.3	1.71	2.40	2.88	3.4	.58	.81	.90
1.4	1.70	2.39	2.73	3.5	.21	.29	-.18
1.5	.23	.33	.04	3.6	.21	.29	.42
1.6	.46	.64	1.09	3.7	-.08	-.11	.22
1.7	.88	1.24	1.08	3.8	-.41	-.58	-1.12
1.8	.82	1.15	-.13	3.9	-1.16	-1.63	-.45
1.9	.00	.00	.73	3.10	-.67	-.95	-.97
				3.11	-.05	-.07	-.02
				3.12	-.43	-.61	-1.07
2.1	-.08	-.11	-.08	4.1	-.06	-.08	.72
2.2	1.17	1.64	.70	4.2	-.80	-1.13	-.14
2.3	-1.45	-2.04	-1.27	4.3	-.30	-.43	-.63
2.4	-.37	-.51	-.09	4.4	.58	.81	.76
2.5	-.96	-1.35	-1.28	4.5	.21	.29	-.39
2.6	1.11	1.56	.61	4.6	.21	.29	.68
2.7	.55	.77	.48	4.7	-.08	-.11	.15
2.8	-.60	-.84	-.37	4.8	-.41	-.58	.29
2.9	-.34	-.48	-.86	4.9	-1.16	-1.63	-1.36
2.10	-.21	-.29	-.69	4.10	-.67	-.95	-.93
2.11	.74	1.04	.54	4.11	-.05	-.07	-.26
2.12	.53	.74	.09	4.12	-.43	-.61	-.13

Note. LLTM<sup>1</sup> = variance adjusted parameters; RM = Rasch Model

### LLTM for the Instruction Group

The calculation of the LLTM was also separately carried out for the experimental group that received an explanation of the transformation rules prior to the test ( $N = 249$ ). Again all 44 items were taken into account and the

Table 5.16: LLTM: Basic Parameter Estimates, Total Sample (LPCM-Win)

Operation	Basic parameter	SE	z-value	p
sp	.29	.04	7.31	.00
sm	.29	.04	7.35	.00
rr	-1.19	.06	20.75	.00
rl	-1.16	.06	21.08	.00
r180	-.54	.06	9.79	.00
rfx	-1.91	.06	32.31	.00
rfy	-1.42	.06	23.60	.00
sqp	-.96	.04	23.05	.00
sqm	-.60	.04	14.52	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

nine basic parameters were estimated. The LLTM total likelihood resulted in -6037.52 and the Rasch model total likelihood in -5738.32. With a  $\chi^2$ -statistic of 598.40 ( $df=34$ ) the estimated LLTM and Rasch item parameters (Table 5.18) significantly differed, thus the LLTM again provided a significant worse fit to the data compared to the Rasch model. However, the Pearson correlation between parameters of the Rasch model and the LLTM was  $r = .83$ , indicating that the item design variables accounted for a quite large proportion of explained variance. Table 5.17 presents the estimates of the basic parameters by the LLTM. Like for the data of the total sample, the transformations size plus and size minus were equally easy and represent the transformations that make items easier. The transformation reflection x was again the most difficult one followed by reflection y, rotation right, rotation left, sequence plus, rotation 180°, and sequence minus.

Table 5.17: LLTM: Basic Parameter Estimates, Instruction Group (LPCM-Win)

Operation	Basic parameter	SE	z-value	<i>p</i>
sp	.20	.06	3.51	.00
sm	.20	.05	3.63	.00
rr	-1.30	.08	15.66	.00
rl	-1.29	.08	16.17	.00
r180	-.67	.08	8.41	.00
rfx	-2.04	.08	24.09	.00
rfy	-1.60	.09	18.76	.00
sqp	-.97	.06	16.15	.00
sqm	-.61	.06	10.64	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

### LLTM for the Non-Instruction Group

For the data of the non-instruction group ( $N = 235$ ) the LLTM provided a significant worse fit compared to the Rasch model (LLTM total likelihood: -5741.59, Rasch model total likelihood : -5480.11,  $\chi^2 = 522.97$ ,  $df = 34$ ). However, the Pearson correlation between LLTM and Rasch item parameters presented in Table 5.20 was  $r = .82$ . Looking at the basic parameters (Table 5.19) the operations referring to size were again the easiest transformations with size plus being slightly easier than size minus. The transformations sequence minus and rotation 180° changed positions, but the order of all other operation corresponded to the rank of the transformations of the instruction-group with reflection on the x-axis being the most difficult transformation and thus decreasing the probability of solving an item the most.

Table 5.18: LLTM vs. Rasch Parameter: Instruction Group (LPCM-Win)

Item	LLTM (rescaled)	LLTM <sup>1</sup>	RM	Item	LLTM (rescaled)	LLTM <sup>1</sup>	RM
1.1	-.43	-.58	-.50	3.1	-.09	-.12	-.05
1.2	.31	.42	.13	3.2	-.84	-1.13	-2.19
1.3	1.79	2.42	2.61	3.4	.52	.71	.94
1.4	1.80	2.42	2.86	3.5	.17	.23	-.23
1.5	.29	.40	.08	3.6	.17	.23	.64
1.6	.63	.86	1.99	3.7	-.10	-.14	.04
1.7	.93	1.26	.87	3.8	-.45	-.61	-1.16
1.8	.99	1.34	.09	3.9	-1.20	-1.62	-.48
1.9	.00	.01	.83	3.10	-.76	-1.02	-.96
				3.11	-.09	-.13	.06
				3.12	-.47	-.63	-1.30
2.1	-.02	-.04	-.25	4.1	-.09	-.12	.54
2.2	1.13	1.53	.62	4.2	-.84	-1.13	-.16
2.3	-1.39	-1.88	-1.37	4.3	-.40	-.55	-.70
2.4	-.30	-.41	-.02	4.4	.52	.71	.71
2.5	-.96	-1.29	-1.50	4.5	.17	.23	-.32
2.6	1.19	1.60	.58	4.6	.17	.23	.73
2.7	.51	.69	.51	4.7	-.10	-.14	.27
2.8	-.60	-.81	-.30	4.8	-.45	-.61	.20
2.9	-.29	-.39	-.84	4.9	-1.20	-1.62	-1.35
2.10	-.23	-.32	-.59	4.10	-.76	-1.02	-1.11
2.11	.83	1.13	.41	4.11	-.09	-.13	-.37
2.12	.49	.67	.15	4.12	-.47	-.63	-.11

Note. LLTM<sup>1</sup> = variance adjusted parameters; RM = Rasch Model

Table 5.19: LLTM: Basic Parameter Estimates, Non-Instruction Group (LPCM-Win)

Operation	Basic parameter	SE	z-value	<i>p</i>
sp	.40	.06	6.95	.00
sm	.38	.06	6.74	.00
rr	-1.07	.08	13.55	.00
rl	-1.05	.08	13.66	.00
r180	-.42	.08	5.41	.00
rfx	-1.79	.08	21.40	.00
rfy	-1.24	.08	14.65	.00
sqp	-.96	.06	16.16	.00
sqm	-.59	.06	9.95	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

To analyze whether the parameter estimates differed in size, a test concerning the differences between the parameter estimates obtained by the instruction group and the non-instruction group was conducted. Results are presented in Table 5.21 and show that apart from the transformation rules sqp and sqm parameter estimates differed between the instruction group and the non-instruction group.

### LLTM Including Random Item Effects

As shown when calculating the LLTM above, accurate predictions of item difficulties by item properties can hardly be yielded due to strict assumptions of the LLTM. The linear logistic test model with random item effects, however, represents a more realistic approach to validating the cognitive structure (as described in chapter 3).

