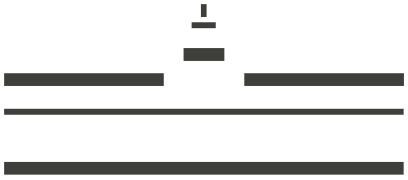


Westfälische
Wilhelms-Universität
Münster

Psychologie

Rule-based item construction: Analysis with and comparison of linear logistic test models and cognitive diagnostic models with two item types

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

vorgelegt von

Nina Zeuch
aus Stadthagen

Münster, 2010

Tag der mündlichen Prüfung: 20. April 2011
Dekan: Prof. Dr. Christian Pietsch
Erstgutachter: Prof. Dr. Heinz Dieter Holling
Zweitgutachter: PD Dr. Günther Gediga

Contents

List of Tables

List of Figures

Danksagung

Bei mehreren wichtigen Personen, die einen nicht unerheblichen Anteil zum Gelingen der Arbeit beigetragen haben, möchte ich mich an dieser Stelle in aller Form bedanken.

Vor allem danke ich meinem Doktorvater Herrn Prof. Dr. Heinz Holling, der mir immer ein wertvoller Ratgeber war und mir erst ermöglicht sowie mich ermutigt hat, die vorliegende Arbeit zu verfassen. Große Anteile meines methodischen Interesses und meiner wissenschaftlichen Arbeitsweise habe ich ihm zu verdanken.

Herrn PD Dr. Gediga danke ich sehr für seine Hilfestellungen vor allem in methodischen und softwaretechnischen Fragen und auch für seine Bereitschaft, als Zweitgutachter der Arbeit tätig zu werden. Des Weiteren möchte ich Herrn Dr. Jörg-Tobias Kuhn und Herrn Dipl.-Psych. Jonas Bertling für ihre durchweg hilfreichen und wertvollen Ratschläge und Inspirationen danken.

Nicht zuletzt gilt mein Dank auch den zahlreichen Praktikanten und studentischen Hilfskräften, die stets hervorragende Arbeit bei Datenerhebungen, Testauswertung und Dateneingabe geleistet haben.

Mein besonderer Dank gilt außerdem meinem Mann und meinen Eltern, die mich stets ermutigt und unterstützt haben und mir an jedem einzelnen Tag eine wundervolle Hilfe waren.

Nina Zeuch

1 Summary

Item construction for diagnostic and research issues requires careful theoretical preparation, accurate item writing techniques, empirical evaluation and application of adequate statistical models. During the last years, important developments have been made in item construction, partly due to more serious requirements through high-stakes testing, large-scale test settings, computerized and adaptive testing and internet based testing, and partly due to new statistical developments.

Rule-based item construction constitutes an exact, safe and sophisticated way to generate valid, reliable and verifiable items based on difficulty-generating basic parameters for tests in many areas. Furthermore, rule-based item construction can be an important help for automatic item generation, item cloning and adaptive testing. To investigate item quality and adequacy of theoretical basis as well as its operationalization in test items, linear logistic test models (LLTMs) and cognitive diagnostic models (CDMs) can be applied.

The current work focuses on demonstration and evaluation of rule-based item construction and application as well as comparison of LLTMs and CDMs as statistical analysis methods. Since both model classes lend themselves to analysis of rule-based constructed items while implying totally different statistical concepts, a direct comparison of these models including empirical application and interpretational analogies and contrasts promises practical and theoretical proceedings regarding model choice and item construction.

Several sets of two different item types (figural reasoning items and mathematical word problems) are constructed rule-based and tested with three German school student samples. Additionally, an item cloning approach as well as item construction for a longitudinal study and application of the noted statistical models to these special requirements are demonstrated. Results show Rasch scalability of items, confirm the importance of the chosen basic parameter sets and demonstrate precise item construction and analysis processes.

It is shown how LLTM and its variants can contribute substantial insights into cognitive solution processes and composition of item difficulty in relational reasoning

and mathematical word problems and also for item cloning and longitudinal data. However, CDM application detects severe modeling problems and misfit. Application hints regarding test item construction as well as statistical model application and interpretation of results for practitioners and researchers are pointed out.

It can be concluded that LLTMs provide great means to analyze rule-based constructed items which are flexible and versatile instruments with well documented software implementations. CDMs turn up to be more restrictive than LLTMs and not adequate for the current item types. However, CDMs should not be neglected in research and practice as they provide useful insights into item construction and examinee behavior given the model assumptions are met.

2 Zusammenfassung

Aufgabenkonstruktion für diagnostische und Forschungszwecke erfordert sorgfältige theoretische Vorbereitung, exakte Aufgabengenerierungstechniken, empirische Überprüfung und Anwendung geeigneter statistischer Modelle. Während der letzten Jahre gab es wichtige Fortschritte in der Aufgabenkonstruktion, teilweise durch die Entwicklung neuer statistischer Analysemethoden, teilweise auch durch höhere Anforderungen in groß angelegten Studien, aber auch durch computerbasiertes und adaptives sowie internetbasiertes Testen.

Regelgeleitete Aufgabenkonstruktion bietet eine genaue, sichere und technisch ausgefeilte Methode, valide, reliable und überprüfbare Aufgaben aus schwierigkeitsgenerierenden Basisparametern für Tests in vielen verschiedenen Bereichen zu erstellen. Des Weiteren kann regelgeleitete Aufgabenkonstruktion eine wichtige Ausgangsbasis für automatische Aufgabengenerierung, Aufgabencloning und adaptives Testen bilden. Linear-logistische Testmodelle (LLTMs) sowie kognitive Diagnosemodelle (CDMs) können dabei zur Überprüfung der Aufgabenqualität sowie der Angemessenheit der theoretischen Grundlagen und ihrer Operationalisierung durch die Testaufgaben herangezogen werden.

Die vorliegende Arbeit behandelt die Darstellung und Evaluierung regelgeleiteter Aufgabenkonstruktion sowie die Anwendung und den Vergleich von LLTMs und CDMs als statistische Analysemethoden. Da beide Modellklassen sich für die Analyse regelgeleitet konstruierter Aufgaben anbieten, aber auf ganz unterschiedlichen statistischen Grundannahmen basieren, verspricht ein direkter Vergleich dieser Modelle, der Vergleich ihrer empirischen Anwendung sowie Interpretationsparallelen und -unterschiede praktisch und theoretisch bedeutsame Erkenntnisse bezüglich der Modellwahl und der Aufgabenkonstruktion.

Mehrere Sets unterschiedlicher Aufgabentypen (figurale Reasoning-Aufgaben und mathematische Textaufgaben) werden regelgeleitet konstruiert und an drei Stichproben deutscher OberstufenschülerInnen getestet. Außerdem werden ein Aufgabencloning-Ansatz und die Aufgabenkonstruktion für eine longitudinale Studie sowie die Anwendung der genannten statistischen Modelle auf diese speziellen Erfordernisse dargestellt. Die Ergebnisse zeigen Rasch-Skalierbarkeit der

Aufgaben, bestätigen die Bedeutung der gewählten Basisparametersets und demonstrieren einen präzisen Aufgabenkonstruktions- und Analyseprozess.

Es wird gezeigt, wie LLTM-Varianten wichtige Einblicke in kognitive Lösungsprozesse und in die Zusammensetzung der Aufgabenschwierigkeit im Bereich des relationalen Reasoning und mathematischer Textaufgaben genauso wie für Aufgabencloning und longitudinale Datenstrukturen liefern können. Dagegen zeigen sich in der CDM-Anwendung ernste Modellierungsprobleme und Unangemessenheit des Ansatzes für die vorliegenden Aufgabenbeispiele. Anwendungshinweise bezüglich der Aufgabenkonstruktion, der statistischen Modelle und der Interpretation der Ergebnisse für Anwender und Forscher werden herausgestellt.

Zusammenfassend bieten LLTMs hervorragende Möglichkeiten, regelgeleitet konstruierte Aufgaben zu analysieren. LLTMs sind sehr flexible und vielseitige Instrumente mit gut dokumentierten Softwareumsetzungen. CDMs, zumindest das hier verwendete DINA-Modell, erscheinen aber restriktiver als das LLTM und sind offensichtlich für die hier verwendeten Aufgabentypen nicht anwendbar. Jedoch sollten CDMs in Forschung und Anwendung nicht vernachlässigt werden, da sie ebenfalls wertvolle Einblicke in die Aufgabenkonstruktion und das Probandenverhalten bieten, solange die Modellannahmen zulässig sind.

3 Introduction

Testing of intelligence and competencies has a long history (cf. Spearman, 1904; McClelland, 1973). Today, application of ability and achievement tests is almost obligatory in selection settings for job applicants or university as well as in school. One can distinguish between tests of (fluid or crystallized) intelligence and tests of competencies in single content areas. Whereas intelligence tests often consist of broad and general item content as figural, verbal or numerical material, competence testing concentrates on relatively narrow content areas and is often preceded by analyses of this area concerning typical desired knowledge and abilities (one example are school exams). While intelligence tests aim at assessment of more general (cognitive) abilities to process new experiences and draw helpful conclusions, or to handle new situations adaptively, competence tests focus on knowledge necessary to fulfill special tasks and job demands.

Recent developments in the testing industry show a concentration on high-stakes testing and large-scale test settings with hundreds or thousands of examinees (cf. US university admissions tests as the *Graduate Record Examination* (GRE) or the *Test of English as a Foreign Language* (TOEFL), and the *Programme for International Student Assessment* (PISA)), requiring large amounts of test items with robustness against recognition and faking. Moreover, tests should be as short and informative (of the examinees' abilities) as possible at the same time. These developments require major changes in test development and item characteristics. Not only is a sound cognitive theory necessary to lay the basis for item and test construction, but also this theory has to be verifiable by inspection of empirical results in the sense that carefully constructed items allow for valid and reliable conclusions about the test taker's abilities. In order to improve item construction with regard to these new needs in testing, several new developments have been made. On the one hand, techniques of rule-based item generation and automatic item generation (AIG) help researchers to construct items accurately to meet the requirements of theory and practice, that is, controlling for item characteristics and validity as well as reliability. On the other hand, there is a variety of statistical models that allow conclusions about item characteristics and quality of the construction process as

well as test takers and their abilities along with possible gaps in competencies or intelligence.

The current developments include IRT (item response theory) modeling like the linear logistic test model (LLTM) and cognitive diagnosis models (CDMs). These models provide powerful instruments for analysis of item and test taker properties and gain more and more scientific interest. While the LLTM and its variations focus on the item side, that is on basic parameters that influence item difficulty while assuming unidimensional concepts, CDMs focus on the test taker and so-called ability classes with certain solution patterns and are based on mixture assumptions. These two perspectives differ conceptually but help to investigate the structure of the analyzed test domain with the same aim of improving testing technology.

The current work focuses on rule-based item construction in two domains, namely the Latin Square Task (LST) as a measure of fluid intelligence, relational complexity and working memory (first developed by Birney, Halford, & Andrews, 2006), and mathematical word problems as a test of mathematical competence. Inspection of these contents is derived from their role in testing: Working memory is a central concept in intelligence theory (for example, Kyllonen & Christal, 1990; Wilhelm, 2000) and paramount to the fluid part of intelligence. Tests of fluid intelligence often include figural material in order to make them as independent as possible from any cultural content. The LST provides a promising domain-independent measure of intelligence and working memory. Of special interest for often exercised mass and repeated testing and theoretical inspections of intelligence properties are learning and practice effects. For this reason, parallel versions of the LST are tested and analyzed in a longitudinal study.

Mathematical word problems play a great role in school and university lessons as they measure logical, creative and mathematical abilities in parallel (cf. Dimitrov, 1996; Jonassen, 2003). The problem content in the current work is limited to statistical concepts as statistics are very important for everyday life and also for school and scientific work, especially for the social sciences. Unfortunately, statistics are often omitted during school lessons which results in poorly prepared university and job applicants. This is why measurement of statistical competencies is of special interest for both school and university settings.

The current work is built of a theoretical background part to lay the basis for understanding the content focuses and statistical modeling, followed by three main

studies considering LST, a longitudinal analysis of LST, statistical word problems, and a general discussion in order to integrate the results from all studies and draw conclusions about application of the shown methods. It is shown how the item construction process is conducted, which LLTM and CDM results emerge and how these results can be compared and integrated to make item construction more efficient and to improve item characteristics and test validity and reliability.

4 Theoretical background

This chapter provides the reader with general theoretical outcomes and important concepts for the understanding of the current work. The special content areas of working memory, learning effects and word problems are displayed in the particular studies (chapters 5, 6 and 7).

4.1 Rule-based item construction, item cloning and automatic item generation

In times of large-scale testing, the role of testing as important decision criterion and possible preparation for tests via internet, items are needed that are valid, reliable and precise with enough robustness against cheating and enough variety to avoid recognition effects by participants, especially when taking the same test several times. Additionally, items should allow for testing of particular abilities and competencies, depending on application contexts for example in schools, universities or in research. Moreover, adaptive testing gains more and more interest in the scientific community and requires special techniques of item construction and item selection. Powerful means to meet all these requirements are the techniques of rule-based item construction (Embretson, 1999; Fischer, 1973; Freund, Hofer, & Holling, 2008), item cloning and automatic item generation (Bejar, 1993; Glas & van der Linden, 2003; Irvine & Kyllonen, 2002) which are described in the following.

Rule-based item construction is a process that involves several parts. Usually, one determines so-called basic cognitive components or basic parameters that are supposed to affect an item's difficulty significantly. These components require specific cognitive operations to solve an item and so makes specific demands on information processing of test takers. This choice of parameters can be grounded on a purely theoretical basis which helps to identify possible difficulty-generating factors or even better also on empirical findings that propose certain basic parameters. Some helpful parameter identifications stem from inspection of existing

tests whose items were analyzed for construction principles that affect solution results (e.g. Enright, Morley, & Sheehan, 2002). These (theoretically based or empirically grounded) basic parameters are put into a design or Q-matrix with as many rows as items and as many columns as basic parameters. Cells contain a 0 for the parameters which are not required to solve an item and a 1 for the parameters that are required to solve an item. Based on this matrix, items with various combinations of basic parameters can be built. When constructing a Q-matrix, it should be taken care that as few and as disjoint basic parameters as possible are used to ensure statistical modeling to be efficient (few items with many basic parameters may lead to non-converging algorithms, non-significant parameter estimates and multi-collinearity and point to poorly defined cognitive theory behind the basic parameters). Thus, careful considerations are needed before designing the Q-matrix to guarantee for parsimony. Rupp and Templin (2008a) investigated Q-matrix-misspecification effects for a cognitive diagnostic model (DINA, cf. section 4.2.2) and show that item-specific over- and underestimations occur depending on too few or too many attributes in the Q-matrix. Baker (1993) investigated Q-matrix-misspecifications under the linear logistic test model (LLTM, cf. section 4.2.1). He concludes that a low proportion of non-zero elements in the Q-matrix leads to large root mean squares volumes and that this effect is much larger than the effect of sample size. These results clearly show the necessity to put lots of careful work in the definition of the Q-matrix and basic parameter choice. However, even the greatest efforts do not guarantee for sufficient and satisfactory explanation of variance by the Q-matrix.

There are several manifestations of such basic parameters, depending on test type and content. Often basic parameters describe particular parts of cognitive processes that are necessary to solve the items. In the current work, two item types are constructed and tested. For LST, the basic parameters describe necessary cognitive steps with a special complexity level and the number of the steps. For the word problems, basic parameters are typical statistical concepts that one has to understand and manipulate in order to solve the items.

So it can be distinguished between components that affect an item's difficulty (the mentioned basic parameters) and components that affect only surface characteristics. Irvine and Kyllonen (2002) call these two types radicals and incidentals. Radicals affect item difficulty significantly whereas incidentals only describe surface characteristics that should have only negligible impact on item difficulty. Radicals or basic parameters and their combinations are used to compose items

with certain difficulties and certain cognitive requirements. Incidentals are very helpful to vary the items' appearance which is particularly important for item cloning, generation of item banks and adaptive testing.

There are several great advantages of rule-based item construction: As the basic parameters are known, the validation process is refined because results can be analyzed with regard to the underlying basic parameters. Therefore, item validation can be made very efficient and systematic. The item construction process is coherent and structured from the beginning (theoretical basis) to the end (final valid and reliable items with known constituting difficulty structure). As soon as the impact of the basic parameters has been determined statistically and empirically, items consisting of new combinations of these basic parameters do not have to be calibrated again. This opens the possibility for large item banks without the necessity to calibrate every single item. Items with unique characteristics to test particular abilities or sub-abilities can be produced. Additionally, typical errors made during item writing by humans (for example spelling mistakes, ambiguity and extra solution paths, cf. Freund et al., 2008 for a similar argument) can be avoided.

Rule-based item construction can also be helpful for item cloning and automatic item generation. In item cloning, items with particular demands on information processing are chosen as "parent items". By changing surface characteristics (or incidentals), item siblings of these parent items are cloned. All item siblings with the same demands on information processing (but with possibly differing surface characteristics) belong to the same item family. All items of a family should be (at least nearly) stochastically independent, that is, the test taker's reaction does not depend on the surface characteristics of items with the same cognitive demands when working on them serially. For example, to change the surface information of a figural item, circles instead of triangles can be used. For a verbal item, context stories, single names or numerical information can be varied. Glas and van der Linden (2003) and Geerlings, Glas, and van der Linden (in press) propose examples of item cloning and statistical models that are able to capture the special features of this approach (cf. section 4.2.1 for statistical properties). Bejar (1993) and Roid and Haladyna (1981) give elaborate overviews about item cloning techniques.

Cognitive demands on information processing can, in a special case, also be represented by basic parameters or radicals. This links item cloning to rule-based item construction and makes the validation process more straightforward as item difficulties within item families can then be ascribed to certain basic parameters

and their combinations. In fact, rule-based item construction can be considered a special case of item cloning: Assuming that only basic parameters or radicals affect item difficulty as it is done in rule-based construction is very similar to item cloning in which the general item structure and content is held constant and only surface characteristics are varied to avoid recognition. Thus in rule-based item construction, items with the same combination of basic parameters which only differ in incidentals can be considered item siblings belonging to the same family.

Automatic item generation can then be conducted by either feeding the software with basic parameters and let it produce a number of different combinations of these parameters, or by feeding it with item families and pools of surface information features that are used to vary surface information (Glas & van der Linden, 2003; Geerlings et al., in press).

Not only in rule-based item construction but also (and especially) in item cloning, discrimination between radicals and incidentals is essential for item writing and generation. If the test writer (or maybe software program) uses supposed incidentals which in fact are radicals and do influence item difficulty (that is, change demands on information processing) to create an item clone, validity is severely threatened and test results may not reflect the test takers' abilities.

The advantages of item cloning and automatic item generation lie in the efficient item generation process. As soon as the original items (ideally rule-based constructed) are calibrated and proved to be of high quality, large amounts of items can be produced in a very cost- and time-effective way. Efficiency is maximized, risks as mistakes in item writing, recognition and undesired learning effects are minimized through cloning and automatic generation. In principle, a unique test can be built for every examinee while keeping the comparableness of results.

These procedures of rule-based item construction, item cloning and automatic item generation have a relatively short history in research. The first systematic approaches to rule-based item construction are the ones of Bejar (1990), who constructed mental rotation test items and investigated the effect of construction principles on item difficulty, and of Embretson (1999), who considered design principles for item construction. Glas and van der Linden (2003) and Sinharay, Johnson, and Williamson (2003) investigated item cloning statistically while Zeuch, Geerlings, Holling, van der Linden, and Bertling (2010) show successful applications of item cloning procedures with statistical word problems. Automatic item generation was implemented, for example, by Arendasy (2005) who generated

figural matrices automatically using item generators based on cognitive theory or by Freund et al. (2008) who automatically generated 25 figural matrix items to study the influence of different task parameters on the degree of difficulty in matrix items.

4.2 Statistical modeling

This section describes the statistical models which are important for and partly used in the current work. This part makes no claim to be complete and for further statistical issues beyond this demonstration appropriate literature should be considered.

4.2.1 IRT models

To start with the most general model, the two-level item cloning model (ICM) of Glas and van der Linden (2003) is described: At the lower level, items in families are described through a three parameter logistic (3PL) model (described by Lord, 1980), whereas at the higher level the item parameters in the same family follow a distribution that demonstrates the variability within families. Both persons and items are regarded as random samples from a person or item population, respectively. The person parameter θ is assumed to be standard normally distributed, the vector of the item parameters is regarded a realization of a random vector and assumed to be multivariate normally distributed. This general ICM is described in equation (4.1) and (4.2) (please note that, for sake of uniformity, the logit link version of the equation is used here which is not the case in the original work by Glas & van der Linden, 2003).

Level 1 (3PL model):

$$P(X_{ij} = 1 \mid \theta_j, a_{i_f}, \gamma_{i_f}, \sigma_{i_f}) = \gamma_{i_f} + (1 - \gamma_{i_f}) \frac{\exp\left[a_{i_f}\left(\theta_j - \sigma_{i_f}\right)\right]}{1 + \exp\left[a_{i_f}\left(\theta_j - \sigma_{i_f}\right)\right]} \tag{4.1}$$

Level 2 (family parameters):

$$\xi_{i_f} \equiv \left(a_{i_f}, \sigma_{i_f}, \text{logit}\, \gamma_{i_f}\right) \tag{4.2}$$

with

$$\xi_{i_f} \sim \text{MVN}(\mu_f, \Sigma_f) \tag{4.3}$$

with θ_j the ability parameter for person j, σ_i the difficulty parameter for item i, α_i the discrimination parameter for item i and γ_i the guessing parameter for item i. Equations (4.2) and (4.3) show that the model assumes item parameters with family specific means and variance-covariance matrices, indicated by index f.

An extension of this model was developed by Geerlings et al. (in press) (see also Zeuch et al., 2010). According to a design or Q-matrix, the items are scored on stimulus features: q_{ik} is the score of item i on stimulus feature k in the cognitive complexity model. Estimates include η_k , the weight of stimulus feature k in item difficulty and θ_j, the ability of person j. This yields a description of item difficulty as an additive function of basic parameters. Following this decomposition, the item cloning linear model (ICLM) takes into account the radicals or basic parameters:

$$\sigma_{i_f} = \sum_{k=1}^{K} q_{fk}\eta_k + \epsilon_{i_f} \tag{4.4}$$

ϵ_{i_f} denotes a family specific random error term for the items which accounts for variance in item difficulty that is not captured by the basic parameters. It will be shown that restriction of these general models can result in the linear logistic test model (LLTM, Fischer, 1973) as well as the 2PL (Birnbaum, 1968) and Rasch model (RM; Rasch, 1960). Developed for statistical modeling of results gained from items that were designed and generated using an item cloning approach, these ICMs take into account the special features of cloned items, i.e., the family membership and the covariances within and between item families. Although item cloning ideas were first developed during the 1960s and 1970s, only few research groups have been statistically concerned with it yet. The most famous works have been conducted by Glas and van der Linden (2003) and Sinharay et al. (2003). Sinharay et al. (2003) differentiate item cloning models according to their siblings treatment. The resulting models are called Unrelated Siblings (in which a separate, unrelated item response function for all items is assumed, ignoring their family membership), Identical Siblings (which assumes the same item response function for all items in the same family, ignoring only the variation between siblings) and Related Siblings Model (which resembles the hierarchical model of Glas & van der

Linden, 2003, and assumes a separate response function for each item but uses a hierarchical component to relate the siblings within the same family). Additionally, they propose the Family Expected Response Functions (FERF) that describe the probability that an examinee with ability 0 correctly responds to an item randomly selected from one item family and provide a graphical summary of item families.

These item cloning models are able to account for item parameters (difficulty, discrimination and guessing) as well as for variation between items in the same family (this within-family variation should only be caused by the incidentals that differ between the items and thus should be low compared to between-family variation).

The ICLM is estimated by Bayes algorithm (Gibbs sampler) and assumes random effects on item and examinee level. It provides difficulty modeling on family and basic parameter level as well as discrimination indices on family level. For statistical details please consider the original articles by Glas and van der Linden (2003) and Sinharay et al. (2003).

Item cloning models are a great mean to analyze cloned items and take into account all their features. Nevertheless, many assumptions have to be made to enable the Bayes estimation of these models to work, and the practical outcome of the estimation is sometimes doubtful. Moreover, often estimations only become stable enough with very large samples (cf. Geerlings et al., in press) which narrows the application for smaller studies and explorative purposes. Often simulation studies are more efficient than empirical prestudies for main study planning because in simulation studies important benchmarks can be found that help one to design one's empirical prestudies optimally. In adaptive testing and for large item pool calibration, these models are well-suited. But for smaller data sets and for exploratory purposes, models like LLTM or CDM derivatives (see below) seem to be more easy to apply and interpret (nevertheless, samples should not be too small for LLTM and CDM analyses; cf. Baker, 1993 or Green & Smith, 1987).

The word problems in study 3 (see chapter 7) are constructed using an item cloning approach, but it is shown that LLTM analyses reveal interpretable results, too. It is assumed that application of item cloning models would not lead to important incremental insights into item and basic parameter structure in this case.

By restricting the assumptions and parameter distributions of the ICM and ICLM, respectively, all subsequent models can be considered special cases of these general ICMs. By ignoring the family structure of the included parameters in the ICM

(that is, assuming that influence of incidentals is ignored or only one single family is considered), the 3PL model of Lord (1980) is obtained:

$$P(X_{ij} = 1 \mid \theta_j, \sigma_i, \alpha_i, \gamma_i) = \gamma_i + (1 - \gamma_i)\frac{\exp(\alpha_i(\theta_j - \sigma_i))}{1 + \exp(\alpha_i(\theta_j - \sigma_i))} \tag{4.5}$$

with θ_j the ability parameter for person j, σ_i the difficulty parameter for item i, α_i the discrimination parameter for item i and γ_i the guessing parameter for item i. Omitting γ_i leads to the 2PL model (Birnbaum, 1968):

$$P(X_{ij} = 1 \mid \theta_j, \sigma_i, \alpha_i) = \frac{\exp(\alpha_i(\theta_j - \sigma_i))}{1 + \exp(\alpha_i(\theta_j - \sigma_i))} \tag{4.6}$$

Omitting both γ_i and α_i leads to the 1PL or Rasch model (RM; the probably best known IRT model described by Rasch, 1960). The RM only accounts for item difficulty and states the probability that person j answers item i correctly as follows:

$$P(X_{ij} = 1 \mid \theta_j, \sigma_i) = \frac{\exp(\theta_j - \sigma_i)}{1 + \exp(\theta_j - \sigma_i)} \tag{4.7}$$

Ignoring the family specific parameter structure and the guessing and discrimination parameter as well as ϵ_{i_f} in the ICLM finally reveals the LLTM (Fischer, 1973) which will be, together with CDMs, the focus of the following analyses:

$$P(X_{ij} = 1 \mid \theta_j, q, \eta) = \frac{\exp\left(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k\right)}{1 + \exp\left(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k\right)} \tag{4.8}$$

The LLTM differs from RM only through decomposition of σ_i:

$$\sigma_i = \sum_{k=1}^{K} q_{ik}\eta_k \tag{4.9}$$

Person effects in the LLTM are usually regarded as random and item effects as fixed. Nevertheless, if one wants to consider the abilities of some specific persons instead of the difficulties of specific items, the person effects can be regarded as fixed and the item effects as random (cf. van den Noortgate, de Boeck, & Meulders, 2003). LLTM allows for consideration of basic parameter impact on item difficulty

and thus for evaluation of rule-based item design. Great advantages of LLTM (and its variants, see below) are its parsimony, accurateness, efficiency and well-proved implementation in various software packages (maximum likelihood estimation).

As can be easily seen in equation (4.8), there is no error term included in the original LLTM. Therefore it assumes that the included basic parameters can explain the whole variance in item difficulty. Because this is a rather unrealistic assumption, an extension of the LLTM was made by van den Noortgate et al. (2003) and by Janssen, Schepers, and Peres (2004) who added a random error term with respect to the items to the function in (4.8). This random error term takes into account an estimate of variance in item difficulty that is not accounted for by the complexity model, i.e., by the basic parameters. While van den Noortgate et al. (2003) call this model a cross-classification multilevel logistic model, another common denotation is Random Effects LLTM (shortly RE-LLTM, but do not mix up this model with the Random Weights LLTM of Rijmen & de Boeck, 2002) which will be used in this work. The RE-LLTM assumes that a random error term can be defined for both items and persons, with responses nested within persons and within items, if items and persons are regarded as random samples from a population of items and a population of persons, respectively (van den Noortgate et al., 2003). Item difficulty can still be described by an additive function of basic parameters, with an error term added :

$$\sigma_i = \sum_{k=1}^{K} q_{ik}\eta_k + \epsilon_i \tag{4.10}$$

This is identical to the ICLM item difficulty decomposition except for the explicit family structure. In LLTM identical item difficulties for identical basic parameter combinations are assumed and implicitly variation due to incidentals is ignored. Thus LLTM resembles ICLM except for the fact that ICLM allows for consideration of within-family variation instead of ignoring it.

The assumption that items of one set can be regarded as a random sample from a population has become more prominent during the last years and expanded the statistical viewpoints in item construction and empirical testing. For a sophisticated discussion about random item parameters also see de Boeck (2008). The author also mentions that adding random error terms usually leads to higher standard errors for parameter estimates but that this has to be accepted as long as adding the error term results in better model fit and more precise mapping of data.

Additionally, estimates which are significant in the LLTM may not be significant in the RE-LLTM due to higher standard errors. However, this allows for more careful decisions about significant effects of basic parameters: The ones which are still significant in RE-LLTM most often have robust and reliable influence on item difficulty. The ones which are not significant any more in the RE-LLTM may be ignored and the random error term captures the remaining variance in item difficulty which was partly but wrongly allocated to the now non-significant basic parameters before (i.e., in the LLTM).

If one wants to consider effects beyond the basic parameters, i.e., second level effects that may affect item difficulty and account for variance that is not captured by the basic parameters, the Latent Regression LLTM (LR-LLTM; de Boeck & Wilson, 2004) provides another alternative within the LLTM family. It decomposes the person parameter θ into an additive function and allows for investigation of person predictors on the second level. The LR-LLTM states the probability that person j passes item i as follows:

$$P(X_{ij} = 1 \mid \theta_p, q, \eta, a) = \frac{\exp\left(\sum_{p=1}^{P} a_{ip}\theta_p + \epsilon_p - \sum_{k=1}^{K} q_{ik}\eta_k\right)}{1 + \exp\left(\sum_{p=1}^{P} a_{ip}\theta_p + \epsilon_p - \sum_{k=1}^{K} q_{ik}\eta_k\right)} \tag{4.11}$$

with θ_p the person predictors (second level predictors) and a_{ip} the score of item i on the person predictor p.

Typically, model fit can be compared by Likelihood Ratio tests (LR-tests, for identical data sets and as far as Maximum Likelihood (ML) estimation is conducted), Akaike information criterion (AIC) and Bayesian information criterion (BIC). Please see, for example, Burnham and Anderson (2004) for these indices as well as Raftery (1995) for considerable differences for model selection.

This choice of models from the LLTM family allows for detailed inspection of item and person parameters and for mapping empirical results in order to derive hints for item construction or cognitive model improvement. In principle, more variants of the LLTM can be built, but in the current work the mentioned variants are sufficient for investigation of the research questions (for longitudinal modeling, the described models have to be extended by learning parameters, however).

Successful applications of the LLTM are shown by several authors. Dimitrov (1996) analyzed university examinations and found nine and thirteen significant basic parameters for statistical and algebraic tasks, respectively. Cisse (1995)

investigated addition and subtraction word problems and showed that the full cognitive model containing six basic parameters outperformed two separated models with only three basic parameters. Further investigated content areas are reading comprehension (e.g. Gorin, 2005; Sonnleitner, 2008) and reasoning (cf. Hahne, 2008). Kubinger (2009) shows LLTM applications beyond the purpose of basic parameter identification and describes analysis of item position and speeded presentation effects, content-specific learning effects and effects of item response format using LLTM features.

There have to be noted some special characteristics of LLTM parameter estimation and interpretation. First, absolute LLTM parameter estimates are not interpretable. This means that estimates for one parameter can only be judged in comparison with other estimates of the same data set and estimation procedure. The higher a parameter value compared to other values, the higher is its influence on item difficulty. Random effects estimates are reported as variance.

However, these models only consider the items. What about the persons, apart from second level predictors? The next section describes another model type which enables the researcher to gain information about the examinees and the examinees to gain information about their abilities and shortcomings.

4.2.2 Cognitive diagnostic models

Rather than simply ordering test takers along a continuous latent dimension as it is the case in uni- or multidimensional Rasch models, cognitive diagnostic models (CDMs) focus on the more complex goal of classification of test takers into latent classes that are defined by latent skill profiles. These skill profiles resemble a list of cognitive attributes that the examinee might possess or might not possess, depending on the performance on specific tasks (cf. Junker & Sijtsma, 2001). From a statistical point of view, CDMs are confirmatory factor models with categorical latent variables. As explained in the sections before, cognitive operations necessary to solve an item are coded in a design matrix (Q-matrix) which then contains a set of basic parameters. CDMs are able to map these basic parameter structures as defined in a Q-matrix. CDMs produce skill profiles for examinees and allow for multiple criterion oriented conclusions rather than normative interpretations. This offers the possibility of detailed feedback about mastered and non-mastered skills for every single examinee to enable him or her to practice non-mastered skills more carefully.

While the LLTM focuses on item difficulty and item characteristics that affect item difficulty, CDMs focus on the examinees and their skill profiles. These two kinds of models allow for detailed feedback for the researcher concerning the item construction process as well as underlying cognitive theory. Further, the examinee can find out which skills he or she has to practice and which are already well mastered. Moreover, teachers can benefit from this feedback concerning skills that should be practiced with students. Nevertheless, CDM results can be aggregated in order to provide population-based interpretations. Thus, LLTM and CDM analysis provide helpful instruments in the field of educational research and testing. The theoretical analysis and empirical affirmation of the underlying cognitive principles defined in the Q-matrix can serve as starting point not only for LLTM analysis but also for CDM application. However, as described in more detail in the next section, CDMs and LLTMs differ extremely in their assumptions about the latent variable(s) investigated. But before these differences are described, CDMs and the example used in the current work have to be introduced.

There has been a development of a variety of CDMs during the recent years. Sophisticated summaries about CDMs can be found in diBello, Roussos, and Stout (2007) and Leighton and Gierl (2007). CDMs can be classified according to their flexibility, their compensatory characteristics (i.e., if the skills can compensate for each other or not), their ability to cope with dichotomous or polytomous variables or their method to treat guessing and slipping. Recent developments consider computer adaptive attribute testing (Gierl & Zhou, 2008) or focus on estimation methods, discrimination indices and model robustness (e.g., Henson, Douglas, Roussos, & He, 2008; Rupp & Templin, 2008b; Templin, Henson, Templin, & Roussos, 2008). McGlohen and Chang (2008) suggest combining computerized adaptive testing (CAT) and cognitive diagnostic modeling and show that estimating both the classical person parameter from IRT and the attribute mastery vector from CDM yields an advantage regarding information efficiency compared to estimating only one of these two.

While there are several studies investigating these technical and statistical issues of CDMs, most applications of CDMs concern simulated data and only few applications of CDMs to empirical data are described. Empirical applications (for example, de la Torre, 2008, de la Torre & Douglas, 2004 and Henson, Templin, & Willse, 2008) are mainly based on the empirical fraction-subtraction data set first collected and developed by Tatsuoka (1990). Another empirical example stems from Templin and Henson (2006) who investigate CDM analysis of the underlying

factors contributing to pathological gambling. Additionally, de la Torre (2008) provides an application of his Q-matrix validation procedure to 2003 NAEP 8th grade mathematics data but has to admit that model-data-fit is very poor in this case. Simulation studies are very helpful to answer questions of statistical model definition and software algorithm specifications, but they are not sufficient to show the applicability of CDMs to empirical data. The current work will provide results from CDM analysis of empirical data for two different item types tested in samples of sufficient size to ensure reliable results.

For CDM application, but also for other ends, the Q-matrix should be as parsimonious as possible. The more basic parameters are chosen, the higher becomes the number of possible mastery classes and the higher the number of items and examinees should be for considerable results and model convergence. For this reason, restrictions which exclude particular mastery classes (because these classes contradict theoretical concepts or emerged to be of no considerable size in earlier analyses) before analysis starts can be implemented in CDM software.

Because there is such a variety of CDMs, only the DINA (deterministic inputs, noisy and-gate) model (cf. diBello et al., 2007) will be explained in detail and used in the current work. Unfortunately, software routines for CDMs are very rare and there is no program which can estimate all CDM variants (see also Rupp & Templin, 2008b). The DINA model is used because it is relatively parsimonious while allowing for comparisons of LLTM and CDM modeling in principle.

The DINA model is defined as follows:

$$P(X_{ij} = 1 \mid \xi_{ij}) = (1 - s_i)^{\xi_{ij}} g_i^{1-\xi_{ij}} \tag{4.12}$$

with

$s_i = P\left(X_{ij} = 0 \mid \xi_{ij} = 1\right)$, $g_i = P\left(X_{ij} = 1 \mid \xi_{ij} = 0\right)$ and $\xi_{ij} = \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}$.

i denotes the item number, j the latent classes of attribute mastery patterns or the persons, respectively, k the attribute or basic parameter, α_{jk} the mastery of the basic parameters or abilities as defined in the Q-matrix. α_{jk} is 1 if a person in class j has ability k, 0 otherwise. q is based on the Q-matrix entries and is 1 if basic parameter k is needed for item i, 0 otherwise.

DINA requires mastery of all necessary attributes in one item to solve it:

$$P(X_{ij} = 1) = \begin{cases} 1 - s_i & \text{if } \sum_{k=1}^{K} q_{ik}\alpha_{jk} = \sum_{k=1}^{K} q_{ik} \\ g_i & \text{else} \end{cases} \tag{4.13}$$

DINA is a so-called non-compensatory model because every attribute of an item has to be mastered by the examinee to solve this item (mastery of one attribute cannot compensate for non-mastery of another attribute). s is also called the slipping parameter which indicates the probability that a respondent fails to solve an item although having mastered the required attributes. Likewise, g is called the guessing parameter and indicates the probability that a respondent solves an item without having mastered the required attributes.

The compensatory equivalent of DINA is the DINO model (deterministic inputs, noisy or-gate). It assumes that at least one attribute involved in an item has to be mastered to solve it and that mastery of one attribute can compensate for non-mastery of another attribute. Comparison of DINA and DINO results allows for insight into the basic parameter relations, i.e., can mastery of one attribute compensate for non-mastery of another (then the DINO model would be more appropriate), or is mastery of all attributes included in one item required to solve the item (then the DINA model would be the first choice)? Because in all item sets of the current study it is very clear that all attributes included in an item have to be mastered, only DINA results are reported and interpreted. It is not necessary to compare DINA and DINO fit because the rule-based construction processes of the item sets only include basic parameters that cannot compensate for each other.

Considering guessing and slipping parameters can also help to judge item and model fit in an informal way. Following de la Torre and Douglas (2004), both parameters should lie below .20 for each item. High guessing and slipping parameters indicate that the supposed basic parameter structure seems to map actual testee behavior poorly. However, de la Torre (2008) mentions that small slipping and guessing parameters are a sufficient, but not a necessary condition for establishing model-data-fit as sometimes items based on a particular set of attributes can show high guessing or slipping parameters, but model-data-fit improvement can only be reached by employment of a different set of attributes. Another possibility to judge model fit is provided by the sum of the means of s and g (de la Torre, 2008 denotes a sum of approximately .25 as reasonable good fit).

However, to interpret guessing and slipping parameters seriously, the above mentioned limits seem not to be adequate. Items with slipping parameters of .20 seem

to be of little practical value as 20 percent slipping already strongly questions item quality. On the other hand, guessing parameters may be higher than .20 without pointing to inadequate item quality (during the following section, it will be described why these limits may not be adequate in CDM application). So, in the current work it is recommended to use a more liberal criterion for guessing and slipping parameters: $1 - s - g$ should not be lower than .50. Items with parameters beyond these limits point to Q-matrix misspecifications and suboptimal item quality and are hardly interpretable. Additionally, slipping parameters should be as low as possible and guessing parameters should never exceed .50. Otherwise, results o not provide reliable information about the examinees' abilities. The difference between (1-slipping) and guessing is also called a kind of CDM item discrimination by Templin and Ivie (2006) which is comparable with the classical test theory item discrimination with regard to sizing (.50 means moderate discrimination, below that value low discrimination, above high discrimination; cf. Templin & Ivie, 2006).

Most software applications for CDMs provide AIC and BIC to evaluate model fit. Item fit can be evaluated by the Mean Absolute Difference (MAD) between the observed and model-predicted item response function. The mean of MAD can be used to judge fit of the whole model rather than of single items. MADprop denotes the mean absolute deviation for solution probabilities, MADcor the mean absolute deviation for pairwise item correlations (should be as low as possible), and MADLOR the mean absolute deviation for pairwise log odds ratios of item i and i' (if MADLOR is extremely high for one item compared to the others, this item does not fit well). These MAD variants are not very sophisticated and well-proved criteria, however. Thus, judging item and model fit for DINA and DINO is somehow poor and should be treated carefully. DINA and DINO results therefore have more explorative ends, but new software implementations as well as development and implementation of reliable fit statistics are strongly needed. Some software routines provide the RMSEA statistics which is more sophisticated and better known, but such routines only fit specific models differing from DINA and DINO. So far, there seems to be no agreement about model and item fit indices and criteria or ranges for good or bad fit.

This short description of CDMs and the specific DINA realization used in the current work show that CDMs sometimes may serve explorative and descriptive ends rather than providing solid parameter and skill estimates. Nevertheless, these explorative and descriptive results can serve as an important resource of information

by providing individual skill profiles and hints for the nature and dimensionality of the underlying constructs to be measured. Further developments of CDMs are under construction and probably will eliminate the shortcomings of so far CDM analysis implementations as lacking reliable fit indices and interpretational confusions.

4.2.3 Parallels and differences between LLTM and CDM

LLTM represents a log-linear approach and assumes one population and one latent dimension to be measured. More basic parameters are supposed to increase item difficulty, parameters are linked linearly. There is a probabilistic linkage between item solution and parameter impact.

In CDM (here: DINA), which represents a mixture approach, in principle several populations as well as several dimensions are allowed. Test takers can apply different strategies as CDMs can handle different attribute mastery patterns explicitly. DINA is deterministic with regard to mastery or non-mastery of attributes. Deviations from this deterministic constraint are only possible through guessing and slipping parameters. Thus, DINA is more adequate for distinct learnable and trainable skills than for unidimensional constructs (see below). So, basic parameters in LLTM and skills in CDM look similar in the Q-matrix but the Q-matrix is based on different concepts. DINA results both show impact of basic parameters and provide basis for individual feedback about skill profiles.

To illustrate the different concepts of LLTM and CDM analysis, imagine a set of seven items, constructed on the basis of three basic parameters (table 4.1.

Table 4.1: Design matrix example

Item	BP1	BP2	BP3
I1	1	0	0
I2	0	1	0
I3	0	0	1
I4	1	1	0
I5	1	0	1
I6	1	1	1
I7	0	1	1

Notes: BP = basic parameter.

All possible states which are in principle allowed by DINA and its theoretical assumptions are shown in table 4.2. These states build the basis for the mixture and are distributed to examinees. Thus examinees can in principle create each of these states, but following theoretical assumptions some states are supposed to be created more often than others, for example because of differing difficulty of skills.

Table 4.2: Possible states CDM

State	I1	I2	I3	I4	I5	I6	I7
1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0
3	0	1	0	0	0	0	0
4	0	0	1	0	0	0	0
5	1	1	0	1	0	0	0
6	1	0	1	0	1	0	0
7	0	1	1	0	0	0	1
8	1	1	1	1	1	1	1

To visualize possible ways for item solution, look at figure 4.1: Examinees who possess no skill will not solve any item. Those who master skill 1 will solve item 1, those who master skill 2 will solve item 2, examinees who master skill 1 and 3 will solve items 1, 3 and 5, those who possess all skills will solve all items, including item 6. Keep in mind that DINA is completely deterministic, that is if a skill is mastered, the corresponding item will be solved. Deviations from this deterministic connection are only possible through guessing and slipping parameters. Some examinees will create patterns which do not correspond to this deterministic linkage. These patterns can only be explained by guessing and slipping parameters. The approach of CDMs with their mastery classes is similar in this point to the concept of knowledge spaces (see Doignon & Falmagne, 1999 and Falmagne, Koppen, Villano, Doignon, & Johannesen, 1990, for example).

Now let the three basic parameters or skills be of different difficulty, that is, B3 is more difficult than B2 and B2 more difficult than B1. Still all states are possible to occur (as far as no a priori limitations are made), but given the differing difficulty, some states (states 1, 2, 5 and 8) are more likely to occur than others as it is more probable that examinees who master more difficult skills also master the easier ones. However, different strategies or different knowledge may enable some persons to master actually difficult skills but not actually easy ones which can

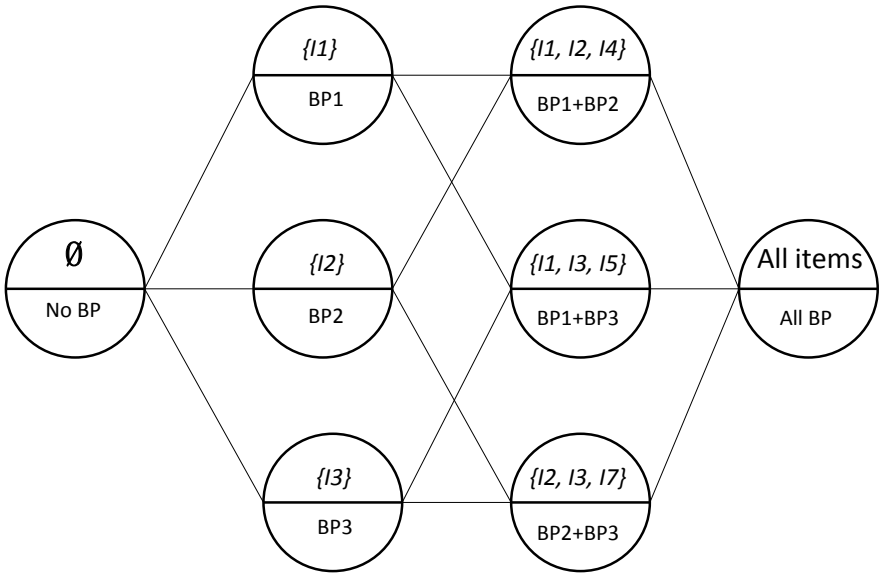


Figure 4.1: Possible skill combinations and item solutions (upper parts: sets of items to be solved; lower parts: skills to be mastered)

result in realization of unexpected states 3, 4, 6 and 7. These deviating patterns can only be explained by guessing and slipping within the DINA framework.

A special fact of CDMs, in this case of DINA, is that in principle a Q-matrix can completely explain an empirical data set if this matrix is only detailed enough (in extreme, but not too seldom, cases, assigning one basic parameter or skill to each item). Restricted Q-matrices thus are always somehow misspecified. However, this sets up a problem of Q-matrix definition: Which deviation from perfect explanation can be accepted and when is the deviation too high? This again points out the acceptable limits of guessing and slipping parameters. The lower one wants the parameters to be, the more detailed the Q-matrix has to be as then empirical results are mostly explained by the true mixture basis and not by guessing and slipping. Thus claiming $1 - g - s$ to be higher than .95, for example, would probably result in the requirement of a huge Q-matrix with almost no reduction of basic parameter number compared to item number. Therefore in the current study the very generous (and probably more realistic) criterion of $1 - g - s$ being higher than .50 is applied.

In LLTM, it is supposed that the basic parameters represent item rules and belong to one single dimension. The more rules have to be applied and the more difficult

these rules are, the higher is item difficulty, the higher are the requirements of this item on cognitive abilities and the lower is the probability that an examinee solves this item.

While the Q-matrix can look very similar or identical for both CDM and LLTM, the implications differ extremely. If examinees are aware of the skills beyond the rules, one-rule problems should be solved correctly. However, if rules have to be combined, this may set limits for solving items. Thus, if only rule combinations, but not the rules themselves, provide difficulties for problem solving, a unidimensional model assumption would be a better choice for modeling than multidimensional skill assumptions. In this case, especially items with only one rule or basic parameter will provide difficulties for CDMs: These problems will be solved nearly perfectly, and extreme high guessing and low slipping parameters would be observed in empirical application. The opposite would then be the case for items with many rules: As the combination causes examinees to reach the limit of their capacity, CDM analysis will reveal high slipping and low guessing parameters. This phenomenon was already mentioned in a slightly different way by Templin and Ivie (2006) who found high guessing and low slipping parameters for easy items and the opposite for difficult items. If skill application itself sets limits for individual capacity, the unidimensional approach of the LLTM would not be appropriate.

This means that CDMs are best qualified if there is a set of skills some of which are mastered and others are not. This often appears in educational contexts, especially in learning new concepts, which makes CDMs very interesting for school settings and teachers.

These differences between assumptions of LLTM and CDM as well as their consequences for Q-matrix design and interpretation of modeling results in empirical application will be explicated in the following chapters. The two different item types provide the possibility to investigate underlying cognitive concepts and adequacy of model assumptions for LLTM and CDM.