Table 5.22 presents the estimates of the parameters. Apart from the op-

Table 5.20: LLTM vs. Rasch Parameter: Non-Instruction Group (LPCM-Win)

Item	LLTM (rescaled)	LLTM <sup>1</sup>	RM	Item	LLTM (rescaled)	LLTM <sup>1</sup>	RM
1.1	-.55	-.81	-.55	3.1	-.02	-.03	.12
1.2	.18	.27	-.36	3.2	-.76	-1.11	-2.00
1.3	1.63	2.38	3.16	3.4	.62	.91	.85
1.4	1.61	2.35	2.61	3.5	.23	.35	-.13
1.5	.15	.23	-.01	3.6	.23	.35	.19
1.6	.27	.40	.43	3.7	-.05	-.07	.39
1.7	.81	1.19	1.27	3.8	-.36	-.54	-1.07
1.8	.64	.94	-.38	3.9	-1.11	-1.62	-.42
1.9	-.00	-.01	.61	3.10	-.58	-.84	-.97
				3.11	-.00	.00	-.11
				3.12	-.39	-.58	-.83
2.1	-.14	-.20	.08	4.1	-.02	-.03	.89
2.2	1.19	1.74	.77	4.2	-.76	-1.11	-.11
2.3	-1.51	-2.20	-1.15	4.3	-.19	-.28	-.55
2.4	-.43	-.63	-.17	4.4	.62	.91	.79
2.5	-.96	-1.40	-1.04	4.5	.23	.35	-.46
2.6	1.04	1.52	.63	4.6	.23	.35	.61
2.7	.58	.85	.45	4.7	-.05	-.07	.01
2.8	-.59	-.86	-.44	4.8	-.36	-.54	.37
2.9	-.40	-.59	-.87	4.9	-1.11	-1.62	-1.38
2.10	-.17	-.26	-.80	4.10	-.58	-.84	-.74
2.11	.65	.96	.65	4.11	-.00	.00	-.13
2.12	.55	.81	.01	4.12	-.39	-.58	-.15

Note. LLTM<sup>1</sup> = variance adjusted parameters; RM = Rasch Model

Table 5.21: Comparison of Parameter Estimates: Instruction Group vs. Non-Instruction Group

Operation	I-G	NI-G	Delta	SE (I-G)	SE (NI-G)	S (pooled)	Z
sp	.20	.40	.20	.06	.08	.07	2.84
sm	.20	.38	.18	.05	.08	.07	2.72
rr	-1.30	-1.07	.23	.08	.08	.08	2.88
rl	-1.29	-1.05	.24	.08	.08	.08	3
r180	-.67	-.42	.25	.08	.08	.08	3.13
rfx	-2.04	-1.79	.25	.08	.08	.08	3.13
rfy	-1.6	-1.24	.36	.09	.08	.09	4.22
sqp	-.97	-.96	.01	.06	.08	.07	.14
sqm	-.61	-.59	.02	.06	.08	.07	.28

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus, I-G = parameter estimates for the Instruction Group, NI-G = parameter estimates for the Non-Instruction Group

erations size plus and size minus, the estimates of all operations were significant. Since the LLTM with random item effects was calculated with the software SAS, the LLTM was again calculated using SAS to compare the estimates of the LLTM with estimates of the LLTM with random effects. The estimates of the basic parameters for the LLTM are presented in Table 5.23. For the LLTM calculated by SAS, algebraic signs are reversed and have to be multiplied by  $-1$  to check them against the estimates of the LLTM with random item effects.

Comparing the results of both analyses, the standard errors of the model with random item effects were considerably larger than the ones of the model with fixed item effects due to the additional error term. Comparison of the parameter estimates showed - as expected - nearly perfect coherence ( $r = .999$ ,  $p = .000$ ). However, the parameters of the random effects model

were larger due to the additional random effects that improved the description of the data and led to “larger estimates of the other effects” (Janssen et al., 2004, p. 204). According to Janssen et al. (2004) small differences between parameter estimates of the models indicate that item difficulty can be well described by the item parameters. Thus, analysis of the data by the linear logistic test model with random effects was effective, and implied exceeding advantages: Referring to the linear logistic test model with random effects is less restrictive than the LLTM with fixed item effects and therefore generally much more realistic. The model with random effects defies the idealistic premise and accounts for more pragmatic assumptions.

The linear logistic test model with random effects was also calculated separately for the instruction group and the non-instruction group. The parameter estimates for the instruction group showed almost perfect correlation with the parameter estimates of the total sample ( $r = .996, p = .000$ ). Again both size transformations contributed to the easiness of an item, but not significantly. A highly significant correlation was also obtained for the parameter estimates of the non-instruction group with parameter estimates of the total sample ( $r = .995, p = .000$ ). For the non-instruction group size again did not significantly contribute to item difficulty, and the transformation  $180^\circ$  did not have significant impact on item difficulty either. Regarding the instruction group and the non-instruction group, a correlation of  $r = .984 (p = .000)$  of the parameter estimates was obtained. Tables of the results of the LLTM analysis with random effects for the instruction group and the non-instruction group are presented in Appendix A (Tables A.1 and A.3). The corresponding analyses of the LLTM conducted by SAS



are also presented in Appendix A (Tables A.2 and A.4).

Table 5.22: LLTM with Random Item Effects: Basic Parameter Estimates, Total Sample (SAS)

Operation	Basic parameter	SE	t-value	p
constant	1.49	.29	5.15	.00
sp	.34	.23	1.51	.13
sm	.30	.22	1.37	.17
rr	-1.36	.32	-4.25	.00
rl	-1.34	.31	-4.35	.00
r180	-.70	.31	-2.27	.02
rfx	-2.10	.32	-6.58	.00
rfy	-1.58	.34	-4.71	.00
sqp	-1.05	.23	-4.59	.00
sqm	-.71	.23	-3.10	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

### LLTM with Random Item Effects for Rasch Conform Data

The linear logistic test model with random effects was also calculated for a model containing only Rasch conform items. The Rasch conform items were selected according to the  $\chi^2$ -fit statistic (Table 5.8). According to the  $\chi^2$ -fit statistic, 12 items were not conform to the Rasch model and were thus excluded from the analysis. Table 5.24 shows the results for the analysis with the 32 remaining Rasch conform items. Parameter estimates for the transformation rules as well as standard errors for the estimates and significance parameters are presented. Parameter estimates for the 32 Rasch conform items and estimates for the 44 test items analyzed according to the LLTM with random effects (Table 5.22) significantly correlated ( $r = .997$ ,  $p$

Table 5.23: LLTM: Total Sample (SAS)

Operation	Basic parameter	SE	t-value	p
sp	-.29	.04	-7.40	.00
sm	-.29	.04	-7.39	.00
rr	1.18	.06	20.94	.00
rl	1.16	.05	21.28	.00
r180	.54	.05	9.86	.00
rfx	1.91	.06	32.30	.00
rfy	1.42	.06	23.82	.00
sqp	.96	.04	23.58	.00
sqm	.60	.04	14.88	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

= .000). The impact of both size transformations was again not significant. Further, the cognitive operation rotation 180° did not yield significant impact on item difficulty either.

Looking at the parameter estimates of the LLTM with random effects for the Rasch conform items, both size operations proved almost identical estimates, as did the transformations 90° rotation right and 90° rotation left. Thus, an alternative, more parsimonious model was considered. The two size transformations were integrated and a new parameter *size*, combining the transformations size plus and size minus, was generated. The rotation transformations 90° left and 90° right were subsumed and a joint transformation *rotation* was created. The transformation 180° rotation was not subjoined, since its parameter estimate was too different from both 90° rotation transformations, indicating a difference in execution of this operation. The cognitive operations reflection about the x-axis and reflection

Table 5.24: LLTM with Random Item Effects for Rasch Conform Items ( $k = 32$ ,  $N = 484$ ) (SAS)

Operation	Basic parameter	SE	$t$ -value	$p$
constant	1.26	.34	3.77	.00
sp	.28	.29	.98	.33
sm	.30	.24	1.23	.22
rr	-1.12	.37	-3.04	.00
rl	-1.12	.35	-3.18	.00
r180	-.66	.35	-1.86	.06
rfx	-1.64	.37	-4.43	.00
rfy	-1.35	.36	-3.69	.00
sqp	-1.00	.28	-3.52	.00
sqm	-.58	.29	-2.02	.04

*Note.* sp = size plus, sm = size minus, rr = rotation  $90^\circ$  right, rl = rotation  $90^\circ$  left, r180 = rotation  $180^\circ$ , rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

about the y-axis were combined to *reflection* and it was further not distinguished between sequence plus and sequence minus operations, since the transformation *sequence* accounted for both. Thus only five parameters (instead of nine parameters) were included in the model, analyzing the 32 Rasch conform items. Table 5.25 presents the results. The size operation still did not yield significant impact on item difficulty, but the estimate of the  $180^\circ$  rotation-transformation significantly influenced difficulty. All other parameter estimates were significant, and the cognitive operation reflection again contributed to difficulty the most.