In the current study, the two used item types provide different starting bases. While LST as a figural working memory test is supposed to increase working memory load by (more) basic parameters of higher complexity, word problems are a test of statistical competencies. Given the possibility that increasing working memory load can be regarded as linear process, more difficult basic parameters are supposed to increase working memory load for all individuals, no matter

if they have a high or low capacity. In word problems, however, there remains the possibility that more complex statistical concepts are mastered while less complex are not mastered. This should not be the normal case but cannot be excluded completely. Therefore, DINA results may be better for word problems than for LST as in case of word problems (which can rather be regarded distinct skills) DINA provides the advantage of mapping possible different strategies and knowledge states. Additionally, in word problems more theoretically unexpected mastery classes can emerge (that is, mastery of difficult statistical concepts with simultaneous non-mastery of easy ones) which should not be the case in LST.

Although LLTM and CDM mixture assumptions are not too far away from each other in fact, results for skill / rule difficulty do not have to be completely identical in order since both model classes use different estimation algorithms. While in LLTMs a strict linear order of basic parameters is assumed, in CDMs the skill probabilities are computed from aggregation of skill mastery classes. These classes can reveal different "learning paths" or mastery orders with the same right to exist (some examinees may master a skill of moderate difficulty before another while for some other examinees the opposite order can emerge). However, the main point is which model class better explains the underlying structure and cognitive requirements. It is not possible that DINA and LLTM both map data adequately. One of both will be more appropriate, thereby also providing information about the cognitive dimension(s) underlying item difficulty or skill mastery.

4.3 Application hints and knowledge gain

LLTMs and CDMs can help the researcher to answer different, but complementary question areas, for example about definition of the Q-matrix and item construction processes (LLTMs) or about response patterns, latent classes of participants with the same attribute mastery profile and compensatory relations between skills (CDMs). To my knowledge, neither a direct comparison nor a simultaneous application of LLTMs and CDMs to empirical data have been conducted until now for the current item types. For this reason, it will be shown how selected LLTM and CDM variations can be applied to rule-based items with verbal and figural content, and how they can help resolve the question how basic parameter impact can change during a longitudinal study of figural items. Additionally, it will be investigated which model class with its inherent assumptions is more appropriate to grasp the underlying cognitive structure of the empirical data. It

will be demonstrated how a careful item construction process can be conducted from theory-based conceptions over rule-based generation, empirical application and statistical modeling to helpful interpretation.

Rule-based item construction, item cloning, adaptive testing and statistical modeling of these areas seem to be well-studied regarding many fields, but much work is left for others. It is assumed that test application in educational and scientific contexts will require more and more efficient test construction and administration. Rule-based item construction prepares the basis for efficient item writing and administration techniques up to AIG and CAT. Perhaps the greatest challenge is the definition of the basic parameters and the construction of items in a way that Q-matrix specification is definite and adequate. Otherwise, statistical results may be not reliable and even misleading. Therefore, theoretical foundation of a test and its Q-matrix is inevitable. As soon as the Q-matrix and principal item construction have been mastered, the way is cleared for AIG and CAT, using large item pools generated via item cloning. In this way, the described techniques are connected to and built on each other and can help to make large scale testing as well as smaller empirical studies more efficient and easy to handle and interpret.

In the three studies it will be shortly described which models and model variants will be used, but for more detailed information as equations and estimation features please consider this chapter and mainly section 4.2.

5 Latin Square Task

The current study demonstrates rule-based construction of figural items and a first application of the item construction principles and statistical models described in chapter 4.

5.1 Introduction

The Latin Square Task (LST) was developed by Birney et al. (2006) and represents a non-domain specific, language-free operationalization of relational complexity (RC) theory. The current study investigates the basic cognitive parameters and structure of LST as defined by RC theory, using the IRT-based linear logistic test models as well as cognitive modeling by cognitive diagnostic models. 850 German school students completed 26 rule-based constructed LST items. Results support the notion of Rasch-scalability. LLTM analyses reveal that both operation complexity as well as number of operations affect item difficulty. CDM analyses suggest ordered classes of mastery of different complexity levels. It is shown how LLTM and its variants as well as CDMs can contribute substantial insights into cognitive solution processes and composition of item difficulty in relational reasoning.

Today, the terms of working memory and reasoning play a great role in the field of psychologic research. These concepts led to a wide range of research activities yielding interesting outcomes. Working memory is an essential component of almost every intelligent achievement. It is very clear that working memory capacity is limited and that these limits are often responsible for mistakes and lacks. Unfortunately, working memory capacity is often defined by somewhat coarse concepts and measured by a variety of tasks that seem to provide solid information about the capacity limits, but are not based on a reliable cognitive theory for capacity limits.

5.2 Background

This section introduces the theoretical basis of the items used in the current study as well as the underlying concepts of intelligence and working memory.

5.2.1 Working memory and reasoning

The concept of intelligence plays an important role in the field of psychometric research. Intelligence is a reliable predictor of school grades, work success, learning speed and other achievement areas (cf. Schmidt & Hunter, 1998; Watkins, Lei, & Canivez, 2007). One popular theory is the CHC-Theory of intelligence (developed in 2000 by Cattell, Horn and Carroll). In this theory, one general factor g is responsible for all intelligent achievements. Then there are several second and third components, ordered hierarchically and ranging from broad to narrow abilities. On the second stage there are (between others) g_f (fluid intelligence) and g_c (crystallized intelligence). g_f is often supposed to be at the core of intelligence and to be responsible for most thinking and learning processes (see also McGrew, 2005). Essential components in this area are reasoning and working memory. As McGrew (2005) and Wilhelm (2000) mention, g_f and reasoning are closely connected. Reasoning is important for making decisions, planning and problem solving, for example. One can differentiate between inductive and deductive reasoning. Another important factor in intelligent achievements is the concept of working memory. Working memory, which is necessary for composition and manipulation of mental representations, is closely related to reasoning. There seems to be a very strong relationship between g_f, reasoning and working memory (cf. Kyllonen & Christal, 1990; Wilhelm, 2000), but causality is not clear yet.

Baddeley (1986) presented a model of working memory in which a phonological loop and a visuo-spatial sketch-pad are short-term-stores for new verbal or visuo-spatial contents. The third part supposed to belong to working memory is the central executive which is responsible for higher cognitive processes and which plans and conducts processing of information. Colom, Abad, Quiroga, Shih, and Flores-Mendoza (2008) could show that short-term storage seems to explain the connection between working memory and intelligence whereas mental speed, updating and control of attention were not consistently related to working memory and intelligence.

Awh and Vogel (2008) showed that probably activity of the basal ganglia is connected to working memory activity and that differences in the efficiency of filtering relevant from irrelevant stimuli could account for interindividual differences in working memory. Barrouillet, Lépine, and Camos (2008) found out that effect of working memory capacity on high-level cognition is mediated by the impact of a basic general-purpose resource which seems to affect each basic little step of cognition. Kessler and Meiran (2008) show that there seem to be two updating processes in working memory: One global process which prevents working memory from interfering input which depends on the total number of items in working memory (comparable to some kind of storage process) and one local process that provides flexibility and is sensitive to the number of items which are modified at one time. Oberauer, Süß, Wilhelm, and Sander (2007) claim that working memory capacity is the best single predictor of reasoning ability and accounts for about the half of systematic variance in reasoning or fluid intelligence tests. The most important component seems to be temporary binding of representations which have to be maintained simultaneously. In line with this point of view, Oberauer, Süß, Wilhelm, and Wittmann (2008) question the storage and central executive view of working memory and claim that working memory should rather be seen as a system which builds relational representations through temporary bindings between representations of several components. Diverging from both the storage and executive point of view and the relational representations focus, Unsworth and Engle (2007) suggest that interindividual differences in working memory at least partially emerge from the ability to maintain information in primary memory and to search for information in secondary memory.

A very sophisticated new working memory model was developed by Oberauer (2009) which takes into account specific requirements for a working memory system, an analytical and associative processing mode and the differentiation between a declarative and a procedural part. Oberauer (2009) underlines the role of working memory in combining pieces of information into new common schemes and the importance of relations between pieces of information.

As can easily be seen from these mostly converging, but sometimes inconsistent concepts of working memory, there seems to be a variety of theories and empirical results accounting for one or another concept. There is no agreement about a global working memory concept. However, almost all theories and trends point out that some kind of relational integration of information as well as the existence of a capacity limit of working memory cannot be questioned. The limitation of

working memory is of special interest for intelligence because working memory's performance is essential for long-term storage and manipulation of information. The greater the capacity of working memory, the more information can be manipulated and integrated and the more and better intelligent achievement can be expected.

Additionally, reasoning and working memory were often linked to personality factors as impulsiveness, sensation seeking and lack of fear (e.g. Colom, Escorial, Shih, & Privado, 2007) and seem to predict school grades (cf. Krumm, Ziegler, & Bühner, 2008) especially for science-related courses (reasoning) and language courses (verbal components of working memory tasks).

5.2.2 Cognitive complexity and RC theory

But how can capacity of working memory be measured? Capacity depends on the complexity of a task. Hence, systematic manipulation of task complexity is very helpful for assessment of working memory capacity. There have been several attempts to measure complexity, including a posteriori measurement as pieces of information, g-load or processing speed. Holzman, Pellegrino, and Glaser (1983) and Carpenter, Just, and Shell (1990) presented an a priori measurement of working memory capacity for number series completion and for the Raven Matrices. However, these proposals are task specific. More helpful would be an a priori measurement which is not task-specific. One very interesting proposal was made by Halford, Wilson, and Phillips (1998). They display complexity as relations, in line with many results and concepts ascribing relations great importance in working memory. This resulted in the RC theory (relational complexity theory). It indicates that number and complexity of the relations between the pieces of information that have to be processed are responsible for the complexity of a task rather than the number of pieces of information itself or processing speed. This relational complexity is a non-domain-specific, a priori metric for complexity which has proved promising for several research areas. The rules of relational complexity are stated as follows (Halford et al., 1998; p. 805):

> The *complexity* of a cognitive process is the number of interacting variables that must be represented in parallel to implement that process. Processing complexity also can vary over time within one task, hence the critical value is the complexity of the most complex step. Tasks can vary in the number of steps they require, but this does not necessarily

> affect processing load because a task with many steps might impose only a low demand for resources at any one time (e.g., counting peas in a box).
>
> The *processing complexity* of a task is the number of interacting variables that must be represented in parallel to perform the most complex process involved in the task, using the least demanding strategy available to humans for that task.

Following this definition, the complexity of a relation $R(a_1, a_2, \ldots, a_n)$ is defined by the number of its arguments n. A unary relation as class membership has one argument: Animal(cat). A binary relation as comparing size has to arguments: Bigger-as(elephant, mouse). A ternary relation has three arguments and so on. However, there are possibilities to reduce capacity requirements especially for higher-dimensional tasks by chunking (recoding and summarizing of concepts into fewer dimensions) and segmentation (serial processing of a relation by splitting it into several lower-dimensional steps).

Halford, Cowan, and Andrews (2007) explicitly link working memory to reasoning and state that working memory and reasoning share common capacity limits which directly points to assumptions of RC theory. Halford et al. (2007) state that the capacity limits can be quantified according to the number of items kept in working memory and according to the number of relations which can be processed in parallel in reasoning. RC theory has proven to be an excellent concept for explaining empirical results in several contexts, for example for transitive inference, the Tower-of-Hanoi problem and for cognitive development of children, and several studies could confirm predictions and characteristics of RC theory empirically (e.g. Birney & Halford, 2002; Halford & Andrews, 2002; Andrews & Halford, 2001).

The next section describes a relatively new task type which was developed to be a domain non-specific, knowledge-free and easy to understand operationalization of the RC theory.

5.2.3 The Latin Square Task

The Latin Square Task (LST, Birney et al., 2006) is a new task type that is supposed to measure the effects of complexity independent of knowledge or processing strategies. Latin Squares consist of several cells containing non-meaningful symbols. One cell contains a question mark and the examinee has to decide which

symbol has to be placed into this cell. The only rule to follow is that every symbol must occur exactly once in every row or column, respectively, and this rule remains the same for every level of complexity. The operationalization of complexity levels of RC theory is defined as follows: A binary relation requires considering only one line or column to find out the correct solution, a ternary relation requires considering one line and one column simultaneously, and a quarternary relation requires considering several lines and columns simultaneously. One item can require to manage several processing steps of differing complexity levels (for example, one binary and two ternary) serially (i.e. content of empty non-target cells has to be determined before resolving the target cell).

Birney et al. (2006) could show that this task type seems to represent an adequate operationalization of the RC theory. They tested a school and university student sample with 18 LST items. Results demonstrated Rasch scalability of the items and highlighted relations between complexity manipulation and item difficulty and response times, respectively. RC was a significant predictor of item difficulty. Statistically controlling for RC, the number of processing steps (i.e. the sum of all cognitive steps to be conducted serially for item solution, no matter if binary, ternary or quarternary) became a significant predictor as well. Hence, two independent pieces of information, task complexity and number of steps, were of central importance in explaining item difficulty. However, the reliability of the LST in Birney et al. (2006) was below .80, which may partly have been due to test length.

Bowman (2006) describes the construction of Greco-Latin Squares. For the solution of these items, one has to consider the symbols as well as their color which results in an extra complexity level named quinary. However, it is questionable in my opinion if this item type can provide insight into construction principles and RC theory operationalization because there seems to be lots of ambiguity in the items. This ambiguity is due to lack of classification clarity and relatively complicated solution paths. The former pure type of LST items looks more promising to me regarding gain in information about task characteristics and underlying construction and theoretical principles. New results concerning complexity manipulation and LST application can be viewed in Birney and Bowman (2009).

The current study describes the rule-based construction of a new LST test version and investigates its psychometric properties via analysis with LLTM and CDM application to get insight into the basic construction principles and possible related person characteristics and into the adequacy of RC theory operationalization

through LST.

5.3 Method

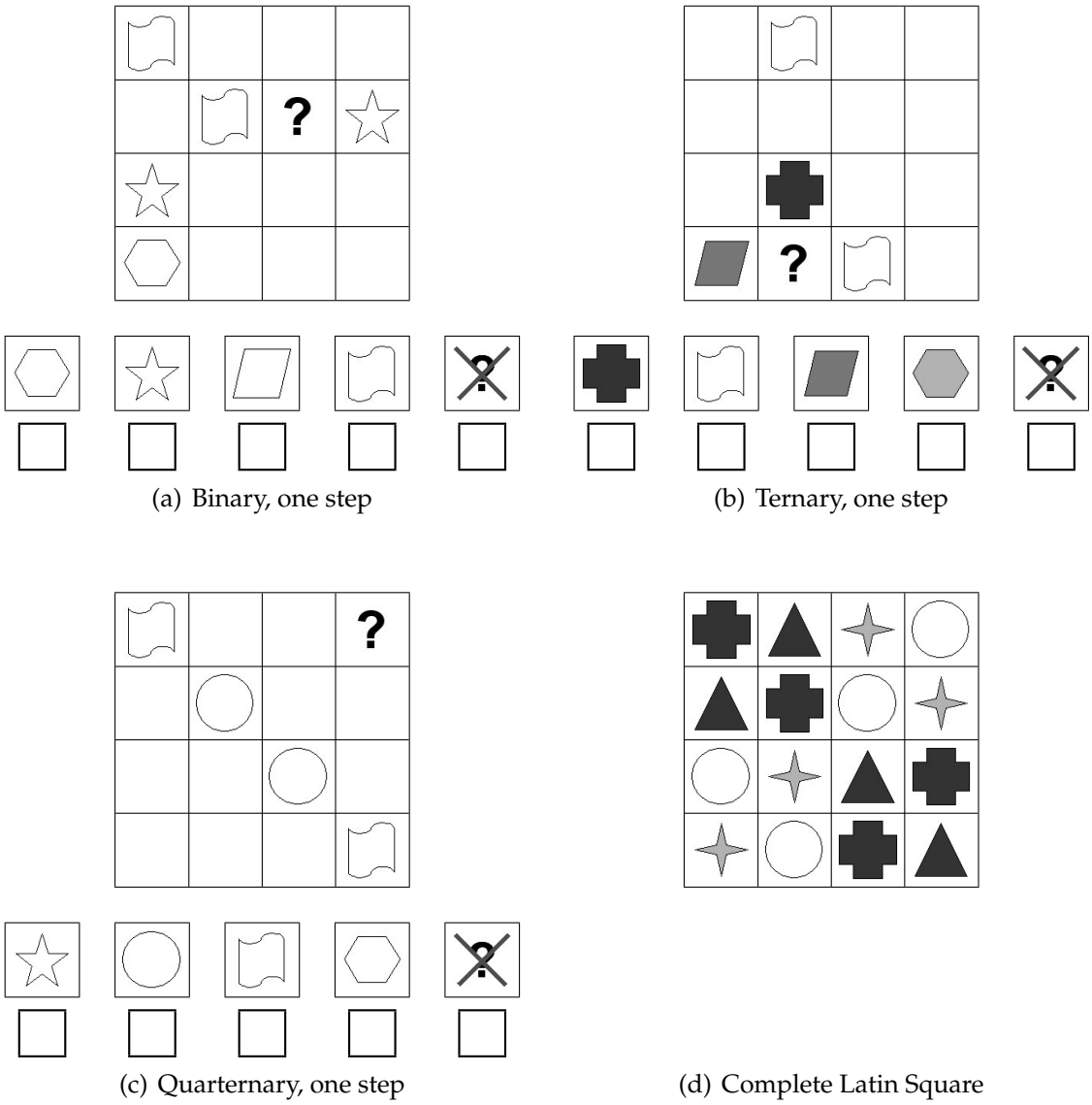
In the following sections, the methods of the current study are described in detail.

5.3.1 Item construction and design

LST items were constructed on the basic principles of RC theory, following the work of Birney et al. (2006). The complexity levels are defined by the number of lines and columns that have to be considered simultaneously. This resulted in a new test version containing 30 test items. Figure 5.1 depicts three (very easy) sample items to illustrate complexity level classification and phenotype of the items as well as a completed Latin Square for demonstration purposes. Solution of these example items according to RC theory principles and notation can be described as follows (let the columns called A, B, C and D from left to right and the lines 1, 2, 3 and 4 from top to bottom; underlining represents the separate chunks, see also Birney et al., 2006):

- Item a (two steps binary):
  1. AND(A1(wave), A3(star), A4(hexagon)) → A2(rhombus)
  2. AND(A2(rhombus), B2(wave), D2(star)) → C2(hexagon)
- Item b (one step ternary):
  AND(A4(rhombus), C4(wave), B3(cross)) → B4(hexagon)
- Item c (one step quarternary):
  AND(A1(wave), B2(circle), C3(circle)) → D1(circle)

The items are answered by marking the box below the symbol that fits into the cell with the question mark. Items are unambiguously solvable, there is only one right solution. Four items (contradictional items 8, 9, 24 and 29) were not solvable due to violation of the rules (that is, there is no possible solution which does not break the rule that every symbol must occur exactly once in every line and column). These four contradictional items were excluded from further analyses because the inherent cognitive processes are not clear enough to test the predictions of the RC theory, i.e., it is not clear how individuals prove the contradiction and



(a) Binary, one step (b) Ternary, one step

(c) Quarternary, one step (d) Complete Latin Square

Figure 5.1: Examples of Latin Squares with different complexity levels

if the cognitive steps involved in the proof are equivalent to the ones involved in solution processes of solvable items. These items were used to explore the cognitive steps and to allow for an extra solution alternative, the crossed question mark, in order to reduce guessing probability and to force examinees to prove their solutions cognitively. If examinees know that there are unsolvable items, they have to check their solution from the first to the last step regarding rule violation and solution alternatives even more carefully. Participants thus have to verify their answers as they have to monitor possible rule violation to exclude the crossed question mark as correct alternative. Otherwise, examinees might solve items only

half the way and guess the answer when arriving at a 50:50 chance decision.

Item construction followed a Q-matrix that was generated through optimal design and defines the number and complexity of processing steps based on RC theory. Items were constructed manually as paper-pencil test and checked for unambiguity by several scientific and student coworkers. Although it cannot be guaranteed for complete unambiguity, the items are well protected against alternative solution strategies. This is quite essential for further analyses because ambiguity would provide interpretational problems.

Birney et al. (2006) classified items according to their most complex step. However, considering only the overall number of the remaining processing steps means that much information is ignored because every item is labeled only with one complexity level instead of defining and analyzing all included processing steps. Thus in the current study all involved complexity levels are coded. This classification procedure was chosen to directly investigate the contribution of operations belonging to different complexity levels to item difficulty as well as the interaction between operations of different complexity levels and person characteristics instead of aggregating complexity levels. Additionally, "true" mental operations conducted by examinees when solving an LST item are not really clear yet. That is, considering the single complexity steps within one item rather than a merely supposed overall complexity classification seems to promise more information about the true effect of these steps on item difficulty and avoids drawing false conclusions due to non-expected solution ways. Additionally, more steps are involved in the items of the current study compared to Birney et al. (2006) who used at most only three steps. The items thus should be more difficult, what is very important for the selected samples of gymnasium pupils in this work to avoid ceiling effects.

For statistical analyses, the Q- matrix only contains the most necessary information. This ought to guarantee for parsimony and estimation efficiency while capturing the complexity information and not losing too much information content. Additionally, the contradictional items were left out. Table 5.1 shows the Q-matrix of the current study. Items are coded with respect to the number of different complexity steps involved (B1 means that one binary step is involved, B2 means that two binary steps are involved and so on). Please consider the appendix (section A.1 on page 164 ff.) for design details of the current version as well as of the longitudinal versions (as size of items or involved symbols).

Table 5.1: Q-matrix LST

Item	B1	B2	T1	T2	T3	T4	Q
Item 1	0	1	0	0	0	0	0
Item 2	0	0	1	0	0	0	0
Item 3	1	0	1	0	0	0	0
Item 4	0	1	1	0	0	0	0
Item 5	1	0	0	1	0	0	0
Item 6	1	0	0	1	0	0	0
Item 7	0	1	0	1	0	0	0
Item 10	1	0	0	0	1	0	0
Item 11	0	0	0	0	0	0	1
Item 12	0	0	0	0	0	0	1
Item 13	1	0	0	0	1	0	0
Item 14	0	0	0	0	0	0	1
Item 15	0	0	0	0	0	0	1
Item 16	0	1	0	1	0	0	0
Item 17	0	0	0	1	0	0	0
Item 18	0	1	1	0	0	0	0
Item 19	0	0	1	0	0	0	1
Item 20	0	1	0	1	0	0	0
Item 21	1	0	0	0	0	0	1
Item 22	0	1	0	0	1	0	0
Item 23	1	0	0	0	0	0	1
Item 25	0	1	0	0	1	0	0
Item 26	0	1	0	0	0	1	0
Item 27	0	0	0	1	0	0	0
Item 28	0	1	0	1	0	0	0
Item 30	0	0	1	0	0	0	1

Notes: B = binary, T = ternary, Q = quarternary. B1 = one binary step, B2 = two binary steps etc.

5.3.2 Selected statistical models

For analysis of LST results, several models were chosen (cf. section 4.2). LLTM, RE-LLTM and LR-LLTM results were investigated to gain information about the cognitive steps involved and about operationalization of RC theory requirements. Additionally, DINA as example of cognitive diagnostic models was applied to explore appropriateness of DINA for the current item type and get information about examinees' characteristics as attribute mastery classes and skill probabilities that allow for further conclusions about cognitive processes involved in solution of the LST items.

5.3.3 Research questions

It will be investigated which model class, LLTM or CDM, is more appropriate to explain empirical results for the current item type. Additionally, given the assumption that LLTM turns out to be the better choice, LLTM variants will be considered if they provide information about influence of person characteristics and basic parameter influence tendencies on item difficulty. Assuming that DINA is better suited, mastery classes and skill probabilities will be investigated for further information about item and person characteristics. Implications for RC theory operationalization will also be explained.

5.3.4 Test procedure

Participants were tested in groups of 20 to 30 persons. They were given a very generous time limit of 60 minutes for the whole test, so that the test was almost a real power instead of a speeded test. Participants were given an instruction introducing the item type and the rules (cf. appendix, section B.1 on page 178 ff.) and were allowed to ask questions before starting the test. They were not allowed to make any notes but were told to solve the items exclusively in mind. Examinees were asked to fill in a questionnaire about gender, age, school type (gymnasium or vocational school), school grades (math and German) and Sudoku experience. A subsample of 569 examinees received additional tests of fluid intelligence (CFT 20; Weiß, 1998), personality (NEO-FFI; Borkenau & Ostendorf, 1993), interests (AIST; Bergmann & Eder, 1999) and motivation (FAM; Rheinberg, Vollmeyer, & Burns, 2001).

Every examinee received a detailed feedback about her or his achievements in all finished tests a few weeks after testing (cf. section D in the appendix; however, note that this is the feedback example for the longitudinal study which is identical to the feedback of this study except for the results of BIS, d2, IST-2000-R and of the three additional LST versions).

5.4 Results

First, the sample of the current study is described. Then item characteristics and fit of items to the Rasch model are investigated, followed by application of LLTM and DINA and the results from these analyses.

5.4.1 Sample

930 German school students participated in the study. Individuals with solution patterns indicating non-serious working (like obvious abandonment of working after two thirds of all items or very fragmentary patterns) were excluded from further analyses, resulting in a total of 850 examinees for statistical analysis. Tables 5.2 and 5.3 show the demographic characteristics of the sample.

Table 5.2: Demographics part 1 LST sample

	Mean	SD	Min	Max
Age	17.94	1.10	16	26
Math grade	2.72	1.06	0.70	6.00
German grade	2.68	0.79	0.70	5.00
LST score	17.33	4.79	5	26

Notes: SD = Standard deviation, Min = minimum, Max = maximum.

Tables 5.4 and 5.5 show the demographic characteristics and test results for the subsample of 569 examinees who completed the additional tests CFT 20, NEO-FFI and AIST and FAM.

5.4.2 Item characteristics, dimensionality and item fit

Item characteristics from classical test theory and item fit indices for all 26 items (based on the whole 850 examinee sample) are shown in table 5.6. Cronbach's

Table 5.3: Demographics part 2 LST sample

	Number	Percent
Gender		
Male	335	39
Female	515	61
School type		
Gymnasium	713	84
Vocational school	137	16
Class		
11	235	28
12	286	33
13	329	39
Sudoku		
Experience	539	63
No experience	311	37

Alpha is .80. Item difficulty indicates that the test was relatively easy for the sample, discrimination indices can be regarded sufficient for most items.

Then, data were checked for dimensionality and RM item fit. Results from Winmira (Q-index, Rost & von Davier, 1994) revealed several misfitting items. For interpretation of the Q-index it has to be noted that usually p-values of the Q-indices imply RM underfit if they are smaller than .05 or .01, respectively, and RM overfit if they are greater than .95 or .99, respectively. RM underfit means that the ICC is too flat (item fits too bad), RM overfit means that the ICC is too steep (near to Guttman scale, item fits too good). Neither in Rost (2004) nor in Rost and von Davier (1994), definite advice is given if RM overfit is as problematic as RM underfit and which p-values for Q-indices should definitely lead to item exclusion. Moreover, it is not mentioned if Q-index depends on sample size, but my empirical experience implies clear dependence of Q-index on sample size: p-values indicating significant deviation from RM fit are more common for higher sample sizes. Due to this interpretational uncertainty, in the current work it is decided to exclude only items with RM underfit from analyses (p smaller than .01 because of relatively high sample sizes) because it can be assumed that RM underfit is more problematic than overfit and that too strict thresholds for p-values would lead to too conservative item exclusions.

Items 5, 13, 25, and 28, which had a bad Q-index (and additionally poor CTT

Table 5.4: Demographics part 1 subsample additional tests LST

	Mean	SD	Min	Max
Age	18.14	1.17	16	26
Math grade	2.77	1.09	0.70	6.00
German grade	2.64	0.79	0.70	5.00
LST score	17.90	4.76	5	26
CFT	115.49	8.23	87.33	131.33
NEO-FFI				
N	97.62	8.54	78.48	134.75
E	102.32	8.06	73.44	119.33
O	94.32	10.97	67.12	123.21
A	104.00	10.41	63.81	125.03
C	102.47	9.07	74.39	123.33
AIST				
R	96.91	9.63	70.00	129.00
I	98.42	9.47	70.00	126.00
A	101.47	9.59	70.00	128.00
S	101.71	9.75	70.00	128.00
E	103.41	9.58	72.00	130.00
C	101.55	9.86	70.00	130.00
FAM				
F	2.99	1.27	1.00	7.00
S	5.01	0.98	1.50	7.00
I	4.58	1.39	1.00	8.40
C	5.21	0.98	1.25	7.00

Notes: SD = Standard deviation, Min = minimum, Max = maximum. NEO-FFI: N = neuroticism, E = extraversion, O = openness, A = agreeableness, C = conscientiousness. AIST: R = realistic, I = investigative, A = artistic, S = social, E = enterprising, C = conventional. FAM: F = anxiety of failure, S = probability of success, I = interest, C = challenge.

Table 5.5: Demographics part 2 subsample additional tests LST

	Number	Percent
Gender		
Male	240	42
Female	329	58
School type		
Gymnasium	473	83
Vocational school	96	17
Class		
11	84	15
12	169	30
13	316	55
Sudoku		
Experience	419	74
No experience	150	26

discrimination indices), were excluded. Item 26 shows a significant Q-index (Rost & von Davier, 1994; Rost, 2004), but for 850 examinees the 5 percent level maybe too rigid and thus item 26 is not excluded. Table 5.6 shows item fit indices for all 26 items. After excluding these four misfitting items, Cressie Read (Read & Cressie, 1988; p = .09) and Pearson χ^2 (Plackett, 1983; p = .26) indicated no substantial deviation any more in the Winmira bootstrap statistics with 300 iterations. Additionally, Andersen Likelihood-Ratio-Test (Andersen, 1973) shows no significant examinee group differences (Andersen χ^2 = 26.16, df = 21, $p > .05$, groups defined by gender) and Martin-Löf-Test (Verhelst, 2001) shows no significant item group differences (Martin-Löf-statistics = 98.55, df = 120, $p > .05$, groups defined by even and odd item numbers). It can be concluded that the RM fit of the remaining items is sufficient to allow LLTM analyses.

5.4.3 LLTM results

Several LLTM variants were investigated. LLTM, RE-LLTM and LR-LLTM parameter estimates are given for the whole sample of 850 examinees. Additionally, AIC and BIC as well as the Log-Likelihood for LR tests are reported. These results are shown in table 5.7.

Note that for comparison of complexity impact only B1, T1 and Q can be compared

Table 5.6: Item difficulty, discrimination indices and Q-index LST

Item	Item difficulty (SD)	Item discrimination	Q-index	p Q-index
Item 1	.96 (0.20)	.13	0.25	.43
Item 2	.94 (0.24)	.07	0.32	.09
Item 3	.85 (0.36)	.41	0.13	.99
Item 4	.78 (0.41)	.40	0.17	.94
Item 5	.70 (0.46)	.22	0.28	.02
Item 6	.69 (0.46)	.29	0.25	.17
Item 7	.71 (0.45)	.28	0.25	.19
Item 10	.51 (0.50)	.36	0.22	.47
Item 11	.70 (0.46)	.54	0.11	.99
Item 12	.74 (0.44)	.36	0.20	.78
Item 13	.57 (0.50)	.22	0.29	.00
Item 14	.76 (0.43)	.40	0.18	.93
Item 15	.70 (0.46)	.55	0.11	.99
Item 16	.61 (0.49)	.44	0.17	.94
Item 17	.78 (0.42)	.22	0.26	.09
Item 18	.68 (0.47)	.33	0.23	.37
Item 19	.31 (0.46)	.37	0.20	.68
Item 20	.66 (0.48)	.39	0.19	.81
Item 21	.64 (0.48)	.44	0.17	.95
Item 22	.63 (0.48)	.40	0.20	.77
Item 23	.70 (0.46)	.25	0.27	.05
Item 25	.57 (0.50)	.24	0.28	.01
Item 26	.40 (0.49)	.26	0.26	.04
Item 27	.69 (0.46)	.42	0.17	.94
Item 28	.56 (0.50)	.16	0.32	.00
Item 30	.52 (0.50)	.34	0.22	.38

Notes: SD = standard deviation. Item difficulty and item discrimination are indices from classical test theory.

Table 5.7: LST parameter estimates for LLTM variants (N=850)

Parameter	LLTM (SE)	RE-LLTM (SE)	LR-LLTM (SE)	RE-LR-LLTM (SE)
Fixed effects				
Constant	3.91 (0.10)**	4.22 (0.33)**	3.24 (0.15)**	3.53 (0.35)**
B1	-0.46 (0.05)**	-0.53 (0.21)*	-0.46 (0.05)**	-0.53 (0.21)*
B2	-0.62 (0.06)**	-0.76 (0.23)**	-0.62 (0.06)**	-0.76 (0.23)**
T1	-1.70 (0.06)**	-1.78 (0.24)**	-1.70 (0.06)**	-1.78 (0.24)**
T2	-2.54 (0.08)**	-2.75 (0.30)**	-2.54 (0.08)**	-2.75 (0.30)**
T3	-3.02 (0.09)**	-3.22 (0.35)**	-3.02 (0.09)**	-3.22 (0.35)**
T4	-3.80 (0.11)**	-3.98 (0.43)**	-3.80 (0.11)**	-3.98 (0.43)**
Q	-2.67 (0.08)**	-2.95 (0.29)**	-2.67 (0.08)**	-2.94 (0.29)**
Gender			-0.19 (0.08)*	-0.19 (0.08)*
School type			0.46 (0.10)**	0.47 (0.11)**
Sudoku			0.63 (0.08)**	0.65 (0.08)**
Random effects				
Person	1.08 (0.07)	1.11 (0.08)	0.97 (0.07)	1.00 (0.07)
Item		0.11 (0.04)		0.11 (0.04)
Fit statistics				
LL (df)	-10016.49 (9)	-9886.56 (10)	-9977.72 (12)	-9847.36 (13)
AIC	20050.97	19793.13	19979.45	19720.72
BIC	20121.50	19871.49	20073.48	19822.59
$\Delta\chi^2$ to LLTM	-	259.86**	77.54**	338.26**

Notes: * p < .05, ** p < .01. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. RE-LLTM = Random Effects LLTM, LR-LLTM = Latent Regression LLTM, RE-LR-LLTM = combined Random Effects and Latent Regression LLTM. Gender: 0 = male, 1 = female; School type: 0 = vocational school, 1 = gymnasium; Sudoku: 0 = no experience, 1 = experience.

Table 5.8: LST parameter estimates for LLTM variants (N=569)

Parameter	LLTM (SE)	RE-LLTM (SE)	LR-LLTM (SE)	RE-LR-LLTM (SE)
Fixed effects				
Constant	3.98 (0.12)**	4.36 (0.40)**	2.92 (0.27)**	3.28 (0.47)**
B1	-0.52 (0.06)**	-0.63 (0.26)*	-0.52 (0.06)**	-0.63 (0.26)*
B2	-0.73 (0.07)**	-0.92 (0.28)**	-0.73 (0.07)**	-0.92 (0.28)**
T1	-1.71 (0.08)**	-1.80 (0.29)**	-1.71 (0.08)**	-1.80 (0.29)**
T2	-2.37 (0.10)**	-2.62 (0.36)**	-2.37 (0.10)**	-2.62 (0.36)**
T3	-2.86 (0.11)**	-3.09 (0.43)**	-2.87 (0.11)**	-3.09 (0.43)**
T4	-3.57 (0.13)**	-3.77 (0.53)**	-3.58 (0.13)**	-3.77 (0.53)**
Q	-2.55 (0.10)**	-2.88 (0.36)**	-2.55 (0.10)**	-2.88 (0.36)**
Math grade			-0.12 (0.05)**	-0.13 (0.05)**
Sudoku			0.44 (0.12)**	0.45 (0.12)**
CFT			0.30 (0.06)**	0.31 (0.06)**
FAM I			0.14 (0.04)**	0.14 (0.04)**
Random effects				
Person	1.17 (0.10)	1.22 (0.10)	0.93 (0.08)	0.98 (0.08)
Item		0.17 (0.06)		0.17 (0.06)
Fit statistics				
LL (df)	6536.30 (9)	-6422.24 (10)	-6484.44 (13)	-6369.70 (14)
AIC	13090.59	12864.48	12994.89	12767.39
BIC	13157.51	12938.83	13091.54	12871.48
$\Delta\chi^2$ to LLTM	-	228.12**	103.72**	333.20**

Notes: * p < .05, ** p < .01. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. RE-LLTM = Random Effects LLTM, LR-LLTM = Latent Regression LLTM, RE-LR-LLTM = combined Random Effects and Latent Regression LLTM. Math grade ranging from 1 to 6; Sudoku: 0 = no experience, 1 = experience; CFT = CFT 20 score (normalized and transformed to z-values); FAM I = FAM scale "Interest".

to each other as otherwise complexity and serial processing indicated by number of steps were confounded. To explore impact of serial processing, B1 can be compared to B2 and T1 can be compared to T2, T3 and T4, respectively.

Estimates for the basic parameters are completely in line with RC theory in every model variant: B1 has lower impact on item difficulty than T1, and T1 has lower impact than Q. More steps lead to more difficult items (T4 has higher impact than T3 which has higher impact than T2 and so on). The RE-LLTM fits significantly better than the LLTM ($\Delta\chi^2(1)$ = 259.86, p<0.001) which indicates that there must be some impact on item difficulty in the items which is not accounted for by the basic parameters defined in the Q-matrix. LR-LLTM fits significantly better than the LLTM and reveals effects of the second level person predictors gender, school type and Sudoku experience on item difficulty which results in decrease of the constant (because basic item difficulty is explained partly by examinee characteristics) and person variance (because differences between examinees are explained partly by some of their characteristics). Person variance (random effect id) is still high, however, which shows the necessity to include more second level person predictors. This is conducted with the 569 examinee sample in which examinees received additional tests (see below). In order to investigate second level predictors while modeling items as random effects, a combination of RE- and LR-LLTM was computed (RE-LR-LLTM) which has the best model fit of all specified models for 850 examinees.

To investigate the impact of more second level person variables (as additional test results) on item difficulty, analyses were repeated with the subsample of 569 examinees to allow for Likelihood-Ratio-Tests (LR-test) between models (LR-tests are only allowed if there is an identical number of observations in both models which is not the case in the 850 examinee sample because not every examinee received the additional tests). These results are shown in table 5.8.

The results again confirm the role of the basic parameters as in the sample with 850 examinees. Again the RE-LLTM fits significantly better than the LLTM, and so does the LR-LLTM. These results are not surprising because this is what should have been expected from the whole 850 examinee sample. Regarding additional second level predictors, only CFT score and the interest scale of the FAM have significant impact on item difficulty, all further FAM scales as well as all NEO-FFI and AIST scales are no significant second level predictors. Gender and school type are no significant predictors in this sample, perhaps because CFT score accounts for variance components shared with gender and/or school type. Math grade is

also a significant second level predictor in this sample.

Correlation between basic parameter estimates in all models for both samples is 1, i.e., order and relation of basic parameters remains stable across all model specifications.

To investigate possible interactions between basic parameters as well as between basic parameters and person characteristics, several models with interactions were specified again for both the full sample of 850 examinees and the subsample of 569 examinees. However, interaction effects did not lead to incremental explanation of variance in addition to person characteristics.

To evaluate explained variance by LLTM basic parameters, item location parameters were reconstructed from basic parameter estimates (simply sum up the products from one Q-matrix line per item with basic parameter estimates with inversed signs from LLTM analysis) and compared with Rasch item locations (from Winmira). Correlation between Rasch and LLTM item locations is .93 for the whole 850 examinee sample and .90 for the 569 examinee subsample. This means that about 87 percent of the whole variance in item difficulty is explained by the basic design parameters for the 850 examinee sample and 81 percent of the whole variance in item difficulty is explained for the 569 examinee subsample. Rasch and LLTM item locations and standard errors for both samples are summarized in table 5.9.

However, although correlation between parameter sets is described as proved way to obtain explained variance proportions by several authors (cf. Embretson, 1998; Freund et al., 2008; Preckel, 2003), this may lead to misinterpretations as variation of item parameters has also to be adequate. For this reason, absolute differences between Rasch and LLTM item locations are computed and standardized: LLTM item locations are sum-normalized and subtracted from Rasch parameters. The difference is then divided by Rasch SEs. The results are shown in table 5.10. Obviously, there are severe deviations between LLTM and Rasch item locations. This suggests that there have to be further important sources of variance in item difficulty apart from basic parameters or that the Q-matrix is misspecified.

All LLTM analyses were repeated with the partial triangular design matrix (see table 5.11 in the following section). Results are almost identical. The only difference is that B2, T2, T3 and T4 describe the gain in difficulty compared to the next lower step of the same complexity level.

Table 5.9: Rasch and reconstructed LLTM item location parameters and standard errors for LST

	N = 850				N=569			
Item	Loc. Rasch	SE Rasch	Loc. LLTM	SE LLTM	Loc. Rasch	SE Rasch	Loc. LLTM	SE LLTM
Item 1	-2.52	0.17	-3.30	0.007	-2.31	0.20	-3.25	0.011
Item 2	-2.07	0.15	-2.21	0.005	-2.39	0.21	-2.27	0.008
Item 3	-1.00	0.10	-1.75	0.005	-0.64	0.12	-1.75	0.008
Item 4	-0.50	0.09	-1.60	0.004	-0.24	0.11	-1.54	0.006
Item 6	0.04	0.08	-0.91	0.004	-0.08	0.11	-1.09	0.006
Item 7	-0.08	0.09	-0.76	0.003	-0.21	0.11	-0.88	0.005
Item 10	1.01	0.08	-0.43	0.005	0.91	0.10	-0.60	0.008
Item 11	-0.00	0.08	-1.24	0.003	0.00	0.11	-1.43	0.004
Item 12	-0.25	0.09	-1.24	0.003	-0.45	0.11	-1.43	0.004
Item 14	-0.33	0.09	-1.24	0.003	-0.56	0.12	-1.43	0.004
Item 15	0.01	0.08	-1.24	0.003	0.17	0.10	-1.43	0.004
Item 16	0.49	0.08	-0.76	0.003	0.49	0.10	-0.88	0.005
Item 17	-0.49	0.09	-1.37	0.004	-0.67	0.12	-1.61	0.006
Item 18	0.11	0.08	-1.60	0.004	0.20	0.10	-1.54	0.006
Item 19	2.06	0.09	0.46	0.004	2.13	0.10	0.28	0.006
Item 20	0.25	0.08	-0.76	0.003	0.38	0.10	-0.88	0.005
Item 21	0.36	0.08	-0.78	0.004	0.44	0.10	-0.91	0.006
Item 22	0.36	0.08	-0.28	0.005	0.47	0.10	-0.39	0.008
Item 23	0.01	0.08	-0.78	0.004	-0.08	0.11	-0.91	0.006
Item 26	1.56	0.08	0.50	0.009	1.52	0.10	0.32	0.013
Item 27	0.05	0.08	-1.37	0.004	0.14	0.10	-1.61	0.006
Item 30	0.94	0.08	0.46	0.004	0.80	0.10	0.28	0.006

Notes: Loc. = location, SE = standard error. LLTM item location parameters and standard errors reconstructed from basic parameter estimates and variances and covariances of estimates. Rasch results from Winmira, LLTM results from Stata.

Table 5.10: Absolute differences between Rasch and LLTM item locations for LST

	N = 850			N=569		
Item	LLTM sum normalized	Rasch	Diff. / Rasch-SE	LLTM sum normalized	Rasch	Diff./ Rasch-SE
Item 1	-2.29	-2.52	-1.32	-2.12	-2.31	-0.95
Item 2	-1.20	-2.07	-5.98	-1.14	-2.39	-5.96
Item 3	-0.74	-1.00	-2.52	-0.62	-0.64	-0.23
Item 4	-0.59	-0.50	1.02	-0.41	-0.24	1.48
Item 6	0.10	0.04	-0.70	0.04	-0.08	-1.20
Item 7	0.25	-0.08	-3.81	0.25	-0.21	-4.21
Item 10	0.58	1.01	5.45	0.53	0.91	3.81
Item 11	-0.23	-0.00	2.74	-0.30	0.00	2.81
Item 12	-0.23	-0.25	-0.25	-0.30	-0.45	-1.32
Item 14	-0.23	-0.33	-1.17	-0.30	-0.56	-2.27
Item 15	-0.23	0.01	2.82	-0.30	0.17	4.49
Item 16	0.25	0.49	3.02	0.25	0.49	2.39
Item 17	-0.36	-0.49	-1.41	-0.48	-0.67	-1.62
Item 18	-0.59	0.11	8.46	-0.41	0.20	5.87
Item 19	1.47	2.06	6.88	1.41	2.13	6.86
Item 20	0.25	0.25	0.00	0.25	0.38	1.22
Item 21	0.23	0.36	1.57	0.22	0.44	2.10
Item 22	0.73	0.36	-4.52	0.74	0.47	-2.69
Item 23	0.23	0.01	-2.64	0.22	-0.08	-2.87
Item 26	1.51	1.56	0.64	1.45	1.52	0.67
Item 27	-0.36	0.05	4.86	-0.48	0.14	5.90
Item 30	1.47	0.94	-6.63	1.41	0.80	-6.27

Notes: Diff. = difference; Diff./Rasch-SE = difference divided by Rasch-SE; SE = standard error.

5.4.4 CDM results

For DINA, a different Q-matrix has to be assumed. Because one could argue that results are not comparable due to different Q-matrices, as mentioned in the preceding section LLTM analyses were repeated with the partial triangular matrix structure and no differences emerged regarding LLTM results. Thus results are comparable despite different Q-matrices.

In DINA, thinking of different complexity levels as different skills (note the difference compared to LLTM which assumes one dimension, in this case working memory capacity) leads to a partial triangular structure: If more steps of the same complexity level are mastered, also less steps of the same complexity level have to be mastered. That means, number of steps cannot play a role in DINA as different numbers of steps of the same complexity level still belong to the same dimension of that complexity level. As explained in section 4.2.3, for DINA only skills themselves should impose difficulties, not number of skills. Table 5.11 shows this partial triangular structure.

CDM results can be shown in terms of estimated class probabilities, i.e. the probability that one specific attribute mastery class occurs in the sample. Additionally, skill probabilities are given, i.e. the probability that one specific skill is mastered in the whole sample. Model fit indices AIC and BIC are also mentioned. These results can be seen in table 5.12.

Class probabilities are mostly as one would expect from the theoretical concept and the prior analyses of the test. In DINA, the biggest class represents mastery of all attributes but T4, which can be explained by the easiness of the test and the relative difficulty of T4. The next class in probability order is the one in which all skills are mastered. The next class masters all skills but T4 and Q. The remaining classes are seldom to very seldom (less than about five percent). This class ordering shows that almost no class of considerable size consists of mastery of difficult attributes and non-mastery of easy attributes. That is, no class of considerable size masters T1 but not B1, or Q but not T1 and B1. Skill probabilities also overall confirm LLTM results which could have been expected as DINA smoothes the mixture model on one dimension and this one dimension smoothes the LLTM. The only difference between LLTM and DINA results so far is the order of B1 and B2 and the order of T3 and Q which is reversed for DINA compared to LLTM results (more examinees master B2 than B1, and more examinees master T3 than Q).

Table 5.11: Q-matrix DINA with triangular structure

Item	B1	B2	T1	T2	T3	T4	Q
Item 1	1	1	0	0	0	0	0
Item 2	0	0	1	0	0	0	0
Item 3	1	0	1	0	0	0	0
Item 4	1	1	1	0	0	0	0
Item 5	1	0	1	1	0	0	0
Item 6	1	0	1	1	0	0	0
Item 7	1	1	1	1	0	0	0
Item 10	1	0	1	1	1	0	0
Item 11	0	0	0	0	0	0	1
Item 12	0	0	0	0	0	0	1
Item 13	1	0	1	1	1	0	0
Item 14	0	0	0	0	0	0	1
Item 15	0	0	0	0	0	0	1
Item 16	1	1	1	1	0	0	0
Item 17	0	0	1	1	0	0	0
Item 18	1	1	1	0	0	0	0
Item 19	0	0	1	0	0	0	1
Item 20	1	1	1	1	0	0	0
Item 21	1	0	0	0	0	0	1
Item 22	1	1	1	1	1	0	0
Item 23	1	0	0	0	0	0	1
Item 25	1	1	1	1	1	0	0
Item 26	1	1	1	1	1	1	0
Item 27	0	0	1	1	0	0	0
Item 28	1	1	1	1	0	0	0
Item 30	0	0	1	0	0	0	1

Notes: B = binary, T = ternary, Q = quarternary

Table 5.12: LST model estimates for DINA

Class probabilities	
Class	DINA
0111101	0.010
1110101	0.011
1111011	0.012
0111100	0.013
1111001	0.016
1111011	0.016
1111000	0.024
1111010	0.024
1110000	0.030
1110100	0.030
1110010	0.030
1110110	0.030
0100010	0.035
1100010	0.035
1111100	0.071
1111101	0.297
1111111	0.272
Else	<0.01
Skill probabilities	
Skill	DINA
B1	0.91
B2	0.94
T1	0.91
T2	0.78
T3	0.78
T4	0.45
Q	0.62
Fit indices	
Index	DINA
AIC	19640
BIC	20451
Mean MADprop	0.06

Notes: Class probability: Probability that class of attribute mastery pattern occurs in sample. Attribute order in class: B1, B2, T1, T2, T3, T4, Q. Skill probability = probability that skill is mastered in the whole sample.

The estimation algorithm also presents item fits as well as slipping and guessing parameters for every item. These results are shown in table 5.13. Unfortunately, many guessing and slipping parameters are very high. Only eight out of 22 items reach even the liberal criterion of $1 - s - g > .50$, and of these, three have a slipping parameter of about .20 or higher.
MADprop statistics indicate a very good to somewhat good fit for most of the items except for items 7, 17 and 23. MADcor and MADLOR show some misfitting items (items 3, 4, 11, 15 and 27) compared to the remaining ones.

Table 5.13: LST item parameter and fit estimates for DINA

Item	Guess (SE)	Slip (SE)	(1-slip) -guess	MAD LOR	MAD cor	MAD prop
Item 1	0.88 (0.03)	0.03 (0.01)	0.09	0.63	0.04	0.04
Item 2	0.91 (0.03)	0.06 (0.01)	0.03	0.71	0.04	0.06
Item 3	0.19 (0.02)	0.04 (0.01)	0.77	1.63	0.16	0.01
Item 4	0.10 (0.01)	0.10 (0.01)	0.80	1.33	0.16	0.06
Item 6	0.49 (0.03)	0.23 (0.02)	0.28	0.67	0.11	0.08
Item 7	0.51 (0.03)	0.21 (0.02)	0.28	0.61	0.10	0.11
Item 10	0.23 (0.02)	0.34 (0.02)	0.44	0.84	0.13	0.02
Item 11	0.30 (0.05)	0.06 (0.01)	0.64	1.18	0.17	0.04
Item 12	0.48 (0.05)	0.10 (0.01)	0.42	0.91	0.14	0.07
Item 14	0.45 (0.05)	0.06 (0.01)	0.49	0.81	0.11	0.08
Item 15	0.28 (0.05)	0.05 (0.01)	0.67	1.19	0.18	0.01
Item 16	0.21 (0.03)	0.24 (0.02)	0.55	0.99	0.16	0.05
Item 17	0.65 (0.04)	0.18 (0.02)	0.17	0.36	0.04	0.16
Item 18	0.14 (0.01)	0.22 (0.02)	0.64	0.96	0.14	0.09
Item 19	0.11 (0.02)	0.57 (0.02)	0.32	0.92	0.05	0.06
Item 20	0.28 (0.03)	0.20 (0.02)	0.53	0.98	0.15	0.05
Item 21	0.38 (0.04)	0.20 (0.02)	0.41	1.11	0.16	0.06
Item 22	0.32 (0.03)	0.19 (0.02)	0.49	1.11	0.19	0.06
Item 23	0.59 (0.04)	0.23 (0.02)	0.18	0.36	0.05	0.12
Item 26	0.28 (0.02)	0.29 (0.05)	0.43	0.47	0.07	0.04
Item 27	0.30 (0.03)	0.17 (0.02)	0.54	1.18	0.20	0.05
Item 30	0.36 (0.03)	0.38 (0.02)	0.27	0.59	0.10	0.02
Mean	0.38 (0.03)	0.19 (0.02)	0.43	0.89	0.12	0.06

Notes: Guess = guessing parameter, slip = slipping parameter, SE = standard error, (1-slip)-guess = "CDM discrimination", MAD = Mean absolute difference, LOR = log odds ratio, cor = correlation, prop = proportion.

Additionally, classification probabilities and skill mastery probabilities for every single examinee are given which is very helpful for feedback and learning purposes on the individual level. As already mentioned in section 4.2.2, this is one special feature of cognitive diagnostic models. However, demonstration of these results would go beyond the scope and space of this work and therefore are not reported in detail here.

5.5 Discussion

The LST represents a relatively new operationalization of the RC theory which is able to measure working memory capacity and fluid intelligence as well as reasoning abilities while being rather parsimonious, non-verbal and non-domain specific. The current work investigates LST modeling with LLTM and CDM variants with regard to its complexity characteristics and psychometric properties. 850 German school students participated in the study and completed a LST test version which was constructed rule-based. Design principles and person characteristics influencing item difficulty are investigated by means of linear logistic test models and cognitive diagnostic models.

5.5.1 LLTM results

Complexity manipulation conducted by combination of processing steps of varying complexity can be regarded as successful in the current LST test version. LLTM analyses show that basic parameters which are supposed to be less complex than others also have less impact on item difficulty. That is, binary steps show the least impact on item difficulty, ternary steps have more impact and quarternary steps which are supposed to be the most complex steps in the current test have the highest effect on item difficulty (only comparing the same number of steps of one complexity level, that is B1 with T1 and Q, B2 with T2). As expected, more steps of one complexity level have higher impact on item difficulty than less steps of the same complexity level.

Reconstruction of LLTM item locations shows that correlation between LLTM and Rasch item parameters is .93 (for the 850 examinee sample) and .90 (for the 569 examinee subsample). This means that basic parameters explain 87 and 81 percent of variance in item difficulty, respectively. This can be regarded a good result

compared to usual findings for intelligence tests (cf. Embretson, 1998; Freund et al., 2008; Preckel, 2003). However, as already mentioned, absolute differences between Rasch and LLTM item locations indicate severe deviations for LLTM reconstruction. The fact that Rasch and LLTM item locations are not identical is not surprising as LLTM locations are only based on included basic parameters and their estimates for every item. This means that items with the same combination of basic parameters always get the same LLTM item location and that variations beyond these effects cannot be mapped. Obviously there are additional factors with impact on item difficulty apart from the investigated basic parameters. These additional effects are not investigated which results in deviations between Rasch and LLTM item locations. Another possible reason can be a misspecified Q-matrix. Deviations from the ideal Q-matrix reproducing item difficulty perfectly can thus also result in lack of explained variance.

However, these deviations should be interpreted with caution: As the LLTM splits item difficulty into a few basic parameters and so uses fewer parameters than the Rasch model, it will probably never be able to reconstruct Rasch item locations perfectly without any deviations. Please note also that the perfect Q-matrix rarely exists. In many cases, the Q-matrix which reproduces data perfectly, would have to assign an own basic parameter to every item which in turn would mean that LLTM application is senseless in these cases. To sum up the above mentioned ideas, it can be stated that correlation between Rasch and LLTM item locations indicates a general good localization, but for many items there are high absolute differences. As for a couple of items these differences are very high, it can be concluded that there have to be additional important impact factors beyond basic parameters which affect item difficulty or that the Q-matrix suffers from severe misspecifications. However, regarding basic parameter impact, the fixed LLTM application provides almost the same conclusions as the RE-LLTM given that the Q-matrix is not misspecified. That is, for practitioners who only want to identify basic parameter impact, it does in principle not matter if they consider LLTM or RE-LLTM results.

Including additional random effects for items leads to better model fit which is a rather common finding (cf. de Boeck, 2008) and completely in line with the deviations between Rasch and LLTM item locations. The random item effect captures variance which is not accounted for by the basic parameters. This shows (together with the absolute differences between LLTM and Rasch item locations) that there seems to be additional variance in item difficulty beyond that explained

by the basic parameters. However, the random item effect is small compared to the random person effect and therefore may also account for some kind of random noise in the data rather than for systematic design shortcomings (compare also the high correlation between item locations for Rasch and LLTM indicating a general good localization). The correlation between fixed effects for the basic parameters between LLTM and RE-LLTM is approximately 1, i.e., regarding interpretation of basic parameter effect on item difficulty it does not matter if one considers LLTM or RE-LLTM parameters. The fact that the RE-LLTM fits better therefore provides rather theoretical than practical insight.

Including second-level person predictors in the LR-LLTM reveals several variables with significant influence on item difficulty (gender, school type, Sudoku experience, math grade, CFT scores and interest measured by FAM). Including these predictors leads to a decreasing variance of the random person effect as well as to a lower constant. This result together with the improved model fit shows that person predictors contribute considerably to explanation of item solution processes. The significant person predictors can be explained in line with known results from other studies which show that intelligence (represented by the CFT in the current study) and school grades (math grade in this case) are closely connected to reasoning and working memory (e.g. Ackerman, Beier, & Boyle, 2005; Colom et al., 2008, 2007). The higher CFT value and the better the math grade, the higher the person's cognitive abilities and thus the higher item solution probability for an examinee. The significant school type effect favoring gymnasium students is not surprising, too, because it can be assumed that gymnasium students in general will show higher cognitive abilities than vocational school students (when CFT is included as predictor, school is not a significant impact factor any more). Sudoku experience has the highest second order predictor effect and points out the necessity to measure familiarity with a task type to avoid misinterpretations of test scores due to different levels of practice of similar or identical tasks. Interest as measured by the FAM has also significant impact. The higher the interest score, the higher is item solution probability for an examinee. Interest as operationalized in FAM may include facets of need for cognition and striving for knowledge as well as collecting new experiences. This may lead to intensified efforts in working on LST items.

Gender as a significant effect was not expected since there are no explicit hints that men do better in working memory tasks and intelligence tests (e.g. Colom & García-López, 2002; Lynn & Irwing, 2002). However, as was shown by McGlone,

Aronson, and Kobrynowicz (2006) and Su, Rounds, and Armstrong (2009), there are gender differences in specific interest domains and interest is linked to recall and learning rate (Schiefele & Krapp, 1996). Thus, perhaps male examinees developed more interest in the item type or in the general domain of such puzzles as Sudoku and therefore scored higher in LST. This effect emerged although already controlling for Sudoku experience and interest in models. However, this gender effect was only significant (and rather low compared to other effects) in the 850 examinee sample and was not existent in the 569 examinee sample which shows that this effect might not be stable.

Taking into account the measurement scale of the person predictors (gender, school type and Sudoku experience are dichotomous, but CFT, math grade and interest have nearly interval scales with different ranges and standard deviations), most of them have still lower impact on item difficulty than the chosen basic parameters.

No further effects of personality (NEO-FFI), interest facets (AIST) or motivation (FAM) were found, possibly due to the pure and carefully constructed item type which mainly captures cognitive abilities. The fact that motivational, personality and further person characteristics have no or only little impact on item difficulty (compared to basic parameters) shows that the theoretical concept is well operationalized and empirically mapped by the current items. However, note the high absolute differences between Rasch and LLTM item locations and the extreme mode fit improvement when modeling items as random effects. Thus there seem to be still parts of variance in item difficulty which are not explained by the selected parameters. The complete RE-LR-LLTM has the best fit of all variants for the described item type and sample and shows a detailed picture of impact factors for item difficulty. However, it has to be pointed out that the main purpose of LLTM analysis is and remains identification and confirmation of basic parameters. For rule-based item construction this is the main information. The described additional variables help to resolve the question which characteristics apart from basic parameters make items difficult but are not essential for evaluating the item construction process. In principle, all described LLTM variants provide enough information to evaluate the construction process.

However, changes in random item and person variance as well as changes in constant values for the full models can be interpreted in terms of different parameter effects: Inclusion of person characteristics leads to reduction of random person variance as well as of the constant. This shows that the chosen person characteristics explain parts of interindividual differences between examinees

and of basic item difficulty. This means that person characteristics are identified which cause items to be more or less difficult for specific groups of examinees with certain characteristics.

Additionally, model fit improvement by inclusion of further parameters allows for conclusions about relations between item and person characteristics concerning their impact on item difficulty: Including random item effects leads to extreme model fit improvement and including person characteristics leads to far less model fit improvement. Hence it can still be concluded that in LST much random item variance is left which cannot be completely explained by the current study design and the included variables. A possible explanation could be that random item and person variance in LST simply reflect impact of surface characteristics or different ability levels in reasoning and working memory capacity or pattern recognition which are not captured by any other variables. Another possible reason could be severe misspecification of the Q-matrix.

5.5.2 CDM results

Cognitive diagnostic models focus on the examinees rather than on the items and provide attribute mastery classes of examinees with certain skill profiles which indicate which requirements or attributes of a task, again specified in a design matrix, have been mastered and which not.

CDM results for the DINA model from the current study at first sight mainly confirm LLTM results regarding attribute mastery. Typical attribute mastery patterns show that there are mainly classes of examinees who can handle all of the basic parameters or who can handle less complex steps but not more complex steps at the same time. Overall skill mastery probabilities indicate that most examinees master binary steps, less master ternary steps and the smallest skill mastery probability is found for quarternary steps (only comparing B1, T1 and Q). Comparing B1 and B2 as well as T2 and T3 suggests that B2 is mastered by more examinees than B1, and T2 is mastered by as many examinees as T3. However, because of the partial triangular structure of the Q-matrix this is in fact an impossible result. This possibly indicates problems inherent in the software routine used here.

However, guessing and slipping parameters in the current study are very high. From an interpretational point of view, for DINA guessing means the probability

of giving a correct answer although less than all attributes required to solve an item have been mastered. Assuming that model application leads to interpretable results, one can conclude that if guessing parameters are too high for many items, examinees may have reached their score through guessing rather than through mastery of the basic parameters. If slipping parameters are too high, too many examinees who have mastered all attributes fail although they should be able to solve the item.

It can be stated that ideally, both slipping and guessing parameters should be low. Applying a rather liberal criterion of 1-g-s being higher than 0.50 shows that only eight of 22 items reach this criterion. Among these, three items show slipping parameters of .20 or higher. At first sight, the high guessing and sipping parameters do not make sense: Especially the items which are very easy (indicated by CTT and IRT, for example items 1, 2 and 3) show extremely high guessing parameters. As it can be assumed that these very easy items are not only mastered by guessing strategies (the included cognitive requirements are very low and DINA results themselves indicate mastery of the included easy attributes), this result cannot be interpreted seriously.

But how can the current high parameters be explained? Tatsuoka (1990) analyzed a sample of 2144 examinees who completed 30 items, and guessing and slipping parameters shown in the analysis of this data set by de la Torre and Douglas (2004) are better than for the current data, but far from good values. Only 18 of 30 items reach the liberal criterion, and of these 18 items 10 show too high slipping parameters. Templin and Henson (2006) used a data set with 593 examinees and 41 items, and there are also very high guessing and slipping parameters shown in their results. 28 of 41 items reach the liberal criterion, and of these 28 items, 22 have too high slipping parameters.

The estimation algorithm used in the current study was validated by comparing results from the used R macro with the results of de la Torre (2008), both result sets based on the same data set. The R macro reproduced the parameters nearly without errors. Thus, software and implementation explanations can be excluded.

Do the high guessing and slipping parameters imply that the chosen cognitive basic parameter set is not adequate since generally high guessing and slipping parameters indicate that the chosen set leads to poor mapping of examinee behavior? In the current case, for several reasons it can be assumed that the chosen basic parameters are adequate. First, LLTM analyses show definitely the impact of

the chosen basic parameters with more than sufficient significance. Additionally, LLTM parameters explain about 87 percent of the variance in Rasch item difficulties. However, note that this only indicates that the chosen basic parameters are important factors influencing item difficulty. As Rasch parameters show higher variance than LLTM parameters and this variance is included into estimation of person parameters, person parameters are not explained as well by the chosen basic parameters. Moreover, CDM class and skill probabilities confirm the theoretical assumptions of item construction principles and RC theory in most points. MAD indices and CDM discrimination indices are quite good, too.

Templin and Ivie (2006) analyzed the Raven's Progressive Matrices (RPM; tested with a sample of N = 1364) with DINA and got similarly poor guessing and slipping parameters compared to the LST items. The authors point out that there seems to be a correlation between item difficulty and guessing as well as slipping parameters: They report a correlation of -.95 between (1-s) and item difficulty and conclude from their results that both parameters seem to depend on item difficulty: Guessing is high if item difficulty is low, slipping is high if item difficulty is high. In fact, these two parameter types are not independent of each other which results in slipping being rather high and guessing rather low for one item and vice versa. In the current study, correlation between guessing parameters and Rasch item difficulty is .73 (slipping .76). So, one possible reason for the high parameters of the current test can be seen in item difficulty (mainly too easy items, cf. also CTT and Rasch results in tables 5.6 and 5.9).

Another, and probably more logical, reason for the bad guessing and slipping parameters which can be deducted from the ideas in the theoretical background chapter (cf. section 4.2.3) is that the current item type is based on a strictly unidimensional construct and thus CDM application is not appropriate. CDM application is adequate for concepts and skills which are learnable and trainable, as CDMs are discrete models mapping the structure of contents itself without including a resource dimension on an intermediate level. LLTMs are unidimensional models with content only emerging by assignments to resources. Extreme high guessing parameters are observed for especially easy items with few or only one rule, extreme high slipping parameters are observed for the more difficult items combining several rules. This pattern is typical if only rule combinations, but not the rules themselves, provide difficulties for problem solving. Guessing and slipping parameters thus reveal inadequacy of DINA for LST items. This may also be the reason for the parameter values found by Templin and Ivie (2006).

This reason may also apply to the above mentioned studies which also found high guessing and slipping parameters. Even for the fraction-subtraction data which can be regarded using concepts which are learn- and trainable and thus could in principle represent single skills, guessing and slipping parameters of many items are not satisfying. As these problems seem to occur often in empirical application of DINA and point to inherent problems of the model rather than of the items, its qualification and usefulness for empirical application in these certain contexts may be questioned in general.

5.5.3 Comparison of LLTM and CDM

High guessing and slipping parameters of DINA imply inappropriateness of DINA application for the current data (confer the above discussed reasons). As stated in section 4.2.2, poor guessing and slipping parameters do not automatically mean that CDM application is not adequate or does not map examinee behavior adequately. However, it can be concluded that for the current item type DINA is not appropriate as guessing and slipping parameters do not provide seriously interpretable results and because LST is based on a unidimensional construct and rule combination rather than the rules themselves sets limits for item solving. The result that guessing and slipping parameters are that high although attributes are mastered by a high proportion of examinees shows a severe contradiction within model assumptions. The consequence of the strictly deterministic linkage in DINA is absorption of uncertainty in guessing and slipping parameters. For highly automated processes and speeded tests as well as well-defined learn- and trainable skills, this phenomenon may be reasonable (examinees master all attributes but tend to slip because of time pressure, for example, or they have learned one skill but not the other), but in the current study neither of these cases applies. Thus it has to be concluded that DINA fails to explain the current data adequately and LLTM has to be preferred.

LLTM rationale shows up to be the more reasonable rationale in data explanation for LST: Basic parameter inclusion makes items more difficult. Model and item fit indices as well as explained variance underline adequacy of LLTM application in this case. LLTM results demonstrate satisfying explanation of item difficulty and detect important processes for item solution. Thus it has to be reasoned that LLTM is superior and preferable to DINA in this study.

5.5.4 RC theory, item construction and operationalization by LST

In general, operationalization of RC theory by the current LST version seems to be satisfying and in line with the results of Birney et al. (2006). The current study expands the so far findings about LST to influence of random effects, second level person predictors and cognitive diagnosis. Based on the results of former studies and the current work, this task type is a promising tool to measure working memory capacity, reasoning abilities and fluid intelligence while being an excellent non-verbal and non-domain specific operationalization, successful implementation and application of RC theory. Rule-based generation and Q-matrix design can in general be regarded successful as basic parameters are confirmed to be significant predictors of item difficulty for LST items. Additionally, basic parameters remain stable difficulty influencing factors even when additional test results and person characteristics as well as interactions with these other variables are included.

5.5.5 Limitations and prospects

Regarding statistical modeling, it has to be noted that only DINA as one example of CDMs was applied. Thus the stated conclusions cannot be transferred to other members of the CDM-family. However, the discussed points show that CDMs in general may not be adequate to model data from item types as LST.

Limitations of LST can be seen in the possible ambiguity. Although several rounds of quality and unambiguity verification were conducted, it cannot be guaranteed that all subjects applied identical solution strategies. Moreover, automatic and software-based construction of LST is complicated at the moment because of the relative complicated effect of each step on each other step when combined in an item. For example, a former ternary step can become binary by including another ternary step which provides one additional element for the first ternary step. Construction has to be accomplished really carefully and it is definitely necessary to check each hand-written item several times for the actual complexity steps. Nevertheless, compared to several other item types (e.g., Raven matrices, Raven, 1938, 1962), LST is a parsimonious, transparent, efficient, and relatively easy to construct item type. Software applications which provide automatic generation, cloning and adaptive presentation of LST are currently under construction.

Since LST items of the current study are relatively easy, especially for gymnasium

students, construction of more difficult items would be desirable. A possible way of increasing item difficulty is to use grids with more lines and columns as number and combination possibilities of complexity steps is limited in grids with four or five lines and columns as described in the current study. Successful construction of bigger grids is described by Gold (2008) who used six lines and columns and could show that item difficulty increases while item quality is still good.

Expansion studies about LST are desirable, for example culture comparing studies to investigate cultural fairness. Another important point can be the field of learning and training effects which could help to shed light on the debate if fluid intelligence can be trained, and if gains in intelligence tests which are g-loaded actually reflect a gain in g (for these subjects confer Reeve & Lam, 2005 or te Nijenhuis, van Vianen, & van der Flier, 2007, for example). The second empirical study of the current work (see chapter 6) will concentrate on another facet of learning effects in LST and investigates the basic parameters in longitudinal LLTMs. This will enable us to evaluate the design and operationalization as well as the stability of parallel LST versions over time. Moreover, complexity level specific learning effects can be investigated.

6 Longitudinal modeling of the Latin Square Task

After demonstrating rule-based item construction as well as LLTM and CDM modeling for LST in the preceding study, the current study is concerned with longitudinal modeling of LST items.

6.1 Introduction

Learning effects are of great importance for test and item construction with regard to construct stability, validity and score-based decisions. Repeated testing as it is often allowed in selection settings can provide severe problems if learning effects obscure true ability. On the other hand, learning effects provide helpful insights into solution processes as well as into the theoretical basis of a test and its operationalization especially in the case of rule-based item construction. Longitudinal LLTM versions do exist, but do not allow parameter specific modeling. Thus a longitudinal LLTM variant which provides insight into parameter specific learning processes is constructed by specific learning parameters and applied to longitudinal data from 304 German school students who received four rule-based constructed LST versions in a longitudinal test setting. Results show Rasch scalability of all four test versions. LLTM analyses show interesting parameter specific learning effects as well as effects of several person characteristics and additional test results on intercept but not on slopes of learning curves. DINA application leads to the conclusion that again DINA is not qualified to model LST data and that the longitudinal results are not interpretable at all. Application hints regarding repeated testing and test construction for practitioners and researchers are pointed out.

Learning effects are very important for psychological assessment, especially for intelligence and achievement tests. Learning effects regarding test diagnosis play a role not only in science, but also in personnel psychology and employee selection as well as in admission contexts for school, university and apprenticeship.

On the one hand, learning effects provide important insights into cognitive processing and solution strategies. Additionally, cognitive theories underlying tests can be investigated with regard to their validity and operationalization. On the other hand, taking into account possible learning effects is practically necessary for personnel selection issues if applicants are allowed to repeat admission or selection tests.

6.2 Background

Learning effects can be classified into effects which appear across several test trials with the same or slightly different test materials which are repeatedly given to the examinees, and effects which describe adaptation to changes within test materials. Adaptation gains more and more interest in research. Especially in working world, employees have to handle changes in their complex work environments. Adaptation to such dynamic processes is of great importance for success and (monetary) output. Individual differences which enhance successful adaptation to changes in work environments have thus become an important topic in research (cf. Lang & Bliese, 2009). Adaptation to change is mostly measured by the task-change paradigm in which examinees are confronted with novel or complex tasks. Aspects of these tasks change unexpectedly while the examinee is occupied with skill acquisition to reach a predefined degree of task mastery. So adaptation to these changes is required (for examples and an overview see Lang & Bliese, 2009). Additionally, learning during working on a test without the above mentioned changes can occur. In this case, some (or all) items affect solution of the following ones within the same test.

However, these kinds of learning effects are not in the focus of the current study. The current study is mainly occupied with practice and training effects across and not within test versions which will be explained in the following sections.

6.2.1 Practice and training effects

Another important difference within learning effects is the one between training effects which denote a desired better achievement after a purposeful training of, for example, specific cognitive components or test contents, and practice effects which denote rather undesired higher raw scores through simple test exposure effects.

Practice effects

Ironically, undesired practice effects, that is score gains in intellectual and achievement tests, occur quite often while purposeful trainings often do not produce stable and generalizable desired outcomes. It is widely known that raw scores rise for many test types if an examinee takes a test more than once. Such raw score gains were observed for different time intervals, ranging from a few days up to several months, and for several test types, for example the General Aptitude Test Battery GATB (cf. Jensen, 1998 and te Nijenhuis et al., 2007), Scholastic Assessment Test SAT and the American College Testing Program Assessment ACT (Coyle, 1998), and the Raven Matrices (Bors & Vigneau, 2003; Nkaya, Huteau, & Bonnet, 1994). Kulik, Bangert-Drowns, and Kulik (1984) included 40 studies into their meta-analysis and report effect sizes up to 0.89. Score gains were higher for identical practice and criterion tests, more practice sessions and high-performers. The meta-analysis of Hausknecht, Halpert, Di Paolo, and Moriarty Gerrard (2007) shows that practice effects for identical tests are almost twice as high as for parallel tests. Moreover, the longer the learning phase, the higher score gains. Cliffordson (2004) investigated score gains through repeated testing of the Swedish Scholastic Aptitude Test (SweSAT) and claims that score gains are due to practice and training effects as well as growth (maturation) between test sessions. She also illustrates detailed self-selection and cohort effects and concludes that practice effects have greater impact than growth.

Training effects

Training of intellectual achievement has been studied for many decades. If it was possible to train intelligence, cognitive strategies and intellectual achievement, these trainings would be promising for hundreds and thousands of people who want to improve their school, university and job career chances as well as for many children who suffer from social and educational disregard. Given the high interrelation of intelligence and many outcome variables like school and job performance as well as learning rate (see the frequently cited meta-analysis of Schmidt & Hunter, 1998), training of intelligence seems to be a promising tool. Many extensive training programs were conceptualized during the past, often concentrated on training of cognitive strategies and adjustment of educational disadvantages. Relatively successful examples for chance adjustment of disadvantaged children are the Head Start Programme (Levitt & Dubner, 2005), the Milwaukee-Project

(Garber, 1988) and the Abecedarian Project (Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002), although results are ambivalent. Unfortunately, often outcomes were not generalizable and stable over time (see Coyle, 1998; Jensen, 1998; Reeve & Lam, 2005; te Nijenhuis et al., 2007). However, there are some examples for successful training effects and transfer of these effects: Jaeggi, Buschkuehl, Jonides, and Perrig (2008) present evidence for transfer of training effects for working memory tasks to g_f. Effects of working memory training in preschool children are reported by Thorell, Lindqvist, Bergman Nutley, Bohlin, and Klingberg (2009). They show transfer effects from trained tests to non-trained tests of spatial and verbal working memory and attention. However, no significant effects on working memory or attention were found for inhibition training tasks.

Can g be learned?

Stable general gains of IQ scores (as a measure of g) through so far developed trainings and generalization of gains seem not to be possible. Jensen (1998) declares that targeted trainings do not lead to higher IQ scores. However, there seem to occur large spontaneous inter- and intraindividual fluctuations in IQ. Facing the lack of reliable predictors for these fluctuations, Jensen (1998) assumes the IQ to be more elastic than it was thought before and fluctuations to be genetically determined. In fact, some of the above mentioned training programs for children show higher IQ scores of trained children and young adults. However, these gains are very low taking into account the IQ scale and seem to fade during the years (for example Campbell et al., 2002). It has to be noted that the projects partly included medical and nutritional actions and thus IQ score gains, if substantial ones are observable at all, cannot be ascribed exclusively to the cognitive training programs. Interestingly, te Nijenhuis et al. (2007) note that biological interventions as nutrition and health care could be more effective for IQ score increase than psychological or educational activities. Moreover, they suppose some kind of barrier between the first and second stratum (cf. Three-Stratum-Theory, Carroll, 2005; or broad and narrow abilities in the CHC-Theory, cf. McGrew, 2005) which is responsible for the phenomenon that training effects do not pass the level of specific abilities.

As described above, learning effects are a common observation in testing practice. However, it is not clear yet what exactly is learned during training or practice. If g, or g_f was learn- or trainable, effects should be transferable to many or almost

all achievement and content areas and should be relatively stable. The above mentioned studies mainly describe score gains through training as well as practice, but no or little substantial generalization and stability of effects. Therefore, it can be concluded that score gains and thus learning effects usually are limited to test-specific abilities (Coyle, 1998; Jensen, 1998; Reeve & Lam, 2005) and mostly cannot be transferred from one test to another (te Nijenhuis et al., 2007). Jensen (1998) concludes that learning effects remain on the specific ability level and fade the more the hierarchy is climbed up until in g they are almost not existent at all. These test specific gains can be called "empty" with regard to g.

Problems caused by learning effects

Given the fact that score gains do not automatically reflect true ability improvements, more or less extensive score gains during repeated testing can be a serious threat to test validity, especially in selection situations. Often, applicants are allowed to repeat a test if they did not pass it the first time. Simple practice effects through material exposure then can lead to wrong decisions in selection processes as the examinee who only scores high because of test exposure effects is not appropriate for the job, but is selected based on his or her test result. Another problem in this context are specific training programs which are offered internet based or in form of training groups or other materials and aim at preparation of university or job applicants for specific tests used during the selection process. This can cause higher scores which do not longer indicate the true ability level of the examinees.

Slightly different to judge are programs and trainings which aim at elimination of construct irrelevant disturbing factors as nervousness or at compensation of disadvantages caused by social factors. For example, Agbor-Baiyee (2009) describes a training procedure for the Medical College Admission Test (MCAT) to improve the chances of disadvantaged and underrepresented minority groups. Additionally, providing applicants with pre-test information and preparation materials leads to improved perceptions of fairness and satisfaction with the test process among individuals who do not pass the test, but not to better overall pass rates or more positive overall examinee reactions (Burns, Siers, & Christiansen, 2008). Hausknecht et al. (2007) did not find evidence for reduced validity (concerning criteria as school grades) through repeated tests, and Reeve and Lam (2005) show that neither the factorial structure nor reliability and criterion oriented validity changes in repeated testing. Thus repeated testing does not seem to affect the

quality of measurement instruments the authors apply. However, te Nijenhuis et al. (2007) point out that g-loading (in common IQ-tests they examined) decreases during repeated testing which in turn indicates a different construct being measured compared to former test sessions.

For practitioners, allowing repeated testing should be considered carefully (cf. Hausknecht, Trevor, & Farr, 2002). Even if applicant order regarding achievement level does not change through repeated testing, a cut-off can be reached by applicants who do not truly possess the abilities their score indicates because they just profit by learning effects, no matter if "simple" practice or intentional training. This is even more dangerous if personnel selection procedures are based only on such test results (fortunately, usually they are not). Additionally, applicants who take the test the second or third time have an unfair advantage compared to those who take the test for the first time.

Therefore, it is important for existing and newly constructed tests of intellectual achievement to investigate possible learning effects. This can help to gain insights into cognitive processes and theoretical conceptualization of the test as well as to identify and eliminate possible sources of undesired practice effects. Additionally, corrective actions for results from repeated testing can be developed. Possible ceiling effects and validity problems through repeated testing or test training have to be investigated to make sure that the test maintains sufficient validity and difficulty level, for example.

Learning from learning

The problems caused by learning effects described in the previous section are only one side of the veil. Practice effects which emerge during repeated testing provide helpful insights into cognitive processes the examinees are engaged in during the tests as well as important hints for theoretical issues and hypothesis testing: Which parts of an item are responsible for learning effects and how can this be interpreted against the background of theoretical test basis and former results from other studies? If it is known what and how individuals learn, test construction and administration can be supervised according to this knowledge and be made more robust against undesired effects.

Application of rule-based item construction, item cloning and automatic item generation can help to reduce undesired practice and memory effects through

repeated testing. Ceiling effects occurring after repeated testing can be avoided by including an appropriate range of item difficulties and perhaps by taking into account the practice level of an examinee.

Methods like rule-based item construction, item cloning and automatic item generation promise at least partly relief from these dangers. Because items can theoretically be generated uniquely for each examinee as soon as the item generation process has been implemented and guarantees for valid and reliable items, practice and training effects can considerably be reduced. However, certain basic similarities between unique items will probably not be avoidable because the underlying construct has to be mapped adequately in all generated items. Moreover, Bors and Vigneau (2003) found evidence for learning with regard to the underlying test and item type rather than for single specific items. Thus, learning effects probably cannot totally be avoided during repeated testing of the same abilities.

Also an analysis of mistakes can provide helpful insights into learning processes as Bors and Vigneau (2003) show for the Raven Matrices: Learning effects seem to be due to fluctuation in answers as items which had been solved during the first session were not solved in the subsequent one and vice versa, leading to low reliability on item and high reliability on test level.

A common problem in longitudinal studies are drop-outs and self-selection effects. Usually, not all individuals who participate in the first trial also attend all of the following ones (either they completely quit participation after one of the trials, or they are absent in one or more trials because of illness, appointment problems or something else). That means statistical models have to cope with missing data or the data set is reduced by the number of individuals who did not participate in all sessions which would mean a waste of data sets (and of working time and money, too). Another practical and interpretational problem is self-selection: Drop-outs often are systematically due to specific person variables (e.g. abilities, achievement level, conscientiousness, motivational issues). This results in (often undesired) sample selection effects and reduced variance at least in some person variables. Examinees who continue participation in all sessions thus often constitute a selected sample. Excluding missing data cases is often necessary because of analytical and statistical requirements, but leads to selected sample problems. This should be kept in mind when planning, executing, analyzing and interpreting longitudinal studies and their results.

Statistical and methodological issues in longitudinal data analysis

If one is interested in careful investigation of learning effects, simple pre-post measurements, usually analyzed by analysis of variances, are not sufficient. As Cliffordson (2004) mentions, learning often does not stop after the second test session, but still occurs between the second and third trial. However, after the third test session, learning effects often seem to diminish extremely. Thus, at least three measurement points should be scheduled. Additionally, non-linear slopes (which is a common finding given the described fact of diminishing learning effects from the first to the third or fourth test session) can only be detected if more than two points are available. In this case, longitudinal models are very helpful for analysis of two or more than two measurement points. Examples are longitudinal multilevel models (Byrne & Crombie, 2003; Curran, Bauer, & Willoughby, 2004; Rogosa, Brandt, & Zimowski, 1982; Singer & Willett, 2003) or longitudinal Rasch models (Glück & Indurkhya, 2001; Pastor & Beretvas, 2006; Rijmen, de Boeck, & van der Maas, 2005). Longitudinal multilevel models allow detailed mapping and modeling of longitudinal data and identification of predictors and wave forms for arbitrary numbers of measurement points and lengths of time intervals. Additionally, they provide adequate consideration of within-person correlation of data points across different time points.

There are also variants of the LLTM which allow longitudinal analysis of data. The LLTM with relaxed assumptions (LLRA, cf. Fischer, 1989; Formann & Spiel, 1989; Glück & Spiel, 1997, 2007) was developed explicitly for measurement of change and for longitudinal comparisons between control and experimental groups (Formann & Spiel, 1989) and does not require all items to be unidimensional. Equation (6.1) describes the probability that person j solves item i on $t1$ and $t2$, respectively:

$$\begin{aligned} P(X_{ij1} = 1) &= \frac{\exp(\theta_{ij})}{1 + \exp(\theta_{ij})} \\ P(X_{ij2} = 1) &= \frac{\exp(\theta_{ij} + \delta_j)}{1 + \exp(\theta_{ij} + \delta_j)} \end{aligned} \tag{6.1}$$

with θ_{ij} the ability of person j with regard to the dimension item i measures, and δ_j the total sum of all changes of person j between measurement points $t1$ and $t2$. This means one change parameter each is set for every person and every item.

However, the LLRA cannot be applied if changes are mainly unidirectional (indicating a clear tendency in one direction). For such cases, the Hybrid LLRA (Formann & Spiel, 1989; Glück & Spiel, 1997) can be used. The Hybrid LLRA requires for both measurement points items of different difficulties which are arranged pairwise (both items of a pair measuring the same latent dimension, i.e., pairwise Rasch fit is required). Then the changes in ability on $t2$ can be compensated by the higher item difficulties and ceiling effects as well as parameter estimation divergency are avoided. The Hybrid LLRA states the probability that person j solves item i on $t1$ and $t2$ as following:

$$\begin{aligned} P(X_{ij1} = 1) &= \frac{\exp(\theta_{ij} - \sigma_i)}{1 + \exp(\theta_{ij} - \sigma_i)} \\ P(X_{ij2} = 1) &= \frac{\exp(\theta_{ij} - \sigma_i' + \delta_j)}{1 + \exp(\theta_{ij} - \sigma_i' + \delta_j)} \end{aligned} \tag{6.2}$$

with σ_i the difficulty of item i presented on time point one, and σ_i' the difficulty of the parallel item i presented on time point two (Formann & Spiel, 1989).

LLRA as well as Hybrid LLRA can easily be extended to more than two groups and measurement points. For both LLRA and Hybrid LLRA, Formann and Spiel (1989) demonstrate a decomposition of δ_j into $\delta_j = \sum_{v=1}^{m} q_{jv}\eta_v + \tau$ with η_v the (unknown) effect of treatment T_v, q_{jv} the (known) amount of treatment T_v given to person P_j, and τ the trend. This decomposition demonstrates the similarity to the original LLTM equation (4.8) on page 16: One model parameter (in LLRA δ, in LLTM σ) is decomposed into a linear combination of underlying basic parameters. Here we find the difference in the main focus of LLTM and LLRA: In LLTM, item difficulty is of main interest, in LLRA group differences as well as treatment and time effects are of main interest. So, LLRA and Hybrid LLRA only provide global instead of basic parameter specific change parameters and thus they are not suited to gain knowledge about parameter-specific learning processes.

As shown in this section, all above mentioned models do not offer the opportunity to model possible longitudinal change of single basic parameters. However, investigating longitudinal properties of basic parameters as used in rule-based item construction would probably provide insights into the solution processes and could help to answer the question what exactly is learned during repeated testing.

6.2.2 LST and learning effects

Learning effects in LST have been investigated first by Hoffmann (2007), but only for sum scores. Significant learning effects were detected, the highest score gains were observed between the first and second session, lower gains between the second and third session, and no significant score gains emerged between the third and fourth trial. In addition to a significant time predictor, several second level variables were found to have significant impact on total scores. While several variables seem to affect the intercept (school type, intelligence, figural memory, Sudoku experience, math grade, interest, concentration), almost no predictors were found to affect slopes (concentration has little impact). It can be concluded that learning effects occur still after the second trial and slopes are almost not affected by the measured person variables.

No linear logistic modeling has been applied to LST before this thesis and thus the data gathered by Hoffmann (2007) are now analyzed with linear logistic models and DINA. This will provide more detailed insight into solution processes and learning effects by decomposing effects into basic parameter specific parts. The results will help to gain information in order to answer the question what exactly is learned when examinees are tested more than once with LST.

6.3 Method

In detail, item construction, description of test sessions and the general approach as well as first longitudinal modeling examples are described in Hoffmann (2007). Therefore, the following sections contain only the most important information.

6.3.1 Item construction and design

Item construction and design can be retrieved in principle from chapter 5. In the current study, three additional test versions were used, resulting in four parallel test versions overall. All versions are structurally identical regarding basic parameters (cf. table 5.1) but differ in phenotype (symbols, colors and arrangement of symbols). Design details of all four test versions can be found in the appendix in tables A.1 to A.5. In principle, this can be seen as an item cloning approach as described in chapter 7, with the basic parameters as radicals and the phenotypes consisting of symbols and color as incidentals.

All versions were controlled by several student and academic coworkers to eliminate alternative solution paths. Despite this careful construction and controlling process, single ambiguities cannot be fully excluded. Several prestudies as well as diploma theses could show that surface characteristics as color or symbols seem not to influence LST item difficulty to an important extent and that parallel versions of LST are of sufficient equality. For example, Pauls (2009) investigated two LST test versions and used an equating approach to prove task interchangeability. He also found that basic parameters provide satisfying explanation of item difficulty.

6.3.2 Selected statistical models: LLTM with learning parameters

As summarized in section 6.2.1, with common longitudinal models no mapping of basic parameter specific longitudinal changes is possible. Thus, the current study has to apply another procedure. It is adapted from Rost (2004) who defines learning parameters for within-test learning modeled by LLTM application: Every item which exerts a learning effect on the following counts for the learning parameter. One can define several impact variants, depending on the hypotheses about learning effects. In the current study, this approach is exactly applied to basic parameter specific learning effects across test versions. The defined learning parameters are described in section 6.4.3.

Models with time predictors will then be compared to models without time predictors. First, the RM in equation (4.7) can be rewritten as

$$P(X_{ij\tau} = 1|\theta_j, \sigma_i) = \frac{\exp(\theta_j - \sigma_i)}{1 + \exp(\theta_j - \sigma_i)} \tag{6.3}$$

This model assumes that no learning effects occur. Then it is extended to a longitudinal RM (L-RM) which assumes that general learning effects occur:

$$P(X_{ij\tau} = 1|\theta_j, \sigma_i) = \frac{\exp(\theta_j - \sigma_i + \tau)}{1 + \exp(\theta_j - \sigma_i + \tau)} \tag{6.4}$$

Similarly, the LLTM in equation (4.8) can be rewritten as

$$P(X_{ij\tau} = 1 \mid \theta_j, q, \eta) = \frac{\exp\left(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k\right)}{1 + \exp\left(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k\right)} \tag{6.5}$$

Again this model assumes no learning effects but decomposes item difficulty into basic parameters. Then it is extended to the longitudinal LLTM for basic parameter specific learning across test versions (LBP-LLTM) which adds basic parameter specific learning parameters (described in section 6.4.3):

$$P(X_{ij\tau} = 1 \mid \theta_j, q, \eta) = \frac{\exp\left(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k + \tau_k\right)}{1 + \exp\left(\theta_j - \sum_{k=1}^{K} q_{ik}\eta_k + \tau_k\right)} \tag{6.6}$$

To investigate if basic parameters together with a general learning effect explain the empirical results better than basic parameter specific learning effects, a simple longitudinal LLTM (L-LLTM) is computed:

$$P(X_{ij\tau} = 1 \mid \theta_j, q, \eta) = \frac{\exp\left(\theta_j + \tau - \sum_{k=1}^{K} q_{ik}\eta_k\right)}{1 + \exp\left(\theta_j + \tau - \sum_{k=1}^{K} q_{ik}\eta_k\right)} \tag{6.7}$$

τ acts as a time predictor and denotes time impact. In the L-RM and the L-LLTM, τ describes general time impact, in the LBP-LLTM it helps to describe basic parameter specific time impact (and thus basic parameter specific learning effects). Comparing the RM to the L-RM with time predictor helps to identify general learning effects, that is, if item difficulty increases or decreases as a whole (or, differently speaking, if person ability increases or decreases in general across time). Comparing the LBP-LLTM to the original LLTM and the simple L-LLTM yields insights into basic parameter specific developments across time. Thus stepwise inclusion of time predictor, basic parameters and basic parameter specific learning parameters will provide information about importance and aspects of learning effects in LST.

For these longitudinal models, the xtmelogit procedure in Stata provides a helpful tool: By the inherent multilevel mixed effects structure, an extra definition of time or panel variables is not necessary. The person id serves as indicator for data grouping within persons, the time variable(s) can be included as usual fixed or random effects. This means the intercept can be modeled independently of time variables. Slopes can only be modeled by separate learning parameters.

CDM modeling with DINA imposes several difficulties for the current longitudinal data structure. There is no possibility to directly include time variables or to specify the longitudinal data structure in another way to apply DINA in

a longitudinal manner. However, restrictive assumptions about local stochastic independence across time allow re-computation of skill probabilities from individual person classification and can provide first insights into skill probability development across time. This procedure is only a little help but will show which effects can be detected by DINA or longitudinal LST. Single CDMs for all measurement points would be no correct solution since intercorrelation of data within persons cannot be taken into consideration and thus leads to deficient estimation results. Additionally, sample size is not sufficient to reach stable and reliable CDM estimation results. As long as no save longitudinal CDM application exists, seriously interpretable results cannot be expected.

Because there is no direct possibility to take into account the longitudinal data structure in DINA, only skill probabilities can be shown for the longitudinal LST data. For this purpose, skill probabilities were recomputed from person classifications. This procedure is possible based on the assumption that item parameters remain constant and changes can only be conducted by changes from one state to another for one person. Then we can describe

$$P(X_{i(j\tau)} = 1|\xi_{i(j\tau)}) \tag{6.8}$$

with

$$\xi_{i(j\tau)} = \prod_{k=1}^{K} \alpha_{(j\tau)k}^{q_{ik}} \tag{6.9}$$

with $\alpha_{(j\tau)k} = 1$ if person j possesses skill k at time point τ. Then $\alpha_{(j\tau)k}$ describes all allowed changes, and repeated measures are local stochastic independent in this modification of the model. The DINA model can now be described similarly to equation (4.12):

$$P(X_{i(j\tau)} = 1 \mid \xi_{i(j\tau)}) = (1 - s_i)^{\xi_{i(j\tau)}} g_i^{1-\xi_{i(j\tau)}} \tag{6.10}$$

with

$s_i = P\left(X_{i(j\tau)} = 0 \mid \xi_{i(j\tau)} = 1\right)$, $g_i = P\left(X_{i(j\tau)} = 1 \mid \xi_{i(j\tau)} = 0\right)$ and $\xi_{i(j\tau)} = \prod_{k=1}^{K} \alpha_{(j\tau)k}^{q_{ik}}$.

A posteriori person classification can then be used to recompute skill probabilities for every time point.

6.3.3 Research questions

It will be investigated how definition of learning parameters in LLTM application can help to resolve learning effects in LST. For this purpose, several (sets of) learning parameters are defined and compared to each other in order to gain information about the nature of the underlying learning effects. Additionally, reconstruction of skill probabilities from DINA application is used to investigate possibilities of longitudinal DINA application.

6.3.4 Time schedule and test procedure

It was decided to use four test versions and thus four measurement points for the longitudinal study because this allows for adequate investigation of learning effects: Two measurement points would have been not sufficient because learning often still occurs after the second test session, more than four measurement points would have been exaggerated because in general it can be assumed that no substantial learning takes place after the fourth session (cf. Cliffordson, 2004; Hoffmann, 2007). Three to four measurement points allow to investigate possible non-linear slopes and ceiling effects as well as a statistically appropriate mapping of learning effects.

Overall, three test sessions of different duration took place with a time distance of six to eight days between sessions. During the first two sessions, one LST version was tested with the examinees (versions one and two), in the third session, LST was given twice (versions three and four). A fourth session was abstained from to avoid higher drop-outs and motivational problems. Because of the different time intervals between the trials (only about half an hour between trial three and four), the fourth trial has to be considered with caution: Learning effects between the third and fourth trial are probably not of the same quality as the remaining effects between the other trials. If analyses reveal inadequate results for the fourth trial, this trial can be excluded from analysis while still keeping enough time points to detect non-linear effects. For statistical modeling, different intervals do not cause any difficulties. Different time variables can be specified (cf. section 6.4.4). Additional tests were administered once during the three sessions:

- Fluid intelligence: CFT 20 (Weiß, 1998)
- Figural memory: BIS (Jäger, Süß, & Beauducel, 1997) and I-S-T 2000 R subtests (Amthauer, Beauducel, Brocke, & Liepmann, 2001)

- Concentration: Test d2 (Brickenkamp, 2002)
- Personality: NEO-FFI (Borkenau & Ostendorf, 1993)
- Interests: AIST (Bergmann & Eder, 1999)
- Motivation: FAM (Rheinberg et al., 2001)
- Demographics: Questionnaire (age, gender, school form, school class level, Sudoku experience, math grade)

Giving all additional tests in each session was not possible due to test economic reasons. Moreover, growth effects can almost be excluded for most of the tests during the given overall time interval of about two weeks and thus it can be assumed that test results only would differ because of physical and mental state or similar random influences on the day of testing. Additionally, only LST and no other learning effects are the focus of the described study. The time schedule of the study can be seen in figure 6.1.

Examinees were tested in groups of 20 to 40 individuals. They received detailed instructions for LST (see appendix, section B.1) as well as for all other tests (see test manuals of additional tests). Again, examinees were not allowed to make any notes but were told to solve items exclusively in mind for LST. Examinees were offered a test training with positive training effects for typical selection and test settings as well as individual detailed feedback of all test results (cf. appendix, section D). At the end of session three, a strategy questionnaire was given to all examinees. In this questionnaire, they were asked to describe their strategies (if existent) in solving LST items and to fill in the order of steps they conducted in four example items. Additionally, an elaborate debriefing took place and all examinees were encouraged to ask questions about the test sessions and contents.

6.4 Results

First, the sample of the current study is described, followed by item characteristics and RM fit of items. Specification of time variables and learning parameters is explained, and then LLTM and DINA results are reported.