Two linear logistic models with random item effects were thus obtained for the Rasch conform items. The first model incorporated all nine transformation rules as model parameters (cf. Table 5.24) and the second model resulted by combining parameters of the first model to a more parsimo-

nious model with only five parameters. A likelihood ratio test (LRT) was applied for comparison of the likelihood of the data given the parsimonious model with the likelihood of the data given the original model. The LRT was significant ( $\chi^2 = 18.3$ ,  $df = 4$ ,  $p = .001$ ) and the more parsimonious model with five parameters only could thus not explain the data as well as the complex model with nine parameters.

Table 5.25: LLTM with Random Item Effects: Parsimonious Model for Rasch Conform Items ( $k = 32$ ,  $N = 484$ ) (SAS)

Operation	Basic parameter	SE	<i>t</i> -value	<i>p</i>
constant	1.29	.33	3.90	.00
size	.24	.22	1.09	.28
r180	-.72	.35	-2.06	.04
rotation	-1.10	.31	-3.55	.00
reflection	-1.52	.32	-4.78	.00
sequence	-.78	.25	-3.13	.00

Note. r180 = rotation 180°

#### 5.4.6 Impact of Elements on Item Difficulty

Besides analyzing the influence of cognitive operations on item difficulty, a closer look was also taken on the impact of the different elements. It was assumed that letters and digits had different impact on item difficulty and, depending on the type of element chosen, inferring the relation between the A term and the B term might be unequally difficult.

Having applied two categories of elements in the items (letters and digits), four ways of combining the A & B terms with the C & D terms were generated. Letters of the A & B terms could be combined with letters of the C & D terms (ll). Further, digits of the first two terms could be com-

bined with digits of the C & D terms (dd). The restriction imposed on the application of elements necessitated that elements in A and B, and in C and D had to be of the same category, but categories *between* A & B and C & D could differ. Therefore items could also contain digits in the A & B terms and letters in the C & D terms (dl), or letters in the first terms and digits in the latter (ld).

Like in the LLTM-analysis of the transformations, the program LPCM-Win 1.0 (Fischer & Ponocny-Seliger, 1998) was applied to analyze the impact of elements on item difficulty. As can be seen in Table 5.26, inferring the transformation rule from letters and applying the rule to letters (ll) significantly increased item difficulty. Item difficulty was also significantly increased when digits of the A & B terms were combined with letters of the C & D terms (dl). Further, the combination of letter-digit elements (ld) did not significantly contribute to item difficulty. The probability to solve an item was significantly increased, and items thus facilitated, when the transformation rule had to be inferred from digits in the A & B terms and applied to digits in the C & D terms (dd).

Table 5.26: LLTM: Combination of Letters and Digits (LPCM-Win)

Element	Basic parameter	SE	z-value	<i>p</i>
ll	-.13	.03	4.0791	.00
dd	.43	.03	12.5037	.00
dl	-.24	.03	7.0972	.00
ld	-.05	.03	1.6509	.09

*Note.* ll = letter-letter combination of A- and C-elements, dd = digit-digit combination of A- and C-elements, dl = digit-letter combination of A- and C-elements, ld = letter-digit combination of A- and C-elements

Table 5.27: LLTM: Orientation of A-Elements (LPCM-Win)

Element orientation	Basic parameter	SE	z-value	p
A normally oriented	.20	.04	5.01	.00
A 90° left rotated	.29	.04	8.15	.00
A reflected on x-axis	.37	.08	4.52	.00
A 180° rotated	-.03	.04	0.72	.47
A 90° right rotated	-.40	.04	11.17	.00
A reflected on y-axis	-.43	.04	10.24	.00

Further, the impact of the initial position or orientation of elements was of particular interest. For this purpose the orientation of the A-elements was examined and recorded. In the existing test version, A-elements were either “normally” orientated, rotated 90° to the left or right, or 180° rotated. A-elements were also presented as reflected elements about their x-axis or y-axis. Results are shown in Table 5.27. Apart from when the element was presented in a 180° rotated position, all orientations of elements were significantly related to item difficulty. Normally orientated elements, 90° left rotated elements, and elements rotated about the x-axis made items easier. Elements orientated 90° to the right and elements reflected on the y-axis increased item difficulty and therefore reduced the probability of solving an item.

The format in which C-elements were displayed was analyzed likewise. When digits were applied as elements in the C term, item difficulty significantly decreased whereas when letters were applied as elements, item difficulty significantly increased (Table 5.28). Every orientation of the C-element applied in the test significantly contributed to item difficulty. Normally angled C-elements, 90° right rotated elements, as well as elements

reflected on the x-axis, were positively related to item difficulty, thus increased difficulty. Elements that were presented 90° rotated to the left or 180° rotated contributed to the easiness of an item, thus facilitated items.

Table 5.29 presents summarized findings of the analysis on elements when A-elements and C-elements were jointly examined. All basic parameters analyzed yielded significant impact on item difficulty. However, consistent as well as inconsistent findings concerning the impact of A- and C-elements on item difficulty are reported. Corresponding findings apply to the 90° rotations: left rotated A-elements and left rotated C-elements both increased the probability of solving an item, and right rotated elements, in the A term as well as the C term decreased the probability of solving an item. Looking at the values of these parameters, left rotated A-elements influenced difficulty to the same degree as did left rotated C-elements. The value of the parameter for right rotated A-elements was more than twice as high as the value of the parameter for the right rotated C-elements. Opposed findings on the impact of elements on item difficulty were revealed for all other parameters: Normally oriented A-elements eased items whereas normally oriented C-elements impeded items. The same finding could be observed for elements reflected on the x-axis where A-elements decreased item difficulty and C-elements increased item difficulty. Regarding the elements presented in a 180° rotated position, A-elements were positively related to item difficulty whereas C-elements facilitated items.

Finally, impact of elements on item difficulty and impact of transformation rules on item difficulty were jointly analyzed by means of the linear logistic test model. A review of the parameter estimates of the elements

Table 5.28: LLTM: C-Elements (LPCM-Win)

Element	Basic parameter	SE	z-value	<i>p</i>
C letter	-.14	.04	3.90	.00
C digit	.14	.04	3.90	.00
C normally oriented	-.22	.04	5.89	.00
C 90° left rotated	.28	.04	7.50	.00
C reflected on x-axis	-.27	.04	7.07	.00
C 180° rotated	.35	.03	10.79	.00
C 90° right rotated	-.14	.03	4.11	.00

Table 5.29: LLTM: Elements A &amp; C (LPCM-Win)

Element	Basic parameter	SE	z-value	<i>p</i>
A normally oriented	.21	.05	4.23	.00
A 90° left rotated	.41	.04	9.67	.00
A reflected on x-axis	.79	.10	8.22	.00
A 180° rotated	-.50	.06	8.61	.00
A 90° right rotated	-.44	.04	11.28	.00
A reflected on y-axis	-.48	.05	9.11	.00
A letter	-.32	.04	8.32	.00
A digit	.32	.04	8.32	.00
C letter	-.10	.04	2.90	.00
C digit	.10	.04	2.90	.00
C normally oriented	-.23	.05	4.90	.00
C 90° left rotated	.42	.05	8.59	.00
C reflected on x-axis	-.36	.04	8.31	.00
C 180° rotated	.33	.04	8.88	.00
C 90° right rotated	-.17	.04	4.05	.00



of the A and C terms (cf. Table 5.29) and the cognitive transformations applied in the test (cf. Table 5.16) should thus be obtained. To reduce the number of parameter estimates, it was not further distinguished between the degree of element rotation or reflections on the x-axis or y-axis but combined parameters (A rotated, A reflected, C rotated and C reflected) resulted. Looking at the results of the LLTM analysis (Table 5.30) it is noticeable that the absolute value of the estimates of the elements was generally by far lower than the absolute values of the transformation parameter estimates (apart from the size operations). Among the A-element estimates, reflected A-elements, as well as letters and digits as elements contributed to item difficulty whereas rotated A-elements contributed to item easiness. Rotated C-elements as well as reflected C-elements and digits enhanced the easiness of items whereas normally oriented C-elements and letters contributed to item difficulty. Among the transformation rules, estimates showed that reflection of elements on their x-axis increased item difficulty the most, followed by rotation right, reflection on the y-axis, and rotation left. Then, with less impact in terms of the absolute value, but still related to the difficulty of an item, sequence plus, sequence minus, and rotation  $180^\circ$  followed. Again, as expected, both size operations made items easier.