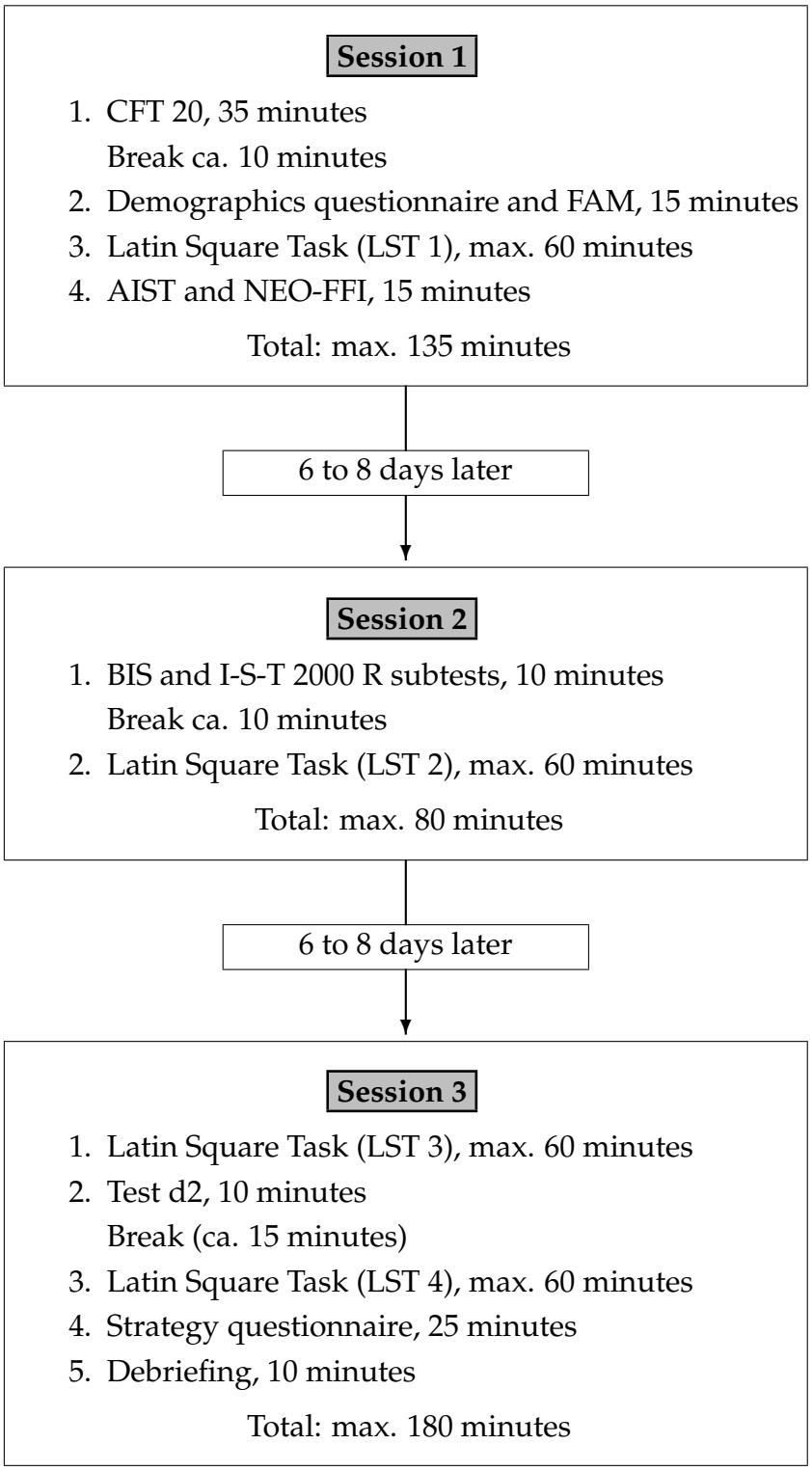


Figure 6.1: Time schedule of the longitudinal LST study

6.4.1 Sample

342 German school students participated in the study. Individuals who did not participate in all three sessions were excluded from further analyses to make estimation more efficient and stable, resulting in a total of 304 examinees for statistical analysis. There were no significant differences between excluded and included examinees. Tables 6.1 and 6.2 show the demographic characteristics of the sample and the additional test results.

Table 6.1: Demographics LST longitudinal study

	Number	Percent
Gender		
Male	130	43
Female	174	57
School type		
Gymnasium	216	71
Vocational school	88	29
Class		
11	51	17
12	102	33
13	151	50
Sudoku		
Experience	218	72
No experience	86	28

6.4.2 Item characteristics, dimensionality and item fit

Item characteristics from classical test theory and item fit indices for all 26 items and four test versions are shown in table 6.3 (again, contradictional items were left out from further analyses). Cronbach's Alpha is .76 for version one, .79 for version two, .73 for version three and .78 for version four. Item difficulty indicates that all test versions were relatively easy for the sample, discrimination indices can be regarded too low for several items, perhaps due to test easiness. Items one to three do not provide any information because of their extreme low difficulty. All subsequent analyses were repeated without these items which revealed no substantial differences in results Therefore items one to three are kept for subsequent analyses and explanations. There are some unexpected results

Table 6.2: Demographics and additional test results LST longitudinal study

	Mean	SD	Min	Max
Age	18.33	1.31	16	26
Math grade	2.75	1.16	0.70	6.00
LST 1 score	18.68	4.22	5	26
LST 2 score	20.51	4.09	4	26
LST 3 score	21.97	3.14	11	26
LST 4 score	21.75	3.43	6	26
CFT	121.90	12.57	81.00	147.00
d2 KL	109.58	11.79	75.00	130.00
BIS	95.33	7.51	76.33	121.00
IST	108.73	7.24	83.00	118.00
NEO-FFI				
N	97.60	8.46	78.48	126.10
E	102.07	8.35	73.44	119.33
O	94.29	10.55	68.72	120.00
A	103.23	10.45	63.81	125.03
C	101.40	9.12	77.04	120.69
AIST				
R	97.70	8.97	75.00	129.00
I	98.31	9.08	73.00	121.00
A	102.29	8.82	74.00	126.00
S	102.30	9.30	70.00	126.00
E	104.18	8.82	82.00	129.00
C	102.73	9.47	75.00	130.00
FAM				
F	3.02	1.26	1.00	7.00
S	5.12	0.99	1.50	6.75
I	4.54	1.25	1.00	7.00
C	5.23	0.96	1.25	7.00

Notes: SD = Standard deviation, Min = minimum, Max = maximum. NEO-FFI: N = neuroticism, E = extraversion, O = openness, A = agreeableness, C = conscientiousness. AIST: R = realistic, I = investigative, A = artistic, S = social, E = enterprising, C = conventional. FAM: F = anxiety of failure, S = probability of success, I = interest, C = challenge.

regarding development of item difficulty: Between the third and fourth LST session, there is an increase in difficulty for many items. This is assumed to indicate severe motivational and concentrational problems within examinees (see also later argumentation and results). Additionally, some items show unexpected developments in difficulty: For item 17 and 26, difficulty decreases between session two and three. For item 22, there is no difficulty change between session two and

three. And for item 30, there is no change in difficulty between session one and two, but a high decrease between session two and three. These developments may indicate impact of factors beyond basic parameters and learning parameters, for example surface characteristics. However, inspection of surface characteristics revealed no abnormalities.

Results from Winmira show no misfitting items with regard to the Q-index (Rost & von Davier, 1994; Rost, 2004). Andersen Likelihood-Ratio-Test (Andersen, 1973) shows only significant examinee group differences for school type (time 3 and 4) as well as for Sudoku experience (time 2). Martin-Löf-Test (Verhelst, 2001) shows no significant item group differences for several item grouping methods (even-odd, quarternary, first ten vs. remaining items for practice effects within test versions), see also table 6.4 for Andersen and Martin-Löf results.

6.4.3 Time variables and learning parameters

To investigate the longitudinal structure and learning effects, several sets of learning parameters were chosen. These sets can be divided into two groups: General time effects and trends.

General time variables allow conclusions about general learning across all test versions while trend variables help to model the intervals between test sessions and thus provide information about changes occurring between single test versions. Additionally, general time variables only provide linear developments while trend variables help to identify non-linear changes. The three trend variables indicating changes between test versions can be described as follows:

- LST session one: trend1 = 0, trend2 = 0, trend3=0
- LST session two: trend1 = 1, trend2 = 0, trend3=0
- LST session three: trend1 = 1, trend2 = 1, trend3=0
- LST session four: trend1 = 1, trend2 = 1, trend3=1

The general time variable is 0 for LST version one, 1 for LST version two, 2 for LST version three and 3 for LST version four.

In addition to these general time and trend parameters which are defined independently from basic parameters, basic parameter specific learning parameters are identified. Comparison of models using these different learning parameter

Table 6.3: Item difficulty, discrimination indices and Q-index LST longitudinal

	LST 1				LST 2				LST 3				LST 4			
Item	Dif (SD)	Dis	Q	p Q	Dif (SD)	Dis	Q	p Q	Dif (SD)	Dis	Q	p Q	Dif (SD)	Dis	Q	p Q
Item 1	.94 (0.23)	.14	0.26	.44	.96 (0.20)	.18	0.24	.48	.96 (0.20)	.08	0.31	.42	.96 (0.20)	.16	0.29	.44
Item 2	.94 (0.24)	.04	0.34	.20	.91 (0.28)	.02	0.35	.10	.93 (0.26)	-.06	0.39	.14	.94 (0.24)	-.06	0.45	.09
Item 3	.96 (0.20)	.21	0.20	.61	.94 (0.24)	.23	0.20	.54	.96 (0.19)	.14	0.26	.50	.95 (0.21)	.23	0.22	.54
Item 4	.86 (0.35)	.15	0.28	.23	.89 (0.31)	.15	0.27	.23	.92 (0.27)	.21	0.20	.54	.95 (0.22)	.10	0.31	.35
Item 5	.72 (0.45)	.29	0.22	.47	.85 (0.36)	.23	0.23	.34	.88 (0.33)	.25	0.18	.54	.80 (0.40)	.11	0.29	.10
Item 6	.74 (0.44)	.28	0.23	.42	.86 (0.35)	.36	0.17	.65	.91 (0.28)	.21	0.21	.50	.92 (0.28)	.30	0.18	.59
Item 7	.72 (0.45)	.30	0.22	.47	.76 (0.43)	.38	0.17	.59	.79 (0.41)	.32	0.17	.55	.84 (0.37)	.31	0.19	.48
Item 10	.60 (0.49)	.27	0.24	.27	.70 (0.46)	.20	0.25	.11	.73 (0.45)	.24	0.20	.37	.73 (0.44)	.30	0.19	.44
Item 11	.83 (0.38)	.42	0.14	.89	.88 (0.32)	.36	0.15	.71	.92 (0.27)	.42	0.09	.78	.95 (0.22)	.43	0.09	.76
Item 12	.78 (0.42)	.40	0.16	.83	.87 (0.34)	.41	0.13	.78	.93 (0.25)	.39	0.10	.75	.96 (0.20)	.40	0.10	.74
Item 13	.59 (0.49)	.32	0.21	.52	.64 (0.48)	.21	0.26	.08	.66 (0.47)	.20	0.23	.24	.69 (0.46)	.34	0.19	.45
Item 14	.80 (0.40)	.42	0.15	.87	.84 (0.37)	.48	0.12	.84	.94 (0.23)	.36	0.12	.71	.92 (0.27)	.37	0.16	.66
Item 15	.77 (0.42)	.46	0.13	.93	.86 (0.35)	.46	0.12	.83	.93 (0.25)	.39	0.10	.76	.96 (0.20)	.43	0.07	.77
Item 16	.69 (0.46)	.32	0.21	.55	.78 (0.42)	.33	0.18	.51	.83 (0.38)	.27	0.19	.50	.82 (0.39)	.34	0.17	.59
Item 17	.79 (0.41)	.26	0.23	.44	.90 (0.30)	.22	0.23	.42	.88 (0.32)	.29	0.17	.60	.93 (0.25)	.22	0.24	.46
Item 18	.77 (0.42)	.25	0.25	.34	.84 (0.37)	.32	0.20	.49	.88 (0.33)	.19	0.23	.38	.88 (0.32)	.34	0.16	.61
Item 19	.36 (0.48)	.36	0.19	.64	.62 (0.49)	.41	0.15	.65	.86 (0.35)	.35	0.14	.70	.56 (0.50)	.23	0.22	.21
Item 20	.77 (0.42)	.21	0.26	.26	.79 (0.41)	.38	0.16	.64	.87 (0.34)	.31	0.17	.60	.87 (0.34)	.43	0.12	.77
Item 21	.73 (0.45)	.35	0.19	.68	.81 (0.40)	.55	0.08	.93	.83 (0.37)	.24	0.22	.41	.93 (0.26)	.37	0.15	.67
Item 22	.68 (0.46)	.33	0.20	.59	.77 (0.42)	.28	0.21	.34	.77 (0.42)	.32	0.16	.60	.76 (0.43)	.22	0.23	.24
Item 23	.68 (0.47)	.32	0.20	.57	.79 (0.41)	.45	0.13	.81	.87 (0.34)	.35	0.14	.68	.84 (0.37)	.46	0.11	.80
Item 25	.61 (0.49)	.27	0.24	.30	.73 (0.45)	.31	0.20	.37	.77 (0.42)	.20	0.23	.28	.77 (0.42)	.30	0.19	.44
Item 26	.42 (0.50)	.28	0.23	.30	.53 (0.50)	.36	0.18	.45	.49 (0.50)	.19	0.24	.20	.54 (0.50)	.35	0.17	.50
Item 27	.78 (0.42)	.33	0.20	.65	.84 (0.37)	.28	0.21	.43	.92 (0.28)	.27	0.17	.62	.91 (0.29)	.27	0.20	.53
Item 28	.56 (0.50)	.19	0.28	.10	.57 (0.50)	.24	0.24	.13	.66 (0.47)	.20	0.22	.28	.67 (0.47)	.31	0.19	.42
Item 30	.59 (0.49)	.26	0.23	.35	.59 (0.49)	.47	0.13	.77	.87 (0.34)	.33	0.15	.66	.71 (0.46)	.43	0.14	.70

Notes: Dif = difficulty, SD = standard deviation, Dis = discrimination, Q = Q-index, p Q = p Q-index. Item difficulty and item discrimination are indices from classical test theory.

Table 6.4: Overview Andersen and Martin-Löf results for longitudinal LST

	And. Chi^2	df	p	M.-L.-stat.	df	p
	Gender			**Even odd**		
LST session 1	32.40	25	$> .05$	95.27	168	$> .05$
LST session 2	31.12	25	$> .05$	82.45	168	$> .05$
LST session 3	25.46	25	$> .05$	59.48	168	$> .05$
LST session 4	26.39	25	$> .05$	61.74	168	$> .05$
	Age			**First 10 items**		
LST session 1	15.44	25	$> .05$	107.26	159	$> .05$
LST session 2	20.15	25	$> .05$	70.71	159	$> .05$
LST session 3	12.06	25	$> .05$	47.85	159	$> .05$
LST session 4	19.87	25	$> .05$	64.47	159	$> .05$
	School type			**Quarternary**		
LST session 1	27.76	25	$> .05$	120.96	143	$> .05$
LST session 2	29.32	25	$> .05$	141.31	143	$> .05$
LST session 3	43.02	25	$< .05$	122.41	143	$> .05$
LST session 4	58.62	25	$< .01$	139.94	143	$> .05$
	Sudoku					
LST session 1	25.66	25	$> .05$			
LST session 2	38.24	25	$< .05$			
LST session 3	32.34	25	$> .05$			
LST session 4	27.85	25	$> .05$			

Notes: And. = Andersen, M.-L.-stat. = Martin-Löf-statistics. Age: Lower or equal 18 vs. older than 18; School type: Gymnasium vs. vocational school; Sudoku: Experience vs. no experience.

definitions help to investigate if general learning effects or basic parameter specific learning effects can better account for the empirical data gathered in the current study.

To answer the question which learning effects can be identified specifically for the chosen basic parameters, the following sets of learning parameters were defined: General time variables which define impact of items including B1 on subsequent items containing B1, impact of items requiring T1 on subsequent items requiring T1 and so on across test versions. Additionally, to investigate possible additional within-test version learning, these parameters were also defined not only across test versions but also both continuously within test versions and across test versions (that means it makes no difference between test versions but models continuous learning across all items). To illustrate these defined learning parameters,

table 6.5 shows the principles exemplarily.

In the same manner, trend parameters are defined specifically for basic parameters. Additionally, interactions between time variables and person characteristics as demographic variables and test results from the described additional tests are investigated.

Table 6.5: Examples for learning parameters, general time effects

		Basic parameters			**Learning parameters**		
Version	Item	B1	T1	Q	LB1	LT1	LQ
LST 1	1	1	1	0	0	0	0
	2	1	0	0	0	0	0
	3	1	1	1	0	0	0
	4	0	0	1	0	0	0
LST 2	1	1	1	0	1	1	0
	2	1	0	0	1	0	0
	3	1	1	1	1	1	1
	4	0	0	1	0	0	1
LST 3	1	1	1	0	2	2	0
	2	1	0	0	2	0	0
	3	1	1	1	2	2	2
	4	0	0	1	0	0	2
LST 4	1	1	1	0	3	3	0
	2	1	0	0	3	0	0
	3	1	1	1	3	3	3
	4	0	0	1	0	0	3

Notes: LB1 = learning parameter for B1, LB2 = learning parameter for B2 and so on.

6.4.4 Longitudinal LLTM results

First of all, repeated measures ANOVA was computed to investigate general score differences between test versions. Results show that there are differences between groups defined by test versions ($F = 190.02$, $df = 3$, $p < .01$) and that these differences can be found in detail between version one and two and between version two and three (1 vs. 2: $F = 170.94$, $df = 1$, $p < .01$; 2 vs. 3: $F = 60.34$, $df = 1$, $p < .01$; 3 vs. 4: $F = 0.01$, $df = 1$, *n.s.*). These results confirm the results found by Hoffmann (2007).

For a first overview about basic parameter developments, four separate LLTMs are specified for each LST session. Table 6.6 shows the results for the basic parameters separately for all four sessions. Of course, this is no correct and exact approach because of intraindividual correlations of scores between sessions. Keep also in mind that absolute parameter values cannot be interpreted directly and that the amount of the constant is also important. Thus only coarse trends can be identified. B1 and B2 influence seems to remain on a relatively stable low level. T1, T3, T4 and Q impact becomes lower from session one to three, with T4 having greater impact than T3, T3 having greater impact than T2 and so on, and with Q impact ranging between T2 and T3. From session three to four there seems to occur an extreme rise in all basic parameter estimates. This may be due to two reasons: Constant development shows that in session four, basic item difficulty is much lower than in the sessions before and thus basic parameter estimates are higher. Additionally, examinees were probably suffering from motivational and concentrational problems during the fourth test session, resulting in lower item solution rates (see also table 6.3).

Table 6.6: Basic parameter estimates separated for all four LST sessions

	Session 1	**Session 2**	**Session 3**	**Session 4**
Parameter	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
Fixed effects				
Constant	4.26 (0.17)**	4.69 (0.19)**	4.25 (0.21)**	6.06 (0.24)**
B1	-0.47 (0.08)**	-0.48 (0.11)**	-0.65 (0.11)**	-0.79 (0.12)**
B2	-0.66 (0.09)**	-0.83 (0.11)**	-0.87 (0.12)**	-0.95 (0.13)**
T1	-1.56 (0.11)**	-1.61 (0.11)**	-0.75 (0.14)**	-2.44 (0.14)**
T2	-2.70 (0.14)**	-2.50 (0.15)**	-1.67 (0.17)**	-3.34 (0.19)**
T3	-3.13 (0.15)**	-2.94 (0.16)**	-2.31 (0.17)**	-3.95 (0.19)**
T4	-3.96 (0.18)**	-3.70 (0.19)**	-3.40 (0.21)**	-4.91 (0.22)**
Q	-2.76 (0.14)**	-2.54 (0.15)**	-1.40 (0.17)**	-2.89 (0.18)**
Random effects				
Person	0.74 (0.09)	1.09 (0.13)	0.86 (0.12)	1.04 (0.13)
Fit indices				
LL (df)	-4124.42 (9)	-3539.16 (9)	-2988.14 (9)	-2968.15 (9)

Notes: * $p < .05$, ** $p < .01$. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. Session = LST session, Est. = estimate.

To further investigate the occurring learning effects, several LLTM variants were

computed. First of all, a so-called empty model is specified without any predictors which resembles the RM with random person effect. Then one model with time predictors only (L-RM) and one with basic parameters only (LLTM) are computed. After that, general time predictors and basic parameters are included (L-LLTM) and then basic parameters specific learning parameters are included (LBP-LLTM). Results from these analyses can be seen in tables 6.7 for continuous time variables and 6.8 for trend variables.

Table 6.7: Longitudinal LLTM results for time variables, four time points

Parameter	RM (SE) (emp. mod.)	L-RM (SE)	LLTM (SE)	L-LLTM	LBP-LLTM (SE)
Fixed					
Constant	1.56 (0.05)**	1.16 (0.05)**	4.62 (0.11)**	4.24 (0.11)**	4.68 (0.11)**
B1			-0.54 (0.05)**	-0.55 (0.05)**	-0.59 (0.07)**
B2			-0.77 (0.05)**	-0.79 (0.05)**	-0.87 (0.07)**
T1			-1.51 (0.06)**	-1.54 (0.06)**	-1.63 (0.08)**
T2			-2.46 (0.08)**	-2.50 (0.08)**	-2.86 (0.09)**
T3			-2.97 (0.08)**	-3.02 (0.08)**	-3.24 (0.10)**
T4			-3.85 (0.10)**	-3.93 (0.10)**	-4.04 (0.14)**
Q			-2.37 (0.08)**	-2.41 (0.08)**	-3.00 (0.09)**
Time		0.28 (0.01)**		0.31 (0.01)**	
LB1					0.03 (0.04)
LB2					0.06 (0.04)
LT1					0.05 (0.04)
LT2					0.25 (0.04)**
LT3					0.16 (0.05)**
LT4					0.09 (0.07)**
LQ					0.44 (0.03)**
Random					
Person	0.70 (0.07)	0.72 (0.07)	0.84 (0.08)	0.87 (0.08)	0.88 (0.08)
Fit					
LL (df)	-14900.08 (2)	-14672.82 (3)	-13681.98 (9)	-13431.86 (10)	-13410.39 (16)
AIC	29804.16	29351.65	27381.97	26883.73	26852.77
BIC	29820.88	29376.73	27457.22	26967.34	26986.56
$\Delta\chi^2$	-	454.52**	2436.20**	2936.44**	2979.38**

Notes: * p < .05, ** p < .01. Emp. mod. = empty model. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. B = binary, T = ternary, Q = quarternary. LB1 = learning parameter for B1, LB2 = learning parameter for B2 etc. $\Delta\chi^2$ = $\Delta\chi^2$ to empty model.

Since trend3 is not significant, it can be concluded that no significant (learning) effects emerge between LST session three and four (note that the assumption

Table 6.8: Longitudinal LLTM results for trend variables, four time points

Parameter	L-RM (SE)	LLTM (SE)	L-LLTM (SE)	LBP-LLTM (SE)
Fixed				
Constant	1.08 (0.06)**	4.62 (0.11)**	4.17 (0.11)**	4.70 (0.11)**
B1		-0.54 (0.05)**	-0.55 (0.05)**	-0.57 (0.08)**
B2		-0.77 (0.05)**	-0.79 (0.05)**	-0.79 (0.08)**
T1		-1.51 (0.06)**	-1.55 (0.06)**	-1.76 (0.09)**
T2		-2.46 (0.08)**	-2.51 (0.08)**	-3.01 (0.10)**
T3		-2.97 (0.08)**	-3.04 (0.08)**	-3.43 (0.11)**
T4		-3.85 (0.10)**	-3.95 (0.10)**	-4.27 (0.16)**
Q		-2.37 (0.08)**	-2.42 (0.08)**	-3.09 (0.09)**
trend1	0.43 (0.04)**		0.47 (0.04)**	
trend2	0.41 (0.04)**		0.45 (0.05)**	
trend3	-0.07 (0.05)		-0.08 (0.05)	
B1trend1				0.06 (0.11)
B1trend2				-0.28 (0.12)*
B1trend3				0.42 (0.13)**
B2trend1				-0.07 (0.12)
B2trend2				-0.18 (0.13)
B2trend3				0.53 (0.13)**
T1trend1				0.14 (0.11)
T1trend2				0.65 (0.13)**
T1trend3				-0.82 (0.13)**
T2trend1				0.48 (0.11)**
T2trend2				0.54 (0.12)**
T2trend3				-0.39 (0.13)**
T3trend1				0.48 (0.14)**
T3trend2				0.35 (0.15)*
T3trend3				-0.44 (0.15)**
T4trend1				0.59 (0.21)**
T4trend2				0.00 (0.22)
T4trend3				-0.31 (0.22)
Qtrend1				0.51 (0.08)**
Qtrend2				0.85 (0.10)**
Qtrend3				-0.23 (0.11)*
Random				
Person	0.73 (0.07)	0.81 (0.08)	0.88 (0.08)	0.89 (0.08)
Fit				
LL (df)	-14635.09 (5)	-13681.98 (9)	-13389.85 (12)	-13312.60 (30)
AIC	29280.18	27381.97	26803.71	26685.19
BIC	29321.98	27457.22	26904.04	26936.03
$\Delta\chi^2$ to empty	529.98**	2436.20**	3020.46**	3174.96**

Notes: * p < .05, ** p < .01. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. B = binary, T = ternary, Q = quarternary. Empty model see table 6.7.

of motivational and concentrational problems is also supported by significantly negative learning parameter estimates for trend3). Additionally, as described in section 6.4.3, the fourth LST session took place almost directly after the third one and therefore this time interval cannot be compared to the other intervals. Therefore, the models described in tables 6.7 and 6.8 were computed again while only including the first three time points. Results from these analyses are described in tables 6.9 and 6.10. Obviously, estimation results do not change considerably when including only three time points compared to four time points. Model fit cannot be compared directly because estimations are not based on the same data sets (less observations for three time points than for four).

As can be seen from tables 6.7 to 6.10, L-RM fits better than RM and L-LLTM fits better than LLTM, which shows important general learning effects. LLTM fits better than RM and L-RM which confirms the important role of the chosen basic parameters and their impact on item difficulty. The best fitting model is the LBP-LLTM for both time and trend variables which shows that both time predictors and basic parameters contribute substantially to item difficulty and that basic parameter specific learning effects explain the current empirical data better than a global learning effect (LBP-LLTM fits significantly better than all other models for both time and trend variables and for both three and four time points, indicated by LR-test, AIC and BIC). Comparing time and trend models reveals that trend models seem to capture data structure more precisely as model fit is better for trend variables (compare table 6.7 to table 6.8 and table 6.9 to table 6.10 with regard to L-RM, L-LLTM and LBP-LLTM). Trend3 is not significant which indicates that there is no significant change between LST session three and four which was already detected by Hoffmann (2007).

No considerable learning effects for B1, B2 and T1 can be detected for the general time variable across four time points in table 6.7. This can be explained by the trend results in table 6.8: For B1 and B2, significant positive learning rates occur mainly between LST session three and four, negative effects between session two and three, and almost no effects between session one and two. The opposite is the case for T1: While positive effects occur between session one and two and between session two and three, a smaller than 0 estimate result is described between session three and four. Negative effects between session three and four can also be seen for T2, T3, T4 and Q. Altogether these results indicate severe motivational and concentrational deficits during session four. Therefore, results are only further interpreted based on the analyses for three time points. Here (table 6.9) we can

Table 6.9: Longitudinal LLTM results for time variables, three time points

Parameter	RM (SE) (emp. mod.)	L-RM (SE)	LLTM (SE)	L-LLTM	LBP-LLTM (SE)
Fixed					
Constant	1.46 (0.05)**	1.08 (0.05)**	4.30 (0.11)**	3.95 (0.12)**	4.38 (0.12)**
B1			-0.49 (0.05)**	-0.51 (0.06)**	-0.44 (0.07)**
B2			-0.74 (0.06)**	-0.76 (0.06)**	-0.67 (0.08)**
T1			-1.31 (0.06)**	-1.34 (0.07)**	-1.69 (0.08)**
T2			-2.28 (0.09)**	-2.33 (0.09)**	-2.78 (0.10)**
T3			-2.76 (0.09)**	-2.83 (0.09)**	-3.18 (0.11)**
T4			-3.64 (0.11)**	-3.73 (0.11)**	-3.95 (0.15)**
Q			-2.26 (0.08)**	-2.31 (0.09)**	-2.89 (0.10)**
Time		0.42 (0.02)**		0.46 (0.02)**	
LB1					-0.08 (0.06)
LB2					-0.11 (0.06)
LT1					0.37 (0.06)**
LT2					0.49 (0.06)**
LT3					0.40 (0.07)**
LT4					0.28 (0.11)*
LQ					0.64 (0.05)**
Random					
Person	0.67 (0.07)	0.70 (0.07)	0.80 (0.08)	0.84 (0.08)	0.85 (0.08)
Fit					
LL (df)	-11618.95 (2)	-11411.20 (3)	-10748.80 (9)	-10521.53 (10)	-10464.80 (16)
AIC	23241.89	22828.40	21515.60	21063.06	20961.60
BIC	23258.04	22852.62	21588.26	21143.79	21090.78
$\Delta\chi^2$	-	415.50**	1740.30**	2194.84**	2308.30**

Notes: * p < .05, ** p < .01. Emp. mod. = empty model. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. B = binary, T = ternary, Q = quarternary. LB1 = learning parameter for B1, LB2 = learning parameter for B2 etc. $\Delta\chi^2$ = $\Delta\chi^2$ to empty model.

Table 6.10: Longitudinal LLTM results for trend variables, three time points

Parameter	L-RM (SE)	LLTM (SE)	L-LLTM (SE)	LBP-LLTM (SE)
Fixed effects				
Constant	1.08 (0.06)**	4.30 (0.11)**	3.95 (0.12)**	4.37 (0.12)**
B1		-0.49 (0.05)**	-0.51 (0.06)**	-0.49 (0.08)**
B2		-0.74 (0.06)**	-0.76 (0.06)**	-0.69 (0.08)**
T1		-1.31 (0.06)**	-1.34 (0.07)**	-1.61 (0.09)**
T2		-2.28 (0.09)**	-2.33 (0.09)**	-2.77 (0.11)**
T3		-2.76 (0.09)**	-2.83 (0.09)**	-3.20 (0.12)**
T4		-3.64 (0.11)**	-3.73 (0.11)**	-4.05 (0.16)**
Q		-2.26 (0.08)**	-2.31 (0.09)**	-2.84 (0.10)**
trend1	0.43 (0.04)**		0.47 (0.04)**	
trend2	0.41 (0.04)**		0.45 (0.05)**	
B1trend1				0.07 (0.11)
B1trend2				-0.28 (0.12)*
B2trend1				-0.06 (0.11)
B2trend2				-0.17 (0.13)
T1trend1				0.13 (0.11)
T1trend2				0.66 (0.12)**
T2trend1				0.47 (0.11)**
T2trend2				0.53 (0.12)**
T3trend1				0.47 (0.13)**
T3trend2				0.34 (0.15)*
T4trend1				0.58 (0.21)**
T4trend2				0.00 (0.22)
Qtrend1				0.50 (0.08)**
Qtrend2				0.84 (0.10)**
Random effects				
Person	0.70 (0.07)	0.80 (0.08)	0.84 (0.08)	0.85 (0.08)
Fit indices				
LL (df)	-11411.19 (4)	-10748.80 (9)	-10521.50 (12)	-10449.85 (23)
AIC	22830.37	21515.60	21065.01	20945.70
BIC	22862.67	21588.26	21153.82	21131.40
$\Delta\chi^2$ to empty	415.52**	1740.30**	2194.90**	2338.20**

Notes: * p < .05, ** p < .01. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. B = binary, T = ternary, Q = quarternary. Empty model see table 6.9.

see no specific learning effects for the general time variable for B1 and B2, and significant effects for the remaining basic parameters. Comparing learning effects for B1, T1 and Q shows the highest effect for Q, followed by T1 and then by B1. Comparing effects of T1, T2, T3 and T4 reveals the highest learning effect for T2, followed by T3, T1 and T4. Table 6.10 shows how these general effects are distributed between test sessions: Learning effects for T1, T2 and Q are lower between session one and two and are highest between session two and three, while effects for T3 and T4 are highest between session one and two and lower between session two and three (T4trend2 is zero). Comparing B1, T1 and Q learning effects reveals highest effects for Q, lower effects for T1 and lowest effects for B1 for all trend variables.

Several additional test results and person characteristics show significant impact on item difficulty for all four LST sessions (see table 6.11). Again be careful with direct comparisons because of dependent measures and different constant values as already mentioned for table 6.6. However, for identification for possible person predictors in longitudinal analysis, these results are helpful. As table 6.11 shows, size of estimates changes for several variables across LST sessions: In the first session, math grade is not significant, from the second to the fourth session it is (with relatively stable parameter sizes). CFT and BIS effects remain stable across test sessions, too. Also the Enterprising scale of AIST and the Interest scale of FAM have significant albeit fluctuating influence on LST item difficulty across test sessions. Including these person characteristics and additional test results into LLTM analyses leads to improved model fit and extreme reduction of the constant and random person variance compared to the LBP-LLTM in tables 6.7 and 6.8, respectively. The resulting full models for time / trend variables can be seen in table 6.12. As the fourth LST session did not provide additional helpful results and may lead to misinterpretations because of motivational and concentrational deficits, only the results for three time points, that is for LST version one to three, are described.

To investigate possible interactions between person characteristics and learning effects, additional interaction models were specified. While several additional test results and person characteristics show significant impact on item difficulty for all four LST sessions (see table 6.11), only CFT shows significant interactions with continuous time and only CFT and Sudoku experience show significant interactions with trend variables (CFT with trend3 and Sudoku experience with trend1) beyond the described LR-LLTM variables. Higher CFT value leads to

Table 6.11: Effects of person characteristics and test results on LST item difficulty in all four test sessions

Parameter	**Session 1** Est. (SE)	**Session 2** Est. (SE)	**Session 3** Est. (SE)	**Session 4** Est. (SE)
Fixed Effects				
Constant	2.95 (0.31)**	3.33 (0.35)**	3.22 (0.36)**	4.90 (0.40)**
B1	-0.47 (0.08)**	-0.48 (0.10)**	-0.65 (0.11)**	-0.79 (0.12)**
B2	-0.66 (0.09)**	-0.83 (0.11)**	-0.87 (0.12)**	-0.96 (0.13)**
T1	-1.56 (0.11)**	-1.60 (0.11)**	-0.75 (0.14)**	-2.44 (0.14)**
T2	-2.70 (0.14)**	-2.50 (0.15)**	-1.67 (0.17)**	-3.34 (0.19)**
T3	-3.13 (0.18)**	-2.94 (0.16)**	-2.31 (0.17)**	-3.95 (0.19)**
T4	-3.96 (0.18)**	-3.70 (0.19)**	-3.40 (0.21)**	-4.92 (0.22)**
Q	-2.76 (0.14)**	-2.54 (0.15)**	-1.40 (0.17)**	-2.89 (0.18)**
Sudoku	0.25 (0.10)*	0.29 (0.12)*	0.29 (0.12)*	0.26 (0.13)*
Math grade	-0.05 (0.04)	-0.15 (0.05)**	-0.12 (0.05)*	-0.11 (0.05)*
School type	0.48 (0.10)**	0.55 (0.12)**	0.40 (0.11)**	0.57 (0.13)**
CFT	0.34 (0.06)**	0.31 (0.07)**	0.35 (0.06)**	0.41 (0.07)**
BIS	0.25 (0.07)**	0.28 (0.08)**	0.21 (0.08)**	0.19 (0.08)*
AIST E	-0.13 (0.05)*	-0.08 (0.06)	-0.22 (0.06)**	-0.16 (0.06)*
FAM I	0.13 (0.04)**	0.19 (0.04)**	0.12 (0.04)**	0.09 (0.05)*
Random Effects				
Person	0.35 (0.05)	0.50 (0.07)	0.39 (0.07)	0.54 (0.08)
Fit indices				
LL (df)	-4048.17 (16)	-3462.86 (16)	-2919.76 (16)	-2904.55 (16)

Notes: * $p < .05$, ** $p < .01$. Est. = estimate, SE = standard error, LL = Log-Likelihood, df = degrees of freedom. Session = LST session, Est. = estimate. Sudoku experience: 0 = no experience, 1 = experience; math grade ranging from 1 to 6; school type: 0 = vocational school, 1 = gymnasium; CFT and BIS normalized and transformed to z-values. AIST E = AIST scale "Enterprising", FAM I = FAM scale "Interest".

more intense learning effects, especially between session three and four, and Sudoku experience leads to more intense learning effects between session one and two. As these interactions are very small, they seem not to be important and thus are not further reported and interpreted. Three-way interactions between basic parameters, time/trend variables and any additional test results or person characteristics are not significant.

To investigate if itemwise learning explains the current data better than testwise effects (that means learning within the test versions, ignoring the change from one version to another), the above described models were computed again with learning parameters that take into account learning from one item to the subsequent one (which involves the same basic parameter), no matter which test version it belongs to. These models fit significantly worse than the models with learning across test versions. Thus learning seems to take place from test session to test session (both in general and specifically for every basic parameter) and not from item to item.

6.4.5 Longitudinal CDM results

The recomputed skill probabilities are shown in table 6.13. The design matrix for DINA analyses again has the triangular structure described in table 5.11. Development of skill probabilities across LST sessions shows mainly one direction for all basic parameters: Skills are mastered by a bigger proportion of examinees from session one to two and from session two to three. From session three to four, this development can still be seen for most basic parameters except for B1, T1 and T2. However, relation of skill probabilities between basic parameters shows confusing results as Q is mastered by a higher proportion of examinees than B1 and T1. The order would have to be reversed for B1, T1 and Q for similar results compared to LLTM analyses. Guessing and slipping parameters are assumed to be constant across time. However, the non-plausible results for skill mastery indicate a severe lack of stability of guessing and slipping parameters. It can be concluded that this longitudinal application of the DINA model does not work at all.

6.5 Discussion

The current study demonstrates modeling of longitudinal data with linear logistic test models and an approach of longitudinal DINA modeling. Four LST test ver-

Table 6.12: Effects of person characteristics and test results on LST item difficulty for continuous time and trend variables, three time points

Time		**Trend**	
Parameter	Est. (SE)	Parameter	Est. (SE)
Fixed effects		**Fixed effects**	
Constant	3.14 (0.26)**	Constant	3.13 (0.26)**
B1	-0.44 (0.07)**	B1	-0.49 (0.08)**
B2	-0.67 (0.08)**	B2	-0.69 (0.08)**
T1	-1.69 (0.08)**	T1	-1.61 (0.09)**
T2	-2.78 (0.10)**	T2	-2.77 (0.11)**
T3	-3.18 (0.11)**	T3	-3.20 (0.12)**
T4	-3.95 (0.15)**	T4	-4.05 (0.16)**
Q	-2.89 (0.10)**	Q	-2.84 (0.10)**
LB1	-0.08 (0.06)	B1trend1	0.07 (0.11)
LB2	-0.11 (0.06)	B1trend2	-0.28 (0.12)*
LT1	0.37 (0.06)**	B2trend1	-0.06 (0.11)
LT2	0.49 (0.06)**	B2trend2	-0.17 (0.13)
LT3	0.40 (0.07)**	T1trend1	0.13 (0.11)
LT4	0.28 (0.11)*	T1trend2	0.66 (0.12)**
LQ	0.65 (0.05)**	T2trend1	0.47 (0.11)**
Sudoku	0.28 (0.09)**	T2trend2	0.53 (0.12)**
Math grade	-0.10 (0.04)**	T3trend1	0.47 (0.13)**
School type	0.47 (0.09)**	T3trend2	0.34 (0.15)*
CFT	0.35 (0.05)**	T4trend1	0.57 (0.21)**
BIS	0.25 (0.06)**	T4trend2	0.00 (0.22)
AIST E	-0.14 (0.05)**	Qtrend1	0.50 (0.08)**
FAM I	0.14 (0.03)**	Qtrend2	0.84 (0.10)**
		Sudoku	0.28 (0.09)**
		Math grade	-0.10 (0.04)**
		School type	0.47 (0.09)**
		CFT	0.35 (0.05)**
		BIS	0.25 (0.06)**
		AIST E	-0.14 (0.05)**
		FAM I	0.14 (0.03)**
Random effects		**Random effects**	
Person	0.38 (0.04)	Person	0.38 (0.04)
Fit indices		**Fit indices**	
LL (df)	-10366.52 (23)	LL (df)	-10351.55 (30)
AIC	20779.04	AIC	20763.11
BIC	20964.74	BIC	21005.32
$\Delta\chi^2$ to empty model	2504.86**	$\Delta\chi^2$ to empty	2534.80**

Notes: * p < .05, ** p < .01. Est. = estimate, SE = standard error, LL = Log-Likelihood, df = degrees of freedom. B = binary, T = ternary, Q = quarternary. Sudoku experience: 0 = no experience, 1 = experience; math grade ranging from 1 to 6; school type: 0 = vocational school, 1 = gymnasium; CFT and BIS normalized and transformed to z-values. AIST E = AIST scale "Enterprising", FAM I = FAM scale "Interest".

Table 6.13: LST longitudinal DINA results for skill probabilities

Skill	LST1	LST2	LST3	LST4
B1	0.53	0.68	0.78	0.78
B2	0.68	0.76	0.81	0.83
T1	0.62	0.73	0.85	0.79
T2	0.59	0.70	0.81	0.81
T3	0.60	0.68	0.70	0.72
T4	0.53	0.57	0.55	0.58
Q	0.68	0.78	0.91	0.93

Notes: B = binary, T = ternary, Q = quarternary.

sions were tested with 304 German school students. Test versions were generated rule-based (cf. chapter 5) and through variation of surface characteristics, thereby implementing a kind of item cloning (cf. chapter 7). It is shown how learning effects can be investigated regarding specific basic parameters which allows insights into cognitive processes involved in LST as well as into learning processes across test sessions. Moreover, additional tests for fluid intelligence, memory, concentration and motivation / personality as well as person characteristics refine results and their interpretation. Since no explicit LLTM variants exist which allow to model basic parameter specific learning effects, a LLTM variant with learning parameters for items containing basic parameter specific learning effects was built and applied to the data. All items and test versions show sufficient fit to the Rasch model.

6.5.1 General longitudinal results

To map learning effects on both a global and a detailed level, two different operationalizations for time points were chosen (following Hoffmann, 2007). The general time variable describes global (linear) change, trend variables describe detailed change between test sessions. Both time operationalizations are confirmed as significant impact factors for item difficulty and result in better model fit compared to an empty model without any predictors as well as to a basic LLTM without time consideration. Including specific learning parameters for basic parameters again improves model fit significantly. This means that not simply a general learning effects occurs which can be imagined as adding a constant to person ability (or differently speaking, subtracting a constant from item difficulty),

but that basic parameter specific learning effects are existent in the current study.

Trend variables show that learning effects are not completely linear and that greatest changes in item difficulty occur between the first and second LST session, slightly less albeit significant changes between the second and third, and no significant changes occur between the third and fourth LST session. This exactly confirms the results of Hoffmann (2007). The fact that the greatest changes occur between the first and second test session, and that there are still score gains between the second and the third, but not between the third and fourth session are in line with results described by other longitudinal studies (e.g., Cliffordson, 2004).

6.5.2 Basic parameters

Separate models for all four LST sessions in table 6.6 show interesting developments of all basic parameters across sessions. A first interpretation of the separate models is as follows: Binary impact remains at a relatively stable low level across sessions and thus almost no learning effects were found for binary in the subsequent analyses. T1 and Q assimilate to each other in difficulty (which in general decreases from session one to three and increases in session four, probably due to motivational problems). This decrease from session one to three and increase between session three and four is also found for T2, T3 and T4.

This may at least partly be due to the fact that Q allows two strategic possibilities: Examinees could have switched from one application of the rule to the alternative one and can thus alternate their strategy from "negative" exclusion principles to "positive" fill-in principles: Instead of finding out which symbols must not occur in the question mark cell, one can try to fill in symbols (in mind) because every symbol has to occur once. The current results could be a cue that this switch may have taken place mainly between session two and tree.

However, this separate modeling approach is not correct and thus only tendencies can be identified. As already noted in section 6.4.4, absolute values cannot be interpreted directly because the constant has also to be taken into account. In addition, computing separate models is no correct and exact approach because of intraindividual correlations of scores between sessions.

Nevertheless, the above described developments are in principle also recovered in the models with learning parameters which provide an even more sophisticated

(and statistically more appropriate) illustration of learning effects. Learning parameter estimates show that effects for four of seven basic parameters are significant and greater than zero, and therefore important for item difficulty. Thus learning effects occur, but size of learning effects differs between basic parameters: Obviously, learning effects are greatest for Q, smallest for B1 and B2 (not significant) and moderate for T1 for both time and trend variables. This is not astonishing because Q is the most difficulty operation, and B1 and B2 are the easiest ones, which implies that there is more scope for learning effects concerning Q.

Trend variables provide detailed insight into learning effects between LST test sessions rather than into general (linear) effects: Between the third and fourth session, for T1 to T4 and for Q, learning effects are smaller than zero, implying some kind of reverse learning effects or simply inferior achievement for the fourth session compared to the third session regarding these parameters. This decline is probably mainly due to motivational and concentrational issues as the third test appointment lasted rather long and LST versions were very similar, thus probably leading to exhaustion and lack of attention. As the time interval between session three and four is not comparable to the other intervals and therefore results from session four are probably of different quality, only the first three sessions are taken into consideration during the following interpretations.

Comparing learning effects for B1 to T1 and Q shows highest effects for Q, followed by T1 and B1 (both not significant, probably due to their easiness and no sufficient scope for learning effects). This order was already interpreted above: The more difficult the underlying concept is, the higher is scope for learning effects . Additionally, there may be a strategic shift for Q. However, comparing learning effects between T2, T3 and T4 shows that although T4 is the most complex and difficult concept in the current basic parameter set, learning effects for T4 are not as high as for T2 and T3. This can perhaps be explained as follows: At first sight, there is no reason why T4 should not benefit in the same manner from LST test practice as T2, for example. However, taking into account the biologically reasonable capacity limit of working memory (cf. Cowan, 2010), learning effects regarding T4 similar to the other learning effects for the other basic parameters would be an uncommon finding: T4 almost exceeds the capacity limit which can be thought of as more biologically determined general ability. Since learning effects usually do not pass the specific ability level, learning effects especially for T4 would mean a really unusual finding (see Jensen, 1998; te Nijenhuis et al., 2007).

6.5.3 Further tests and person characteristics

Only few additional test results and person characteristics were identified as significant predictors of item difficulty. Although several variables affect LST item difficulty for all four time points in separate analyses (separate models for every LST session), this impact seems to be limited to intercept and does not range over to learning effects (that is, slopes of learning curves) as there are only significant interactions between time / trend and CFT score and Sudoku. However, these interactions are very low and therefore may be ignored. Thus, the current results partly contradict the results of Hoffmann (2007) who found significant interactions for d2 results and interest (FAM) with time. However, her interactions regarding d2 and interest were rather small. Most of the variables influencing intercept (math grade, school type, Sudoku experience, CFT, BIS, interest (FAM)) are identical with the results of Hoffmann (2007). Only the fact that no significant impact of d2 was found is completely not in line with the former results from longitudinal modeling.

Current results can be interpreted as follows: LST items are easier for persons who have more Sudoku experience, better math grades, visit the gymnasium, have higher CFT and BIS scores as well as higher interest scores (FAM) and lower enterprising scores (AIST). These variables seem to explain large parts of interindividual differences (expressed by the random person variance) as well as large parts of basic item difficulty (expressed by the constant) as their inclusion leads to reduction of random person variance and of the constant. Except for Sudoku experience, interest and enterprising, these findings can be explained by possible effects of a third confounding variable, probably general intelligence. Examinees who have already Sudoku experience perhaps apply more efficient strategies from the beginning (chunking or quarternary strategy switch, for example) and do not have to develop an efficient strategy from scratch. Thus items are less difficult for these examinees. Interest in the item type leads to reduction of difficulty, perhaps through more ambitious working. Students who score higher in enterprising may consider LST as too boring and no adequate challenge and thus may suffer from motivational problems to solve items carefully.

Interaction results for continuous time variables show that students who have higher CFT scores show higher learning success for LST. For trend variables, higher CFT scores result in higher learning success between the third and fourth LST session (here perhaps higher general intelligence prevents oversights). Sudoku

experience leads to more intense learning between session one and two (probably due to strategy and familiarity with a similar item type as described above). Hence there are only few interaction results which help to explain the occurring learning effects in LST as it was already the case in the work of Hoffmann (2007). The influence and interaction of CFT scores and Sudoku experience are not surprising as it can be assumed that individuals who have a good fluid intelligence as well as Sudoku experience (i.e., experience with a very similar item type) will be more successful in LST and can benefit more from practice effects (cf. Kulik et al., 1984). Please keep also in mind that learning effects would have been even higher in the case of identical rather than parallel test versions (see Hausknecht et al., 2007).

6.5.4 Longitudinal DINA results

As described above, DINA results can only be reported as recomputed skill probabilities as no adequate longitudinal implementation exists. Results show plausible development of skill probabilities, that is changes by shifts from one state to another, but probability order of skill mastery is totally implausible for basic parameters as B1 and B2 show lowest probabilities and Q highest. The opposite would have to be the case to map the current data. It has to be concluded that DINA also fails in this longitudinal adaptation.

6.5.5 Conclusions

In summary, LLTM application to the current data is more appropriate than DINA application as DINA does not meet the basic requirements for longitudinal application and mapping of the underlying basic parameter structure. For LLTM application, the following may be noted: In all LLTM variants, the basic parameters are significant impact factors for item difficulty which indicates that the basic item structure remains stable during repeated testing with regard to design characteristics. Learning effects seem to occur between the first and third LST session and are greatest for Q, followed by T1 and B1. Only CFT and Sudoku experience show impact on both intercept and learning curve slopes; school type, math grade, BIS score, enterprising (AIST) and interest (FAM) only on intercept.

Hence, different examinee cognitive conditions constitute no serious problem for repeated testing of LST as they seem to affect almost only initial values but not learning which means that no considerable shift of learning success caused by

examinee characteristics occurs and interindividual differences are maintained and still mapped adequately during repeated testing (as long as only examinee order and not cut-off is of interest). The fact that several person characteristics exert influence on initial LST scores is no serious problem because it can be assumed that most of these characteristics themselves are affected by third variables which also affect LST score, for example general intelligence. Only Sudoku can become a problem because it can be assumed that examinees with Sudoku experience apply more efficient strategies which have to be developed primarily by examinees without experience who thus have important disadvantages.

However, absolute learning effects are considerably high which means that intraindividually scores may rise heavily during repeated testing and thus may dilute true examinee abilities and threaten test validity, especially in the case of cut-off decisions. This can become a serious problem if important decisions are based on such test results without taking into consideration these possible disturbing factors. However, the described learning effects provide interesting insights into LST item structure regarding basic parameters and their development across several test sessions and again confirms the fundamental role of the chosen parameters. Moreover, there are no definite hints that the measured construct changes during repeated testing as difficulty ordering of basic parameters and impact of additional person characteristics and test results remain relatively stable during repeated testing. This is further supported by relatively stable correlations between scores and additional test results (cf. Hoffmann, 2007) and rather marginal interactions between time /trend variables and additional test results. At present, obviously it has to be accepted that learning effects occur during repeated testing and can not be avoided, at best they can be minimized by means of rule-based item design and item cloning or automatic item generation to minimize similarity of test items.

Concerning the question what exactly is learned during LST repeated testing, it can be concluded that examinees learn about correct application and efficient handling of the chosen basic parameters. In general, learning effects can be interpreted as increase in θ or decrease in σ. Which of both is truly the case can often not be resolved completely. Since there are no hints that item quality and validity change in the current study, learning effects should be interpreted as increase in basic parameter specific parts of θ. Learning effects differ between basic parameters which means that overall score gains can be traced back to parameter specific learning effects. Therefore, no simple general increase of person ability occurs but different learning effects can be observed for different basic parameters. It can

probably be concluded that strategy and task specific learning restricted to LST occurs rather than transfer of learning effects to g_f or other broader abilities.

6.5.6 Limitations and prospects

The current study suffers from several serious shortcomings. First, the nature of the variance-covariance matrix of errors cannot be varied. This means that intercorrelations between data points for the same person cannot be specified directly to be of specific structure (identity, unstructured etc.). This means that estimations may be based on false assumptions as Stata always sets identity as default. This may result in higher standard errors and biased estimates. There is no possibility to fix this for sake of correct statistical modeling, but general results and tendency should not be affected at all. Comparisons by Hoffmann (2007) show that no general differences emerge in results for several matrices, that is unstructured assumptions lead to better model fit compared to identity, but results do not differ in principle. Thus it can be concluded that effects of specification of the variance-covariance matrix do not in general affect longitudinal results.

Second, possible item cloning effects and possible learning effects are completely confounded. In principle, shifts in item difficulty which could be interpreted as learning effects could also be based on cloning effects, that is impact of surface features. Theoretically, changes in item difficulty therefore cannot solely be explained by learning effects but there is always the alternative explanation of surface characteristics which can cause difficulty in- and decreases. This could be the reason for some counterintuitive developments of item difficulties across test sessions (for example, some items do not show decreasing item difficulties across sessions or do not show systematic changes in difficulty at all). However, most contradictional developments occur between the third and fourth session which can be explained by motivational and concentrational deficits. The fourth LST session in general seems to be hardly interpretable as described in the results section of the current study. Together with this argument, there are several further reasons why the assumption of real learning effects is approvable: As Pauls (2009) describes, parallel versions of the LST seem to show adequate task interchangeability and thus surface characteristics should not affect item difficulty considerably. Moreover, the longitudinal results of Hoffmann (2007) as well as ANOVA results approve the existence of learning effects and systematic difficulty changes in the expected direction, pointing to learning in LST. It is statistically implausible that

changes in item difficulty are not caused by learning effects and yet emerge that clearly and consistently. Therefore it can be concluded that clear learning effects occur in LST despite the confounded cloning and time effects.

Despite the above mentioned shortcomings, analyzing longitudinal data with IRT models as described in the current study is a helpful method to investigate learning effects, group differences and change parameters in a probabilistic framework. The advantages of IRT models compared to classical test theory apply to longitudinal analysis, too. The demonstrated adaptation of the LLTM opens up an extra dimension of basic parameter interpretation, i.e. the longitudinal "behavior" of basic parameters and the accompanied learning effects and changes in cognitive processes involved in solution of items. Multilevel modeling is even more comfortable with meanwhile well developed statistical software (in the current study, Stata and its xtmelogit procedure). Results help to refine item construction and testing practice.

It has to be noted that longitudinal data gathering almost always suffers from motivational and drop out problems. Therefore it is difficult to find out a "true" upper limit for scores. However, in the current study, four test sessions for LST items were obviously enough to detect the limit of learning in LST as between test session three and four no significant learning effects occur. Despite the fact that items were very easy for the current sample and considerable learning effects occurred, no severe ceiling effects emerged. However, for gymnasium and university students, in general more difficult items should be applied.

Hoffmann (2007) identified only d2 and interest (FAM) as impact factors for learning slopes and CFT, BIS, d2, school type, math grade, interest and Sudoku experience for intercept. The current study expands these findings and provides insights into learning slopes by attributing learning effects to basic parameters. However, it is still not completely resolved how slopes in LST learning curves can be explained. This should be the focus of subsequent research in LST.

To make sure that item and test quality and validity are stable even if tests are taken more than once, longitudinal studies are strongly needed. For practitioners (for example, in personnel selection, test writing or diagnostics), there remains the advice to think about repeated testing carefully and to avoid undesired practice and learning effects by means of automated and rule-based item generation and item cloning as well as by longitudinal quality controls.

7 Statistical word problems

After the first demonstration of rule-based item construction for the figural item type LST and application of the described statistical models as well as longitudinal application, this chapter is concerned with word problems and item cloning which imposes new challenges on the item construction process.

7.1 Introduction

Student mathematical competencies can be assessed by a variety of item types. Among these, mathematical word problems are a popular instrument as word problems in this area show a high ecological validity and measure applied and creative as well as logical and mathematical abilities at the same time. Since word problem construction is often time-consuming and cost-intensive, rule-based item construction provides a great advantage to make construction more efficient while keeping item quality on a high level. Additionally, item cloning allows for huge amounts of items without new calibration which helps, for example, to reduce undesired recognition effects. The current study shows rule-based construction of probability word problems using basic concepts from probability theory and implementing an item cloning approach. Results are analyzed with LLTM variants and DINA as CDM example. 741 German school students participated in the study. Results show good Rasch model fit of the items and sound explanation of item difficulty by basic parameters. Additionally, it is shown how the cloning approach can be taken into account by LLTM.

Assessment of competencies as mathematical, language or artificial competencies is probably as important in today testing as measurement of intelligence. Well known competence tests are the ones applied as exams in school and university several times a month or week to measure the students' knowledge and abilities to cope with certain domains. However, competency testing requirements slightly differ from intelligence testing (cf. McClelland, 1973). For example, competency testing is most often criterion oriented and not norm-based: All examinees are

expected, depending on their abilities, to reach an as good as possible test result, but at least to show understanding and knowledge of certain core concepts and domain contents. Additionally, only specific domains are tested, in contrast to assessment of fluid intelligence which is supposed to be measured most accurately by domain-independent tests and test items.

There are many item types which can be used to measure competencies in many areas. One well-known and gladly used item type are word problems. The current study describes rule-based construction of statistical word problems within an item cloning frame and shows useful results and interpretations from statistical analysis of empirical data.

7.2 Background

Word problems provide a quite effective opportunity to measure competencies in educational and psychological contexts. Especially mathematical word problems are often applied for cognitive and educational assessment purposes in school and university settings as they show a high ecological validity and measure applied and creative as well as logical and mathematical abilities at the same time (Jonassen, 2003). Mathematical competencies are of great importance for success in school and university as well as in most modern sciences. Acquisition and hence measurement of mathematical competencies therefore is of special interest in educational and psychological areas. For example, the Programme for International Student Assessment (PISA) makes clear the importance of mathematical competence and defines mathematical literacy as follows:

> Mathematical literacy is an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen. (OECD, 2003, p.24)

Mathematical word problems allow for assessment of mathematical competencies as well as knowledge and skills achieved during lessons and courses (Jonassen, 2003). Requirements of mathematical word problems range from just translating words into equations up to really complex mathematical reasoning and transfer.

7.2.1 Statistical competence

An important but in school unfortunately often disregarded facet of mathematical competence is stochastic or statistical competence. Knowledge and handling of at least basic statistical concepts is crucial for understanding and interpretation of not only empirical outcomes, but also theoretical concepts in modern sciences. Moreover, understanding of statistics is important even for everyday life as it helps to balance advantages and disadvantages, to cope with uncertainty and probabilities or to challenge statistical content in newspapers, newscasts and other parts of daily routine. Not surprisingly, many branches of economy require at least basic application knowledge in statistics and many university courses include statistics lessons. Statistical competencies are also mentioned as one of the four overarching ideas in the PISA guidelines: Uncertainty is strongly related to stochastic content. Unfortunately and neglecting the importance of statistical competence to deal with the mentioned requirements, stochastics are not studied as often as most other mathematical content areas. School lessons often omit statistical contents in favor of other contents or put it into advanced courses which are only attended by few students.

Regarding the importance of statistical competence, this part of mathematical competence deserves more attention and should be assigned more space within research as well as educational contexts. Therefore, measurement of statistical competence is of great interest not only in school and university. In this context, probability and statistical word problems provide information about the competence to deal with statistics and probability theory beyond equations and formulas. This includes a deeper understanding of the relations and the sense behind the numerical expressions.

7.2.2 Algebra and statistical word problems

A large body of research concerning algebra and other mathematical word problems has been made by several authors (e.g., Briars & Larkin, 1984; Cisse, 1995; Dimitrov, 1996; Koedinger & Nathan, 2004; Mayer, 1987; Xin, 2007). The basic work of Mayer (1981) investigates how algebra word problems can be classified. Mayer checked standard algebra textbooks that were used in California secondary schools and proposes eight families of problems based on the source formula of the problem, each subdivided into several problem categories. Sebrechts, Enright,

Bennett, and Martin (1996) presented a widely approved cognitive model for solving algebra word problems which has been proved to be essential in word problem solving and which provides a helpful framework for the conceptualization and evaluation of word problems. It includes the four steps problem translation, problem integration, solution planning and, as a final step, monitoring and solution execution. Holling, Blank, Kuchenbäcker, and Kuhn (2008) provide a great detailed literature review and summarize the most important results from word problem research. They also state that while algebra and arithmetic word problems have been the focus of many studies, probability theory and statistical contents of word problems were investigated by only few researchers (e.g. Arendasy, Sommer, Gittler, & Hergovich, 2006; Cisse, 1995; Dimitrov, 1996).

The model proposed by Sebrechts et al. (1996) is easily applicable to statistical word problems. In fact, statistical and probability theory contents in word problems do not differ conceptually from algebraic content. The core requirements of translating the problem, building a problem representation and finding and validating a solution are quite independent from the specific mathematical domain. However, as Arendasy et al. (2006) have shown, different subtypes of mathematical word problems cannot be described on one common conceptual dimension as they are qualitatively different. Hence, it is necessary to investigate probability and statistical word problems as a separate problem type rather than blindly assigning results from studies concerning algebra word problems to statistical and probability word problems.

7.2.3 Rule-based item construction and item cloning of word problems

Rule-based item construction and item cloning is relatively unproblematic for figural and numerical contents, nevertheless these techniques impose high demands on test construction and control processes (cf. Freund et al., 2008). A serious challenge in the design of word problems is the handling of wording and text. Thus, particularly in rule-based item construction, verbal content holds several serious difficulties. As text almost always provides space for interpretation, only slight differences in wording can affect item properties as validity, difficulty and complexity in a serious and undesired manner (e.g. Cummins, 1991).

A possibility to avoid misinterpretation and misunderstanding and thus threats to validity of word problems is the usage of constant and maximal unambiguous

phrases. These phrases are constructed by mapping the underlying basic parameter structure as defined in the Q-matrix definitely in wording. Thus, particular phrases are assigned unambiguously to single basic parameters or to whole lines (i.e., combinations of basic parameters) within the Q-matrix. The allocation to whole lines seems to be more appropriate as the combination of phrases may not be possible in a pure additive way because one has to keep in mind grammar and case which is not trivial especially in German. Thereby, an unambiguous mapping between underlying mathematical expressions and wording should be maintained, i.e., items which require the same cognitive steps to be solved should consist of the same phrases. The so-constructed phrases can hold space for explication of "free variables" as numerical information, particular variable characteristics or other surface features which are not supposed to affect item difficulty significantly (incidentals, cf. section 4.1). The remaining basic structure of the phrase as definite mapping of Q-matrix-lines or attributes is not allowed to change through these free variable explications and thus is kept constant in order to avoid misinterpretation. The variability of the free variable characteristics provides an excellent opportunity for item cloning: Changing these incidentals produces in principle many item clones from the same item family as defined in one Q-matrix line. This is exactly what is demonstrated in the current study. Section 7.3.1 describes the construction process in detail and shows item examples.

Despite the common application of word problems in educational assessment, often there is no item construction process explicated (probably at least partly due to the above mentioned challenges and difficulties) which guarantees for high item quality and comparableness of assessment results. Often teachers in school and university just construct items by hand, based on their lessons. This does not automatically mean that these items are inadequate for assessment, invalid or not reliable enough. However, facing the huge item writer effect especially for word problems, as described above, more clearly defined rules for item construction are quite desirable. In fact, in large international studies like PISA, there are item writing rules also for word problems. Engaging into an even more careful theory building and item writing as well as statistical analysis process promises the prevention of typical item writer mistakes and item quality shortcomings. Rule-based item generation and analysis with LLTMs and CDMs are great means in this area to reach the above mentioned requirements.

Especially in high stakes testing and large testing programs as well as for class exercises or university course tests (for example, for the bachelor courses dur-

ing which students have to collect points through test taking), item cloning of word problems gains even more impact: As described in section 4.1, item cloning maximizes efficiency of item production and minimizes risks as recognition or undesired learning effects as well as inadequate item quality. Since construction of word problems is relatively cost-intensive and time-consuming while holding numerous pitfalls for mistakes and interpretational ambiguities during item writing per hand, rule-based item generation and item cloning provide excellent instruments of efficient word problem generation. In combination with automatic item generation, the item construction process becomes even more safe, cost- and time-effective.

Arendasy et al. (2006) presented the automatic item generator AGen. AGen provides a successful implementation of rule-based item design and automatic item generation for word problems and generates algebra word problems with mathematical content as distance x rate = time problems. The authors show that the generated items are of high quality and stable with regard to Rasch scalability and validity. The need for constraints and quality control mechanisms is strongly emphasized by the authors. They underline the necessity of recurrent processes of item generation, checking for quality and psychometric properties, readjusting items and generating procedures and again item generation until the level of perfect item generation with items of the desired properties is reached.

However, to my knowledge only one study concerns rule-based construction of statistical word problems: Holling, Bertling, and Zeuch (2009) describe a first implementation of rule-based, half automated item construction of statistical word problems for university students. They show through Rasch-scalability and LLTM application that rule-based item construction in fact can work for word problems. The current study now adds an item cloning approach to rule-based and half-automated statistical word problem construction and shows empirical results from more than 700 examinees. The following section describes the item construction process in detail.

7.3 Method

7.3.1 Item construction and design

It was decided to implement only four basic operations from probability theory for the current word problems, partly based on the pilot study (Holling et al., 2009)

and an extensive literature and mathematical textbook review about important item construction characteristics as wording and typical numbers. These basic concepts were supposed to be well-known from school lessons and therefore were supposed to be not too difficult for the aimed school student sample. Additionally, the current cloning approach should be as parsimonious as possible. The implemented basic concepts are "complement events" (CE, one has to find the complement event by addition or subtraction), "intersection of independent events" (IIE, the probabilities of two sets of variables have to be multiplied), "intersection of dependent events" (IDE, one has to take into account the equation for dependent events and use the concept of conditional probability) and "set union for disjoint events" (SDE, simple additional relation of two probabilities without consideration of the intersection of two sets).

Item construction followed a cloning procedure: A Q-matrix with eight item families was built. Each family consists of a certain combination of basic parameters (or radicals). Additionally, 14 context stories were constructed which serve as incidentals and only vary in surface information. Each item consists of the context story and one question. The context story is identical for all items of one context, only the question defines the corresponding item family the item belongs to. Structure and wording of the questions are identical for all items of the same family to guarantee for as few undesired interpretation and language effects as possible. The context story only defines the vocabulary of the question. This procedure resulted in 112 items altogether (every family combined with every context). Table 7.1 shows the design matrix for all eight item families, figure 7.1 shows one context with the corresponding eight questions, one for each family.

As can be seen from figure 7.1, the context story contains all numerical information. Numerical information is given as absolute frequencies, in every item there are three features with four, three and two shapings each. All frequencies are given, no matter which are necessary to solve the item. At the end of the item there is some information about the dependency relations. A calculator is not necessary to solve the items as the results are quite even frequencies. Table 7.2 shows solution principles for the example item in figure 7.1.

An overview about all context stories and numerical information within the items is given in the appendix: Table A.6 on page 170 shows all used context characteristics, table A.7 on page 171 all details of these characteristics, table A.8 on page 172 the numerical information for all context characteristics (absolute and relative frequencies and joint probabilities), and tables A.9 and A.10 show the solution

Grandma Miller would like to buy a computer game for her grandchild because he brought quite good grades home. She is not pleased with all these "violence games" and walks to a store which advertises offering only violence-free games. Because Grandma Miller has no knowledge about computer games, she decides to select a game by chance.

The store offers 500 games altogether. Of these, 100 games are mainly for beginners, 125 games mainly for advanced players, and the remaining games mainly for skilled persons or for professionals. 50 games are strategy games, 150 games adventures and 300 games jump and run. 300 games are made for Playstation and 200 games are made for PC.

120 games are jump and run and made for Playstation. 120 games are mainly for skilled persons and made for PC. 60 games are mainly for professionals and made for PC.

Level of proficiency and type of hardware are dependent of each other, all other characteristics are independent of each other.

Questions

1. What is the probability for Grandma Miller selecting a game which is mainly for skilled persons, given that this game is made for PC? (family 1)
2. What is the probability for Grandma Miller selecting a game which is both mainly for beginners and an adventure? (family 2)
3. What is the probability for Grandma Miller selecting a game which is not mainly for professionals, given that this game is made for PC? (family 3)
4. What is the probability for Grandma Miller selecting a game which is neither for advanced players nor a jump and run game? (family 4)
5. What is the probability for Grandma Miller selecting a game which is either mainly for skilled persons or mainly for professionals, given that this game is made for PC? (family 5)
6. What is the probability for Grandma Miller selecting a game which is either mainly for advanced players or both mainly for beginners and a strategy game? (family 6)
7. What is the probability for Grandma Miller selecting a game which is not either mainly for skilled persons or mainly for professionals, given that this game is made for PC? (family 7)
8. What is the probability for Grandma Miller selecting a game which is either not mainly for beginners or both mainly for beginners and a jump and run game? (family 8)

Figure 7.1: Word problem item example. Context story with eight item families realized in eight questions

Table 7.1: Q-matrix statistical word problems

Family	CE	IDE	IIE	SDE
1	0	1	0	0
2	0	0	1	0
3	1	1	0	0
4	1	0	1	0
5	0	1	0	1
6	0	0	1	1
7	1	1	0	1
8	1	0	1	1

Notes: Every line defines one item family. CE = "complement events", IIE = "intersection of independent events", IDE = "intersection of dependent events", SDE = "set union for disjoint events". Every family has fourteen items (belonging to fourteen different contexts), resulting in 112 items altogether.

algorithms and correct results of all families.

Item construction was conducted half-automatically: General templates in LaTex2e contain the context stories to introduce the word problem, the basic wording structure for each family which is realized in the questions, and free variables for the numerical information (which is held constant for every context but can be easily varied in principle). For the current study, the cloning procedure is restricted to different context stories (i.e., contexts define clones), but the free variables allow in principle for more item clones by changing numerical information, for example.

7.3.2 Selected statistical models

As in study 1, the LLTM and its variants RE-LLTM and LR-LLTM will be employed to analyze the word problem data. DINA modeling constitutes CDM analyses.

7.3.3 Research questions

In parallel to the first LST study, it will be investigated which model class, LLTM or CDM, is more appropriate to explain empirical results for the current item type. Additionally, given the assumption that LLTM turns out to be the better choice, LLTM variants will be considered if they provide information about influence of person characteristics and basic parameter influence tendencies on item difficulty.

Table 7.2: Solutions for example item

Family	Equation	Solution
1	A3 | C2	120/200 = 0.60
2	A1 ∩ B2	(100/500)*(150/500) = 0.06
3	1-(A4 | C2)	1-(60/200) = 0.70
4	1-(A2 ∩ B3)	1-((125/500)*(300/500)) = 0.85
5	(A3∪A4) | C2	(120/200)+(60/200) = 0.90
6	(A1 ∩ B1) ∪ A2	(125/500)+((100/500)*(50/500)) = 0.27
7	1-((A3∪A4) | C2)	1-((120/200)+(60/200)) = 0.10
8	(A1 ∩ B3) ∪ (1-A1)	(1-(100/500))+((100/500)*(300/500)) = 0.92

Notes: A1 = beginners, A2 = advanced players, A3 = skilled players, A4 = professionals, B1 = strategy, B2 = adventure, B3 = jump and run, C1 = Playstation, C2 = PC.

Additionally, it will be investigated how the item cloning approach can be taken into account. Assuming that DINA is better suited, mastery classes and skill probabilities will be investigated for further information about item and person characteristics. Implications for word problem construction will also be explained. As the current item type can be regarded to be based on learnable skills, in contrast to LST, it can be assumed that DINA results are better for word problems than for LST.

7.3.4 Test procedure

Overall, 14 test versions (booklets) were created. Every test version consists of four context stories with four of the eight corresponding questions for each context story, resulting in 16 items per booklet so that every item family is covered twice (table A.11 on page 176 in the appendix shows the distribution of item clones in all 14 test versions). One page shows the context story including the numerical information on its own to introduce the word problem. The next four pages each consist of one item, that is the repeated context story (so that the numerical information given in the context story does not have to be remembered, but can be retrieved from each page) and one question (out of the eight questions for each context). Four contexts per test version were included in order not to bore the participants too much and to keep them alert. The only exceptions are test versions 13 and 14 which were created for a subsequent data collection: These two versions contain only contexts 13 and 14, i.e., all eight families per context.

Every participant received a short instruction about probability theory including some item examples to become familiar with the item type (see section B.2 in the appendix). Then the booklets with the 16 test items were handed out. The time limit was two minutes for every item. Answers had to be given in an open response format to avoid guessing and to enable the detection of typical errors. Responses where scored using a standardized resolution guideline: "1" was scored if the correct (combination of) equations and the correct numbers were used by the examinee. Small miscalculations were allowed as the intention was not to measure simple calculation competencies but mastery and handling of probability theory. "0" was given if there were mistakes indicating no appropriate understanding of the rules or usage of wrong combinations of equation parts or wrong numerical information parts. The scoring guideline was checked and approved by several student and postgraduate coworkers.

Additionally, examinees answered several demographic questions (age, gender, class, school type, school grades in math and German) as well as questions about possible advanced course participation in math (because it was supposed that these students perhaps had some advantages concerning statistical content) or if probability theory had been treated in school lessons before. After the probability theory test, participants received some extra questions concerning motivation and experience of difficulty of the test items (scale ranging from 1-not at all to 5-absolutely; in brackets short denotation used in the following):

- Extra question 1: Doing the test was fun (short form: fun).
- Extra question 2: The test was difficult (short form: difficult).
- Extra question 3: The item examples in the instruction were helpful to solve the items (short form: examples helpful).

Smaller subgroups of participants also finished the CFT 20-R (Weiß, 2006), the AIST (Bergmann & Eder, 1999) and NEO-FFI (Borkenau & Ostendorf, 1993) as well as the d2 (Brickenkamp, 2002).

Examinees were offered a helpful test training and individual feedback for all test results (cf. section D in the appendix).

7.4 Results

This section shows the results of the current study. First of all, the sample is described followed by demonstration of adequacy of test version aggregation. Then item characteristics, dimensionality and item fit are provided and after that LLTM and DINA results are described.

7.4.1 Sample

741 German school students were tested on the 112 items. Tables 7.3 and 7.4 show the demographic characteristics of the sample. As can be seen in this table, most examinees have covered probability theory in lessons but have no advanced math course. The extra question results show that examinees had only moderate fun in solving items and found items to be relatively difficult.

7.4.2 Aggregation of test versions

Since item cloning in the current case implies that contexts should not matter compared to family impact on item difficulty, the data are to be aggregated as if there had been only one test version for the whole student sample. To justify this approach, several analyses were conducted to investigate possible context effects which would prohibit aggregated treating of test versions.

Table 7.3: Demographics part 1 word problems sample

	Mean	SD	Min	Max
Age	17.52	0.91	15	21
Math grade	2.62	0.93	0.70	5.00
German grade	2.80	0.73	1.00	5.00
Word problem score	6.35	4.10	0	16
Fun	2.44	1.22	1	5
Difficult	3.47	1.02	1	5
Example helpful	3.80	1.13	1	5

Notes: SD = Standard deviation, Min = minimum, Max = maximum. "Fun", "difficult" and "example helpful" concern extra motivational questions.

Table 7.4: Demographics part 2 word problems sample

	Number	Percent
Gender		
Male	312	42
Female	429	58
School type		
Gymnasium	639	86
Vocational school	102	14
Class		
10	1	0.1
11	313	42
12	313	42
13	114	15
Lessons		
No math advanced course	557	75
Math advanced course	184	25
Probability in lessons	573	77
Probability not in lessons	168	23

First of all, aggregated sum scores were investigated separately for each context. ANOVA shows significant differences between groups ($F=5.75$, $p<.01$) which disappear when excluding contexts 13 and 14 ($F=0.64$, $p=.79$). However, this only indicates that for contexts 13 and 14, examinees solved more items than for other contexts. Contexts 13 and 14 were given in one specific school with students of higher ability on average because of the subsequent data collection. All other contexts were distributed evenly across several schools. Thus, the higher mean of scores obviously does not require exclusion of contexts 13 and 14 per se because basic parameter estimates need not to be affected by the higher average ability. To consider if there are systematic deviations of basic parameter estimates for some contexts, LLTM analyses were run separately for every context. Results are shown in table 7.5 (IIE estimates are left out because of collinearity of IIE and IDE; see section 7.4.4). Sample size for every context ranges from 95 to 111. Adding and subtracting one standard deviation from the means for each basic parameter (averaged across contexts, last two lines in table 7.5) reveals that there are some contexts whose single basic parameter estimates differ from this range. However, no context shows systematic deviations for all basic parameter estimates.

Figures 7.2 and 7.3 visualize these relations: Only the constant is considerably

Table 7.5: Basic parameter estimates for each context

Context	Constant (SE)	CE (SE)	IDE (SE)	SDE (SE)	id (var) (SE)
1	1.30 (0.23)	-0.95 (0.19)	-2.44 (0.23)	-0.66 (0.21)	2.83 (0.68)
2	1.10 (0.21)	-1.22 (0.20)	-1.61 (0.20)	-1.55 (0.21)	1.68 (0.46)
3	1.62 (0.25)	-1.39 (0.21)	-2.81 (0.25)	-0.77 (0.21)	3.03 (0.74)
4	1.58 (0.25)	-1.16 (0.20)	-2.02 (0.22)	-1.25 (0.21)	3.29 (0.76)
5	0.80 (0.23)	-1.32 (0.21)	-2.59 (0.26)	-0.94 (0.23)	2.71 (0.70)
6	1.98 (0.26)	-1.57 (0.21)	-2.37 (0.23)	-1.23 (0.21)	3.39 (0.79)
7	1.35 (0.26)	-1.28 (0.21)	-2.44 (0.25)	-0.83 (0.22)	4.37 (0.99)
8	1.65 (0.27)	-1.47 (0.23)	-2.91 (0.28)	-1.18 (0.24)	4.51 (1.04)
9	1.45 (0.24)	-1.21 (0.21)	-2.72 (0.25)	-0.90 (0.22)	3.12 (0.76)
10	1.08 (0.22)	-1.22 (0.20)	-2.40 (0.23)	-0.62 (0.20)	2.34 (0.59)
11	1.63 (0.23)	-1.44 (0.21)	-2.38 (0.24)	-1.58 (0.22)	2.14 (0.57)
12	1.37 (0.25)	-1.32 (0.21)	-2.69 (0.25)	-0.62 (0.22)	3.37 (0.81)
13	3.03 (0.34)	-0.85 (0.21)	-2.77 (0.25)	-0.89 (0.21)	3.29 (0.81)
14	2.86 (0.35)	-1.01 (0.21)	-2.47 (0.25)	-0.77 (0.21)	4.50 (1.09)
Mean est. (SD)	1.63 (0.63)	-1.24 (0.20)	-2.47 (0.34)	-0.99 (0.32)	3.18 (0.85)
Mean SE (SD)	0.26 (0.04)	0.21 (0.01)	0.24 (0.02)	0.22 (0.01)	0.77 (0.18)

Notes: SE = standard error, SD = standard deviation, est. = estimates. CE = "complement events", IDE = "intersection of dependent events", SDE = "set union for disjoint events". ID = Person variance. Mean est. = mean estimate.

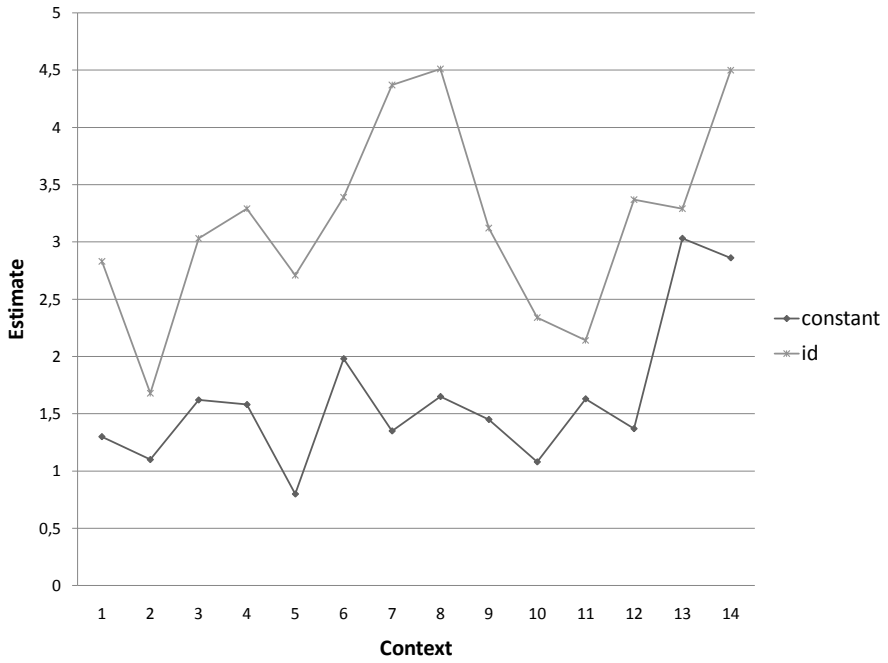


Figure 7.2: LLTM basic parameter estimates for each context (constant and id)

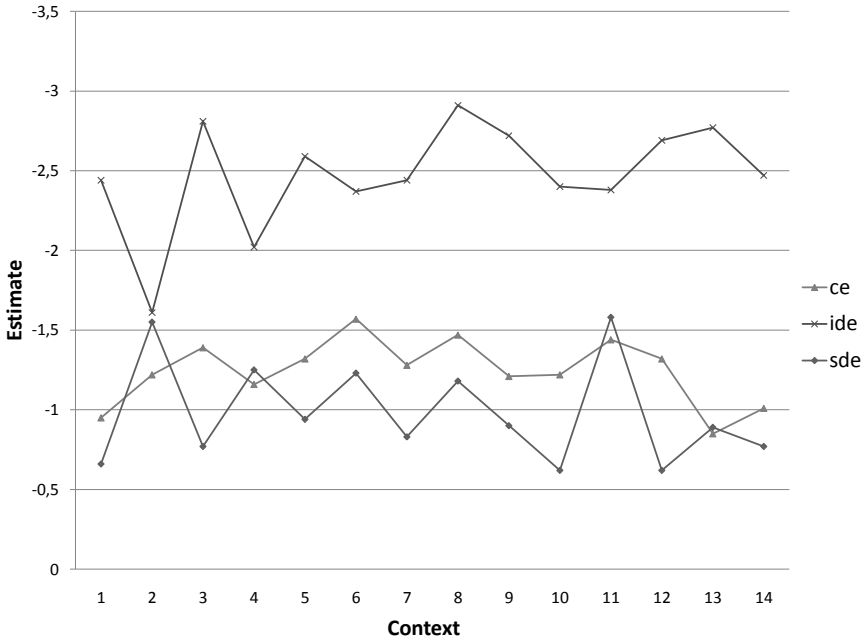


Figure 7.3: LLTM basic parameter estimates for each context (CE, IDE, SDE)

higher for contexts 13 and 14. Since the constant captures some kind of "basic item difficulty", this is not surprising, given the higher average scores for contexts 13 and 14. Since basic parameter estimates seem not to be affected, contexts 13 and 14 are not excluded from further analyses.

Additionally, a LLTM was modeled with the two parameters "context" and "family" as fixed and as random effects. Results show that "context" has far lower albeit significant influence than "family" (fixed effects: family = -0.14 with SE = 0.01, context = 0.03 with SE = 0.01; random effects: family = 2.55 with SE = 1.28, context = 0.33 with SE = 0.15). Excluding several contexts which are suspicious because of the above mentioned separate basic parameter estimates did not lead to non-significant context influence. Modeling "context" and CE, IDE and SDE together reveals again only low impact of context (CE = -1.18 with SE = 0.05, IDE = -2.38 with SE = 0.06, SDE = -1.00 with SE = 0.05, context fixed = 0.03 with SE = 0.01, context random variance = 0.28 with SE = 0.13).

For item cloning, it is also desirable that variances in item difficulties within item families are much lower than between families. This indicates that incidentals' impact on variance is only small. Table 7.6 shows that for the current items this requirement is fulfilled (please compare means of standard deviations across

contexts and families, 0.24 versus 0.10).

Table 7.6: CTT single item difficulties

	Family									
	1	2	3	4	5	6	7	8	Mean	SD
Context										
1	0.26	0.83	0.21	0.50	0.20	0.57	0.19	0.42	0.40	0.23
2	0.36	0.84	0.20	0.39	0.16	0.32	0.13	0.25	0.33	0.23
3	0.29	0.86	0.11	0.47	0.14	0.61	0.20	0.39	0.38	0.26
4	0.33	0.90	0.26	0.54	0.25	0.46	0.22	0.32	0.41	0.23
5	0.24	0.72	0.08	0.28	0.06	0.44	0.12	0.33	0.29	0.22
6	0.39	0.92	0.20	0.53	0.28	0.50	0.15	0.39	0.42	0.24
7	0.26	0.86	0.22	0.42	0.22	0.49	0.19	0.39	0.38	0.22
8	0.31	0.87	0.14	0.42	0.13	0.50	0.15	0.41	0.36	0.25
9	0.28	0.84	0.14	0.48	0.17	0.50	0.15	0.42	0.37	0.24
10	0.23	0.83	0.16	0.36	0.16	0.52	0.14	0.40	0.35	0.24
11	0.36	0.86	0.11	0.50	0.15	0.42	0.11	0.30	0.35	0.25
12	0.20	0.90	0.20	0.39	0.18	0.55	0.18	0.40	0.38	0.25
13	0.54	0.98	0.35	0.76	0.36	0.74	0.36	0.71	0.60	0.23
14	0.51	0.97	0.34	0.72	0.41	0.71	0.41	0.61	0.58	0.21
Mean	0.32	0.87	0.19	0.48	0.20	0.52	0.19	0.41		0.24
SD	0.10	0.06	0.08	0.13	0.09	0.11	0.09	0.12	0.10	

Notes: SD = standard deviation.

Overall, it can be concluded that there seems to be little influence of contexts that can be ignored for further analyses. These results lead to the further proceeding: Items are summarized to item numbers 1 to 16 as given in test, ignoring context variation.

7.4.3 Item characteristics, dimensionality and item fit

Item characteristics from CTT and item fit indices for all 16 items are shown in table 7.7. Cronbach's Alpha is .87. Item difficulty indicates that except for items 1 and 9 the test was of appropriate difficulty for the sample. However, the second half of the test seems to be easier on average. Discrimination indices can be regarded sufficient for most items.

Then, data were checked for dimensionality and item fit. Results from Winmira for Q-index (Rost & von Davier, 1994; Rost, 2004) revealed misfit for item 11 (however,

Table 7.7: Item difficulty, discrimination indices and Q-index for word problems

Item	Item difficulty (SD)	Item discrimination	Q-index	p Q-index
Item 1	.85 (0.36)	.28	0.15	.05
Item 2	.16 (0.36)	.51	0.11	.46
Item 3	.43 (0.50)	.48	0.13	.05
Item 4	.14 (0.34)	.49	0.12	.40
Item 5	.31 (0.46)	.56	0.10	.45
Item 6	.44 (0.50)	.49	0.12	.11
Item 7	.15 (0.36)	.55	0.09	.88
Item 8	.35 (0.48)	.50	0.12	.09
Item 9	.88 (0.32)	.29	0.12	.24
Item 10	.23 (0.42)	.63	0.07	.99
Item 11	.53 (0.50)	.45	0.14	.03
Item 12	.27 (0.44)	.65	0.06	.99
Item 13	.34 (0.47)	.64	0.06	.99
Item 14	.60 (0.49)	.50	0.10	.60
Item 15	.23 (0.42)	.64	0.06	.99
Item 16	.46 (0.50)	.53	0.10	.52

Notes: SD = standard deviation. Item difficulty and item discrimination are indices from classical test theory.

as already mentioned, 1 percent level may be too rigid for the current sample size) and marginal misfit for items 1 and 3. Taking into account the cloning procedure, the misfit of item 1 (belonging to family 2) is not confirmed by misfit of item 9 (also family 2). Since item 9 fits well, item 1 can be regarded sufficient for further analyses. However, items 3 and 11 both belong to family 4. In order to find out if there is something wrong with family 4, Q-indices were computed separately for items 1 to 8 and for items 9 to 16 (i.e., separated according to families). Only the Q-index for item 11 is statistically significant, but not for items 1 and 3. Additionally, analyzing all single item clones (no aggregation over contexts, sample size per item ranges from 47 to 56) reveals no misfitting items.

Additionally, Andersen Likelihood-Ratio-Test (Andersen, 1973) shows no significant examinee group differences (Andersen χ^2 = 17.14, df = 15, $p > .05$, groups defined by age lower/equal or higher than 18) and Martin-Löf-Test (Verhelst, 2001) shows no significant item group differences (Martin-Löf-statistics = 65.45, df = 63, $p > .05$, groups defined by SDE and 74.87, df = 63, $p > .05$, groups defined by CE, respectively), except for split half grouping (Martin-Löf-statistics = 131.41,

df = 63, $p < .01$). This may be due to some kind of practice effect during testing. Familiarity with test material and item type probably leads to slightly less difficult items. Martin-Löf-statistics do not indicate significant item group differences if items 3 and 11 (both family 4, with Q-index-misfit as explained above) are tested against the remaining items (Martin-Löf-statistics = 46.99, df = 27, $p > .05$). Therefore it is decided not to exclude items from further statistical analyses. Excluding items would complicate interpretation issues regarding cloning and seems not to be required by fit analyses. It can be concluded that Rasch fit is sufficient to allow for LLTM analyses.

7.4.4 LLTM results

As in chapter 5, four LLTM variants were investigated: LLTM, RE-LLTM, LR-LLTM, and a combination of LR-LLTM with RE-LLTM. Additionally, AIC and BIC as well as the Log-Likelihood for LR tests are given. Results are shown in table 7.8.

One important remark has to be made: Because of collinearity of IIE and IDE (i.e., items contain either IIE or IDE), only one basic parameter is included for both IIE and IDE which denotes presence or absence of IDE (if IDE is present, IIE is not and vice versa).

Basic parameters as specified in the Q-matrix all are statistically significant. IDE is the most difficult basic parameter, followed by CE and SDE. Of course, RE-LLTM fit is significantly better than LLTM fit. Random item effects are a good choice as it is almost always the case (cf. de Boeck, 2008). In LR-LLTM results, several person characteristics (math grade, German grade, school type, class, extra motivational questions and math advanced course) have significant impact on item difficulty. Taking into account second level person predictors in LR-LLTM results in extreme decrease of person variance and constant which indicates that including these predictors helps explaining parts of person variance and basic difficulty of the items. The same is the case for RE-LR-LLTM. Little albeit significant impact of fluid intelligence measured by CFT, of attention measured by d2 and of the realistic and investigative scale of the AIST was found in the subgroup of examinees who took these tests. However, compared to the other parameters, these results are ignorable and no extra models for the subgroups of examinees who took these extra tests are presented here. No significant effect was found for gender and age as well as for probability in lessons.

Correlation between basic parameter estimates in all models is approximately 1. That is, order and relation of basic parameters remain stable across all model specifications.

Additionally, interactions between basic parameters and between basic parameters and person characteristics were investigated. Table 7.9 shows the interaction models. Interestingly, the random item effect reduces considerably when including interactions between basic parameters (compare random item variance for RE-LR-LLTM in table 7.8 with random item variance for RE-LR-interactions in table 7.9). Thus an important part of variance in item difficulty can obviously be explained by interactions between basic parameters. The full interaction model with random item effect and interactions between basic parameters and person characteristics shows a detailed picture of item difficulty composition.

To summarize the results for interaction and non-interaction LLTMs, it can be stated that stable impact of basic parameters (CE, IDE, SDE) as well as of several person characteristics were found (math and German grade, school type, class, extra questions and math advanced course). Including random item effects improves model fit even more.

As for LST results, item location parameters were reconstructed from basic parameter estimates (simply sum up the products from one Q-matrix line per item with basic parameter estimates with inversed signs from LLTM analysis) and compared with Rasch item locations (from Winmira). Correlation between Rasch and LLTM item locations is .90. Thus about 81 percent of the whole variance in item difficulty is explained by the basic parameters. Rasch and LLTM item locations and standard errors are summarized in table 7.10. Rasch item locations (as CTT item difficulties) show that items of the second test half (items 9 to 16) seem to be easier on average. LLTM item locations do not reflect this fact because they are computed from basic parameter estimates which results in identical locations for the first and second half.

As already explained for LST items, absolute differences between Rasch and LLTM item locations are computed and standardized: Again LLTM item locations are sum-normalized, subtracted from Rasch parameters and the difference is then divided by Rasch SEs. The results are shown in table 7.11. Again there are severe deviations between LLTM and Rasch item locations. As for LST, this suggests that there have to be further sources of variance in item difficulty apart from basic parameters.

Table 7.8: Word problem parameter estimates for LLTM variants

Parameter	LLTM (SE)	RE-LLTM (SE)	LR-LLTM (SE)	RE-LR-LLTM (SE)
Fixed effects				
Constant	1.57 (0.08)**	1.71 (0.36)**	-2.10 (0.99)*	-2.50 (1.12)*
CE	-1.18 (0.05)**	-1.21 (0.35)**	-1.18 (0.05)**	-1.21 (0.35)**
IDE	-2.38 (0.06)**	-2.53 (0.35)**	-2.38 (0.06)**	-2.53 (0.35)**
SDE	-1.00 (0.05)**	-1.06 (0.35)**	-1.00 (0.05)**	-1.06 (0.35)**
Math grade			-0.47 (0.07)**	-0.51 (0.07)**
German grade			-0.20 (0.08)**	-0.20 (0.08)*
School type			1.75 (0.17)**	1.99 (0.18)**
Class			0.27 (0.08)**	0.30 (0.08)**
Fun			0.29 (0.05)**	0.32 (0.05)**
Difficult			-0.30 (0.06)**	-0.31 (0.06)**
Example helpful			0.25 (0.05)**	0.28 (0.05)**
Math adv.			0.45 (0.14)**	0.47 (0.15)**
Random effects				
Person	3.03 (0.22)	3.52 (0.25)	1.54 (0.12)	1.79 (0.14)
Item		0.48 (0.17)		0.48 (0.18)
CE				
IDE				
SDE				
Fit statistics				
LL (df)	-5667.85 (5)	-5397.28 (6)	-5464.65 (13)	-5188.93 (14)
AIC	11345.69	10806.55	10955.30	10405.86
BIC	11382.59	10850.83	11051.25	10509.19
$\Delta\chi^2$ to LLTM	-	541.14**	406.39**	957.83**

Notes: * p < .05, ** p < .01. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. RE-LLTM = Random Effects LLTM, LR-LLTM = Latent Regression LLTM, RE-LR-LLTM = combined Random Effects and Latent Regression LLTM. CE = "complement events", IDE = "intersection of dependent events", SDE = "set union for disjoint events". School type: 0 = vocational school, 1 = gymnasium; Class = school class; Fun = extra question 1; difficult = extra question 2; Example helpful = extra question 3. Math adv.: 0 = no math advanced course participation, 1 = math advanced course participation.

Table 7.9: Word problem parameter interactions

Parameter	Basic inter-actions	Basic RE interactions	RE-LR inter-actions	RE-LR inter-actions full model
Fixed effects				
Constant	-2.58 (0.11)**	-2.70 (0.28)*	-1.51 (1.09)	-0.38 (1.18)
CE	-2.57 (0.10)**	-2.66 (0.35)**	-2.67 (0.35)**	-4.35 (0.94)**
IDE	-3.58 (0.10)**	-3.69 (0.35)**	-3.70 (0.35)**	-3.34 (0.40)**
SDE	-2.31 (0.10)**	-2.39 (0.35)**	-2.40 (0.35)**	-3.22 (0.41)**
Math grade			-0.51 (0.07)**	-0.51 (0.07)**
German grade			-0.20 (0.08)*	-0.20 (0.08)*
School type			1.99 (0.18)**	1.91 (0.21)**
Class			0.30 (0.08)**	0.23 (0.09)*
Fun			0.32 (0.05)**	0.25 (0.06)**
Difficult			-0.31 (0.06)**	-0.31 (0.06)**
Example helpful			0.28 (0.05)**	0.28 (0.05)**
Math adv. course			0.47 (0.15)**	0.11 (0.17)
Fixed Effects				
Interactions				
CE*IDE	1.21 (0.11)**	1.28 (0.40)**	1.28 (0.40)**	1.26 (0.41)**
CE*SDE	1.61 (0.11)**	1.62 (0.40)**	1.63 (0.40)**	1.58 (0.41)**
IDE*SDE	1.03 (0.11)**	1.05 (0.40)**	1.06 (0.40)**	0.95 (0.41)*
IDE * school type				-0.44 (0.20)*
SDE * school type				0.53 (0.19)**
CE * class				0.15 (0.07)*
IDE * math adv. course				0.41 (0.13)**
SDE * math adv. course				0.34 (0.12)**
SDE * Fun				0.13 (0.04)**
Random effects				
Person	3.37 (0.24)	3.53 (0.25)	1.79 (0.14)	1.81 (0.14)
Item		0.15 (0.06)	0.15 (0.06)	0.15 (0.06)
Fit statistics				
LL (df)	-5468.98 (8)	-5388.22 (9)	-5179.94 (17)	-5158.06 (23)
AIC	10953.95	10794.45	10393.87	10362.12
BIC	11013.00	10860.87	10519.34	10531.87
$\Delta\chi^2$ to LLTM	397.74**	559.26**	975.82**	1019.58**

Notes: * p < .05, ** p < .01. SE = standard error, LL = Log-Likelihood, df = degrees of freedom. RE = random effects, LR = latent regression. CE = "complement events", IDE = "intersection of dependent events", SDE = "set union for disjoint events". School type: 0 = vocational school, 1 = gymnasium; Class = school class; Fun = extra question 1; difficult = extra question 2; Example helpful = extra question 3. Math advanced course: 0 = no participation, 1 = participation.

Table 7.10: Rasch and reconstructed LLTM item location parameters and standard errors for word problems

Item	Loc. Rasch	SE Rasch	Loc. LLTM	SE LLTM
Item 1	-3.32	0.14	-1.57	0.01
Item 2	1.91	0.13	1.99	0.01
Item 3	-0.29	0.10	-0.39	0.01
Item 4	2.17	0.14	1.81	0.01
Item 5	0.55	0.10	0.81	0.01
Item 6	-0.37	0.10	-0.58	0.01
Item 7	1.98	0.13	2.99	0.01
Item 8	0.21	0.10	0.61	0.01
Item 9	-3.74	0.15	-1.57	0.01
Item 10	1.18	0.11	1.99	0.01
Item 11	-0.88	0.10	-0.39	0.01
Item 12	0.86	0.11	1.81	0.01
Item 13	0.33	0.10	0.81	0.01
Item 14	-1.32	0.10	-0.58	0.01
Item 15	1.17	0.11	2.99	0.01
Item 16	-0.45	0.10	0.61	0.01

Notes: LLTM item location parameters and standard errors reconstructed from basic parameter estimates and variances and covariances of estimates. Loc. = location, SE = standard error. Rasch results from Winmira, LLTM results from Stata.

Since for CDM analyses IIE has to be included, one LLTM without constant and with all basic parameters was estimated to allow comparison of all basic parameter estimates between LLTM and DINA. Results reveal that IIE is the easiest basic parameter and the only one with a positive effect on item difficulty. Hence including IIE makes items easier. At first sight, this sounds rather strange. However, if IIE is included in one item, IDE (which is the most difficult parameter) is not included in the same item and thus this item becomes easier.