Table 5.30: LLTM: Elements and Transformations (LPCM-Win)

Element/Transformation	Basic parameter	SE	z-value	p
A normally oriented	.03	.05	.60	.55
A rotated	.12	.04	3.49	.00
A reflected	-.15	.05	3.28	.00
A letter	-.17	.04	4.45	.00
A digit	.17	.04	4.45	.00
C normally oriented	-.19	.04	4.40	.00
C rotated	.13	.04	3.66	.00
C reflected	.06	.04	1.28	.20
C letter	-.16	.04	4.30	.00
C digit	.16	.04	4.30	.00
sp	.34	.04	8.38	.00
sm	.37	.04	9.04	.00
rr	-1.55	.07	23.47	.00
rl	-1.19	.06	19.58	.00
r180	-.62	.06	10.60	.00
rfx	-1.90	.06	32.68	.00
rfy	-1.49	.06	23.35	.00
sqp	-.87	.04	19.57	.00
sqm	-.65	.04	15.22	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

# 6

## Summary and Discussion

The aim of this dissertation was the rule-based construction of a figural analogy test, the empirical examination of its psychometric properties, and the validation of the underlying rationale. Figural analogies as subject of the test were chosen because of their indicative function of reasoning ability, a core concept of human intelligence.

Subsequently, results are summarized and discussed in accordance with the stated research questions and hypotheses of this thesis. The first research question referred to constituents of item difficulty and the analysis

of cognitive components. Besides the impact of cognitive transformations applied in the tasks, the impact of different elements on item difficulty was analyzed. The second research question concerned the difference between the test scores of the instruction group and the non-instruction group. The last research question engaged in the relationship of the test to scholastic achievement. In the discussion, references to the research questions and hypotheses are made where appropriate. Concluding remarks of this discussion are on future prospects and related research topics with regard to this study.

For the rule-based test construction the following components were chosen to design the items. Letters and digits were chosen as elements of this test. As elements of the A term they had to be asymmetric in their vertical and horizontal axis to ensure that rules changing the element from A to B could be univocally inferred by the participants of the study. For rules transforming the elements from A to B and thus from C to D, nine different rules were applied. These rules could be assigned to four different categories: Reflection, rotation, size, and sequence. The category reflection referred to reflecting the element about its x-axis or y-axis. The rotation category involved rotations of the elements in different angles: 90° right, 90° left, and 180°. Size transformations were used to increase or decrease elements in their visual appearance. In the category sequence, sequence plus and sequence minus transformations were distinguished. When sequence plus was applied letters were consecutive according to the alphabet and digits were increased according to arithmetic addition. Sequence mi-

nus required to go backwards in the alphabet when letters were applied, and to subtract when digits were applied. The absolute value or alphabetic interval between the C term and the D term was determined by the equivalent value or interval between the A term and B term.

The test was composed of four subtests with a total of 44 items. In the first subtest one transformation rule per item was applied. In the second subtest two rules per item were applied, and items of the third and fourth subtests contained three rules each. Apart from items of subtest 1, all items displayed nine response alternatives plus the option “no solution right”. Items of subtest 1 presented only five distractors plus the “no solution right” - option.

In the main examination of the study 484 students of various German high schools were tested with the analogy test. Testing time was limited for each subtest. Subjects were randomly assigned to two different experimental groups: In the instruction group the transformation rules applied in the test were explained to the subjects on basis of examples. In the non-instruction group subjects only received general test instruction and were not introduced to the cognitive operations applied in the test.

Regarding the evaluation according to the classical test theory for the total sample, item difficulty parameters yielded satisfactory values: Parameters ranged from .11 to .93, and only one item yielded a parameter below .20 and only two items proved parameters above .80. Excluding one item (item 3.2) due to its negative discrimination parameter, item discrimination parameters yielded a mean discrimination power of .23. Interpreting the item

discrimination parameters, the relation between item difficulty and discrimination power has to be taken into account. Very difficult and very easy items can obtain lower item discrimination parameters compared to items of medium difficulty. In this test only two items proved solving probabilities of above 90% and only one item was solved from 11% of the subjects. This was the most difficult item and the only item processed by the total sample that had a negative discrimination parameter. However, the low mean item discrimination parameter of the test cannot only be explained in terms of extreme item difficulties since only these three out of 44 items accumulated at the extreme ends of the difficulty scale. A possible reason for the low discrimination parameters might have been the constraint of testing time. The time limit chosen might have been too tight so that, for example, good but slow participants were not able to process certain items. Thus, extended testing time, generous enough for most of the subjects to process all items could possibly enhance the discrimination power of the items and the test.

Mean comparisons between the analogy scores obtained by the instruction group and the non-instruction group showed that the instruction group significantly outperformed the non-instruction group. Sample comparisons confirmed that subjects of the two groups did not a priori significantly differ, neither in the CFT-20 R test, nor in school performance as measured by composite scores. Difference in performance on the analogy test could therefore be attributed to the different instructions received. Thus subjects that were introduced to the various cognitive operations, that had to be inferred from A to B and applied onto the C term to generate D, score-wise

benefited from this instruction. These findings support the hypothesis of the second research question, assuming that the instructed group performs significantly better than the non-instructed group.

Validity estimates of the figural analogy test were obtained by correlations with the CFT-20 R test scores and a coefficient of  $r = .36$  was calculated. Discriminating between the instruction group and the non-instruction group different validity estimates were indicated. The analogy test scores of members of the non-instruction group correlated only  $r = .24$  with the CFT-20 R test scores whereas the analogy test scores of the instruction group correlated  $r = .54$  with the CFT-20 R test scores.

Further, self-reported school grades were consulted for additional validity estimation by scholastic achievement. For this purpose, composite scores of the reported school grades were established. Correlations between the analogy test score and the composite scores were calculated: Apart from the social science composite score all correlations were highly significant. In accordance with the hypothesis of the third research question, among all composite scores the maths & science score correlated the highest with the analogy test score ( $r = .22$ ). Art & music correlated  $r = .16$  and the language composite score correlated  $r = .10$  with the analogy test score. For each participant an average grade score of all reported grades was calculated and this composite score correlated at  $.18$  with the analogy score. However, the mediocre relationship between scholastic achievement and the analogy test score, that was assumed in the third research question could not be confirmed.

The test was also analyzed according to item response theory models. According to the  $\chi^2$ -fit statistic 12 items showed misfit when parameters were estimated applying the 1PL model. Under the 2PL model, two items showed misfit. Analyzing all items, among the information-theoretic criteria, the BIC- and CAIC- indices suggested the relatively best fit for the 1PL model whereas the AIC suggested the relatively best fit for the 2PL model.

For both, the 1PL model and the 2PL model, test information curves of all items indicated highest test information for medium ability levels and showed that standard errors increased as a function of increasing or decreasing ability levels (increasing or decreasing from the medium ability level). Thus, test information decreased with extreme ability levels.

The main issue of the study referred to the first research question and thus the validation of the cognitive structure of the analogy test and the estimation of the impact of the item components on item difficulty by means of the linear logistic test model.

Ranking the transformation rules according to their difficulty, following order of increasing difficulty was obtained for the total sample: Size plus/size minus, 180° rotation, sequence minus, sequence plus, 90° rotation left, 90° rotation right, reflection about the y-axis, and reflection about the x-axis. Both size operations were equally easy and made items easier. All other transformations were positively related to item difficulty, thus decreased the probability to solve an item. The 90° rotation transformations almost equally contributed to item difficulty, whereas the 180° rotation was relatively easier. Reflection transformations contributed to difficulty



the most.

The linear logistic test model was also separately conducted for the instruction group and the non-instruction group. For the non-instruction group the transformation rules showed the same ranking as for the total sample and the size operations again made items easier. Size plus was slightly easier than size minus and reflection about the  $x$ -axis again increased difficulty the most. For the instruction group both size operations were equally easy and once more the easiest transformations. Sequence minus and the  $180^\circ$  rotation transformation were interchanged in rank with sequence minus being marginally easier than the  $180^\circ$  rotation. The rank order of all other transformations remained the same. Thus, in all analyses every cognitive transformation significantly influenced item difficulty.

The stated hypothesis among this issue suggested that transformations referring to displacement increase item difficulty and transformations referring to distortion decrease item difficulty. Because of their character, the transformations sequence minus and sequence plus could neither be assigned to the displacement type, nor to the distortion type. As the results show, findings here are consistent with other research (eg. Whiteley & Schneider, 1981): Cognitive operations of the category rotation and reflection significantly contributed to item difficulty the most except for the transformation  $180^\circ$  rotation. Both size operations significantly contributed to item facileness as they increased the probability of solving an item. Thus neglecting the transformation  $180^\circ$  rotation the hypothesis of the first research question of the study can be confirmed for the total sample as well as for the instruction group and the non-instruction group. With minor

deviations in the rank order of transformations obtained by the instruction group, the second hypothesis of the second research question falls into line here, assuming that, conducting separate analyses for the instruction group and the non-instruction group, cognitive operations do not differ in their order of difficulty for those groups. However, the hypothesis assumes further that the absolute values of parameter estimates do not differ. To analyze this issue precisely, a test concerning the differences between the parameter estimates obtained by the instruction group and the non-instruction group was conducted. Results showed that experimental variation effected all parameters except  $s_{qp}$  and  $s_{qm}$ . Thus, the assumption of equal parameter estimates cannot be confirmed. However, looking at the *SE* estimates for both groups obtained by the random effect models and conducting a test of differences, no differences between the groups are observed and the hypothesis can therefore confirmed. Thus, depending on the model chosen, the hypothesis can either be rejected or confirmed.

Different cognitive processes involved in the transformations seem to be responsible for the variability in the impact of the transformation categories on item difficulty. In general, reflection of elements is more difficult than rotation of elements, which is more difficult than sequence and size transformations. Different cognitive operations involved in the transformation processes of the elements account for the different parameter estimates: The process of reflecting elements is more complex than the process of rotating elements since reflecting involves an additional abstract process. Performance of sequence processes mainly involves relational requirements and size operations mainly require visual thinking. Thus, the

role of different information-processes involved in various transformation categories and their impact on item difficulty offer a great potential for further research.

Further, the impact of the elements applied in the items processed by the total sample was analyzed by means of the linear logistic test model. First, the impact of the type of element (digit or letter) in the A & B terms and in the C & D terms was studied. Digit-digit combinations (A & B elements and C & D elements are both digits) significantly made items easier and letter-digit combinations (A & B elements represent letters whereas C & D elements represent digits) did not significantly contribute to item difficulty. Letter-letter combinations and digit-letter combinations were both significantly related to item difficulty and decreased the probability of correctly solving an item. In comparison digit-letter combinations were more difficult than letter-letter combinations.

It was also examined if the orientation of the A term elements played a significant role for item difficulty when relations were inferred from A to B. Analysis of only the A-elements showed that presentation of the A-element in a 180° rotated initial position did not yield significant impact on item difficulty. Normally oriented, and 90° left rotated A-elements, and elements reflected about the x-axis significantly decreased item difficulty, thus increased the probability of solving an item. When the A-element was presented in a 90° right rotated orientation or reflected on its y-axis, item difficulty increased.

Analyzing the C-elements, different orientations and types always proved

significant impact on item difficulty. When the C-element was a digit, 90° left or 180° rotated, a contribution to the easiness of an item was made. When the C-element was a letter, normally orientated, reflected on its x-axis, or 90° right rotated, item difficulty was significantly enhanced.

A joint analysis of the A- and C-elements revealed following significant impact: Normally oriented, 90° left rotated, and x-reflected A-elements made items easier. Further, 90° left and 180° rotated C-elements made items easier, too. Among the features increasing item difficulty 180° rotated A-elements contributed to item difficulty the most. Digits chosen as elements for A and C predicted items of minor difficulty whereas letters chosen as A- and C-elements predicted items of high difficulty.

Thus, the impact of element orientation was not always consistent and provided no clear structure. Future research should analyze the impact of single digits and letters applied. Transformation of the element *K*, for example, might have a different difficulty than the transformation of the element *B* or the digit 3. Knowing the impact of single elements and their interaction with element orientation will certainly clarify the above stated findings regarding the impact of element orientation on item difficulty.

Summarized, the pre-hypothesized task structure could be validated. To construct items of enhanced difficulty only transformations and elements positively related to item difficulty should be allowed for, if one aims to construct difficult test items. Hence, it should be considered to apply only such transformations and elements and ignore those that ease items. Providing the participants with information on the transformational rules ap-

plied in the test, enables one to practice these cognitive operations required to correctly solve the items, and enhances the establishment of equal a priori test conditions for all subjects. Only then, differences in test scores can be attributed to real differences in ability and not to differences in practice and test taking experience. The predictive validity of the test for school performance was confirmed but can be improved since correlations were lower than expected.

Critical reflection can be allowed on two further issues: The culture-fairness of the test and the sample comparison of the instruction and non-instruction group. Culture-fair tests aim to capture the subject's ability independent of his or her cultural background. Thus emphasis on knowledge depending on education and cultural background should not be made. However, this study's transformations sequence plus and sequence minus, applied to digits and letters as elements, require the knowledge of the order of numbers and knowledge of the alphabet. Therefore, strictly speaking, this cognitive demand does not correspond to the characteristic of pure culture-fair tests, since participants of different backgrounds might not meet these requirements or do not provide such knowledge. The test can therefore be considered a merely restricted culture-fair reasoning test because of the constraint imposed by the sequence transformations.

The sample comparison conducted only alluded to some of the subjects, and did not include all participants because not all test takers were required to take part in the CFT-20 R test. Thus the statement that the instruction group and the non-instruction group did not differ in the CFT-20 R test

score is inferred from subsamples that provided these measures.

## 6.1 Future Prospects

Effects of learning and training, and the stability of the construct reasoning ability measured by analogies should be examined by retesting samples several times and measure their increase, decrease or invariance of proficiency. To reliably predict performance across time, the test-retest reliability of these tasks have to be explored. Retest-reliability indicates the extent to which test scores can be generalized over a period of time. The greater the reliability, the less susceptible the scores are to random daily variations and fluctuations in subjects' constitution and testing environment.

Another important issue refers to computerized adaptive testing (CAT). IRT-scaled items are requirement for computer adaptive testing since tests are tailored to the ability level of each test taker. Examinees thus receive different sets of items to estimate their proficiency level and IRT allows computing and comparing scores due to its postulate of invariance. Local independence is therefore required to ensure sample-independence of item and ability estimation which in turn is fundamental for adaptive testing.

To adapt the test to the examinee's ability level, item difficulty estimated by the IRT-model must be known to choose items of a difficulty that matches the proficiency level of the test taker. The Rasch model offers good characteristics for CAT as it assumes a continuous scale for item difficulty ( $\sigma$ ) and ability ( $\theta$ ). Selecting items adequate in difficulty, information

can be maximized while minimizing the standard error of measurement (Gershon, 2005). Thus the difference between item difficulty and ability needs to be minimized in order to obtain highest measurement precision.

Advantages of CAT are therefore maximized precision of measurement or ability estimates due to item selection on basis of its information, whilst applying tests that are shorter than most paper and pencil tests. Test motivation can therefore be increased. However, CAT requires large numbers of items, pre-calibrated according to IRT in order to provide sufficient items for each ability level. Large item pools are supposed to ensure that the same items are not overused for same ability levels but different items of similar difficulty levels can be applied. Exposure control algorithms can be applied to ensure test security. Automatic item generation is therefore needed to allocate large item pools necessary for CAT and constitutes an important topic of further research.

Having mentioned computerized testing another related research topic referring to information-processing theories of reasoning tasks and working memory in particular arises. Computerized test versions enable one to additionally measure working memory capacity and its role within analogy and reasoning tasks. The A, B and C terms could be sequentially presented and then masked or faded out and thus the D term would have to be inferred from memory. Participants could then choose either among distractors displayed in the item or self-construct their response by means of drag & drop of presented elements. This test format would enhance further research on information-processing theories of reasoning, the essence of human intelligence.