7.4.5 CDM results

DINA results again can be shown in terms of estimated mastery class probabilities and skill probabilities. Model fit indices AIC and BIC are also reported. These results can be seen in table 7.12.

The most probable class masters all basic parameters except for IDE, of almost identical probability size is the class which masters all basic parameters. The

Table 7.11: Absolute differences between Rasch and LLTM item locations for word problems

Item	LLTM sum normalized	Rasch	Diff. div. by Rasch-SE
Item 1	-2.28	-3.32	-7.44
Item 2	1.28	1.91	4.84
Item 3	-1.10	-0.29	8.09
Item 4	1.10	2.17	7.63
Item 5	0.10	0.55	4.49
Item 6	-1.29	-0.37	9.19
Item 7	2.28	1.98	-2.32
Item 8	-0.10	0.21	3.09
Item 9	-2.28	-3.74	-9.74
Item 10	1.28	1.18	-0.92
Item 11	-1.10	-0.88	2.19
Item 12	1.10	0.86	-2.19
Item 13	0.10	0.33	2.29
Item 14	-1.29	-1.32	-0.31
Item 15	2.28	1.17	-10.10
Item 16	-0.10	-0.45	-3.51

Notes: Diff. = difference; div. = divided; SE = standard error.

third largest class masters only IIE, the fourth largest CE and IIE. The remaining classes of considerable probability size all reach the 5% mark: One class masters only CE, one only SDE, one CE and SDE, and the last one masters none of the basic parameters. All remaining classes are of ignorable size. Concerning the skill probabilities, IIE is the easiest one followed by CE and SDE, and IDE is the most difficult skill. For IIE and IDE, DINA results thus are in line with LLTM results. For CE and SDE, a converse order emerges: SDE is more difficult than CE which is not the case in LLTM results.

Mastery classes show also the possible merit of CDMs in case of skills as realized in the current study: It can be seen from table 7.12 that some mastery classes can be assigned to families. Classes assigned to families including dependent probability (odd families 1, 3, 5, and 7) do almost not occur on their own or in combination with classes assigned to other odd families. However, mastery classes of skills in even families which do not require understanding of dependent probability (families 2, 4, 6, and 8) occur on their own and in combination with other even families.

Table 7.12: Word problem model estimates for DINA

Class probabilities		
Class		DINA
0000		0.05
1000		0.05
0100	Family 1	0.00
0010	Family 2	0.11
0001		0.05
1100	Family 3 + Family 1	0.00
1010	Family 4 + Family 2	0.09
1001		0.05
0110		0.01
0101	Family 5 + Family 1	0.01
0011	Family 6 + Family 2	0.03
1110		0.01
1101	Family 7 + Family 5 + Family 1	0.01
1011	Family 8 + Family 6 + Family 2	0.27
0111		0.00
1111	All families	0.27
Skill probabilities		
Skill		DINA
CE		0.74
IDE		0.30
IIE		0.78
SDE		0.68
Fit indices		
Index		DINA
AIC		10573
BIC		10790
Mean MADprop		0.03

Notes: Class probability: Probability that class of attribute pattern occurs in sample. Attribute order in class: CE, IDE, IIE, SDE. Skill probability = probability that skill is mastered in the whole sample.

Table 7.13: Word problem item parameter and fit estimates for DINA

Item	Guess (SE)	Slip (SE)	(1-slip) -guess	MAD LOR	MAD cor	MAD prop
Item 1	0.54 (0.02)	0.06 (0.03)	0.39	0.99	0.03	0.05
Item 2	0.05 (0.01)	0.56 (0.04)	0.39	0.83	0.06	0.03
Item 3	0.08 (0.01)	0.36 (0.06)	0.57	0.56	0.07	0.04
Item 4	0.03 (0.01)	0.60 (0.03)	0.36	0.75	0.03	0.05
Item 5	0.11 (0.01)	0.24 (0.02)	0.65	0.94	0.11	0.00
Item 6	0.08 (0.01)	0.28 (0.03)	0.65	0.74	0.10	0.04
Item 7	0.02 (0.01)	0.51 (0.03)	0.47	1.26	0.07	0.01
Item 8	0.06 (0.01)	0.39 (0.04)	0.55	0.50	0.07	0.01
Item 9	0.57 (0.02)	0.03 (0.03)	0.40	1.60	0.03	0.04
Item 10	0.03 (0.00)	0.27 (0.02)	0.70	1.36	0.12	0.02
Item 11	0.18 (0.01)	0.27 (0.05)	0.56	0.63	0.08	0.04
Item 12	0.05 (0.01)	0.19 (0.01)	0.76	1.44	0.13	0.01
Item 13	0.09 (0.01)	0.09 (0.01)	0.81	1.46	0.11	0.01
Item 14	0.21 (0.02)	0.11 (0.01)	0.68	0.86	0.08	0.03
Item 15	0.04 (0.01)	0.27 (0.02)	0.70	1.41	0.12	0.02
Item 16	0.10 (0.01)	0.23 (0.03)	0.67	0.87	0.10	0.02
Mean	0.14 (0.01)	0.28 (0.03)	0.58	1.01	0.08	0.03

Notes: Guess = guessing parameter, slip = slipping parameter, SE = standard error, (1-slip)-guess = "CDM discrimination", MAD = Mean absolute difference, LOR = log odds ratio, cor = correlation, prop = proportion.

Table 7.13 shows guessing and slipping parameters as well as MAD and CDM discrimination indices. Guessing and slipping parameters are considerably better than for LST but still quite bad. Applying the liberal criterion of $1 - g - s$ being higher than .50 reveals that 11 out of the 16 items fulfill this criterion, but 8 of these 11 items show too high slipping parameters (much higher than .20). Again high slipping parameters can mainly be found for very difficult items and high guessing parameters mainly for easy items. In general, parameters are better than for LST, but far from good.

7.5 Discussion

The current study demonstrates rule-based design and half-automatic generation of statistical word problems within an item-cloning approach. Altogether, 112

items belonging to eight item families (specified as lines in the Q-matrix) and 14 contexts were constructed and tested with 741 German school students in gymnasium and vocational school. Results show Rasch scalability of the items as well as satisfying item difficulty and discrimination indices. Effects of basic parameters and several person characteristics were investigated by LLTM variants, in which random effects emerge as important factors, and DINA application which again does not reveal satisfying results.

It can be stated that in general rule-based design and item cloning for the constructed statistical word problems were successful. Despite the serious difficulties in word problem construction (cf. section 7.2.3), the current study demonstrates that rule-based design and item cloning can work for word problems when accounting for the numerous pitfalls during the construction process as well as for extreme accurateness during the item writing and cloning procedure.

Several analyses confirmed that the cloning procedure was successful and that context stories (the used incidentals) do not have too much impact compared to family effect so that test versions could be aggregated and treated as one version. This check-up was conducted through a) comparison of basic parameter estimates separated by context, b) LLTM modeling of context and family impact, and c) comparison of CTT means and standard deviations for item clones.

7.5.1 LLTM results

To investigate effects of basic parameters, six LLTM variants were specified. In all models, CE, IDE and SDE have significant impact on item difficulty. IIE is left out for most analyses because of collinearity between IIE and IDE. SDE is (apart from IIE) the easiest basic parameter. SDE requires simple addition of two probabilities and therefore its relative easiness can be explained by the simple underlying operations to solve items which require SDE. IDE is the most difficult operation. To solve items which require handling of IDE, there has to be an understanding and a modeling competency for dependent events and conditional probability which requires much more statistical competencies than the other involved basic parameters. CE, which requires only a subtraction operation to form the complement of an event is a bit more difficult than SDE. The size of CE and SDE is relatively similar, probably because of the similar operations required, and much lower than IDE. Including IIE shows that this is the easiest basic parameter. Two probabilities have to be multiplied which seems to be as

trivial as CE and SDE. However, the dependence of IDE and IIE results into complicated interpretation of IIE impact as every item which requires IIE does not require IDE and thus is easier because IDE is the most difficult basic parameter.

Compared to the results of Holling et al. (2009), the rank order of basic parameter impact is similar: IIE is the easiest, IDE the most difficult operation. However, in the current study CE is more difficult than SDE while in Holling et al. (2009), SDE was much more difficult than CE. It has to be kept in mind that both studies were conducted with different populations: Holling et al. (2009) tested university students, the current study is concerned with school pupils. The university students are supposed to be more familiar with the operations required by the used basic parameters than the school students.

This finding demonstrates that basic parameter estimates from one test in one study cannot directly be transferred to another test construction in another study, especially if different populations are considered. Basic parameter estimates from former works can help to find hypotheses and to select possible difficulty generating operations, but every new test construction and item type has to be investigated again carefully before drawing conclusions about parameter influences.

Including a random item effect improves model fit very much as it was already the case for LST. This is a common finding (cf. de Boeck, 2008) and shows that there are variance parts which are not captured by the basic parameters. Additionally, random effects provide for the assumption of the current item set as sample from an infinite item population. Among further difficulty influencing characteristics of items could be wording and grammar (perhaps not all items are equally easy to read and understand albeit identical basic wording because of the cloning approach) as well as more or less appealing contents (familiar contents can be easier to process than unfamiliar ones) and numerical information (big numbers perhaps are more difficult than lower numbers although all computations could be conducted without calculator and led to smooth results). The impact of German grades on item difficulty (see next paragraphs) supports the assumption that reading comprehension and competencies may affect item difficulty.

LR-LLTM analyses reveal several person characteristics influencing item difficulty significantly in addition to the basic parameters. Results show that better math and German grades lead to reduced item difficulty. This finding confirms the supposed importance of mathematical as well as reading and text comprehension competencies. Math grade has a higher impact than German grade which implies

that the items mainly measure mathematical competencies which was intended. However, the fact that word problems are easier for students who have higher reading and text comprehension skills is not surprising as these students will be able to build up an internal problem model faster and more precisely than students who have more problems with text comprehension and reading (Jonassen, 2003; Nathan, Kintsch, & Young, 1992).

A very high impact was found for school type: For gymnasium students, items are much more easy than for students from vocational school. This is not surprising because both school forms focus on slightly different skills. In gymnasium, students are often taught to find alternative solution paths, to transfer modeling competencies from one subject to another, to find out new rules and to learn and apply them quickly. Additionally, general requirements are higher for gymnasium students. In vocational school, practical and application skills are of more interest than in gymnasium. Vocational school students probably do not learn theoretical modeling competencies and knowledge transfer as intensively as gymnasium students. Additionally, probability theory may not be treated as often as in gymnasium because of differing curricula.

The extra motivational questions show that for students who had fun in solving the items, items were less difficult. The more difficult items were experienced by the examinees, the more difficult items were indeed. For examinees who found the instruction examples helpful, items were less difficult, perhaps because these examinees used the instruction examples more intensely than others during the test. For examinees who took part in a math advanced course, items were easier, probably because of more intrinsic mathematical interest, mathematical practice and skills and perhaps because of more familiarity with the item type and probability theory.

Including these person characteristics results into extremely reduced random person variance which shows that the chosen person characteristics at least partly explain variance parts within the whole structure of the individual item solution processes. The fact that person variance is still greater than zero demonstrates that there have to be further person characteristics which can play a role in explaining individual scores and solution processes. Additionally, the constant is extremely reduced which implies that the included person characteristics help to explain why basic item difficulty is higher for some examinees and lower for others. Taking into account the measurement scale of the person predictors (math grade, German grade, class and extra motivational questions which have nearly interval scales),

most of them have lower impact on item difficulty than CE, IDE and SDE (except for school type). Combining RE-LLTM with LR-LLTM leads to even better model fit. Therefore, additional item characteristics seem to be of importance for item difficulty (as mentioned above).

Furthermore, several significant interactions between person characteristics and basic parameters were found: School type, class, math advanced course and fun show significant interactions with basic parameters (IDE is easier for vocational school students and math advanced course participants, CE is easier for students from higher class levels, and SDE is easier for gymnasium students, math advanced course participants and those who had more fun doing the test) and provide incremental variance explanation in addition to their fixed isolated effects.

Note that including person predictors leads to decrease of random person variance (compare LR-LLTM to LLTM in table 7.8) as it was the case for LST. It has also to be underlined that differences in model fit between RE-LLTM and LR-LLTM are considerably lower for the word problems than for LST. Hence person characteristics and interactions seem to play a more important role for word problems than for LST. This may be due to the supposed importance of wording in items and prior knowledge about probability theory while in LST such possible sources of unexplained variance are not existent.

Furthermore, interaction analyses show that including interactions between basic parameters reduces random item variance (compare RE-LLTM in table 7.8 to RE interaction models in table 7.9). Thus interactions between basic parameters are important factors affecting item difficulty beyond single basic parameters. this demonstrates that interactions between basic parameters mainly influence random item variance and thus have considerable impact on the remaining variance parts in item difficulty beyond basic parameters themselves, while person characteristics (as included in all LR-models) mainly reduce random person variance and the constant. This means that person characteristics affect basic item difficulty and general score and explain interindividual differences of the examinees. Including interactions between basic parameters and person characteristics again reduces the constant but not random person variance and thus only provides incremental explanation of basic item difficulty, but not of interindividual differences.

Including these significant interactions between basic parameters and between person characteristics and basic parameters into the former RE-LR-LLTM leads to better model fit compared to all other models. However, this complex inter-

relationship demonstrates that the best fitting model needs not to be the most helpful one. For sake of parsimony, the RE-LR-LLTM provides elaborated insight into item construction characteristics, basic parameter structure and further item difficulty influencing variables and enough results to draw conclusions about item construction and solution processes.

Although it can be assumed that there are further factors (both item and person characteristics) which affect item difficulty as can be concluded from person variance and random item effect, explanation of item difficulty by the chosen basic parameters is quite good. Correlation between Rasch and reconstructed LLTM item locations is .90, thus about 81 percent of the variance in item difficulties can be explained by the chosen basic parameters. However, as already for LST note that this only indicates that the chosen basic parameters are important factors influencing item difficulty. As Rasch parameters show higher variance than LLTM parameters and this variance is included into estimation of person parameters, person parameters are not explained as well by the chosen basic parameters. Correlation between basic parameters of all specified models is approximately 1. That is, order and relation of basic parameters remain stable across model specifications.

However, as differences between Rasch and reconstructed LLTM item locations show, there are still serious deviations between "true" and reconstructed difficulties. This finding is in line with the model fit improvement by including random effects in the RE-LLTM. Thus Q-matrix misspecification can again not be ruled out in this case. But again one has to keep in mind that LLTM basic parameters probably never will be able to explain empirical item difficulties perfectly. In the preceding argumentation, several additional reasons as interactions between basic parameter effects and impact of person characteristics were mentioned which can help to explain additional difficulty variance and differences between Rasch and LLTM item parameters. Altogether, given the relative complex and problematic item type and the elaborate cloning procedure, these results can be regarded a good outcome.

7.5.2 CDM results

As it was the case for LST, at first sight CDM findings for the current item type overall correspond to LLTM results. Overall attribute mastery probabilities indicate that IIE is the easiest basic parameter, followed by CE, SDE and IDE. Note that CE is mastered by more examinees than SDE which is not the case in LLTM

analyses. However, CE and SDE do not differ much in size in LLTM analyses and 74 versus 68 percent in DINA results is also not an extreme difference. In line with this different rank order of basic parameters, attribute mastery patterns, i.e., class frequencies, show that there are no classes of substantial size in which difficult parameters as IDE are mastered and easier ones are not mastered. The most probable class masters all basic parameters but IDE (the most difficult basic parameter), and the only class which does not match LLTM results is the one in which CE and IIE are mastered (8.6 percent). Following LLTM results, this class should have displayed mastery of IIE and SDE. Additionally, as explained in section 4.2.2, CDM and LLTM do not have to provide identical results because different algorithms are conducted.

Apart from these results again demonstrating inadequacy of DINA for the current item type, assigning families to mastery classes shows advantages of DINA application in principle. It can be clearly concluded which families are mastered better than others: Odd families which include dependent probability are extremely less well mastered than even families which require only independent probability. This direct mapping of mastery of attribute combinations in item families can only be detected by DINA results in this definite way.

7.5.3 Comparison of LLTM and CDM

However, guessing and slipping parameters again show that DINA modeling is not adequate for the constructed word problems. Although guessing and slipping parameters are much better than for LST, the current results clearly lead to the conclusion that DINA again is not able to explain the empirical data adequately. MAD indices show satisfying model and item fit. Again, high guessing parameters are mainly found for easy items and high slipping parameters for difficult items. The proportion of items reaching the $1 - g - s$ higher than .50 criterion together with acceptable slipping parameters is far too low to be interpreted seriously. The reasons for bad guessing and slipping parameters described for LST in chapter 5 also apply to the word problems in this study. Although the word problems are supposed to be based on well learn- and trainable skills rather than on a latent unidimensional concept for cognitive capacity like LST, DINA outcomes do not match the expectations regarding results especially for guessing and slipping parameters. This could be a hint that the underlying construct is unidimensional rather than multidimensional and thus better captured by a log-linear approach

with additive decomposition of item difficulty into basic parameters. Again not skills themselves may set limits but rather combination of skills (cf. LST results in study 5). This imposes the question what exactly is measured by the word problems. Perhaps they simply measure rule application as a unidimensional construct and not mastery of the underlying statistical concepts. General implications of DINA results from LST and word problems will be discussed in chapter 8.

Again, as for LST, LLTM rationale shows up to be the more reasonable rationale in data explanation for LST and model and item fit indices as well as explained variance underline adequacy of LLTM application in this case. LLTM results demonstrate satisfying explanation of item difficulty in word problems and detect important processes for item solution. Thus it has again to be reasoned that LLTM is superior and preferable to DINA in this study.

7.5.4 Limitations and prospects

Collinearity of basic parameters became a problem during LLTM analyses. However, suppressing the constant allows for identification of all parameter effects and omitting IIE in main LLTM analyses in the current study did not impair further analyses and interpretation of results.

The second test half suffers from obvious practice effects as the items of the second half are in general easier than their identically designed siblings of the first half. This is not mapped by LLTM reconstructed item locations (because locations are simply computed from basic parameter estimates and thus identical for the first and second test half) but only by separate Rasch and CTT analyses. This demonstrates inaccurateness of LLTM results, but this is no serious problem for clarification of the question which basic parameters have significant impact on item difficulty. Additionally, absolute differences between Rasch and LLTM item locations are very high and random effects in LLTM lead to extreme model fit improvement, indicating lots of variance in item difficulty left which is not accounted for. Identification of further difficulty-affecting variables (beyond the investigated person characteristics) therefore should be the focus of subsequent research.

Word problems provide much room for automated item generation and item cloning through the shown means of variations in context and numerical information. The half-automatic generation by text templates and free variables is a great

help for more efficient item construction. However, text templates still have to be written and arranged manually. Full automated item generation is still very difficult for word problems (see explanations in section 7.2.3). Grammar, flexion, case demands etc. impose hundreds and thousands entries in a data base for this end and require cooperation with linguists. This effort must not be underestimated, but a full automated generator for word problems could simplify the work for lots of teachers in school and university as well as for researchers and large test settings.

The current word problems can be seen as a prototype to demonstrate the possibility and quality of rule-based and half automated generation and cloning of word problems. The item cloning concept will be extended to other statistical contents. Adaptive implementation and automatic generation of these items are under progress at the moment. They can be extended to other content areas not only in mathematics but also chemistry and physics, for example. A first automatic item generator which provides items with almost identical structure as the items in the current study has been finished already. Further research is urgently needed to pick up and continue the demonstrated results. The perfect final state could be seen in a software tool which automatically generates word problems of high (because proven by LLTM analyses) quality from almost arbitrary domains and presents them adaptively to the examinee. Additional features as tutorials and advanced examples can be included.

8 General discussion

The current work aims at description, demonstration and application of rule-based item generation and statistical methods to analyze and control item quality. Two item types and their rule-based generation as well as their empirical testing are described, including item cloning and longitudinal results. Two different model classes, i.e. linear logistic test and cognitive diagnostic models, are used to analyze empirical results. Both model classes are based on different assumptions: LLTMs are based on a log-linear approach and focus mainly on item difficulties, CDMs are based on a mixture approach and focus mainly on person characteristics and attribute mastery classes. Results from model application provide information about the structure of the items and the constructs under consideration.

8.1 Rule-based construction principles

The demonstrated item generation approach has led to several item sets of two item types: Altogether, four LST test versions as a working memory capacity measure and operationalization of RC theory as well as probability word problems as a measure of statistical competence generated within an item cloning approach were constructed. For both item types, rule-based generation and Q-matrix design can be regarded successful as basic parameters are confirmed to be significant predictors of item difficulty for both word problems and LST items. Additionally, basic parameters remain stable difficulty influencing factors even when additional test results and person characteristics as well as interactions with these other variables are included.

Please note that Q-matrix construction is conducted by rule-based principles for both LLTM and CDM analyses, but that interpretation of models is extremely different: While LLTMs are unidimensional models based on resources in which contents are only realized by occupation of resources, and thus have an intermediate resource level. CDMs are discrete models which shall map the structure of contents themselves without an intermediate resource level.

In LST, significant basic parameters are completely in line with RC theory and confirm the quality of the item construction process consistently with RC theory and the first results of Birney et al. (2006). For probability word problems, all basic parameters were chosen based on a detailed school book review and all were found to have significant impact on item difficulty as defined in the Q-matrix. In longitudinal LST, all basic parameters show considerable effects on item difficulty across all test sessions. Additionally, basic parameter specific learning effects occur across test sessions while basic parameters themselves still remain significant item difficulty influencing factors.

Hence theoretically based construction principles as defined in the Q-matrices were all in all confirmed in the conducted empirical studies. Correlations between LLTM and Rasch item locations show that between 81 and 87 percent of overall item difficulty can be explained by the chosen basic parameters which is quite a good result (cf. Embretson, 1998; Freund et al., 2008; Preckel, 2003). Thus, the described accurate item construction was rewarded by affirming empirical results and can be regarded successful. However, as already discussed in the preceding study-specific discussion parts, absolute differences between Rasch and LLTM item locations are very high and point to lack of explanation of basic parameters. These findings are supported by the strong model fit improvement when including a random error term in LLTM. These results are further discussed in the following sections.

In general, the shown Q-matrix generation and item construction principles establish very efficient and economic item generation which provides great possibilities for large scale testing and prevention from cheating and undesired practice / training effects enabled by, for example, internet distribution of test items. It can be strongly recommended to use the advantages of rule-based item generation whenever possible in constructing new tests.

Careful statistical analyses are crucial for check-up of successful item generation when using rule-based item construction. Theoretical and empirical assumptions and knowledge about difficulty-influencing basic parameters are important in item generation though not sufficient. Only careful statistical analyses with the described model classes ensure confirmation of the role of the chosen basic parameters. Therefore comparing model classes with regard to modeling quality and adequacy will help to gain decision support which models to chose in order to explain empirical data and underlying constructs as good as possible.

8.2 Statistical modeling

Two model classes, LLTMs and CDMs, were used to analyze empirical data from rule-based item generation. LLTM variants turned out to be very versatile and flexible while providing clear and easy to understand and to interpret results. At first sight, DINA results mainly provide similar results concerning skill probabilities and mastery classes compared to LLTM results. However, both model classes are based on completely different assumptions. High guessing and slipping parameters of DINA point to inadequacy of DINA application to the current data and item types. While LLTM variants shaped up as excellent tool for several research questions as basic parameter effects and person characteristics influence on item difficulty as well as for longitudinal modeling, DINA results cannot be used for further interpretation as DINA application seems to be inadequate.

8.2.1 LLTM

Several LLTM variants were applied in the current work. The basic LLTM results simply display the impact of the Q-matrix design parameters. In the studies for LST and statistical word problems, LLTM revealed definite results for the chosen parameters. Order and impact relation of the basic parameters did not change for results from extensions of the basic LLTM. RE-LLTM always fits better than LLTM. This is, together with the higher standard errors, a common finding for RE-LLTM (de Boeck, 2008). This fact can be explained by and confirms the assumption that the chosen set of basic parameters does not capture all the variance in item difficulty. Basic parameters may have significant and rather considerable influence on item difficulty, but there will probably never be the ideal set of parameters which totally captures all variance (cf. Freund et al., 2008). Although a sufficient explanation of item difficulty by basic parameters is desirable, identification of the exhaustive parameter set is probably not the main purpose in rule-based item construction. In fact, identification of a sound parameter set capturing main variance components in item difficulty is more important than a messy huge set of all possible impact factors. In many cases, perfect explanation will only be reached by assigning each item its own basic parameter in the Q-matrix, leading to absurdity of LLTM (as well as CDM) application. Even if an ideal set of parameters can be identified, there remains another problem: Analyzing empirical data always suffers from some noise resulting in random variance within examinee answers. In

RE-LLTM, the random item effect captures variance both from unknown cognitive components involved in item solution which are not specified in the Q-matrix, and random variance. Since one can hardly avoid the former and not avoid the latter, RE-LLTM will probably always fit better than LLTM. Additionally, random effects take into account the assumption that the tested item set is only a random sample from a whole item population.

The rationale from RE-LLTM applies to LR-LLTM, too. Both LLTM variants capture variance in addition to the variance explained by the basic parameters. LR-LLTM helps to identify which additional (person-specific) factors affect item difficulty. For all three data sets, several person characteristics and additional test results were found to affect item difficulty significantly. Based on these results, conclusions can be drawn about factors which make items easier for some individuals but not for other. To investigate which additional characteristics make specific basic parameters easier for some individuals, one has to examine interaction effects. The demonstrated results show that there are only few factors for all data sets which show significant interactions with basic parameters and hence explain incremental variance. Far more characteristics were found to affect general item difficulty, and these characteristics were often related to general intelligence. The explanation that test items from various domains are easier for examinees of higher intelligence is no surprising fact (Kulik et al., 1984).

RE-LLTM shows that there is even more random variance from other sources which were not identifiable by the used study designs and included variables. Including person characteristics as well as random item effects results in reduced random person and item variance which indicates that some parts of this random variance can be captured by additional variables. However, the order of basic parameter effects and their correlation between different model variants indicate that the chosen parameters are stable and well-selected concepts.

Combining all these possibilities to explain item difficulty leads to combinations of random item effects and additional person characteristics in one model. The complete RE-LR-LLTM (together with interactions between some basic parameters and person characteristics for the word problems) has the best fit of all variants for the described items and samples and shows a detailed picture of impact factors for item difficulty. However, it has to be pointed out that the main purpose of LLTM analysis here is and remains identification and confirmation of basic parameters. For rule-based item construction this is the main information. The described additional variables help to resolve the question which characteristics apart from

basic parameters make items difficult but are not essential for evaluating the item construction process. In principle, all described LLTM variants provide enough information to evaluate the construction process.

Changes in random item and person variance as well as changes in constant values for the full models can be interpreted in terms of different parameter effects: Including person predictors reduces random person variance. Inclusion of person characteristics and of interactions between person characteristics and basic parameters led to reduction of random person variance as well as of the constant for both LST and word problems. This shows that the chosen person characteristics explain parts of interindividual differences between examinees and of basic item difficulty. This means that person characteristics are identified which cause items to be more or less difficult for specific groups of examinees with certain characteristics. Furthermore, interactions between basic parameters explain parts of the random item variance for the word problems but not for LST. Inclusion of interactions between basic parameters led to reduction of random item variance for the word problems but not for LST (no considerable interactions).

Additionally, model fit improvement by inclusion of further parameters allows for conclusions about relations between item and person characteristics concerning their impact on item difficulty: While for LST items including random item effects leads to extreme model fit improvement and including person characteristics leads to far less model fit improvement, different results are found for the word problems. For the latter, model fit improvement (compared to basic LLTM) by LR-LLTM and RE-LLTM shows no such great differences as in LST. It can be concluded that in LST much random item variance is left which cannot be completely explained by the current study design and the included variables. For the word problems, person characteristics as well as interactions between basic parameters seem to explain more variance (which can be also seen in the extreme reduction of random item and person variance in the described models) than they do in LST. A possible explanation could be that random item and person variance in LST simply reflect different ability levels in reasoning and working memory capacity which are not captured by any other variables. Thus, the current results demonstrate the multiple interactive structure of the constructed word problems with their multiple influencing factors and the relative unidimensionality of LST with its close parameter influence structure.

The application of the longitudinal LLTM is intended to demonstrate another helpful variant to investigate basic parameter development across several test

sessions. As no adequate model variant exists for application in this context, specific learning parameters were used. The described results show impressively the contribution of LLTM to theoretical (how does parameter influence change across test sessions) and practical (what can we learn from these changes for repeated testing, for example in selection settings) research questions. Despite the mentioned problems (no intentional specification of variance-covariance matrix of errors as well as confounding of learning effects and item cloning) results turned out to be very good. Learning effects are mapped adequately and the general learning effect is decomposed into basic parameter specific effects which shows how worthwhile this approach is for understanding learning in repeated testing.

The here demonstrated advantages and diversity of linear logistic modeling promises further application of this model class to a variety of research contexts and questions. Recent empirical studies which apply LLTM and LLTM variants (e.g. Hohensinn et al., 2008; Poinstingl, 2009; Xie & Wilson, 2008) demonstrate the usefulness of linear logistic modeling for a plenty of research questions and empirical applications.

However, there are also some limitations to mention. Absolute differences between Rasch and LLTM item locations are extremely high for both LST and word problems. This meets the findings about model fit improvement by inclusion of random effects and person predictors into the basic LLTM, yielding the RE- and LR-LLTM, respectively. Altogether, these results support the certainty that there have to be further important variables affecting item difficulty. Partly these variables are met in the LR-LLTMs as person characteristics turn out to be of important impact on item difficulty. But still random effects improve model fit, leaving open the question which variables could explain the remaining parts of variance. Another explanation which cannot be ruled out is misspecification of the Q-matrices (see also the following section). If Q-matrices are misspecified, the chosen basic parameters cannot explain for variance. There is always a Q-matrix which is definitely not misspecified: The matrix which assigns one basic parameter to each item. Deviations from this matrix inevitably lead to misspecifications of more or less seriousness. Thus misspecified Q-matrices may also be an explanation for the above mentioned findings of non-explained variance left.

One should keep in mind that the ideal basic parameter set may not exist at all. The aim of LLTM is to decompose item difficulty into basic construction parameters, thereby reducing the number of parameters in the model. Perfect explanation and reconstruction of item locations may only be possible by assigning

each item its own basic parameter - ending up in the Rasch model again, reducing LLTM application to absurdity. The same reasoning applies to CDM application, discussed below.

8.2.2 CDM

The here shown CDM application examples and results provide only little insight into the area of cognitive diagnostic modeling. Despite the theoretically sound basis for cognitive diagnostic modeling, there are only few software implementations which are well-documented. Extensions of software applications as well as implementation of important features like solid fit statistics (for example, embedding likelihood-ratio-tests) are strongly needed.

The resulting high guessing and slipping parameters of the current studies constitute a serious problem for model application and interpretation. High slipping and guessing parameters do not make logical sense but nevertheless often emerge during empirical data analysis (cf. also Templin & Henson, 2006; Templin & Ivie, 2006) which indicates too rigid model assumptions and questions the basic model conception and applicability to several task types at least of DINA. For word problems, slipping and guessing parameters are not as extreme as for LST, but still are really high. This might be due to the more difficult items (word problems compared to LST) as guessing and slipping parameters are often correlated with item difficulty (cf. Templin & Henson, 2006), but still is not tolerable for data explanation purposes. In a strict sense, the assumption that guessing and slipping up to .20 constitute good fit (cf. de la Torre & Douglas, 2004) is actually a bold one as 20 percent guessing and slipping probability should not be regarded as acceptable results for test items. The deterministic linkage in DINA forces uncertainty and noise to be captured in guessing and slipping parameters, thereby "rescuing" moderate to good fit indices and logical mastery patterns. In the current work, the more liberal criterion of $1 - g - s$ being higher than .50 while showing slipping parameters much lower than .20 was applied, but also this criterion was not met by most of the constructed items.

As explained in section 4.2.3, guessing and slipping results can provide insight into the problem structure by which cognitive demands are imposed. If skill application itself does not set real limits for cognitive capacity but combination of skills does, one-rule problems would impose difficulties for DINA as these problems would be solved nearly perfect with resulting high guessing parameters. Items with

many rules then would produce high slipping and low guessing parameters. Both cases were found in the current data sets: For LST as well as word problems, high guessing parameters were found for easy items (mainly consisting of few basic parameter requirements) and high slipping parameters were mainly found for difficult items (based on many rules). These findings contribute to the conclusion that DINA application is not adequate for the current item types.

Another topic concerns dimensionality of the underlying constructs. As DINA emerges to be inadequate to model the data, it can also be assumed that the underlying constructs are better thought of as unidimensional ones rather than as discrete skills one can learn and train. For the latter case, DINA would be supposed to reveal better results, capturing the multidimensional structure by its mixture approach.

During the current work, it arose from the literature and software conditions that CDMs are a relatively complicated and not easy to handle model class. Rigid model assumptions compared to LLTM, relative inflexibility and complicated model equations combined with the software situation may have discouraged many practitioners and also theoreticians from further research and application of CDMs. Additionally, CDM estimation often requires high hardware standard and lasts very long (Rupp & Templin, 2008a). The works of several researchers (e.g., de la Torre, 2008; de la Torre & Douglas, 2004; Henson, Templin, & Willse, 2008; Templin & Henson, 2006) are welcome exceptions which show that CDMs are often wrongly disregarded and deserve more attention in research and practice.

Moreover, CDMs provide an interesting access to investigation of knowledge spaces (Doignon & Falmagne, 1999; Falmagne et al., 1990). They can provide information about possible learning paths of sub-abilities: Which components are learned first, second or in parallel can be retrieved in principle from class probabilities. Classes can reveal that there is no strict linear order of skill acquisition but that some individuals master one skill after another and other individuals master these skills in a reversed order.

Altogether, there seems to be a vicious circle in CDM research: Because of the seemingly complexity and unhandiness as well as inadequacy for probably many item types as shown in the current work, only few empirical applications can be found in the literature (cf. section 4.2.2 in chapter 4) which in turn prevents reasonable development of CDM application to empirical data. Simulation studies show that CDMs are a theoretically sound and functional concept, but empirical

application suffers from the described shortcomings as sample size, software, and debatable model conception leading to high slipping and guessing parameters. The current work again shows the severe problems of at least DINA in explaining empirical data from moderately complex items.

As several other studies report similarly bad guessing and slipping parameters, the question arises if this can be a systematic problem of DINA. The deterministic assumptions forcing any deviating examinee behavior into guessing and slipping parameters perhaps disqualifies DINA for many empirical applications in psychological research. In educational contexts with concepts that are either mastered or not mastered, DINA can still be a relevant statistical approach. However, the described problems even for the word problems, which in principle can be thought of as consisting of learnable skills, reduces hopes of adequate data basis for DINA application. These findings raise the question how items have to be designed to accomplish satisfying DINA results.

Another possible reason for bad DINA outcomes could be a misspecification of the design matrix. However, as described in section 4.2.3, for perfect state reproduction and minimal guessing and slipping parameters, the design matrix would probably often need as many basic parameters or skills as items. There is no clear criterion how large the deviation of the design matrix from the ideal case (of every item getting its own skill) is allowed to be while still providing interpretable results. This means uncertainty about possible misspecification of the here used design matrices as it cannot be decided if this (and not the constructs to be measured) is the reason for the bad guessing and slipping parameters and how the design matrix would have to look like to provide interpretable results.

8.2.3 Model fit and model choice

It has to be pointed out that model fit and significance of parameter results reveal specific theoretical and empirical characteristics of the data set and serve as decision help to choose variables for model specification. However, effects of basic parameters as defined in design matrices are in principle of higher practical meaning. Identification and confirmation of difficulty generating cognitive rules is an important aim in LLTM or CDM analyses and in several cases basic parameter identification or confirmation is of higher interest than finding the (probably non-existing) model with perfect fit. There may be rules which are statistically significant but not of practical importance as their absolute effect is low compared

to other rules. For sake of parsimony such impact factors may be ignored as sometimes they can dilute the whole basic parameter structure. However, even if model fit is not ideal and some design parameters are not statistically significant at all, they can anyhow be of practical value for test construction. For example, basic design parameters without significant impact on item difficulty can be used to vary surface characteristics in item cloning because this causes no shift in item difficulty.

Correlation between basic parameter estimates in different model specifications can help to evaluate if basic parameter estimates are stable and significant across several model specifications. The relation between single parameter estimates in one model should not be altered by defining another model. If parameter relations change between models (and thus correlation is low), one should become suspicious of parameter impact (except for parameters of nearly identical impact whose order can change due to minimal differences in parameter size). The absolute size of parameter estimates is no sufficient criterion to evaluate parameter impact as this size can be dependent of model specification. Additionally, standard errors are dependent of model specification (e.g. SEs in RE-LLTM are usually higher than in LLTM, cf. de Boeck, 2008). The only definite conclusion about parameter effects can be drawn from the relation between and size-order of parameters within the same model and from the steadiness of this order between different model specifications.

The nature of research questions as well as the constructs to be measured help to find the adequate analysis method and not vice versa: If one is only interested in check-up of item construction, random effects and person or group predictors for mainly unidimensional constructs, LLTM variants are the first choice. If one is interested in individual diagnosis, feedback and learning paths or strategy alternatives for distinct skills or learnable concepts, CDMs should be applied. Both model classes may provide similar results with regard to basic parameter impact if basic parameters resemble skills, but modeling results will also help to decide which model is adequate and how the nature of the constructs to be measured is like, as it was the case for both item types of the current studies.

8.2.4 Comparison of LLTM and CDM

The current work leads to the conclusion that LLTM variants are much more easy to apply and to handle and more adequate to explain the current data results

than DINA. LLTM rationale shows up to be the more reasonable rationale in data explanation for both LST and word problems: Basic parameter inclusion makes items more difficult. For longitudinal analysis, DINA did not provide interpretable results at all while LLTM shows flexible modeling of learning effects. Model and item fit indices as well as explained variance underline adequacy of LLTM application in all three studies. LLTM results demonstrate satisfying explanation of item difficulty and detect important processes for item solution. Thus it has to be reasoned that LLTM is superior and preferable to DINA for both LST and word problems in this work.

There remains the more general question if the results one can obtain from CDMs at all legitimate the relative rigidity and unhandiness of this model class, particularly thinking of the limited number of cases in which DINA may be applicable. In practice, for example in school context and selection settings, CDM application is complex and expensive due to the required sample sizes and software applications. Additionally, the here shown problems concerning guessing and slipping parameters question applicability of at least DINA in many contexts. Thus it is stated clearly here that at least for LST and word problems, DINA is not able to explain data accurately which can be supposed to pertain to other item types similar to the ones used in these studies.

No matter which methods are appropriate and applied, the law of parsimony should always be accounted for and one should take care not to lose track of the main questions during analysis. It should be pointed out that at first sight the parsimony of LLTM is attractive and constitutes an advantage against CDMs and other model classes, but that this parsimony may not lead to sufficient mapping of empirical data (cf. Rijmen & de Boeck, 2002). Hence, one has to find a compromise between parsimony and accurateness. This especially applies to the topic of design matrix definition. The absolute differences between Rasch and LLTM item locations as well as extreme model fit improvement by random effects in LLTM shows that there is much variance left which is not accounted for by the chosen basic parameter sets. Similarly, DINA results can also be interpreted to point to possible design matrix misspecifications. However, reproducing item difficulty or person states perfectly is the aim neither of LLTM nor of DINA application. Deciding to decompose item difficulty into basic units of difficulty generating factors which is the case both for LLTM and for DINA means compromises concerning explanation of item difficulty. Though the aim to find the best possible parameter set should always be of top priority, number of parameters cannot be expanded endlessly

without contradicting basic assumptions of both model classes. Thus definite decision criteria indicating tolerable deviations between perfect explanation and adequate decomposition are strongly needed.

8.3 General limitations

One problem in rule-based (and also in automatic) item generation is the limitation of free combinability of basic parameters. The current work demonstrates combination of different rules and item design based on these rules. However, for both LST and probability word problems there are severe constraints regarding combination of rules. As mentioned in chapter 5, LST suffers from possible ambiguities due to failed combination of complexity steps. Every item has to be checked for unambiguity carefully because simply adding one step, no matter of which complexity, can disturb the unambiguity. In probability word problems, IIE and IDE cannot be combined easily because these two basic parameters are based on contrary concepts (there cannot be dependence and independence at the same time). Also the described facts impose serious problems for automatic item generation and optimal design issues because several constraints in the design matrix have to be taken into account.

However, in research this issue is often simply ignored (as it is the case in the current work). Perfect free combinability can probably hardly be reached at all. Especially when many basic parameters are used, there will probably always be some constraints preventing free combinability. Interactions between basic parameters further complicate advanced item construction procedures and require perhaps more than one calibration cycle or systematic specification of parameter combinations in the design matrix to test possible interaction effects of basic parameters directly. At least a minimum of human responsibility and input will thus be necessary for rule-based design and automatic generation as well as item cloning for many item types until considerable progress in artificial intelligence.

Concerning statistical modeling, one important limitation is that with DINA only one representative of CDMs was applied. Nevertheless many of the mentioned points of criticism apply to other examples of the CDM class and demonstrate serious reasons to think of and apply CDMs carefully.

It has also to be noted that the samples tested in this work consist exclusively of school pupils. This means that rather selected samples were tested, probably

reducing variance of outcomes. However, for the current item types this sample selection is rather adequate as both cognitive capacity and statistical competence should be of sufficient intensity in school pupils of an age between 16 and 20 years to allow for measurement of these constructs.

8.4 Future work

The current work and the described results provide a good basis for future developments. First, enlarged sample sizes can be involved to confirm stability of the described results and to extend the generalizability (not only school and university students, for example). Word problems can and will be extended by including more statistical concepts. A broader range of basic parameters would provide even more possibilities to apply these word problems in practical contexts as school and university (for example, exam preparation or selection settings). The cloning approach can be easily extended by more context stories and variation of numerical information, sentence structure or linguistic variations.

Another important development is the automated generation of both LST and word problems which is presently under construction. The software applications will provide versatile possibilities for the test administrator to define Q-matrices and content areas to test with a lot of free adjustment possibilities (for example, feedback of success and person ability or recording of response times). Moreover, the programs will be extended to adaptive testing and thus provide an extremely efficient testing method.

The CDM review of Rupp and Templin (2008b) states that there are several future assignments to be fulfilled in CDM development and application as well as software implementation. However, sophisticated software documentations and freeware implementations are strongly needed to make sure practitioners can easily use CDM software applications in order to gain more information about possible inherent problems of DINA in empirical application. For LLTM analysis, interesting and well-documented freeware implementations exist already in R (lme4 package, see lme4 documentation and Doran, Bates, Bliese, & Dowling, 2007) and are supposed to be extended at the moment.

Altogether, it can be expected that rule-based item construction, item cloning, automated item generation and the necessary statistical methods will gain more

and more importance in research (cf. also Freund et al., 2008). For example, large-scale test settings like PISA employ more and more probabilistic methods and careful item construction principles and will surely extend this in future research, leading to more efficient, flexible and economic testing procedures.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*, 30-60.

Agbor-Baiyee, W. (2009). A study of cognitive achievement in a special premedical program. *College Student Journal*, *43*, 36-44.

Amthauer, R., Beauducel, A., Brocke, B., & Liepmann, D. (2001). *Intelligenz-Struktur-Test 2000 R [Intelligence structure test] (I-S-T 2000 R)*. Göttingen: Hogrefe.

Andersen, E. B. (1973). A goodness of fit test for the Rasch Model. *Psychometrika*, *38*, 123-140.

Andrews, G., & Halford, G. S. (2001). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*, 153-219.

Arendasy, M. (2005). Automatic generation of Rasch-calibrated items: Figural matrices test GEOM and Endless-Loops test E-super(c). *International Journal of Testing*, *5*, 197-224.

Arendasy, M., Sommer, M., Gittler, G., & Hergovich, A. (2006). Automatic generation of quantitative reasoning items: A pilot study. *Journal of Individual Differences*, *27*, 2-14.

Awh, E., & Vogel, E. K. (2008). The bouncer in the brain. *Nature Neuroscience*, *11*, 5-6.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, *17*, 201-210.

Barrouillet, P., Lépine, R., & Camos, V. (2008). Is the influence of working memory capacity on high-level cognition mediated by complexity or resource-dependent elementary processes? *Psychonomic Bulletin & Review*, *15*, 528-534.

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, *14*, 237-245.

Bejar, I. I. (1993). A generative approach to psychological and educational mea-

surement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (p. 323-357). Hillsdale, NJ: Erlbaum.

Bergmann, C., & Eder, F. (1999). *Allgemeiner Interessen-Struktur-Test [General interest structure test] (AIST)*. Göttingen: Beltz Test.

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.* Reading, MA: Addison-Wesley.

Birney, D. P., & Bowman, D. B. (2009). An experimental-differential investigation of cognitive complexity. *Psychology Science Quarterly*, *51*, 449-469.

Birney, D. P., & Halford, G. S. (2002). Cognitive complexity of suppositional reasoning: An application of the relational complexity metric to the knight-knave task. *Thinking and Reasoning*, *8*, 109-134.

Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the influence of complexity on relational reasoning. The development of the Latin Square Task. *Educational and Psychological Measurement*, *66*, 146-171.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar [Neo five factor inventory] (NEO-FFI)*. Göttingen: Hogrefe.

Bors, D. A., & Vigneau, F. (2003). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences*, *13*, 291-312.

Bowman, D. B. (2006). *An investigation of the determinants of cognitive complexity and individual differences in fluid reasoning ability*. Unpublished doctoral dissertation, University of Sydney, Australia.

Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition and Instruction*, *1*, 245-296.

Brickenkamp, R. (2002). *Aufmerksamkeits-Belastungs-Test [Attention stress test d2] (Test d2)*. Göttingen: Hogrefe.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261-304.

Burns, G. N., Siers, B. P., & Christiansen, N. D. (2008). Effects of providing pre-test information and preparation materials on applicant reactions to selection procedures. *International Journal of Selection and Assessment*, *16*, 73-77.

Byrne, B. M., & Crombie, G. (2003). Modeling and testing change: An introduction to the Latent Growth Curve Model. *Understanding Statistics*, *2*, 177-203.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science*, *6*, 42-57.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures:

A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404-431.

Carroll, J. B. (2005). The Three-Stratum Theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (p. 69-76). New York: Guilford Press.

Cisse, D. (1995). *Modeling children's performance on arithmetic word problems with the linear logistic test model*. Unpublished doctoral dissertation, University of Alberta, Canada.

Cliffordson, C. (2004). Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment*, *20*, 192-204.

Colom, R., Abad, F. J., Quiroga, M. A., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, *36*, 584-606.

Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences*, *42*, 1503-1514.

Colom, R., & García-López, O. (2002). Sex differences in fluid intelligence among high school graduates. *Personality and Individual Differences*, *32*, 445-451.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*, 51-57.

Coyle, T. R. (1998). Test-retest changes on scholastic aptitude tests are not related to *g*. *Behavioral and Brain Sciences*, *21*, 803-865.

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. *Cognition and Instruction*, *8*, 261-289.

Curran, P. J., Bauer, D. J., & Willoughby, M. T. (2004). Testing main effects and interactions in latent curve analysis. *Psychological Methods*, *9*, 220-237.

de Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533-559.

de Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333-353.

diBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sin-

haray (Eds.), *Handbook of statistics 26* (p. 979-1030). Amsterdam, Netherlands: Elsevier.

Dimitrov, D. M. (1996). Cognitive item subordinations in linear logistic test modeling. *Dissertation Abstracts International*, *57(6-A)*, 2452.

Doignon, J. P., & Falmagne, J. C. (1999). *Knowledge Spaces*. Berlin: Springer.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*, 1-18.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380-396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407-433.

Enright, M. K., Morley, M., & Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation on item statistical characteristics. *Applied Measurement in Education*, *15*, 49-74.

Falmagne, J. C., Koppen, M., Villano, M., Doignon, J. P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test and search them. *Psychological Review*, *97*, 201-224.

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

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*, 599-624.

Formann, A. K., & Spiel, C. (1989). Measuring change by means of a hybrid variant of the linear logistic model with relaxed assumptions. *Applied Psychological Measurement*, *13*, 91-103.

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.

Garber, H. L. (1988). *The Milwaukee Project: Preventing mental retardation in children at risk*. Washington, D.C.: American Association of Mental Retardation.

Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (in press). Modeling rule-based item generation. *Psychometrika*.

Gierl, M. J., & Zhou, J. (2008). Computer adaptive-attribute testing – a new approach to cognitive diagnostic assessment. *Zeitschrift für Psychologie / Journal of Psychology*, *216*, 29-39.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247-261.

Glück, J., & Indurkhya, A. (2001). Assessing changes in the longitudinal salience

of items within constructs. *Journal of Adolescent Research*, *16*, 169-187.

Glück, J., & Spiel, C. (1997). Item response models for repeated measures designs: Application and limitations of four different approaches. *Methods of Psychological Research*, *2*, 1-19.

Glück, J., & Spiel, C. (2007). Using item response models to analyze change: Advantages and limitations. In A. D. Ong & M. H. M. van Dulmen (Eds.), *Oxford handbook of methods in positive psychology* (p. 349-361). Oxford: University Press.

Gold, B. (2008). *Stabilität von psychometrischer Intelligenz - Wechselwirkungen mit Testängstlichkeit und selbsteingeschätzter Intelligenz [Stability of psychometric intelligence - interactions with test anxiety and self-rated intelligence]*. Unpublished diploma thesis, University of Münster, Germany.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*, 351-373.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, *12*, 369-381.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, *50*, 379-390.