## References

- Alexander, P. A., Haensly, P. A., Crimmins-Jeanes, M., & White, C. S. (1986). Analogy training: A study of the effects on verbal reasoning. *Journal of Educational Research, 80*, 77-80.
- Amelang, M., & Bartussek, D. (1997). *Differentielle Psychologie und Persönlichkeitsforschung [Differential psychology and personality research]* (4th ed.). Stuttgart, Ger: Kohlhammer.
- Anastasi, A. (1982). *Psychological testing* (2nd ed.). New York: Macmillan.
- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.
- Baron-Boldt, J., Funke, U., & Schuler, H. (1989). Prognostische Validität von Schulnoten. Eine Metaanalyse der Prognose des Studien- und Ausbildungserfolgs [Prognostic validity of school grades: A meta-analysis]. In R. S. Jäger, R. Horn, & K. Ingenkamp (Eds.), *Tests und Trends* (p. 11-39). Weinheim, Ger: Beltz.
- Baron-Boldt, J., Schuler, H., & Funke, U. (1988). Prädiktive Validität von Schulabschlussnoten: Eine Metaanalyse [Predictive validity of school



## References

---

- grades: A meta-analysis]. *Zeitschrift für Pädagogische Psychologie*, 2, 79-90.
- Bergmann, C., & Eder, F. (1999). *Allgemeiner Interessen Struktur Test [General interests structure test]*. Göttingen, Ger: Hogrefe.
- Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence*, 8, 205-238.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (p. 395-479). Reading, Mass: Addison-Wesley.
- Bisanz, J., Bisanz, G. L., & LeVevre, J.-A. (1984). Interpretation of instructions: A source of individual differences in analogical reasoning. *Intelligence*, 8, 161-177.
- Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) [NEO-five-factor inventory]*. Göttingen, Ger: Hogrefe.
- Boulanger, F. D. (1981). Ability and science learning: A quantitative synthesis. *Journal of Research in Science Teaching*, 18, 113-121.
- Briel, J. B., O'Neill, K., & Scheuneman, J. D. (1993). *GRE technical manual*. Princeton, NJ: Educational Testing Service.
- Bundesen, C., & Larsen, A. (1975). A visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 214-220.
- Carroll, J. (1989). Factor analysis since Spearman: Where do we stand? What do we know? In R. Kanfer, P. L. Ackerman, & R. Cudeck

## References

---

- (Eds.), *Abilities, motivation, and methodology* (p. 43-67). Hillsdale, NJ: Erlbaum.
- Carroll, J. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, MA: Cambridge University Press.
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.
- Costa, P. T., & McCrae, R. R. (1992). *NEO-PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.
- Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13-21.
- Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175-186.
- Evans, T. G. (1968). A program for the solution of geometric analogy intelligence test questions. In M. Minsky (Ed.), *Semantic information processing* (p. 271-353). Cambridge, Mass: MIT Press.
- Facaoaru, C. (1985). *Kreativität in Wissenschaft und Technik. Operationalisierung von Problemlösefähigkeiten und kognitiven Stilen [Creativity in science and technology]*. Bern, Switzerland: Huber.
- Fischer, G. H. (1972). Conditional maximum-likelihood estimations of item parameters for a linear logistic test model. *Research Bulletin, 9*. Vienna: Department of Psychology, University of Vienna.
- Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.
- Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling. Handbook of the usage of LPCM-Win 1.0*. Groningen: PROGAMMA.

## References

---

- Fleming, M. L., & Malone, M. R. (1983). The relationship of student characteristics and student performance in science as viewed by meta-analysis research. *Journal of Research in Science Teaching*, 20, 481-495.
- Freund, P. A. (2008). *Practice and training effects on measures of cognitive abilities*. Berlin, Ger: Weißensee.
- Freund, P. A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, 32, 195-210.
- Freund, P. A., Holling, H., & Preckel, F. (2007). A multivariate, multilevel analysis of the relationship between cognitive abilities and scholastic achievement. *Journal of Individual Differences*, 28, 188-197.
- Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, 6, 109-127.
- Goos, P. (2002). *The optimal design of blocked and split-plot experiments*. New York: Springer.
- Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, 8, 179-203.
- Gustafsson, J.-E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (p. 35-71). Hillsdale, NJ: Erlbaum.
- Gustafsson, J.-E., Lindström, B., & Björck-Åkesson, E. (1981). A general model for the organization of cognitive abilities. Report from the Department of Education, University of Göteborg.
- Hattie, J. A., & Hansford, B. C. (1982). *Personality and intelligence: What relationship with achievement?* Paper presented at the Annual Conference

- of Australian Association for Research in Education, Brisbane.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.
- Heller, K. A., Gaedike, A.-K., & Weinläder, H. (1985). *Kognitiver Fähigkeits-Test für 4. bis 13. Klassen (KFT 4-13+) [Cognitive ability test for class level 4 to 13]*. Weinheim, Ger: Beltz.
- Heller, K. A., Kratzmeier, H., & Lengfelder, A. (1998). *Matrizen-Test-Manual, Band 2. Ein Handbuch mit deutschen Normen zu den Advanced Progressive Matrices von Raven [Manual with German norms for the Raven's APM]*. Göttingen, Ger: Beltz-Test.
- Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior, 8*, 176-184.
- Holling, H., Preckel, F., & Vock, M. (2004). *Intelligenzdiagnostik [Diagnosing intelligence]*. Göttingen, Ger: Hogrefe.
- 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.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 189-212). New York: Springer.
- Jensen, A. R. (1980). *Bias in mental testing*. New York: The Free Press.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT:

Praeger.

- Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multimodal classification of intelligence performance. Experimentally controlled development of a descriptive intelligence structure model]. *Diagnostica*, 28, 195-226.
- Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., et al. (2006). *Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB) [Berlin structure of intelligence test for youth: Assessment of talent and giftedness]*. Göttingen, Ger: Hogrefe.
- Johnson, D. M. (1962). Serial analysis of verbal analogy problems. *Journal of Educational Psychology*, 53, 86-88.
- Klauer, K. J. (2001). *Handbuch Kognitives Training [Handbook cognitive training]* (2nd ed.). Göttingen, Ger: Hogrefe.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, 86, 148-161.
- LeFevre, J.-A., & Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction*, 3, 1-30.
- Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse [test construction and test analysis]* (6th ed.). Weinheim, Ger: Beltz.
- Lindley, R. H., Smith, W. R., & Thomas, J. T. (1988). The relationship between speed of information processing as measured by timed paper-and-pencil tests and psychometric intelligence. *Intelligence*, 12, 17-25.

## References

---

- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Luo, D., Thompson, L. A., & Detterman, D. K. (2003). The causal factor underlying the correlation between psychometric g and scholastic performance. *Intelligence, 31*, 67-83.
- Martin-Löf, P. (1973). *Statistica modeller: Anteckningar från seminarier lasåret 1969-70 utarbetade av rolf sundberg*. [Statistical models: Notes from seminars 1969-1970, prepared by Rolf Sundberg]. Stockholm: Institutet för Försäkringsmatematik och Matematisk Statistik vid Stockholms Universitet.
- Miller, W. S. (1960). *Technical manual for the Miller Analogies Test*. New York: The Psychological Corporation.
- Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology, 12*, 252-284.
- Novick, L. R., & Tversky, B. (1987). Cognitive constraints on ordering operations: The case of geometric analogies. *Journal of Experimental Psychology: General, 116*, 50-67.
- Oswald, W., & Roth, E. (1978). *Der Zahlen-Verbindungs-Test (ZVT) [The number combination test]*. Göttingen, Ger: Hogrefe.
- Otis, A. S. (1918). An absolute pointscale for the group measurement of intelligence. *Journal of Educational Psychology, 9*, 239-261, 333-348.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Raven, J. C. (1936). *Standard Progressive Matrices, Sets A, B, C, D, E*. London: Lewis.

## References

---

- Raven, J. C. (1958). *Advanced progressive matrices*. London: Lewis.
- Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement, 26*, 271-285.
- Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence, 32*, 573-589.
- Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2nd ed.). Bern: Huber.
- Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement, 18*, 171-182.
- Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology, 5*, 1-28.
- Schoppe, K. (1975). *Verbaler Kreativitätstest (VKT) [Verbal creativity test]*. Göttingen, Ger: Hogrefe.
- Shepard, R. N. (1975). Form, formation, and transformation of internal representations. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (p. 87-122). Hillsdale, NJ: Erlbaum.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293.
- Spearman, C. (1923). *The nature of intelligence and the principles of cognition*. London: Macmillan.
- Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan.
- Süß, H.-M. (2001). Prädiktive Validität der Intelligenz im schulischen und außerschulischen Bereich [Predictive validity of intelligence in school-related and external domains]. In E. Stern & J. Guthke (Eds.),

## References

---

- Perspektiven der Intelligenzforschung [Perspectives on intelligence research]* (p. 109-136). Lengerich, Ger: Pabst.
- Süß, H.-M., & Beauducel, A. (2005). Faceted models of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Understanding and measuring intelligence* (p. 313-332). London: Sage.
- Steinkamp, M. W., & Maehr, M. L. (1983). Affect, ability, and science achievement: A quantitative synthesis of correlational research. *Review of Educational Research*, 53, 369-396.
- Sternberg, R. J. (1977a). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.
- Sternberg, R. J. (1977b). Component processes in analogical reasoning. *Psychological Review*, 84, 353-378.
- Sternberg, R. J. (1986). Toward a unified theory of human reasoning. *Intelligence*, 10, 281-314.
- Strand, S. (2006). Comparing the predictive validity of reasoning tests and national end of Key Stage 2 tests: Which tests are the 'best'? *British Educational Research Journal*, 32, 209-225.
- Thomas, S., & Mortimore, P. (1996). Comparison of value-added models for secondary school effectiveness. *Research Papers in Education*, 11, 5-33.
- Thorndike, R. L., & Hagen, E. (1974). *Cognitive abilities test*. New York: Houghton-Mifflin.
- Thorndike, R. L., Hagen, E., & France, N. (1986). *Cognitive Abilities Test Second Edition: Administration Manual*. Windsor: nferNelson.



## References

---

- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- Undheim, J. O. (1976). Ability structure in 10–11-year-old children and the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 68*, 411-423.
- 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 den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369-386.
- von Davier, M. (2001). *Winmira 2001 [computer software]*. Groningen, NL: ProGamma.
- Weiß, R. H. (2006). *Grundintelligenztest Skala 2–Revision (CFT 20–R) [Reasoning Test Scale 2–Revision ]*. Göttingen, Ger: Hogrefe.
- White, C. S., & Alexander, P. A. (1986). Effects of training on four-year-old's ability to solve geometric analogy problems. *Cognition and Instruction, 3*, 261-268.
- White, C. S., & Caropreso, E. J. (1989). Training in analogical reasoning processes: Effects on low socioeconomic status preschool children. *Journal of Educational Research, 83*, 112-118.
- Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement, 5*, 383-397.

## References

---

- Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Understanding and measuring intelligence* (p. 373-392). London: Sage.
- Woodcock, R. W. (1973). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.
- Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [computer software]*. Chicago, IL: Scientific Software International.



## **Tables and Figures**

Table A.1: LLTM with Random Item Effects: Instruction Group (SAS)

Operation	Estimate	SE	t-value	p
constant	1.80	.31	5.76	.00
sp	.22	.24	.89	.38
sm	.21	.24	.88	.38
rr	-1.49	.34	-4.35	.00
rl	-1.46	.33	-4.43	.00
r180	-.84	.33	-2.55	.01
rfx	-2.24	.34	-6.55	.00
rfy	-1.78	.36	-4.96	.00
sqp	-1.03	.25	-4.18	.00
sqm	-.71	.25	-2.87	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

Table A.2: LLTM: Instruction Group (SAS)

Operation	Estimate	SE	t-value	p
sp	-.20	.06	-3.51	.00
sm	-.20	.05	-3.68	.00
rr	1.30	.08	16.09	.00
rl	1.29	.08	16.41	.00
r180	.66	.08	8.48	.00
rfx	2.04	.08	24.30	.00
rfy	1.60	.08	18.87	.00
sqp	.97	.06	16.86	.00
sqm	.61	.06	10.71	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

Table A.3: LLTM with Random Item Effects: Non-Instruction Group (SAS)

Operation	Estimate	SE	t-value	p
constant	1.19	.30	3.98	.00
sp	.45	.23	1.93	.05
sm	.38	.23	1.69	.09
rr	-1.24	.33	-3.78	.00
rl	-1.22	.32	-3.88	.00
r180	-.57	.32	-1.82	.07
rfx	-1.97	.33	-5.99	.00
rfy	-1.39	.35	-4.05	.00
sqp	-1.04	.24	-4.42	.00
sqm	-.70	.24	-2.96	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

Table A.4: LLTM: Non-Instruction Group (SAS)

Operation	Estimate	SE	t-value	p
sp	-.40	.06	-7.02	.00
sm	-.38	.06	-6.83	.00
rr	1.07	.08	13.50	.00
rl	1.05	.08	13.64	.00
r180	.42	.08	5.47	.00
rfx	1.79	.08	21.31	.00
rfy	1.24	.08	14.73	.00
sqp	.96	.06	16.47	.00
sqm	.59	.06	10.33	.00

*Note.* sp = size plus, sm = size minus, rr = rotation 90° right, rl = rotation 90° left, r180 = rotation 180°, rfx = reflection about x-axis, rfy = reflection about y-axis, sqm = sequence minus, sqp = sequence plus

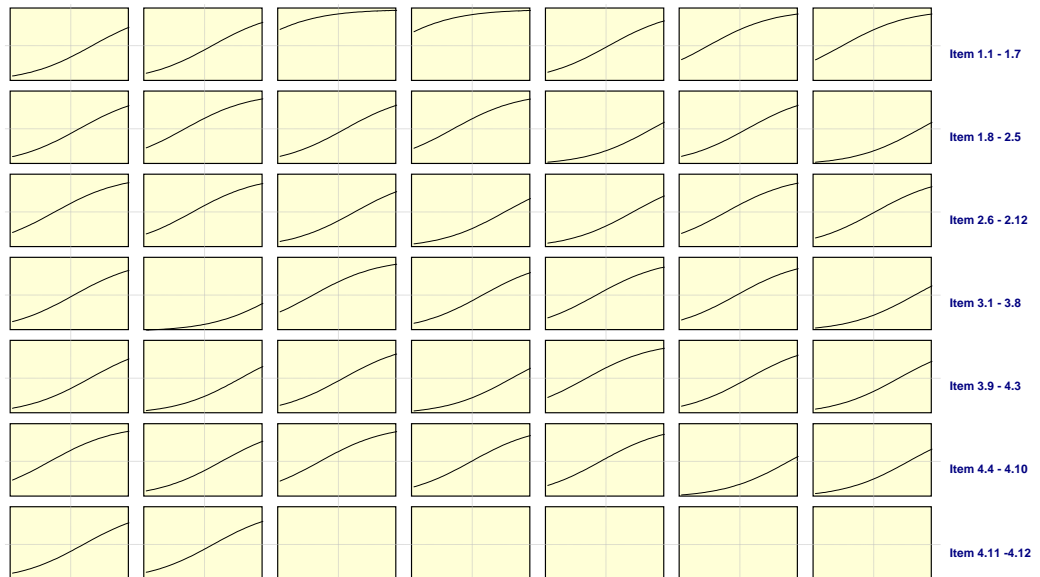


Figure A.1: Item Characteristic Curves of the Figural Analogy Test Items

# B

## **Test Materials**

The German instructions of the test versions A and B are presented in the following sections.

## Testversion A

### Figurale Analogien






Sie sehen in den folgenden Aufgaben zwei Elemente vor dem Gleichheitszeichen. Diese zwei Elemente stehen in einer bestimmten Beziehung zueinander. Hinter dem Gleichheitszeichen stehen ein anderes Element und ein Fragezeichen. Rechts daneben finden Sie verschiedene Antwortalternativen.

Bitte wählen Sie die Antwortalternative aus, die Ihrer Meinung nach anstelle des Fragezeichens einzusetzen ist, damit zwischen den Elementen hinter dem Gleichheitszeichen dieselbe Gesetzmäßigkeit besteht wie zwischen den Elementen vor dem Gleichheitszeichen.

Aufgabe ist es also, zuerst die Regeln zu erkennen, nach denen sich das erste Element zum zweiten verhält. Diese Regeln sollen dann zur Ermittlung des vierten Elementes angewandt werden.

Jede Antwortalternative ist mit einem Buchstaben gekennzeichnet. Bitte streichen Sie den Buchstaben unter der richtigen Lösung durch.

In dem folgenden Beispiel ist die Antwortalternative *b* richtig.

$3:3 = 7:?$						Keine Lösung richtig
	a	b	c	d	e	f

Sollten sie bei manchen Aufgaben der Meinung sein, dass die richtige Lösung nicht unter den Antwortalternativen ist, so streichen sie bitte den Buchstaben unter dem Kästchen "Keine Lösung richtig" durch.



Im Folgenden sind vier Tests mit unterschiedlich vielen und unterschiedlich schwierigen Aufgaben zu bearbeiten. Für die Bearbeitung der einzelnen Tests steht Ihnen nur begrenzte Zeit zur Verfügung. Jeder Test besteht aus zwei Seiten.

Bitte blättern sie innerhalb eines Tests selbständig um und warten Sie nach jedem Test mit dem Umblättern bis Sie dazu aufgefordert werden!