Halford, G. S., & Andrews, G. (2002). Young children's performance on the balance scale: The influence of relational complexity. *Journal of Experimental Child Psychology*, *81*, 417-445.

Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Science*, *11*, 236-242.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803-865.

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.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*, *87*, 243-254.

Henson, R., Douglas, J., Roussos, L., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement*, *32*, 275-288.

Henson, R., Templin, J. L., & Willse, J. T. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191-210.

Hoffmann, N. (2007). *Lateinische Quadrate: Psychometrische Eigenschaften und Lerneffekte [Latin Square Task: Psychometric properties and learning effects]*. Unpublished diploma thesis, University of Münster, Germany.

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, *50*, 391-402.

Holling, H., Bertling, J. P., & Zeuch, N. (2009). Probability word problems: Automatic item generation and LLTM modeling. *Studies in Educational Evaluation*, *35*, 71-76.

Holling, H., Blank, H., Kuchenbäcker, K., & Kuhn, J.-T. (2008). Rule-based item design of statistical word problems: A review and first implementation. *Psychology Science Quarterly*, *50*, 363-378.

Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*, 603-618.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *105*, 6829-6833.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In De Boeck, Paul and Wilson, Mark (Ed.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 189-210). New York: Springer.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport: Prager.

Jäger, A. O., Süß, H. M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test [Berlin test of intelligence structure] (BIS)*. Göttingen: Hogrefe.

Jonassen, D. H. (2003). Designing research-based instruction for story problems. *Educational Psychology Review*, *15*, 267-296.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258-272.

Kessler, Y., & Meiran, N. (2008). Two dissociable updating processes in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1339-1348.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, *13*, 129-164.

Krumm, S., Ziegler, M., & Bühner, M. (2008). Reasoning and working memory as predictors of school grades. *Learning and Individual Differences*, *18*, 248-257.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, *69*, 232-244.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*, 179-188.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity ?! *Intelligence*, *14*, 389-433.

Lang, J. W. B., & Bliese, P. D. (2009). General mental ability and two types of adaptation to unforeseen change: Applying discontinuous growth models to the task-change paradigm. *Journal of Applied Psychology*, *94*, 411-428.

Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and application*. Cambridge: University Press.

Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics: A rogue economist explores the hidden side of everything*. New York: Morrow.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lynn, R., & Irwing, P. (2002). Sex differences in general knowledge, semantic memory and reasoning ability. *British Journal of Psychology*, *93*, 545-556.

Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science*, *10*, 135-175.

Mayer, R. E. (1987). Learnable aspects of problem solving: Some examples. In D. E. Berger, K. Pezdek, & W. P. Banks (Eds.), *Applications of cognitive psychology: Problem solving, education, and computing* (p. 109-122). Hillsdale, NJ: Erlbaum.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence". *American Psychologist*, *28*, 1-14.

McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, *40*, 808-821.

McGlone, M. S., Aronson, J., & Kobrynowicz, D. (2006). Stereotype threat and the gender gap in political knowledge. *Psychology of Women Quarterly*, *30*, 392-398.

McGrew, K. S. (2005). The Cattell-Horn-Carroll Theory of cognitive abilities: Past, present and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (p. 136-181). New York: Guilford Press.

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and Instruction*, *9*, 329-389.

Nkaya, H. N., Huteau, M., & Bonnet, J. (1994). Retest effects on cognitive performance on the Raven-38 Matrices in France and in the Congo. *Perceptual and Motor Skills*, *78*, 503-510.

Oberauer, K. (2009). Design for a working memory. *The psychology of learning and motivation*.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (p. 49-75). New York: Oxford University Press.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*, 641-652.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch Modeling in the context of psychotherapy outcome assessment. *Applied Psychological Measurement*, *30*, 100-120.

Pauls, F. (2009). *Theory-driven item construction and IRT equating of parallel test forms for measuring reasoning ability*. Unpublished diploma thesis, University of Münster, Germany.

Plackett, R. (1983). Karl Pearson and the Chi-squared Test. *International Statistical Review*, *51*, 59-72.

Poinstingl, H. (2009). The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly*, *51*, 123-134.

Preckel, F. (2003). *Diagnostik intellektueller Hochbegabung. Testentwicklung zur Erfassung der fluiden Intelligenz [Assessment of intellectual giftedness: Test development for the assessment of fluid intelligence]*. Göttingen: Hogrefe.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.

Raven, J. C. (1938). *Progressive Matrices: A perceptual test of intelligence, 1938, sets a, b, c, d, and e*. London: H. K. Lewis.

Raven, J. C. (1962). *Advanced Progressive Matrices, set II*. London: H. K. Lewis.

Read, T., & Cressie, N. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York: Springer.

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, *33*, 535-549.

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [A questionnaire for recording actual motivation in learning and achievement situations]. *Diagnostica*, *47*, 57-66.

Rijmen, F., & de Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 271-285.

Rijmen, F., de Boeck, P., & van der Maas, H. L. J. (2005). An IRT model with a parameter-driven process for change. *Psychometrika*, *70*, 651-669.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726-748.

Roid, G., & Haladyna, T. (Eds.). (1981). *A technology of test-item writing*. New York: Academic Press.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion [Textbook test theory - test construction]* (2nd ed.). Bern, Göttingen: Huber.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement*, *18*, 171-182.

Rupp, A. A., & Templin, J. (2008a). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78-96.

Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219-262.

Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text. *Learning and Individual Differences*, *8*, 141-160.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research. *Psychological Bulletin*, *124*, 262-274.

Sebrechts, M. M., Enright, M., Bennett, R. E., & Martin, K. (1996). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. *Cognition and Instruction*, *14*, 285-343.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. Oxford:

University Press.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics*, *28*, 295-313.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, *50*, 345-362.

Spearman, C. E. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A metaanalysis of sex differences in interests. *Psychological Bulletin*, *135*, 859-884.

Tatsuoka, K. (1990). Toward an integration of item response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (p. 453-488). Hillsdale, NJ: Erlbaum.

te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on *g*-loaded test: No *g*. *Intelligence*, *35*, 283-300.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, *32*, 559-574.

Templin, J. L., & Ivie, J. L. (2006, April). *Analysis of the Raven's Progressive Matrices (RPM) scale using skills assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME).

Thorell, L. B., Lindqvist, S., Bergman Nutley, S., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science*, *12*, 106-113.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*, 104–132.

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.

Verhelst, N. (2001). Testing the unidimensionality assumption of the Rasch model. *Methods of Psychological Research Online*, *6*, 231-271.

Watkins, M. W., Lei, P., & Canivez, G. L. (2007). Psychometric intelligence and

achievement: A cross-lagged panel analysis. *Intelligence*, *35*, 59-68.

Weiß, R. H. (1998). *Grundintelligenztest Skala 2 [Test of basic intelligence scale 2] (CFT 20)*. Göttingen: Hogrefe.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 [Test of basic intelligence scale 2] (CFT 20-R Revision)]* (4. ed.). Göttingen: Hogrefe.

Wilhelm, O. (2000). *Psychologie des schlussfolgernden Denkens [The psychology of reasoning]*. Hamburg: Kovac.

Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly*, *50*, 403-416.

Xin, Y. P. (2007). Word problem solving task in testbooks and their relation to students performance. *The Journal of Educational Research*, *100*, 347-360.

Zeuch, N., Geerlings, H., Holling, H., van der Linden, W. J., & Bertling, J. (2010). Regelgeleitete Konstruktion von statistischen Textaufgaben: Anwendung von linear logistischen Testmodellen, Itemcloning und Optimal Design [Rule-based item generation of statistical word problems based upon linear logistic test models for item cloning and optimal design]. In E. Klieme, D. Leutner, & M. Kenk (Eds.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes. 56. Beiheft der Zeitschrift für Pädagogik [German journal of education]* (p. 52-63). Weinheim: Beltz.

Appendix

A Design details

The following section show the design details of the described studies. First, design details of LST versions are presented, then details of word problem construction are described.

A.1 LST version 1 to 4

Table A.1: Detailed design matrix LST 1

Item	Size	B	T	Q	Steps	Contrad.	DR	col	KR	col	KU	col	PL	col	QT	col	SE	col	ST	col	TR	col	WE	col
1	0	2	0	0	2	0											1		2		3		4	
2	0	0	1	0	1	0					1	4					4	2			3	3	2	1
3	0	1	1	0	2	0	3				2						4						1	
4	0	2	1	0	3	0			1	3	4	2			2	4							3	1
5	0	1	2	0	3	0	2	1					1	4							4	2	3	3
6	0	1	2	0	3	0			4		2								3		1			
7	0	2	2	0	4	0	4		2						1						3			
8	0	0	0	1	1	1	2	1	4	4	3	3	1	2										
9	0	1	1	0	2	1	1		3				4						2					
10	0	1	3	0	4	0	1	2							3	1	2	3	4	4				
11	0	0	0	1	1	0			3				4		2		1							
12	0	0	0	1	1	0			2								4		1				3	
13	0	1	3	0	4	0			1	3							2	4			3	1	4	2
14	0	0	0	1	1	0					1				3				4		2			
15	0	0	0	1	1	0					4	2	2	4			3	1			1	3		
16	0	2	2	0	4	0							2	4	1	3			3	1			4	2
17	1	0	2	0	2	0	3								5		2				1		4	
18	1	2	1	0	3	0	1	4			5	2			4	3	2	1	3	5				
19	1	0	1	1	2	0					3		5		1				2		4			
20	1	2	2	0	4	0	2				1		4				3				5			
21	1	1	0	1	2	0	4	5	2	1	3	3			1	2	5	4						
22	1	2	3	0	5	0	5				4								2		3		1	
23	1	1	0	1	2	0			5				3				4		1		2			
24	1	1	1	0	2	1	2	4			1	2	4	3			3	1			5	5		
25	1	2	3	0	5	0			5	2			1	5	4	3			2	1			3	4
26	1	2	4	0	6	0			4		2						5				3		1	
27	1	0	2	0	2	0			1	5					3	2			5	4	2	3	4	1
28	1	2	2	0	4	0	4	3	2	1			3	4	5	2					1	5		
29	1	1	1	0	2	1					5		2		4		3						1	
30	1	0	1	1	2	0	4	5	2	1	3	4			1	3	5	2						

Notes. Size: 0 = 4 lines ×4 columns, 1 = 5 lines ×5 columns; B = binary, T = ternary, Q = quarternary; Steps = sum of steps; Contrad. = contradictional (0 = not contradictional, 1 = contradictional, excluded from further analyses). DR = triangle, KR = circle, KU = cross, PL = plus, QT = square, SE = hexagon, ST = star, TR = trapezoid, WE = wave; col = color (1 = white, 2 = light grey, 3 = dark grey, 4 = black, 5 = hatched).

Table A.2: Detailed design matrix LST 2

Item	Size	B	T	Q	Steps	Contrad.	DR	col	KR	col	KU	col	PL	col	QT	col	SE	col	ST	col	TR	col	WE	col
1	0	2	0	0	2	0											1		2		3		4	
2	0	0	1	0	1	0					1	4					4	2			3	3	2	1
3	0	1	1	0	2	0					4		2				3				1			
4	0	2	1	0	3	0	4	2	2	4					1	1					3	3		
5	0	1	2	0	3	0			1	3	4	2			2	4							3	1
6	0	1	2	0	3	0			4		2								3		1			
7	0	2	2	0	4	0							2		1				3				4	
8	0	0	0	1	1	1	2	1	4	4	3	3	1	2										
9	0	1	1	0	2	1	3				2						4						1	
10	0	1	3	0	4	0			3	2			4	1	2	3	1	4						
11	0	0	0	1	1	0	1								3		2		4					
12	0	0	0	1	1	0			2								4		1				3	
13	0	1	3	0	4	0			1	3							2	4			3	1	4	2
14	0	0	0	1	1	0	1		3				4						2					
15	0	0	0	1	1	0	2	1					1	4							4	2	3	3
16	0	2	2	0	4	0					1	4			3	3			4	1	2	2		
17	1	0	2	0	2	0			5				1		4				2				3	
18	1	2	1	0	3	0			5	5			3	1			4	3	1	2	2	4		
19	1	0	1	1	2	0	3								5		2				1		4	
20	1	2	2	0	4	0					3		5		1				2		4			
21	1	1	0	1	2	0	1	4			5	2			4	3	2	1	3	5				
22	1	2	3	0	5	0	2				1		4				3				5			
23	1	1	0	1	2	0			1						3				5		2		4	
24	1	1	1	0	2	1	2	4	3	3	4	1	1	5									5	2
25	1	2	3	0	5	0	4	3	2	1			3	4	5	2					1	5		
26	1	2	4	0	6	0	5				4								2		3		1	
27	1	0	2	0	2	0			4	2	2	5					5	3			3	1	1	4
28	1	2	2	0	4	0					5	4	2	3	4	5	3	2					1	1
29	1	1	1	0	2	1	4		2		3				1		5							
30	1	0	1	1	2	0	4	5					5	1			3	4	1	3			2	2

Notes. Size: 0 = 4 lines ×4 columns, 1 = 5 lines ×5 columns; B = binary, T = ternary, Q = quarternary; Steps = sum of steps; Contrad. = contradictional (0 = not contradictional, 1 = contradictional, excluded from further analyses). DR = triangle, KR = circle, KU = cross, PL = plus, QT = square, SE = hexagon, ST = star, TR = trapezoid, WE = wave; col = color (1 = white, 2 = light grey, 3 = dark grey, 4 = black, 5 = hatched).

Table A.3: Detailed design matrix LST 3

Item	Size	B	T	Q	Steps	Contrad.	DR	col	KR	col	KU	col	PL	col	QT	col	SE	col	ST	col	TR	col	WE	col
1	0	2	0	0	2	0											1		2		3		4	
2	0	0	1	0	1	0					1	4					4	2			3	3	2	1
3	0	1	1	0	2	0	3				2						4						1	
4	0	2	1	0	3	0			1	3	4	2			2	4							3	1
5	0	1	2	0	3	0	2	1					1	4							4	2	3	3
6	0	1	2	0	3	0			4		2								3		1			
7	0	2	2	0	4	0	4		2						1						3			
8	0	0	0	1	1	1	2	1	4	4	3	3	1	2										
9	0	1	1	0	2	1	1		3				4						2					
10	0	1	3	0	4	0	1	2							3	1	2	3	4	4				
11	0	0	0	1	1	0			3				4		2		1							
12	0	0	0	1	1	0			2								4		1				3	
13	0	1	3	0	4	0			1	3							2	4			3	1	4	2
14	0	0	0	1	1	0					1				3				4		2			
15	0	0	0	1	1	0					4	2	2	4			3	1			1	3		
16	0	2	2	0	4	0							2	4	1	3			3	1			4	2
17	1	0	2	0	2	0	3								5		2				1		4	
18	1	2	1	0	3	0	1	4			5	2			4	3	2	1	3	5				
19	1	0	1	1	2	0					3		5		1				2		4			
20	1	2	2	0	4	0	2				1		4				3				5			
21	1	1	0	1	2	0	4	5	2	1	3	3			1	2	5	4						
22	1	2	3	0	5	0	5				4								2		3		1	
23	1	1	0	1	2	0			5				3				4		1		2			
24	1	1	1	0	2	1	2	4			1	2	4	3			3	1			5	5		
25	1	2	3	0	5	0			5	2			1	5	4	3			2	1			3	4
26	1	2	4	0	6	0			4		2						5				3		1	
27	1	0	2	0	2	0			1	5					3	2			5	4	2	3	4	1
28	1	2	2	0	4	0	4	3	2	1			3	4	5	2					1	5		
29	1	1	1	0	2	1					5		2		4		3						1	
30	1	0	1	1	2	0	4	5	2	1	3	4			1	3	5	2						

Notes. Size: 0 = 4 lines ×4 columns, 1 = 5 lines ×5 columns; B = binary, T = ternary, Q = quarternary; Steps = sum of steps; Contrad. = contradictional (0 = not contradictional, 1 = contradictional, excluded from further analyses). DR = triangle, KR = circle, KU = cross, PL = plus, QT = square, SE = hexagon, ST = star, TR = trapezoid, WE = wave; col = color (1 = white, 2 = light grey, 3 = dark grey, 4 = black, 5 = hatched).

Table A.4: Detailed design matrix LST 4

Item	Size	B	T	Q	Steps	Contrad.	DR	col	KR	col	KU	col	PL	col	QT	col	SE	col	ST	col	TR	col	WE	col
1	0	2	0	0	2	0											1		2		3		4	
2	0	0	1	0	1	0					1	4					4	2			3	3	2	1
3	0	1	1	0	2	0					4		2				3				1			
4	0	2	1	0	3	0	4	2	2	4					1	1					3	3		
5	0	1	2	0	3	0			1	3	4	2			2	4							3	1
6	0	1	2	0	3	0			4		2								3		1			
7	0	2	2	0	4	0							2		1				3				4	
8	0	0	0	1	1	1	2	1	4	4	3	3	1	2										
9	0	1	1	0	2	1	3				2						4						1	
10	0	1	3	0	4	0			3	2			4	1	2	3	1	4						
11	0	0	0	1	1	0	1								3		2		4					
12	0	0	0	1	1	0			2								4		1				3	
13	0	1	3	0	4	0			1	3							2	4			3	1	4	2
14	0	0	0	1	1	0	1		3				4						2					
15	0	0	0	1	1	0	2	1					1	4							4	2	3	3
16	0	2	2	0	4	0					1	4			3	3			4	1	2	2		
17	1	0	2	0	2	0			5				1		4				2				3	
18	1	2	1	0	3	0			5	5			3	1			4	3	1	2	2	4		
19	1	0	1	1	2	0	3								5		2				1		4	
20	1	2	2	0	4	0					3		5		1				2		4			
21	1	1	0	1	2	0	1	4			5	2			4	3	2	1	3	5				
22	1	2	3	0	5	0	2				1		4				3				5			
23	1	1	0	1	2	0			1						3				5		2		4	
24	1	1	1	0	2	1	2	4	3	3	4	1	1	5									5	2
25	1	2	3	0	5	0	4	3	2	1			3	4	5	2					1	5		
26	1	2	4	0	6	0	5				4								2		3		1	
27	1	0	2	0	2	0			4	2	2	5					5	3			3	1	1	4
28	1	2	2	0	4	0					5	4	2	3	4	5	3	2					1	1
29	1	1	1	0	2	1	4		2		3				1		5							
30	1	0	1	1	2	0	4	5					5	1			3	4	1	3			2	2

Notes. Size: 0 = 4 lines ×4 columns, 1 = 5 lines ×5 columns; B = binary, T = ternary, Q = quarternary; Steps = sum of steps; Contrad. = contradictional (0 = not contradictional, 1 = contradictional, excluded from further analyses). DR = triangle, KR = circle, KU = cross, PL = plus, QT = square, SE = hexagon, ST = star, TR = trapezoid, WE = wave; col = color (1 = white, 2 = light grey, 3 = dark grey, 4 = black, 5 = hatched).

Table A.5: Solution positions and free cells for LST 1 to 4

	LST 1				LST 2				LST 3				LST 4			
Item	FrZ	L	LP	LF	FrZ	L	LP	LF	FrZ	L	LP	LF	FrZ	L	LP	LF
Item 1	10	SE	1	7	10	SE	1	7	10	SE	1	7	10	SE	1	7
Item 2	11	SE	4	14	11	SE	4	14	11	SE	4	14	11	SE	4	14
Item 3	10	WE	1	12	10	KU	4	3	10	DR	3	16	10	PL	2	14
Item 4	9	KR	1	5	9	KR	2	4	9	KU	4	13	9	TR	3	9
Item 5	10	WE	3	7	9	WE	3	7	10	PL	1	14	10	KU	4	16
Item 6	9	TR	1	11	11	TR	1	16	10	KR	4	16	10	ST	3	4
Item 7	10	TR	3	11	10	ST	3	7	10	KR	2	6	10	PL	2	7
Item 8	9	wd	fz	4	11	wd	fz	13	10	wd	fz	4	10	wd	fz	11
Item 9	11	wd	fz	3	9	wd	fz	3	11	wd	fz	4	11	wd	fz	7
Item 10	10	DR	1	8	10	PL	4	2	10	SE	2	14	10	KR	3	15
Item 11	12	PL	4	9	11	QT	3	14	12	QT	2	10	12	DR	1	7
Item 12	11	KR	2	4	11	KR	2	4	12	ST	1	14	12	WE	3	6
Item 13	10	SE	2	9	10	WE	4	9	10	TR	3	3	10	KR	1	7
Item 14	12	KU	1	11	11	DR	1	15	11	ST	4	5	11	PL	4	12
Item 15	12	TR	1	7	12	TR	4	1	12	SE	3	11	12	DR	2	2
Item 16	10	WE	4	11	10	KU	1	10	10	PL	2	10	10	ST	4	6
Item 17	16	KR	5	11	18	QT	4	23	18	SE	2	14	18	ST	2	3
Item 18	17	KU	5	10	17	ST	1	24	17	QT	4	16	15	TR	2	19
Item 19	17	QT	1	21	16	DR	3	2	17	ST	2	7	16	WE	4	7
Item 20	17	DR	2	19	17	PL	5	17	17	TR	5	17	17	QT	1	7
Item 21	16	QT	1	2	17	KU	5	9	17	KR	2	7	17	ST	3	23
Item 22	16	DR	5	23	16	SE	3	15	16	DR	5	15	16	KU	1	19
Item 23	16	SE	4	10	17	WE	4	6	17	ST	1	7	17	TR	2	8
Item 24	16	wd	fz	24	18	wd	fz	24	19	wd	fz	19	19	wd	fz	24
Item 25	17	WE	3	12	17	KR	2	14	17	WE	3	18	17	QT	5	8
Item 26	16	KU	2	12	16	ST	2	8	16	TR	3	18	16	KU	4	12
Item 27	19	WE	4	20	16	WE	1	4	19	KR	1	6	19	SE	5	21
Item 28	15	SE	3	9	15	KU	5	7	15	QT	5	17	15	WE	1	7
Item 29	16	wd	fz	22	16	wd	fz	6	16	wd	fz	24	16	wd	fz	2
Item 30	16	DR	4	10	17	PL	5	19	17	KR	2	18	17	PL	5	23

Notes: FrZ = number of free cells, L = solution, LP = solution position, DR = triangle, KR = circle, KU = cross, PL = plus, QT = square, SE = hexagon, ST = star, TR = trapezoid, WE = wave, fz = question mark, wd = contradiction.

A.2 Statistical word problems

This section shows the numerical information within the word problems and the algorithms and solutions for the single item families. As the test material was designed for German school pupils, the original German wording is used below.

Table A.6: Characteristics within contexts for word problems

Context	Characteristics (A, B, C)
1	Psychiatrie (Störung, Schweregrad, Therapieform)
2	Schokotafeln (Kakao, Zucker, Verpackung)
3	Fahrrad (Typ, Farbe, Schaltung)
4	Hotel (Etage, Ausstattung, Ausblick)
5	Studienbewerber (Fakultät, Studieneingangstest, Notenschnitt)
6	Kaffee (Herkunft, Sorte, Koffeingehalt)
7	Computerspiele (Schwierigkeit, Genre, Plattform)
8	Reise (Ziel, Verpflegung, Unterkunft)
9	Stadtfest (Handy: Farbe, Extra, Kamera)
10	Drogen (Substanz, Einstiegsalter, Schäden)
11	Zeitschriften (Thema, Intervall, Format)
12	DJ (Genre, Interpret, Sprache)
13	Hemden (Muster, Material, Ärmellänge)
14	Kino (Genre, Sitzplatz, Sprache)

Table A.7: Details of characteristics within contexts for word problems

Cont.	**A1**	**A2**	**A3**	**A4**	**B1**	**B2**	**B3**	**C1**	**C2**
1	Depress. Stör.	Angststör.	Essstör.	Persönl.-stör.	leicht	mittel	schwer	Gesprächs-therapie	Verhaltens-therapie
2	30% Kakao	40% Kakao	60% Kakao	70% Kakao	Frucht-zucker	Raffinade-zucker	Zucker-ersatz	normales Papier	Schmuck-Karton
3	Mountain-bike	Rennrad	Holland-rad	Senioren-rad	silber	schwarz	weiß	3-Gang	7-Gang
4	1. Stock	2. Stock	3. Stock	4. Stock	Economy	Comfort	Suite	Meerblick	Parkblick
5	wirtsch. Fak.	naturw. Fak.	philol. Fak.	künstl. Fak.	durchschn.	besonders gut	kaum geeig.	schlechter als 3,0	besser als 3,0
6	Brasilien	Guatemala	Equador	Kolumbien	Filter-kaffee	Cappuc-cino	Espresso	mit Koffein	koffein-frei
7	Anfänger	Fortgeschr.	Könner	Profis	Strategie	Adventure	Jump and Run	Playstation	PC
8	Kanaren	Mallorca	Kreta	Sizilien	Früh-stück	Halb-pension	Voll-pension	Hotel	Ferien-wohnung
9	rot	silber	schwarz	bunt	Flatrate	Prepaid	keine Zugaben	integr. Kamera	keine Kamera
10	Ecstasy	Cannabis	Alkohol	LSD	13 Jahre	14 Jahre	15 Jahre	schwerw. Schäden	leichte Schäden
11	Auto-magazin	Mode-magazin	Boulevard-magazin	Computer-magazin	viertel-jährlich	monat-lich	wöchent-lich	DIN A4	DIN A5
12	Techno	Schlager	Pop	Rock	weibl. Gruppe	männl. Gruppe	Solo-künstler	Englisch	Deutsch
13	gestreift	kariert	gepunktet	unifarben	Baum-wolle	Leinen	Synthetik	lange Ärmel	kurze Ärmel
14	Action	Horror	Komödie	Liebesfilm	Loge	Sperrsitz	Parkett	deutsch	OmU

Notes: Cont. = context number, A1 to C2 = characteristics.

Table A.8: Numerical information of characteristic values within contexts for word problems

Context number	A1	A2	A3	A4	B1	B2	B3
1	400 (0.40)	100 (0.10)	300 (0.30)	200 (0.20)	500 (0.50)	400 (0.40)	100 (0.10)
2	20 (0.10)	40 (0.20)	100 (0.50)	20 (0.10)	60 (0.30)	100 (0.50)	40 (0.20)
3	80 (0.20)	100 (0.25)	160 (0.40)	60 (0.15)	40 (0.10)	120 (0.30)	240 (0.60)
4	50 (0.10)	150 (0.30)	125 (0.25)	175 (0.35)	150 (0.30)	100 (0.20)	250 (0.50)
5	800 (0.40)	200 (0.10)	600 (0.30)	400 (0.20)	1000 (0.50)	800 (0.40)	200 (0.10)
6	10 (0.10)	20 (0.20)	50 (0.50)	10 (0.10)	30 (0.30)	50 (0.50)	20 (0.20)
7	100 (0.20)	125 (0.25)	200 (0.40)	75 (0.15)	50 (0.10)	150 (0.30)	300 (0.60)
8	30 (0.10)	90 (0.30)	75 (0.25)	105 (0.35)	90 (0.30)	60 (0.20)	150 (0.50)
9	80 (0.40)	20 (0.10)	60 (0.30)	40 (0.20)	100 (0.50)	80 (0.40)	20 (0.10)
10	100 (0.10)	200 (0.20)	500 (0.50)	100 (0.10)	300 (0.30)	500 (0.50)	200 (0.20)
11	60 (0.20)	75 (0.25)	120 (0.40)	45 (0.15)	30 (0.10)	90 (0.30)	180 (0.60)
12	40 (0.10)	120 (0.30)	100 (0.25)	140 (0.35)	120 (0.30)	80 (0.20)	200 (0.50)
13	40 (0.40)	10 (0.10)	30 (0.30)	20 (0.20)	50 (0.50)	40 (0.40)	10 (0.10)
14	100 (0.20)	125 (0.25)	200 (0.40)	75 (0.15)	50 (0.10)	150 (0.30)	300 (0.60)

	C1	C2	jp idp	jp dep 1	jp dep 2	total
1	400 (0.40)	600 (0.60)	B1 ∩ C2: 300 (0.30)	A3 ∩ C1: 200 (0.20)	A4 ∩ C1: 100 (0.10)	1000
2	140 (0.70)	60 (0.30)	B2 ∩ C1: 70 (0.35)	A4 ∩ C2: 12(0.06)	A3 ∩ C2: 42 (0.21)	200
3	240 (0.60)	160 (0.40)	B3 ∩ C1: 96 (0.24)	A3 ∩ C2: 96(0.24)	A4 ∩ C2: 48 (0.12)	400
4	400 (0.80)	100 (0.20)	B1 ∩ C2: 30 (0.06)	A4 ∩ C1:160 (0.32)	A3 ∩ C1: 80 (0.16)	500
5	800 (0.40)	1200 (0.60)	B1 ∩ C2: 600 (0.30)	A3 ∩ C1: 400 (0.20)	A4 ∩ C1: 200 (0.10)	2000
6	70 (0.70)	30 (0.30)	B2 ∩ C1: 35 (0.35)	A4 ∩ C2: 6 (0.06)	A3 ∩ C2: 21 (0.21)	100
7	300 (0.60)	200 (0.40)	B3 ∩ C1: 120 (0.24)	A3 ∩ C2:120 (0.24)	A4 ∩ C2: 60 (0.12)	500
8	240 (0.80)	60 (0.20)	B1 ∩ C2: 18 (0.06)	A4 ∩ C1: 96 (0.32)	A3 ∩ C1: 48 (0.16)	300
9	80 (0.40)	120 (0.60)	B1 ∩ C2: 60 (0.30)	A3 ∩ C1: 40 (0.20)	A4 ∩ C1: 20 (0.10)	200
10	700 (0.70)	300 (0.30)	B2 ∩ C1: 350 (0.35)	A4 ∩ C2: 60(0.06)	A3 ∩ C2: 210 (0.21)	1000
11	180 (0.60)	120 (0.40)	B3 ∩ C1: 72 (0.24)	A3 ∩ C2: 72 (0.24)	A4 ∩ C2: 36 (0.12)	300
12	320 (0.80)	80 (0.20)	B1 ∩ C2: 24 (0.06)	A4 ∩ C1: 128 (0.32)	A3 ∩ C1: 64 (0.16)	400
13	40 (0.40)	60 (0.60)	B1 ∩ C2: 30 (0.30)	A3 ∩ C1: 20 (0.20)	A4 ∩ C1: 10 (0.10)	100
14	300 (0.60)	200 (0.40)	B3 ∩ C1: 120 (0.24)	A3 ∩ C2:120 (0.24)	A4 ∩ C2: 60 (0.12)	500

Notes: jp = joint probability, idp = idependent, dep = dependent. A1 - C2 = characteristics.

Table A.9: Solution algorithms for all families for word problems

Cont.	Family 1	Family 2	Family 3	Family 4
1	A3 | C1	A1 ∩ B3	1-(A4 | C1)	1-(A2 ∩ B1)
2	A4 | C2	A2 ∩ B1	1-(A3 | C2)	1-(A1 ∩ B2)
3	A3 | C2	A1 ∩ B2	1-(A4 | C2)	1-(A2 ∩ B3)
4	A4 | C1	A2 ∩ B3	1-(A3 | C1)	1-(A1 ∩ B1)
5	A3 | C1	A1 ∩ B3	1-(A4 | C1)	1-(A2 ∩ B1)
6	A4 | C2	A2 ∩ B1	1-(A3 | C2)	1-(A1 ∩ B2)
7	A3 | C2	A1 ∩ B2	1-(A4 | C2)	1-(A2 ∩ B3)
8	A4 | C1	A2 ∩ B3	1-(A3 | C1)	1-(A1 ∩ B1)
9	A3 | C1	A1 ∩ B3	1-(A4 | C1)	1-(A2 ∩ B1)
10	A4 | C2	A2 ∩ B1	1-(A3 | C2)	1-(A1 ∩ B2)
11	A3 | C2	A1 ∩ B2	1-(A4 | C2)	1-(A2 ∩ B3)
12	A4 | C1	A2 ∩ B3	1-(A3 | C1)	1-(A1 ∩ B1)
13	A3 | C1	A1 ∩ B3	1-(A4 | C1)	1-(A2 ∩ B1)
14	A3 | C2	A1 ∩ B2	1-(A4 | C2)	1-(A2 ∩ B3)
Cont.	Family 5	Family 6	Family 7	Family 8
1	(A3∪A4) | C1	(A1 ∩ B2) ∪ A2	1-((A3∪A4) | C1)	(A1 ∩ B1) ∪ (1-A1)
2	(A3∪A4) | C2	(A2 ∩ B3) ∪ A1	1-((A3∪A4) | C2)	(A2 ∩ B2) ∪ (1-A2)
3	(A3∪A4) | C2	(A1 ∩ B1) ∪ A2	1-((A3∪A4) | C2)	(A1 ∩ B3) ∪ (1-A1)
4	(A3∪A4) | C1	(A2 ∩ B2) ∪ A1	1-((A3∪A4) | C1)	(A2 ∩ B1) ∪ (1-A2)
5	(A3∪A4) | C1	(A1 ∩ B2) ∪ A2	1-((A3∪A4) | C1)	(A1 ∩ B1) ∪ (1-A1)
6	(A3∪A4) | C2	(A2 ∩ B3) ∪ A1	1-((A3∪A4) | C2)	(A2 ∩ B2) ∪ (1-A2)
7	(A3∪A4) | C2	(A1 ∩ B1) ∪ A2	1-((A3∪A4) | C2)	(A1 ∩ B3) ∪ (1-A1)
8	(A3∪A4) | C1	(A2 ∩ B2) ∪ A1	1-((A3∪A4) | C1)	(A2 ∩ B1) ∪ (1-A2)
9	(A3∪A4) | C1	(A1 ∩ B2) ∪ A2	1-((A3∪A4) | C1)	(A1 ∩ B1) ∪ (1-A1)
10	(A3∪A4) | C2	(A2 ∩ B3) ∪ A1	1-((A3∪A4) | C2)	(A2 ∩ B2) ∪ (1-A2)
11	(A3∪A4) | C2	(A1 ∩ B1) ∪ A2	1-((A3∪A4) | C2)	(A1 ∩ B3) ∪ (1-A1)
12	(A3∪A4) | C1	(A2 ∩ B2) ∪ A1	1-((A3∪A4) | C1)	(A2 ∩ B1) ∪ (1-A2)
13	(A3∪A4) | C1	(A1 ∩ B2) ∪ A2	1-((A3∪A4) | C1)	(A1 ∩ B1) ∪ (1-A1)
14	(A3∪A4) | C2	(A1 ∩ B1) ∪ A2	1-((A3∪A4) | C2)	(A1 ∩ B3) ∪ (1-A1)

Notes: Cont. = context number. A1 - C2 = characteristics.

Table A.10: Solutions for all families for word problems

C	Solution	
	Family 1	**Family 2**
1	200/400 = 0.50	(400/1000)*(100/1000) = 0.04
2	12/60 = 0.20	(40/200)*(60/200) = 0.06
3	96/160 = 0.60	(80/400)*(120/400) = 0.06
4	160/400 = 0.40	(150/500)*(250/500) = 0.15
5	400/800 = 0.50	(800/2000)*(200/2000) = 0.04
6	6/30 = 0.20	(20/100)*(30/100) = 0.06
7	120/200 = 0.60	(100/500)*(150/500) = 0.06
8	96/240 = 0.40	(90/300)*(150/300) = 0.15
9	60/300 = 0.20	(80/200)*(20/200) = 0.04
10	40/80 = 0.50	(200/1000)*(300/1000) = 0.06
11	72/120 = 0.60	(60/300)*(90/300) = 0.06
12	128/320 = 0.40	(120/400)*(200/400) = 0.15
13	20/40 = 0.50	(40/100)*(10/100) = 0.04
14	120/200 = 0.60	(100/500)*(150/500) = 0.06
	Family 3	**Family 4**
1	1-(100/400) = 0.75	1-((100/1000)*(500/1000)) = 0.95
2	1-(42/60) = 0.30	1-((20/200)*(100/200)) = 0.95
3	1-(48/160) = 0.70	1-((100/400)*(240/400)) = 0.85
4	1-(80/400) = 0.80	1-((50/500)*(150/500)) = 0.97
5	1-(200/800) = 0.75	1-((200/2000)*(1000/2000)) = 0.95
6	1-(21/30) = 0.30	1-((10/100)*(50/100)) = 0.95
7	1-(60/200) = 0.70	1-((125/500)*(300/500)) = 0.85
8	1-(48/240) = 0.80	1-((30/300)*(90/300)) = 0.97
9	1-(20/80) = 0.75	1-((0.20/200)*(100/200)) = 0.95
10	1-(210/300) = 0.30	1-((100/1000)*(500/1000)) = 0.95
11	1-(36/120) = 0.70	1-((75/300)*(180/300)) = 0.85
12	1-(64/320) = 0.80	1-((40/400)*(120/400)) = 0.97
13	1-(10/40) = 0.75	1-((10/100)*(50/100)) = 0.95
14	1-(60/200) = 0.70	1-((125/500)*(300/500)) = 0.85
	Family 5	**Family 6**
1	(200/400)+(100/400) = 0.75	(100/1000)+((400/1000)*(400/1000)) = 0.26
2	(12/60)+(42/60) = 0.90	(20/200)+((40/200)*(40/200)) = 0.14
3	(96/160)+(48/160) = 0.90	(100/400)+((80/400)*(40/400)) = 0.27
4	(80/400)+(160/400) = 0.60	(50/500)+((150/500)*(100/500)) = 0.16
5	(400/800)+(200/800) = 0.75	(200/2000)+((800/2000)*(800/2000)) = 0.26
6	(21/30)+(6/30) = 0.90	(10/100)+((20/100)*(20/100)) = 0.14
7	(120/200)+(60/200) = 0.90	(125/500)+((100/500)*(50/500)) = 0.27
8	(48/240)+(96/240) = 0.60	(30/300)+((90/300)*(60/300)) = 0.16
9	(40/80)+(20/80) = 0.75	(20/200)+((80/200)*(80/200)) = 0.26
10	(210/300)+(60/300) = 0.90	(100/1000)+((200/1000)*(200/1000)) = 0.14

Continuation on following page

Table A.10: Solutions for all families for word problems *(continuation)*

C	Solution	
11	(72/120)+(36/120) = 0.90	(75/300)+((60/300)*(30/300)) = 0.27
12	(64/320)+(128/320) = 0.60	(40/400)+((120/400)*(80/400)) = 0.16
13	(20/40)+(10/40) = 0.75	(10/100)+((40/100)*(40/100)) = 0.26
14	(120/200)+(60/200) = 0.90	(125/500)+((100/500)*(50/500)) = 0.27
	Family 7	**Family 8**
1	1-((200/400)+(100/400)) = 0.25	(1-(400/1000))+((400/1000)*(500/1000)) =0 .80
2	1-((12/60)+(42/60)) = 0.10	(1-(40/200))+((40/200)*(100/200)) = 0.90
3	1-((96/160)+(48/160)) = 0.10	(1-(80/400))+((80/400)*(240/400)) = 0.92
4	1-((80/400)+(160/400)) = 0.40	(1-(150/500))+((150/500)*(150/500)) = 0.79
5	1-((400/800)+(200/800)) = 0.25	(1-(800/2000))+((800/2000)*(1000/2000)) = 0.80
6	1-((21/30)+(6/30)) = 0.10	(1-(20/100))+((20/100)*(50/100)) = 0.90
7	1-((120/200)+(60/200)) = 0.10	(1-(100/500))+((100/500)*(300/500)) = 0.92
8	1-((48/240)+(96/240)) = 0.40	(1-(90/300))+((90/300)*(90/300)) = 0.79
9	1-((40/80)+(20/80)) = 0.25	(1-(80/200))+((80/200)*(100/200)) = 0.80
10	1-((210/300)+(60/300)) = 0.10	(1-(200/1000))+((200/1000)*(500/1000)) = 0.90
11	1-((72/120)+(36/120)) = 0.10	(1-(60/300))+((60/300)*(180/300)) = 0.92
12	1-((64/320)+(128/320)) = 0.40	(1-(120/400))+((120/400)*(120/400)) = 0.79
13	1-((20/40)+(10/40)) = 0.25	(1-(40/100))+((40/100)*(50/100)) = 0.80
14	1-((120/200)+(60/200)) = 0.10	(1-(100/500))+((100/500)*(300/500)) = 0.92

Notes: C = context number.

Table A.11: Distribution of items in test versions for word problems

	Test version													
Item	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1a	c1f2	c3f2	c5f2	c7f2	c9f2	c11f2	c2f2	c4f2	c6f2	c8f2	c10f2	c12f2	c13f2	c14f2
1b	c1f3	c3f3	c5f3	c7f3	c9f3	c11f3	c2f3	c4f3	c6f3	c8f3	c10f3	c12f3	c13f3	c14f3
1c	c1f4	c3f4	c5f4	c7f4	c9f4	c11f4	c2f4	c4f4	c6f4	c8f4	c10f4	c12f4	c13f4	c14f4
1d	c1f5	c3f5	c5f5	c7f5	c9f5	c11f5	c2f5	c4f5	c6f5	c8f5	c10f5	c12f5	c13f5	c14f5
2a	c2f1	c4f1	c6f1	c8f1	c10f1	c12f1	c1f1	c3f1	c5f1	c7f1	c9f1	c11f1	c13f1	c14f1
2b	c2f6	c4f6	c6f6	c8f6	c10f6	c12f6	c1f6	c3f6	c5f6	c7f6	c9f6	c11f6	c13f6	c14f6
2c	c2f7	c4f7	c6f7	c8f7	c10f7	c12f7	c1f7	c3f7	c5f7	c7f7	c9f7	c11f7	c13f7	c14f7
2d	c2f8	c4f8	c6f8	c8f8	c10f8	c12f8	c1f8	c3f8	c5f8	c7f8	c9f8	c11f8	c13f8	c14f8
3a	c3f2	c5f2	c7f2	c9f2	c11f2	c2f2	c4f2	c6f2	c8f2	c10f2	c12f2	c1f2	c14f2	c13f2
3b	c3f3	c5f3	c7f3	c9f3	c11f3	c2f3	c4f3	c6f3	c8f3	c10f3	c12f3	c1f3	c14f3	c13f3
3c	c3f4	c5f4	c7f4	c9f4	c11f4	c2f4	c4f4	c6f4	c8f4	c10f4	c12f4	c1f4	c14f4	c13f4
3d	c3f5	c5f5	c7f5	c9f5	c11f5	c2f5	c4f5	c6f5	c8f5	c10f5	c12f5	c1f5	c14f5	c13f5
4a	c4f1	c6f1	c8f1	c10f1	c12f1	c1f1	c3f1	c5f1	c7f1	c9 f1	c11f1	c2f1	c14f1	c13f1
4b	c4f6	c6f6	c8f6	c10f6	c12f6	c1f6	c3f6	c5f6	c7f6	c9 f6	c11f6	c2f6	c14f6	c13f6
4c	c4f7	c6f7	c8f7	c10f7	c12f7	c1f7	c3f7	c5f7	c7f7	c9 f7	c11f7	c2f7	c14f7	c13f7
4d	c4f8	c6f8	c8f8	c10f8	c12f8	c1f8	c3f8	c5f8	c7f8	c9 f8	c11f8	c2f8	c14f8	c13f8

Notes: Item = item number in test book, c = context, f = family.

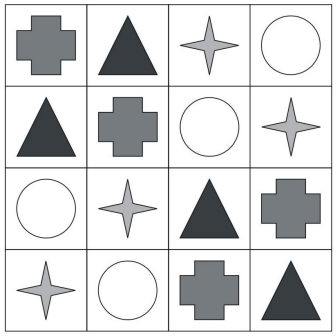
B Instructions

This section shows the original instructions of LST as well as of the word problems. As the test material was designed for German school pupils, the original German wording is used below.

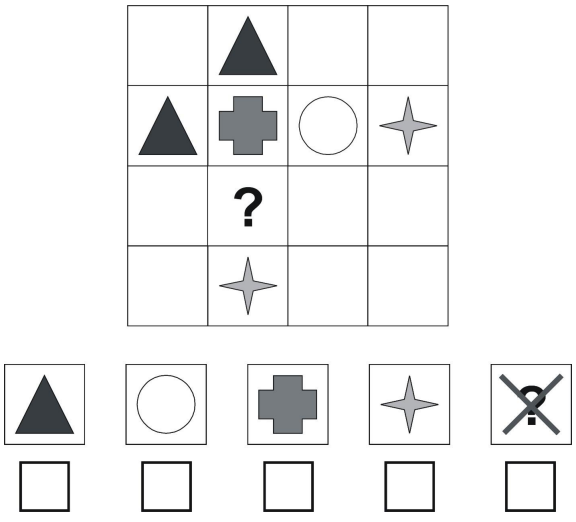
B.1 LST

Lateinische Quadrate

Lateinische Quadrate bestehen aus mehreren Zellen, in denen sich unterschiedliche Elemente befinden. Jedes dieser Elemente darf in jeder Zeile und Spalte genau einmal vorkommen. Die folgende Abbildung zeigt ein vollständiges lateinisches Quadrat:



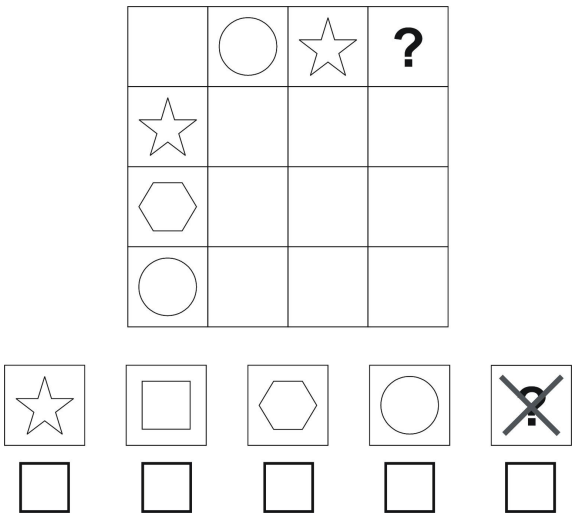
Jede geometrische Figur kommt in jeder Zeile und Spalte genau einmal vor. In diesem Test geht es darum, in unvollständigen lateinischen Quadraten die Inhalte freier Zellen zu erschließen. Eine bestimmte Zelle enthält ein Fragezeichen. Es soll die geometrische Figur ausgewählt werden, die in diese Zelle passt. Ein Beispiel verdeutlicht dies:



Hier ist das Kästchen unter dem Kreis angekreuzt, da er in der zweiten Spalte als einziges Element noch nicht vorkommt. In den leeren Zellen können nur die Figuren vorkommen, die auch als Antwortalternative angegeben sind.

Bitte umblättern!

Manchmal sind zur Lösung einer Testaufgabe mehrere Schritte erforderlich, wie ein weiteres Beispiel illustriert:



Hier kann zunächst (erste Spalte) erschlossen werden, dass in der oberen linken Zelle ein Quadrat stehen muss. Betrachtet man nun im nächsten Schritt die erste Zeile (Kreis + Stern + Quadrat), kann in der Zelle mit dem Fragezeichen nur das Sechseck stehen. Manchmal müssen für die Aufgabenlösung auch Zeilen und Spalten gleichzeitig betrachtet werden.

Die in diesem Test verwendeten unvollständigen lateinischen Quadrate haben zunächst 4 Zeilen und 4 Spalten, später 5 Zeilen und 5 Spalten. Manchmal kann keine der zur Auswahl stehenden Figuren an die Stelle des Fragezeichens gesetzt werden. Dies ist dann der Fall, wenn die Regel "Jede Figur darf in jeder Spalte und Zeile nur einmal vorkommen" verletzt wird. In diesem Fall ist das durchgestrichene Fragezeichen auszuwählen. Für jede Testaufgabe trifft eine korrekte Lösungsalternative zu.

WICHTIGER HINWEIS: Bitte zeichnen Sie nicht in das Testheft. Aufgaben, bei denen in leeren Zellen oder an anderer Stelle außerhalb der Ankreuzkästchen Markierungen vorgenommen wurden, können nicht gewertet werden. Bedenken Sie bitte, dass es bei dieser Aufgabe darum geht, die Lösung ausschließlich im Kopf zu bestimmen. Achten Sie auch darauf, die ausgewählte Antwortalternative im jeweils darunter liegenden Kästchen deutlich erkennbar anzukreuzen.

B.2 Statistical word problems

Instruktion

In diesem Test sollen Sie Aufgaben zur Wahrscheinlichkeitsrechnung lösen. Bitte notieren Sie dabei **auf jeden Fall** Ihren Lösungsweg. Es genügt nicht, das korrekte Ergebnis ohne Lösungsweg aufzuschreiben. Es ist nicht nötig, dass Sie das Ergebnis komplett ausrechnen. Sie können mit Brüchen oder Dezimalzahlen arbeiten. Sie müssen keine Antwortsätze ausformulieren.

Hier sehen Sie eine Beispielaufgabe:

Situation Aus einer Urne wird ein Objekt zufällig gezogen. Insgesamt befinden sich 20 Objekte in der Urne. Davon sind 4 rot, 2 blau, und die restlichen sind gelb oder grün. 10 Objekte haben eine glatte Oberfläche, 5 eine angeraute und 5 eine geriffelte Oberfläche. 8 Objekte sind Kugeln, 12 sind Würfel.
2 Objekte sind rot und glatt. 6 Objekte sind Würfel und grün. 3 Objekte sind Würfel und gelb.
Farbe und Form der Objekte sind abhängig voneinander, alle anderen Merkmale sind unabhängig voneinander.

Fragen a) Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *nicht* rot ist?

Für die Lösung muss berechnet werden, wie groß die **Gegenwahrscheinlichkeit** für ein Ereignis (hier „nicht rot“) ist:

$$P(\text{nicht A}) = 1 - P(\text{A})$$
$$P(\text{nicht rot}) = 1 - P(\text{rot}) = 1 - \tfrac{4}{20} = 1 - 0,20 = 0,80$$

Weiteres Beispiel: Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *nicht* blau ist?

b) Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *entweder* rot *oder* blau ist?

Für die Lösung muss berechnet werden, wie groß die Wahrscheinlichkeit für die **Vereinigungsmenge** zweier Ereignisse (hier „rot oder blau“, die Ereignisse stehen also in einer **Entweder-Oder-Beziehung** zueinander) ist:

$$P(\text{A oder B}) = P(\text{A}) + P(\text{B})$$

$$P(\text{rot oder blau}) = P(\text{rot}) + P(\text{blau}) = \tfrac{4}{20} + \tfrac{2}{20} = 0,20 + 0,10 = 0,30$$

Weiteres Beispiel: Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *entweder* glatt *oder* geriffelt ist?

c) Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *sowohl* geriffelt *als auch* eine Kugel ist?

Für die Lösung muss berechnet werden, wie groß die Wahrscheinlichkeit für die **Schnittmenge** zweier Ereignisse (hier „geriffelt und Kugel“, die Ereignisse stehen also in einer **Sowohl-Als-Auch-Beziehung** zueinander) ist:

$$P(\text{A und B}) = P(\text{A}) \cdot P(\text{B})$$

$$P(\text{geriffelt und Kugel}) = P(\text{geriffelt}) \cdot P(\text{Kugel}) = \tfrac{5}{20} \cdot \tfrac{8}{20} = 0,25 \cdot 0,40 = 0,10$$

Weiteres Beispiel: Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *sowohl* glatt *als auch* ein Würfel ist?

d) Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das grün ist, *vorausgesetzt*, dieses Objekt ist ein Würfel?

Für die Lösung muss man berechnen, wie groß die **bedingte Wahrscheinlichkeit** für ein Ereignis (hier „grün, unter der Bedingung Würfel“) ist. Es werden also bei der Berechnung der Wahrscheinlichkeit für ein grünes Objekt hier *nur die Würfel* betrachtet, und *nicht alle möglichen* Objekte. Außerdem wird *im Zähler* nicht die Gesamtmenge der grünen Objekte betrachtet, sondern die *Schnittmenge* der grünen Objekte und der Würfel:

$$P(\text{A wenn B}) = \frac{P(\text{A und B})}{P(\text{B})}$$
$$P(\text{grün wenn Würfel}) = \frac{\text{Anzahl(grün und Würfel)}}{\text{Anzahl(Würfel)}} = \frac{6}{12} = 0,50$$

Weiteres Beispiel: Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das gelb ist, *vorausgesetzt*, dieses Objekt ist ein Würfel?

Diese Regeln können auch kombiniert werden. Zum Beispiel ist es möglich, die Gegenwahrscheinlichkeit für eine Vereinigungsmenge, eine Schnittmenge oder eine bedingte Wahrscheinlichkeit zu berechnen.

Beispiele:

1. Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *nicht sowohl* geriffelt *als auch* eine Kugel ist?

$$P(\text{nicht(geriffelt und Kugel)}) = 1 - P(\text{geriffelt und Kugel})$$

2. Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *nicht entweder* rot *oder* blau ist?

$$P(\text{nicht(rot oder blau)}) = 1 - P(\text{rot oder blau})$$

3. Wie groß ist die Wahrscheinlichkeit, dass ein Objekt gezogen wird, das *nicht* grün ist, *vorausgesetzt*, dieses Objekt ist ein Würfel?

$$P(\text{nicht(grün wenn Würfel)}) = 1 - P(\text{grün wenn Würfel})$$

C Software and syntax

LLTM: Stata 10

- LLTM: xtmelogit y radical1 radical2... ||id:, var
- RE-LLTM: xtmelogit y radical1 radical2... ||_all:R.item ||id:, var
- LR-LLTM: xtmelogit y radical1 radical2... characteristic1 characteristic2... ||id:, var
- RW-LLTM: xtmelogit y radical1 radical2... ||id: radical1 radical2..., var
- RE-LR-LLTM: xtmelogit y radical1 radical2... characteristic1 characteristic2... ||_all:R.item ||id:, var
- RW-LR-LLTM: xtmelogit y radical1 radical2... characteristic1 characteristic2... ||id: radical1 radical2..., var
- LR-test: lrtest model1 model2, stats

D Feedback examples

This section shows examples of feedback information given in written form to examinees after testing.

Westfälische
Wilhelms-Universität
Münster

Westfälische
Wilhems-Universität
Münster

Psychologisches Institut IV: Statistik und Methoden

CODE: XXXX0000

Nina Hoffman
Horstmarer Landweg 103
48 149 Münster
kognitives_training@web.de

Sehr geehrte Testteilnehmerin, sehr geehrter Testteilnehmer!

Auf den folgenden Seiten finden Sie Ihre Ergebnisse aus dem von Ihnen besuchten Testtraining, das von der Universität Münster durchgeführt wurde, sowie eine ausführliche Erläuterung. Um Missverständnisse zu vermeiden, empfehlen wir Ihnen, vorab die Erläuterungen zu lesen, bevor Sie Ihre Ergebnisse einordnen.

1 Allgemeine Informationen zu den rückgemeldeten Werten

Die Ihnen rückgemeldeten Werte sagen etwas darüber aus, in welchem Ausmaß Sie über ein bestimmtes Persönlichkeits- oder Fähigkeitsmerkmal im Verhältnis zu anderen Personen verfügen. Sie erhalten die Rückmeldung in Form sogenannter *Referenzwerte*. Die Personen, auf die für die Beurteilung Ihrer Ergebnisse in Form von Referenzwerten Bezug genommen wird, werden im folgenden *Referenzgruppe* genannt. Die Referenzgruppe stellt die Gruppe von Personen dar, mit der Ihre Ergebnisse verglichen werden. Bei vielen Tests gibt es verschiedene Referenzgruppen (zum Beispiel getrennt nach Schulbildung, Alter oder Geschlecht). Die hier rückgemeldeten Werte sind also immer bezogen auf die Referenzgruppe, der Sie zuzuordnen sind.

Hinweise zur Interpretation der Referenzwerte: Im Allgemeinen ist es sowohl bei Leistungs- als auch Persönlichkeitsmerkmalen so, dass die meisten Personen eine mittlere Ausprägung besitzen, wenige Personen eine hohe beziehungsweise

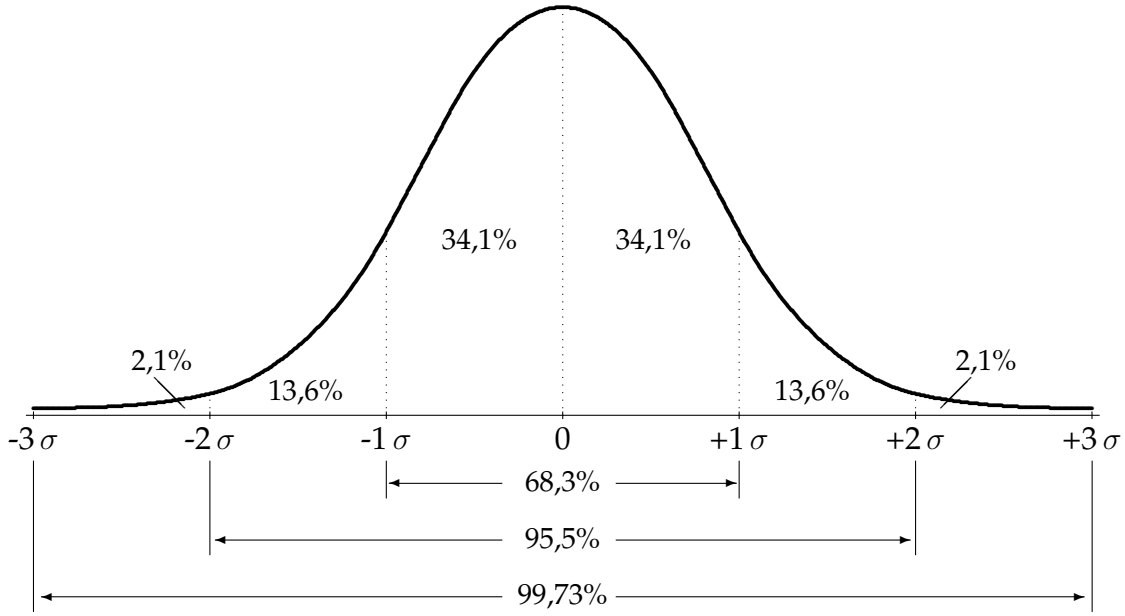


Abbildung 1: Verteilung der Referenzwerte und deren Interpretation als Prozentwerte.

niedrige Ausprägung und sehr wenige Personen haben eine sehr hohe beziehungsweise sehr niedrige Ausprägung. Dieser Sachverhalt wird grafisch durch die Kurve in Abbildung 1 illustriert.

Des Weiteren sehen Sie anhand dieser Abbildung, dass die Referenzwerte auch im Sinne von Prozentwerten interpretiert werden können, und zwar in Prozent von Personen, die einen bestimmten Referenzwert erreichen. Tabelle 1 auf der nächsten Seite gibt Ihnen einen Überblick über eine Reihe von Referenzwerten und den zugehörigen Prozentwerten.

Sie finden zunächst eine Beschreibung der Referenzgruppe und Erläuterungen zu den rückgemeldeten Merkmalen. Die Interpretation Ihrer Ergebnisse sollte nur vor dem Hintergrund dieser Informationen erfolgen!

Bitte beachten Sie bei der Rückmeldung die folgenden Punkte:

- Die Werte Ihrer Rückmeldung sind nicht als Bewertung zu verstehen. Ein geringerer Wert kann daher für Sie, zum Beispiel bei einzelnen Persönlichkeitsmerkmalen, durchaus eine positive Bedeutung haben!
- Bei den Testaufgaben vom Typ „Lateinische Quadrate“ wird Ihre Leistung nicht mit der anderer Probanden verglichen, daher erhalten Sie für diese Aufgaben lediglich eine Angabe darüber, wie viele Aufgaben Sie richtig gelöst haben. Ein Vergleich mit anderen Teilnehmern wäre irreführend, da Studenten und Schüler unterschiedlichen Alters und unterschiedlichen

GB	RW	Bedeutung
+3 σ	130	Wenn Sie bei einem Merkmal diesen Wert haben, hatten 99,9% der Personen der RG keine so hohe Ausprägung.
+2,5 σ	125	99% der Personen der RG hatten keine so hohe Ausprägung.
+2 σ	120	97% der Personen der RG hatten keine so hohe Ausprägung.
+1,5 σ	115	93% der Personen der RG hatten keine so hohe Ausprägung.
+1 σ	110	84% der Personen der RG hatten keine so hohe Ausprägung.
+0,5 σ	105	69% der Personen der RG hatten keine so hohe Ausprägung.
0	100	Hier hatten genau 50% der Personen der RG eine höhere und 50% der Personen der RG eine niedrigere Ausprägung. Dieser Wert entspricht somit einer mittleren Ausprägung.
-0,5 σ	95	31% der Personen der RG hatten keine so hohe Ausprägung.
-1 σ	90	16% der Personen der RG hatten keine so hohe Ausprägung.
-1,5 σ	85	7% der Personen der RG hatten keine so hohe Ausprägung.
-2 σ	80	3% der Personen der RG hatten keine so hohe Ausprägung.
-2,5 σ	75	1% der Personen der RG hatten keine so hohe Ausprägung.
-3 σ	70	0,1% der Personen der RG hatten keine so hohe Ausprägung.

Tabelle 1: Bedeutung der Referenzwerte und zugehörige Prozentangaben. GB = Grafikbezugswert aus Abbildung 1, RW = Referenzwert und RG = Referenzgruppe.

Geschlechts teilgenommen haben.

- Die aus der Tabelle ersichtlichen Persönlichkeitswerte reflektieren eine Zusammenfassung Ihrer Selbstbeschreibung. Sie selbst haben sich anhand der Fragen beschrieben. Die Rückmeldung der Ergebnisse ist daher abhängig davon, wie genau Sie die Fragen beantwortet haben und welches Bild Sie von sich selbst haben.

- Jede Messung psychischer Merkmale ist messfehlerbehaftet. Das heißt, dass wir im Einzelnen Ihre Merkmalsausprägung durchaus über- oder unterschätzt haben können. Je besser Sie bei der Testdurchführung unseren Anweisungen gefolgt sind, desto genauer dürften die Ergebnisse für Sie sein. Insbesondere bei den Persönlichkeitsmerkmalen ist zu bedenken, dass die Ergebnisse von dem „Ausmaß“ der Ehrlichkeit Ihrer Antworten abhängig sind.

- Bedenken Sie, dass es sich vor allem bei den Fähigkeits- und Leistungstests um eine Momentaufnahme handelt, die durch eine Vielzahl von Faktoren

(zum Beispiel Tagesform, Ablenkung, Motivation etc.) beeinflusst wird. Leistungsmerkmale unterliegen damit oft starken Schwankungen und sind keineswegs endgültig.

2 Erläuterungen der durchgeführten Tests

2.1 „Lateinische Quadrate“ (LST)

Diese Aufgaben haben Sie insgesamt viermal bearbeitet, wobei es sich bei jeder Durchführung um ähnliche, aber nicht identische Aufgaben handelte. Dieses neu konstruierte Testverfahren misst die Arbeitsgedächtniskapazität durch sprachfreie, figurale Aufgaben. Der Test befindet sich in der Erprobungsphase, zeigte bisher aber an einer Stichprobe von 222 Personen (Oberstufenschüler, Studenten) sehr gute Eigenschaften. Die mehrfache Verwendung des Tests erlaubt somit Aussagen zur Größe von Übungs- und Trainingseffekten psychologischer Tests auf individueller Ebene.

Für die Lateinischen Quadrate kann Ihnen nur die Anzahl der von Ihnen zu den verschiedenen Zeitpunkten richtig gelösten Aufgaben und kein Referenzwert rückgemeldet werden, da noch keine zuverlässige Referenzgruppe existiert. Maximal konnten zu jedem Zeitpunkt 30 Aufgaben richtig gelöst werden.

2.2 Interessentest „AIST“ (Allgemeiner Interessen-Struktur-Test)

Der AIST wird oft in Zusammenhang mit Berufs- und Laufbahnentscheidungen eingesetzt (zum Beispiel Berufsorientierung, Berufsentscheidung, innerbetriebliche Laufbahn- und Personalentscheidungen). Der AIST ist ein Fragebogen zur Erfassung schulisch-beruflicher Interessen und Tätigkeiten. Er besteht aus 60 Fragen, mit denen sechs Interessendimensionen gemessen werden: Praktisch-technische Interessen, intellektuell-forschende Interessen, künstlerisch-sprachliche Interessen, soziale Interessen, unternehmerische Interessen sowie konventionelle Interessen.

R = Realistischer Typ: Menschen mit dieser Grundorientierung haben eine Vorliebe für Tätigkeiten, die Kraft, Koordination und Handgeschicklichkeit erfordern und zu konkreten, sichtbaren Ergebnissen führen. Charakteristisch ist der formende Umgang mit Materialien und die Verwendung von

Werkzeugen oder Maschinen. Menschen dieses Typs weisen Fähigkeiten und Fertigkeiten vor allem im mechanischen, technischen, elektronischen und landwirtschaftlichen Bereich auf, während sie erzieherische oder soziale Tätigkeiten eher ablehnen. Ihre Werthaltungen sind auf materielle Dinge gerichtet: Geld, Macht und sozialer Status.

I = Intellektueller Typ: Personen mit dieser Grundorientierung haben eine Vorliebe für Aktivitäten, bei denen die symbolische, schöpferische oder beobachtende Auseinandersetzung mit physischen, biologischen oder kulturellen Phänomenen im Vordergrund steht. Sie möchten diese Phänomene verstehen und unter Kontrolle bringen. Gleichzeitig besteht eher eine Abneigung gegenüber überredenden, sozialen oder repetitiven Tätigkeiten. Ihre Fähigkeiten und Fertigkeiten liegen vor allem im mathematischen und naturwissenschaftlichen Bereich, ihre Werthaltungen sind vor allem auf Wissen(schaft) gerichtet.

K = Künstlerischer Typ: Menschen mit dieser Grundorientierung haben vor allem eine Vorliebe für offene, unstrukturierte Aktivitäten, die ihnen den auf künstlerische Selbstdarstellung oder die Schaffung kreativer Produkte gerichteten Umgang mit Material, Sprache oder auch Menschen ermöglichen. Weniger gut liegen ihnen klar abgegrenzte, systematische und geordnete Tätigkeiten. Ihre Fähigkeiten und Fertigkeiten liegen in den Bereichen Sprache, bildende Kunst, Musik, Schauspiel und Schriftstellerei. Sie streben vor allem ästhetische Werte an.

S = Sozialer Typ: Menschen mit dieser Grundorientierung haben eine Vorliebe für Tätigkeiten, bei denen sie sich mit anderen Menschen in Form von Unterrichten, Lehren, Ausbilden, Versorgen oder Pflegen befassen können. Weniger gut liegen ihnen klar abgegrenzte, systematische Tätigkeiten oder der Umgang mit Werkzeugen oder Maschinen. Ihre speziellen Fähigkeiten und Fertigkeiten liegen in den zwischenmenschlichen Beziehungen, insbesondere im sozialen Umgang und im erzieherischen Bereich. Ihre zentrale Werteausrichtung bezieht sich auf soziale und ethische Fragestellungen.

U = Unternehmerischer Typ: Personen mit dieser Grundorientierung haben eine Vorliebe für Tätigkeiten oder Situationen, in denen sie andere - meist um ein organisatorisches Ziel oder einen wirtschaftlichen Gewinn zu erreichen - mit Hilfe der Sprache oder anderen Mitteln beeinflussen, zu etwas bringen, führen, oder auch manipulieren können. Weniger gut liegen ihnen beobach-

tende oder systematische Tätigkeiten. Die spezifischen Fähigkeiten und Fertigkeiten solcher Personen sind ihre Führungs- und Überzeugungsstärke. Ihre zentrale Werthaltung ist der soziale, politische oder ökonomische Erfolg.

C = Konventioneller Typ: Menschen mit dieser Grundorientierung haben eine Vorliebe für den genau bestimmten, geordneten, systematischen Umgang mit Daten: Dokumentationen anlegen, Aufzeichnungen führen, Materialien ordnen, maschinelle Verarbeitung organisatorischer oder wirtschaftlicher Daten. Weniger gut liegen ihnen offene, unstrukturierte Tätigkeiten. Ihre speziellen Fähigkeiten und Fertigkeiten sind rechnerischer, verwaltender und geschäftlicher Art.

Referenzgruppe für den Interessentest: *Die Bezugsgruppe besteht aus einer Stichprobe von ungefähr 4400 14- bis 20-jährigen Probanden.*

2.3 Persönlichkeitstest NEO-FFI (NEO-Fünf-Faktoren-Inventar)

Das NEO-FFI ist ein häufig verwendeter Persönlichkeitstest. Einsatzmöglichkeiten liegen insbesondere in der Schullaufbahn- und Studienberatung, in Berufsberatung und Organisationspsychologie sowie in der psychologischen Forschung. Das NEO-FFI ist ein multidimensionales Persönlichkeitsinventar, das die wichtigsten Bereiche individueller Unterschiede erfasst. Das NEO-FFI erfasst mit seinen insgesamt 60 Fragen diese Dimensionen auf fünf Skalen: Emotionale Labilität, Extraversion, Offenheit für Erfahrung, Verträglichkeit und Gewissenhaftigkeit.

L = Emotionale Labilität: Mit diesen Werten werden individuelle Unterschiede in der emotionalen Stabilität und der emotionalen Labilität zum Ausdruck gebracht. Der Kern dieses Merkmals liegt in der Art und Weise, wie Emotionen, vor allem negative Emotionen, erlebt werden. Personen mit hohen Werten beschreiben sich selbst als nervös, sind leicht aus dem seelischen Gleichgewicht zu bringen, neigen zur Traurigkeit, sind eher ängstlich, unsicher und verlegen. Emotional labile Personen neigen zu unrealistischen Ideen, machen sich Sorgen um ihre Gesundheit und sind weniger in der Lage, auf Stresssituationen angemessen zu reagieren. Emotional stabile Personen dagegen (Personen mit niedrigen Werten) beschreiben sich selbst als ruhig, ausgeglichen, sorgenfrei und geraten in Stresssituationen nicht so schnell aus der Fassung.

E = Extraversion: Personen mit hohen Werten (extravertierte Personen) sind gesprächig, gesellig und aktiv. Sie beschreiben sich als selbstsicher, personenorientiert, herzlich, heiter und optimistisch und bevorzugen Aufregungen und Anregungen. Personen mit niedrigen Werten (introvertierte Personen) sind eher zurückhaltend als unfreundlich, eher unabhängig als folgsam, eher ausgeglichen als unsicher oder phlegmatisch. Wenn ihnen auch nicht die Lebhaftigkeit des Extravierten zu eigen ist, so sind Introvertierte doch nicht unbedingt unglücklich oder pessimistisch.

O = Offenheit für Erfahrungen: Personen, die hohe Werte bezüglich Offenheit für Erfahrungen aufweisen, zeichnen sich durch eine hohe Wertschätzung für neue Erfahrungen aus. Sie beschreiben sich als kreativ, wissbegierig, phantasievoll und machen ihre Urteile nicht von anderen abhängig. Sie sind vielseitig interessiert und bevorzugen Abwechslung. Personen mit niedrigen Werten neigen demgegenüber eher zu konventionellem Verhalten und zu konservativen Einstellungen. Sie ziehen Bekanntes und Bewährtes dem Neuen vor, und ihre emotionalen Reaktionen sind eher gedämpft.

V = Verträglichkeit: Personen mit hohen Werten beschreiben sich als hilfsbereit, mitfühlend, kooperativ, verständnisvoll und wohlwollend. Sie neigen zu zwischenmenschlichem Vertrauen, sind eher nachgiebig und haben ein erhöhtes Harmoniebedürfnis. Personen mit niedrigen Werten beschreiben sich im Gegensatz dazu als egozentrisch und misstrauisch gegenüber den Absichten anderer Menschen.

G = Gewissenhaftigkeit: Mit diesem Merkmal wird eine Art der Selbstkontrolle beschrieben, die sich auf den aktiven Prozess der Planung, Organisation und Durchführung von Aufgaben bezieht. Personen mit hohen Werten sind ordentlich, zuverlässig, hart arbeitend, pünktlich, penibel, diszipliniert, ehrgeizig, systematisch und genau. Personen mit niedrigen Werten beschreiben sich dagegen eher als nachlässig, gleichgültig und unbeständig, sie verfolgen ihre Ziele also mit geringem Engagement.

Referenzgruppe für den Persönlichkeitstest: *Die Bezugsgruppe besteht aus einer bevölkerungsrepräsentativen Stichprobe von ungefähr 2100 Personen.*

2.4 Grundintelligenztest (CFT)

Der CFT erfasst das allgemeine intellektuelle Niveau (Grundintelligenz) im Sinne der Cattell'schen „flüssigen Intelligenz". Diese kann umschrieben werden als Fähigkeit, figurale Beziehungen und formal-logische Denkprobleme mit unterschiedlichem Komplexitätsgrad zu erkennen und innerhalb einer bestimmten Zeit zu verarbeiten. Da dies durch sprachfreie und anschauliche Testaufgaben geschieht, werden Personen mit geringen Kenntnissen der deutschen Sprache nicht benachteiligt. Die zu Beginn der Erhebung durchgeführten Testteile des CFT bestanden aus vier Untertests (Reihenfortsetzen, Klassifikationen, Matrizen und topologische Schlussfolgerungen). In der Tabelle ist in der Rubrik „CFT" der oben beschriebene Referenzwert aufgeführt.

Referenzgruppen: *Verfügbar für verschiedene Altersstufen, wobei die Gruppengröße je nach Gruppe variiert.*

2.5 Gedächtnistests

2.5.1 Berliner Intelligenzstruktur-Test (BIS)

Mit 45 sehr verschiedenen, repräsentativ ausgewählten Aufgabentypen erfasst der BIS-Test eine außergewöhnliche Vielfalt und Breite von Intelligenzleistungen. Die Vielfalt der Anforderungen erhöht die Akzeptanz, die sehr abwechslungsreiche Folge der Aufgaben verstärkt dauerhafte Aufmerksamkeit und Leistungsmotivation in der Durchführung. Aus dem BIS wurden drei figurale Gedächtnistests verwendet:

- BIS 1 (Umrandungen)
- BIS 2 (Gebäudeplan)
- BIS 3 (Weg)

Da die Unterskalen des Berliner Intelligenzstruktur-Tests nicht einzeln interpretiert werden dürfen, sind die dargestellten Werte für diese durchgeführten Untertests nicht als gleichwertig zu den oben beschriebenen Referenzwerten zu verstehen.

2.5.2 Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)

Der I-S-T 2000 R ist ein häufig eingesetzter, ökonomischer Intelligenztest. Im Rahmen des Testtrainings wurde aus dem I-S-T 2000 R der Untertest zur figuralen

Merkfähigkeit verwendet.

Referenzgruppen: *Verfügbar für verschiedene Altersstufen und Schulbildung (Gruppengröße variiert je nach Gruppe).*

2.6 Aufmerksamkeits-Belastungs-Test (d2-Test)

Der Test d2 ist ein Einzel- und Gruppentest zur Untersuchung der individuellen Aufmerksamkeit und Konzentrationsfähigkeit. Er findet Verwendung in nahezu allen psychologischen Arbeitsbereichen. Der Test d2 misst Tempo und Sorgfalt des Arbeitsverhaltens bei der Unterscheidung ähnlicher visueller Reize und ermöglicht damit die Beurteilung individueller Aufmerksamkeits- und Konzentrationsleistungen. Es werden folgende Werte zurückgemeldet:

- GZ: Gesamtzahl der bearbeiteten Zeichen als Maß für die Schnelligkeit der Bearbeitung.
- F%: Anteil der Fehler an der Gesamtzahl als Maß für die Genauigkeit der Bearbeitung. Je weniger Fehler desto besser.
- KL: Zuverlässige fehlerkorrigierte Konzentrationsleistung, die sowohl Genauigkeit als auch Schnelligkeit der Bearbeitung berücksichtigt.

Es ist zu beachten, dass der Anteil der Fehler hier ein Wert in der Rückmeldung ist, der eine günstigere Aussage macht, je niedriger er ausfällt. Das heißt, Personen, die einen niedrigen Referenzwert haben, haben viele Fehler gemacht.

Referenzgruppen: *Verfügbar für verschiedene Altersstufen, Schulbildung, Geschlechter (Gruppengröße variiert je nach Gruppe).*

3 Ihre Ergebnisse

Sollten Teile Ihrer Ergebnisse nicht verwertbar gewesen sein, so entfällt die Rückmeldung für die betroffenen Merkmale.

Test	**LST**				**CFT**	**BIS**			**IST**	**d2**		
	1	**2**	**3**	**4**		**1**	**2**	**3**		**GZ**	**F%**	**KL**
Ihr Wert												

Test	**AIST**						**NEO-FFI**				
	R	**I**	**K**	**S**	**U**	**C**	**L**	**E**	**O**	**V**	**G**
Ihr Wert											

Für Rückfragen hinsichtlich Ihrer Ergebnisse stehen wir Ihnen gerne zur Verfügung. Bei konkreten Rückfragen zu individuellen Ergebnissen schreiben Sie eine E-Mail unter Angabe von Testungsort (zum Beispiel Name der Schule), Testdatum und CODE an Nina Zeuch (E-Mail-Adresse: *kognitives_training@web.de*).

Westfälische
Wilhelms-Universität
Münster

Psychologisches Institut IV: Statistik und Methoden

CODE: XXXX0000

Dipl.-Psych. Nina Zeuch
Psychologisches Institut IV
Fliednerstraße 21
48 149 Münster
n_hoff01@uni-muenster.de

Sehr geehrte Testteilnehmerin, sehr geehrter Testteilnehmer!

Auf den folgenden Seiten finden Sie Ihre Ergebnisse aus dem von Ihnen besuchten Testtraining, das von der Universität Münster durchgeführt wurde, sowie eine ausführliche Erläuterung. Um Missverständnisse zu vermeiden, empfehlen wir Ihnen, vorab die Erläuterungen zu lesen, bevor Sie Ihre Ergebnisse einordnen.

1 Allgemeine Informationen zu den rückgemeldeten Werten

Die Ihnen rückgemeldeten Werte sagen etwas darüber aus, in welchem Ausmaß Sie über ein bestimmtes Persönlichkeits- oder Fähigkeitsmerkmal im Verhältnis zu anderen Personen verfügen. Sie erhalten die Rückmeldung in Form sogenannter *Referenzwerte*. Die Personen, auf die für die Beurteilung Ihrer Ergebnisse in Form von Referenzwerten Bezug genommen wird, werden im folgenden *Referenzgruppe* genannt. Die Referenzgruppe stellt die Gruppe von Personen dar, mit der Ihre Ergebnisse verglichen werden. Bei vielen Tests gibt es verschiedene Referenzgruppen (zum Beispiel getrennt nach Schulbildung, Alter oder Geschlecht). Die hier rückgemeldeten Werte sind also immer bezogen auf die Referenzgruppe, der Sie zuzuordnen sind.

Hinweise zur Interpretation der Referenzwerte: Im Allgemeinen ist es sowohl bei Leistungs- als auch Persönlichkeitsmerkmalen so, dass die meisten Personen eine mittlere Ausprägung besitzen, wenige Personen eine hohe beziehungsweise

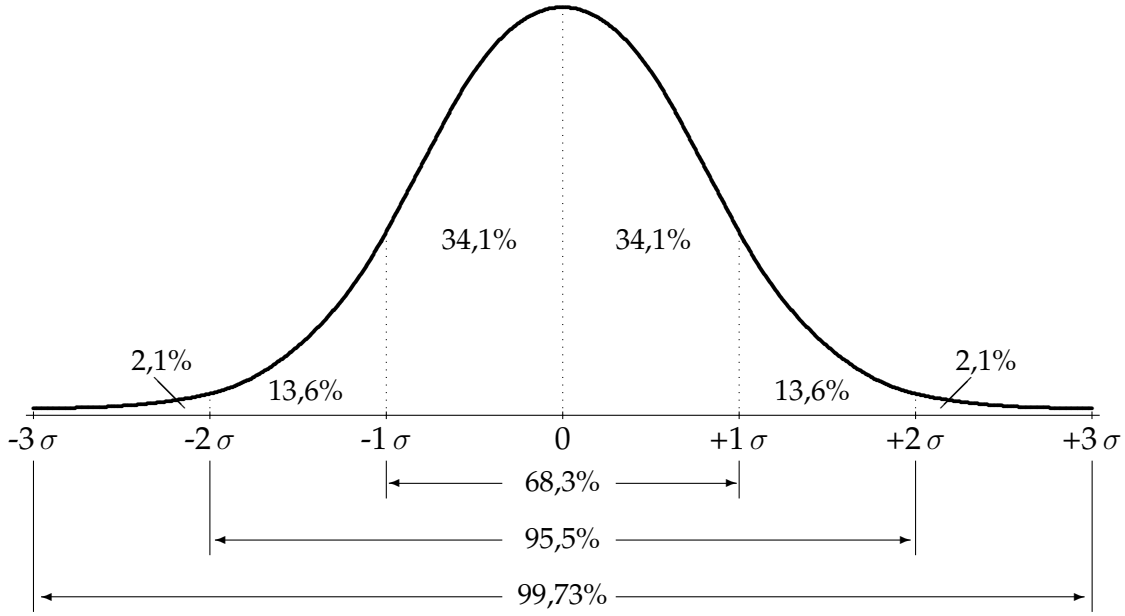


Abbildung 1: Verteilung der Referenzwerte und deren Interpretation als Prozentwerte.

niedrige Ausprägung und sehr wenige Personen haben eine sehr hohe beziehungsweise sehr niedrige Ausprägung. Dieser Sachverhalt wird grafisch durch die Kurve in Abbildung 1 illustriert.

Des Weiteren sehen Sie anhand dieser Abbildung, dass die Referenzwerte auch im Sinne von Prozentwerten interpretiert werden können, und zwar in Prozent von Personen, die einen bestimmten Referenzwert erreichen. Tabelle 1 auf der nächsten Seite gibt Ihnen einen Überblick über eine Reihe von Referenzwerten und den zugehörigen Prozentwerten.

Sie finden zunächst eine Beschreibung der Referenzgruppe und Erläuterungen zu den rückgemeldeten Merkmalen. Die Interpretation Ihrer Ergebnisse sollte nur vor dem Hintergrund dieser Informationen erfolgen!

Bitte beachten Sie bei der Rückmeldung die folgenden Punkte:

- Die Werte Ihrer Rückmeldung sind nicht als Bewertung zu verstehen. Ein geringerer Wert kann daher für Sie, zum Beispiel bei einzelnen Persönlichkeitsmerkmalen, durchaus eine positive Bedeutung haben!

- Bei den „Textaufgaben zur Wahrscheinlichkeitsrechnung“ wird Ihre Leistung nicht mit der anderer Probanden verglichen, daher erhalten Sie für diese Aufgaben lediglich eine Angabe darüber, wie viele Aufgaben Sie richtig gelöst haben. Ein Vergleich mit anderen Teilnehmern wäre irreführend, da Studenten und Schüler unterschiedlichen Alters und unterschiedlichen

GB	RW	Bedeutung
+3 σ	130	Wenn Sie bei einem Merkmal diesen Wert haben, hatten 99,9% der Personen der RG keine so hohe Ausprägung.
+2,5 σ	125	99% der Personen der RG hatten keine so hohe Ausprägung.
+2 σ	120	97% der Personen der RG hatten keine so hohe Ausprägung.
+1,5 σ	115	93% der Personen der RG hatten keine so hohe Ausprägung.
+1 σ	110	84% der Personen der RG hatten keine so hohe Ausprägung.
+0,5 σ	105	69% der Personen der RG hatten keine so hohe Ausprägung.
0	100	Hier hatten genau 50% der Personen der RG eine höhere und 50% der Personen der RG eine niedrigere Ausprägung. Dieser Wert entspricht somit einer mittleren Ausprägung.
-0,5 σ	95	31% der Personen der RG hatten keine so hohe Ausprägung.
-1 σ	90	16% der Personen der RG hatten keine so hohe Ausprägung.
-1,5 σ	85	7% der Personen der RG hatten keine so hohe Ausprägung.
-2 σ	80	3% der Personen der RG hatten keine so hohe Ausprägung.
-2,5 σ	75	1% der Personen der RG hatten keine so hohe Ausprägung.
-3 σ	70	0,1% der Personen der RG hatten keine so hohe Ausprägung.

Tabelle 1: Bedeutung der Referenzwerte und zugehörige Prozentangaben. GB = Grafikbezugswert aus Abbildung 1, RW = Referenzwert und RG = Referenzgruppe.

Geschlechts teilgenommen haben.

- Die aus der Tabelle ersichtlichen Persönlichkeitswerte reflektieren eine Zusammenfassung Ihrer Selbstbeschreibung. Sie selbst haben sich anhand der Fragen beschrieben. Die Rückmeldung der Ergebnisse ist daher abhängig davon, wie genau Sie die Fragen beantwortet haben und welches Bild Sie von sich selbst haben.

- Jede Messung psychischer Merkmale ist messfehlerbehaftet. Das heißt, dass wir im Einzelnen Ihre Merkmalsausprägung durchaus über- oder unterschätzt haben können. Je besser Sie bei der Testdurchführung unseren Anweisungen gefolgt sind, desto genauer dürften die Ergebnisse für Sie sein. Insbesondere bei den Persönlichkeitsmerkmalen ist zu bedenken, dass die Ergebnisse von dem „Ausmaß“ der Ehrlichkeit Ihrer Antworten abhängig sind.

- Bedenken Sie, dass es sich vor allem bei den Fähigkeits- und Leistungstests um eine Momentaufnahme handelt, die durch eine Vielzahl von Faktoren

(zum Beispiel Tagesform, Ablenkung, Motivation etc.) beeinflusst wird. Leistungsmerkmale unterliegen damit oft starken Schwankungen und sind keineswegs endgültig.

2 Erläuterungen der durchgeführten Tests

2.1 Textaufgaben zur Wahrscheinlichkeitsrechnung

Bei den Aufgaben zur Wahrscheinlichkeitsrechung handelt es sich um einen neu entwickelten Aufgabentyp, der sich noch in der Pilotphase befindet. Aus diesem Grund können auch noch keine Normwerte oder sonstige Vergleichsgrößen angegeben werden. Die Ergebnisse werden Ihnen in absoluter Form zurückgemeldet, d.h. wie viele Aufgaben Sie richtig gelöst haben (insgesamt wurden 16 Aufgaben bearbeitet). Die Aufgaben sollen der Kompetenzmessung im Bereich Wahrscheinlichkeitsrechnung dienen. Dies ist vor allem für Lehrer, aber auch für die SchülerInnen selbst relevant, sei es im Zuge des Unterrichts und regelmäßiger Lernstandserhebungen, im Bereich der Verlaufsmessung bei Fördermaßnahmen oder auch für die Vorbereitung auf das Zentralabitur.

2.2 Aufmerksamkeits-Belastungs-Test (d2-Test)

Der Test d2 ist ein Einzel- und Gruppentest zur Untersuchung der individuellen Aufmerksamkeit und Konzentrationsfähigkeit. Er findet Verwendung in nahezu allen psychologischen Arbeitsbereichen. Der Test d2 misst Tempo und Sorgfalt des Arbeitsverhaltens bei der Unterscheidung ähnlicher visueller Reize und ermöglicht damit die Beurteilung individueller Aufmerksamkeits- und Konzentrationsleistungen. Es werden folgende Werte zurückgemeldet:

- GZ: Gesamtzahl der bearbeiteten Zeichen als Maß für die Schnelligkeit der Bearbeitung.
- F%: Anteil der Fehler an der Gesamtzahl als Maß für die Genauigkeit der Bearbeitung. Je weniger Fehler desto besser.
- KL: Zuverlässige fehlerkorrigierte Konzentrationsleistung, die sowohl Genauigkeit als auch Schnelligkeit der Bearbeitung berücksichtigt.

Es ist zu beachten, dass der Anteil der Fehler hier ein Wert in der Rückmeldung ist, der eine günstigere Aussage macht, je niedriger er ausfällt. Das heißt, Personen, die einen niedrigen Referenzwert haben, haben viele Fehler gemacht.

Referenzgruppen: *Verfügbar für verschiedene Altersstufen, Schulbildung, Geschlechter (Gruppengröße variiert je nach Gruppe).*

2.3 Grundintelligenztest (CFT)

Der CFT erfasst das allgemeine intellektuelle Niveau (Grundintelligenz) im Sinne der Cattell'schen „flüssigen Intelligenz". Diese kann umschrieben werden als Fähigkeit, figurale Beziehungen und formal-logische Denkprobleme mit unterschiedlichem Komplexitätsgrad zu erkennen und innerhalb einer bestimmten Zeit zu verarbeiten. Da dies durch sprachfreie und anschauliche Testaufgaben geschieht, werden Personen mit geringen Kenntnissen der deutschen Sprache nicht benachteiligt. Die zu Beginn der Erhebung durchgeführten Testteile des CFT bestanden aus vier Untertests (Reihenfortsetzen, Klassifikationen, Matrizen und topologische Schlussfolgerungen). In der Tabelle ist in der Rubrik „CFT" der oben beschriebene Referenzwert aufgeführt.

Referenzgruppen: *Verfügbar für verschiedene Altersstufen, wobei die Gruppengröße je nach Gruppe variiert.*

2.4 Interessentest „AIST" (Allgemeiner Interessen-Struktur-Test)

Der AIST wird oft in Zusammenhang mit Berufs- und Laufbahnentscheidungen eingesetzt (zum Beispiel Berufsorientierung, Berufsentscheidung, innerbetriebliche Laufbahn- und Personalentscheidungen). Der AIST ist ein Fragebogen zur Erfassung schulisch-beruflicher Interessen und Tätigkeiten. Er besteht aus 60 Fragen, mit denen sechs Interessendimensionen gemessen werden: Praktisch-technische Interessen, intellektuell-forschende Interessen, künstlerisch-sprachliche Interessen, soziale Interessen, unternehmerische Interessen sowie konventionelle Interessen.

R = Realistischer Typ: Menschen mit dieser Grundorientierung haben eine Vorliebe für Tätigkeiten, die Kraft, Koordination und Handgeschicklichkeit erfordern und zu konkreten, sichtbaren Ergebnissen führen. Charakteristisch ist der formende Umgang mit Materialien und die Verwendung von

Werkzeugen oder Maschinen. Menschen dieses Typs weisen Fähigkeiten und Fertigkeiten vor allem im mechanischen, technischen, elektronischen und landwirtschaftlichen Bereich auf, während sie erzieherische oder soziale Tätigkeiten eher ablehnen. Ihre Werthaltungen sind auf materielle Dinge gerichtet: Geld, Macht und sozialer Status.

I = Intellektueller Typ: Personen mit dieser Grundorientierung haben eine Vorliebe für Aktivitäten, bei denen die symbolische, schöpferische oder beobachtende Auseinandersetzung mit physischen, biologischen oder kulturellen Phänomenen im Vordergrund steht. Sie möchten diese Phänomene verstehen und unter Kontrolle bringen. Gleichzeitig besteht eher eine Abneigung gegenüber überredenden, sozialen oder repetitiven Tätigkeiten. Ihre Fähigkeiten und Fertigkeiten liegen vor allem im mathematischen und naturwissenschaftlichen Bereich, ihre Werthaltungen sind vor allem auf Wissen(schaft) gerichtet.

K = Künstlerischer Typ: Menschen mit dieser Grundorientierung haben vor allem eine Vorliebe für offene, unstrukturierte Aktivitäten, die ihnen den auf künstlerische Selbstdarstellung oder die Schaffung kreativer Produkte gerichteten Umgang mit Material, Sprache oder auch Menschen ermöglichen. Weniger gut liegen ihnen klar abgegrenzte, systematische und geordnete Tätigkeiten. Ihre Fähigkeiten und Fertigkeiten liegen in den Bereichen Sprache, bildende Kunst, Musik, Schauspiel und Schriftstellerei. Sie streben vor allem ästhetische Werte an.

S = Sozialer Typ: Menschen mit dieser Grundorientierung haben eine Vorliebe für Tätigkeiten, bei denen sie sich mit anderen Menschen in Form von Unterrichten, Lehren, Ausbilden, Versorgen oder Pflegen befassen können. Weniger gut liegen ihnen klar abgegrenzte, systematische Tätigkeiten oder der Umgang mit Werkzeugen oder Maschinen. Ihre speziellen Fähigkeiten und Fertigkeiten liegen in den zwischenmenschlichen Beziehungen, insbesondere im sozialen Umgang und im erzieherischen Bereich. Ihre zentrale Werteausrichtung bezieht sich auf soziale und ethische Fragestellungen.

U = Unternehmerischer Typ: Personen mit dieser Grundorientierung haben eine Vorliebe für Tätigkeiten oder Situationen, in denen sie andere - meist um ein organisatorisches Ziel oder einen wirtschaftlichen Gewinn zu erreichen - mit Hilfe der Sprache oder anderen Mitteln beeinflussen, zu etwas bringen, führen, oder auch manipulieren können. Weniger gut liegen ihnen beobach-

tende oder systematische Tätigkeiten. Die spezifischen Fähigkeiten und Fertigkeiten solcher Personen sind ihre Führungs- und überzeugungsstärke. Ihre zentrale Werthaltung ist der soziale, politische oder ökonomische Erfolg.

C = Konventioneller Typ: Menschen mit dieser Grundorientierung haben eine Vorliebe für den genau bestimmten, geordneten, systematischen Umgang mit Daten: Dokumentationen anlegen, Aufzeichnungen führen, Materialien ordnen, maschinelle Verarbeitung organisatorischer oder wirtschaftlicher Daten. Weniger gut liegen ihnen offene, unstrukturierte Tätigkeiten. Ihre speziellen Fähigkeiten und Fertigkeiten sind rechnerischer, verwaltender und geschäftlicher Art.

Referenzgruppe für den Interessentest: *Die Bezugsgruppe besteht aus einer Stichprobe von ungefähr 4400 14- bis 20-jährigen Probanden.*

2.5 Persönlichkeitstest NEO-FFI (NEO-Fünf-Faktoren-Inventar)

Das NEO-FFI ist ein häufig verwendeter Persönlichkeitstest. Einsatzmöglichkeiten liegen insbesondere in der Schullaufbahn- und Studienberatung, in Berufsberatung und Organisationspsychologie sowie in der psychologischen Forschung. Das NEO-FFI ist ein multidimensionales Persönlichkeitsinventar, das die wichtigsten Bereiche individueller Unterschiede erfasst. Das NEO-FFI erfasst mit seinen insgesamt 60 Fragen diese Dimensionen auf fünf Skalen: Emotionale Labilität, Extraversion, Offenheit für Erfahrung, Verträglichkeit und Gewissenhaftigkeit.

L = Emotionale Labilität: Mit diesen Werten werden individuelle Unterschiede in der emotionalen Stabilität und der emotionalen Labilität zum Ausdruck gebracht. Der Kern dieses Merkmals liegt in der Art und Weise, wie Emotionen, vor allem negative Emotionen, erlebt werden. Personen mit hohen Werten beschreiben sich selbst als nervös, sind leicht aus dem seelischen Gleichgewicht zu bringen, neigen zur Traurigkeit, sind eher ängstlich, unsicher und verlegen. Emotional labile Personen neigen zu unrealistischen Ideen, machen sich Sorgen um ihre Gesundheit und sind weniger in der Lage, auf Stresssituationen angemessen zu reagieren. Emotional stabile Personen dagegen (Personen mit niedrigen Werten) beschreiben sich selbst als ruhig, ausgeglichen, sorgenfrei und geraten in Stresssituationen nicht so schnell aus der Fassung.

E = Extraversion: Personen mit hohen Werten (extravertierte Personen) sind gesprächig, gesellig und aktiv. Sie beschreiben sich als selbstsicher, personenorientiert, herzlich, heiter und optimistisch und bevorzugen Aufregungen und Anregungen. Personen mit niedrigen Werten (introvertierte Personen) sind eher zurückhaltend als unfreundlich, eher unabhängig als folgsam, eher ausgeglichen als unsicher oder phlegmatisch. Wenn ihnen auch nicht die Lebhaftigkeit des Extravierten zu eigen ist, so sind Introvertierte doch nicht unbedingt unglücklich oder pessimistisch.

O = Offenheit für Erfahrungen: Personen, die hohe Werte bezüglich Offenheit für Erfahrungen aufweisen, zeichnen sich durch eine hohe Wertschätzung für neue Erfahrungen aus. Sie beschreiben sich als kreativ, wissbegierig, phantasievoll und machen ihre Urteile nicht von anderen abhängig. Sie sind vielseitig interessiert und bevorzugen Abwechslung. Personen mit niedrigen Werten neigen demgegenüber eher zu konventionellem Verhalten und zu konservativen Einstellungen. Sie ziehen Bekanntes und Bewährtes dem Neuen vor, und ihre emotionalen Reaktionen sind eher gedämpft.

V = Verträglichkeit: Personen mit hohen Werten beschreiben sich als hilfsbereit, mitfühlend, kooperativ, verständnisvoll und wohlwollend. Sie neigen zu zwischenmenschlichem Vertrauen, sind eher nachgiebig und haben ein erhöhtes Harmoniebedürfnis. Personen mit niedrigen Werten beschreiben sich im Gegensatz dazu als egozentrisch und misstrauisch gegenüber den Absichten anderer Menschen.

G = Gewissenhaftigkeit: Mit diesem Merkmal wird eine Art der Selbstkontrolle beschrieben, die sich auf den aktiven Prozess der Planung, Organisation und Durchführung von Aufgaben bezieht. Personen mit hohen Werten sind ordentlich, zuverlässig, hart arbeitend, pünktlich, penibel, diszipliniert, ehrgeizig, systematisch und genau. Personen mit niedrigen Werten beschreiben sich dagegen eher als nachlässig, gleichgültig und unbeständig, sie verfolgen ihre Ziele also mit geringem Engagement.

Referenzgruppe für den Persönlichkeitstest: *Die Bezugsgruppe besteht aus einer bevölkerungsrepräsentativen Stichprobe von ungefähr 2100 Personen.*

3 Ihre Ergebnisse

Sollten Teile Ihrer Ergebnisse nicht verwertbar gewesen sein, so entfällt die Rückmeldung für die betroffenen Merkmale.

Test	**Textaufgaben**	**d2**			**CFT**
		GZ	**F%**	**KL**	
Ihr Wert					

Test	**AIST**						**NEO-FFI**				
	R	**I**	**K**	**S**	**U**	**C**	**L**	**E**	**O**	**V**	**G**
Ihr Wert											

Für Rückfragen hinsichtlich Ihrer Ergebnisse stehen wir Ihnen gerne zur Verfügung. Bei konkreten Rückfragen zu individuellen Ergebnissen schreiben Sie eine E-Mail unter Angabe von Testungsort (zum Beispiel Name der Schule), Testdatum und CODE an Nina Zeuch (E-Mail-Adresse: *n_hoff01@uni-muenster.de*).