Der Versuchsleiter gibt vor jedem Test das Startsignal und nach Ablauf der Testzeit das Stoppsignal. Notizen sind nicht erlaubt. Sollten Sie noch Fragen zum Test haben, dann stellen Sie diese bitte jetzt!






## **Testversion B**


### Figurale Analogien






Sie sehen in den folgenden Aufgaben zwei Elemente vor dem Gleichheitszeichen. Diese zwei Elemente stehen in einer bestimmten Beziehung zueinander. Hinter dem Gleichheitszeichen stehen ein anderes Element und ein Fragezeichen. Rechts daneben finden Sie verschiedene Antwortalternativen. Bitte wählen Sie die Antwortalternative aus, die Ihrer Meinung nach anstelle des Fragezeichens einzusetzen ist, damit zwischen den Elementen hinter dem Gleichheitszeichen dieselbe Gesetzmäßigkeit besteht wie zwischen den Elementen vor dem Gleichheitszeichen. Aufgabe ist es also, zuerst die Regeln zu erkennen, nach denen sich das erste Element zum zweiten verhält. Diese Regeln sollen dann zur Ermittlung des vierten Elementes angewandt werden. Jede Antwortalternative ist mit einem Buchstaben gekennzeichnet. Bitte streichen Sie den Buchstaben unter der richtigen Lösung durch. Insgesamt gibt es neun Regeln, die die Relation zwischen den Elementen bestimmen. Im Folgenden wird jede Regel anhand von 2 Beispielen erklärt. Sollten sie bei manchen Aufgaben der Meinung sein, dass die richtige Lösung nicht unter den Antwortalternativen ist, so streichen sie bitte den Buchstaben unter dem Kästchen "Keine Lösung richtig" durch.

Regel 1: Vergrößerung

3:3 = 7:?






a  b  c  d  e  f Keine Lösung richtig




R:R =  :?






a  b  c  d  e  f Keine Lösung richtig

Regel 2: Verkleinerung

5:5 = 6:?






a  b  c  d  e  f Keine Lösung richtig




 :  =  :?






a  b  c  d  e  f Keine Lösung richtig

Regel 3: 90° Rotation rechts

2:2 = 7:?






a  b  c  d  e  f Keine Lösung richtig


 :  =  :?






a  b  c  d  e  f Keine Lösung richtig

Regel 4: 90° Rotation links

4:4 = J:?

a  b  c  d  e  f Keine Lösung richtig

2:2 =  :?

a  b  c  d  e  f Keine Lösung richtig

Regel 5: 180° Rotation

R : Y = L : ?

a b c d e f

Keine Lösung richtig

L : 1 = 7 : ?

a b c d e f

Keine Lösung richtig

Regel 6: Spiegelung an der x-Achse

G : @ = 7 : ?

a b c d e f

Keine Lösung richtig

U : Q = U : ?

a b c d e f

Keine Lösung richtig

Regel 7: Spiegelung an der y-Achse

1 : 3 = 2 : ?

a b c d e f

Keine Lösung richtig

3 : E = F : ?

a b c d e f

Keine Lösung richtig

Regel 8: Addition Zahlen/Buchstaben

1 : 3 = B : ?

a b c d e f






Keine Lösung richtig






M : Σ = S : ?

a b c d e f

Keine Lösung richtig

Regel 9: Subtraktion Zahlen/Buchstabe

$C : A = 5 : ?$						Keine Lösung richtig
	a	b	c	d	e	f

$F : E = \text{Ⓞ} : ?$						Keine Lösung richtig
	a	b	c	d	e	f

Im Folgenden sind vier Tests mit unterschiedlich vielen und unterschiedlich schwierigen Aufgaben zu bearbeiten. Für die Bearbeitung der einzelnen Tests steht Ihnen nur begrenzte Zeit zur Verfügung. Jeder Test besteht aus zwei Seiten.

Bitte blättern sie innerhalb eines Tests selbständig um und warten Sie nach jedem Test mit dem Umblättern bis Sie dazu aufgefordert werden!

Der Versuchsleiter gibt vor jedem Test das Startsignal und nach Ablauf der Testzeit das Stoppsignal. Notizen sind nicht erlaubt. Sollten Sie noch Fragen zum Test haben, dann stellen Sie diese bitte jetzt!



## Zusammenfassung

Schlussfolgerndes Denken oder *Reasoning* ist in zahlreichen alltäglichen Situationen unabdingbar. In der Wissenschaft spielt Reasoning in vielen einflussreichen Intelligenzmodellen eine große, oft zentrale Rolle (z.B. Thurstone, 1938; Cattell, 1971; Carroll, 1993). Analogieaufgaben gelten im Rahmen der Erfassung der analytischen Intelligenz als prototypische Aufgabenart und wurden zum Gegenstand für die Testentwicklung der vorliegenden Arbeit bestimmt. Sternbergs Modell der Informationsverarbeitung (Sternberg 1977a, 1977b) wurde als zugrundeliegendes Prozessmodell für das Lösen von Analogieaufgaben ausgewählt und ausführlich referiert.

Des Weiteren konnte die prädiktive Validität von Intelligenzleistungen im Allgemeinen und von Analogieaufgaben im Besonderen bezüglich akademischer Erfolgskriterien aufgezeigt werden, auch um Nutzen und Implikationen für die Praxis zu verdeutlichen.

Im Rahmen der Analyse kognitiver Strukturen in Testaufgaben wurden Vorteile der regelgeleiteten Testkonstruktion aufgeführt. Voraussetzung der regelgeleiteten Testentwicklung ist die Spezifikation der Aufgabenparameter, also der kognitiven Operationen, die am Lösungsprozess der Aufgabe beteiligt sind. Items werden somit nicht willkürlich konstruiert, sondern werden auf Basis einer zugrundeliegenden Rationalen und prä-experimentellen Hypothesen zur Aufgabenstruktur generiert. Durch Validierung dieser kognitiven Struktur und Erklärung der Aufgabenschwierigkeit durch die Aufgabenparameter entstehen substantielle psychometrische Vorteile. Beispielsweise können durch Generierung von Items mit adäquaten Schwierigkeiten hinsichtlich des Fähigkeitslevels der Testperson verminderte Stan-

dardschätzfehler realisiert werden.

Im Rahmen der vorliegenden Arbeit wurde ein regelgeleiteter Test zur Erfassung des schlussfolgernden Denkens entwickelt. Da es sich um ein neu konstruiertes Testverfahren handelt, wurde der Test zuerst in mehreren Vorstudien hinsichtlich seiner psychometrischen Eigenschaften untersucht. Außerdem verfolgten die Voruntersuchungen das Ziel, erste Aussagen zum Einfluss der im Test verwandten Regeln auf die Aufgabenschwierigkeit zu machen. In der anschließenden Hauptuntersuchung wurden 484 Schülerinnen und Schüler der gymnasialen Oberstufe getestet.

Der figurale Analogietest bestand aus vier Untertests, die auf effizienten Designs beruhten. Die resultierenden Daten der Testitems wurden mit Analyseverfahren der klassischen Testtheorie und der probabilistischen Testtheorie ausgewertet. Im Rahmen der klassischen Testtheorie wurden zunächst deskriptive Statistiken der Items dargestellt und Reliabilitätsanalysen durchgeführt. Die Validität des Tests wurde über den Zusammenhang zum Intelligenztest CFT-20 R (Weiß, 2006) bestimmt. Außerdem wurde zur Prüfung der prädiktiven Validität für Schulerfolg der Zusammenhang des Tests mit verschiedenen Schulnoten berechnet. Anschließend wurden die Itemparameter im Rahmen der probabilistischen Testmodelle geschätzt. Die vergleichende Passung der Modelle auf den Datensatz und spezifische Modellgeltungstests für das eindimensionale Rasch Modell wurden durchgeführt.

Die Überprüfung des Konstruktionsansatzes wurde nicht nur mittels des Linear Logistischen Testmodells vorgenommen, sondern auch durch Anwendung des Linear Logistischen Testmodells mit zufälligen Itemeffekten,



um realistischere Annahmen zu gewährleisten. Die prä-experimentell angenommene Aufgabenstruktur konnte so validiert werden, da die im Test angewandten Regeln zur Erklärung der Itemschwierigkeiten herangezogen werden konnten. In der Diskussion wurden diese Ergebnisse kritisch reflektiert und diskutiert. Ein Ausblick auf die Möglichkeit des Computer adaptiven Testens und automatischer Itemgenerierung bildete den Abschluss der Arbeit.



