

Westfälische Wilhelms-Universität Münster
Fachbereich Psychologie und Sportwissenschaften

Measuring Reasoning Ability: Applications of Rule-Based Item Generation

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

vorgelegt von
Jonas Pablo Bertling
aus Dinslaken

2012

Tag der mündlichen Prüfung: 05.07.2012

Dekan: Prof. Dr. Christian Pietsch

Referent: Prof. Dr. Heinz Holling

Koreferent: PD Dr. Günther Gediga

Danksagung

Mein Dank gilt allen, die mich bei der Konzeption, Durchführung und Fertigstellung dieser Arbeit unterstützt haben. Die hier dargestellten Studien haben nicht nur meine Forschungsinteressen, sondern auch mein Leben in einer für mich wichtigen Phase entscheidend geprägt.

Zu allererst danke ich meinem Doktorvater Prof. Dr. Heinz Holling sowie meinem Zweitgutachter PD Dr. Günther Gediga für die hervorragende fachliche Betreuung, das Vertrauen und die stetige Förderung über den gesamten Verlauf meines Studiums hinweg. Die Zeit als Mitglied in Prof. Hollings Arbeitseinheit wird mir immer in Erinnerung bleiben, nicht nur als ein prägender Lebensabschnitt sondern auch als eine Zeit, in der ich mein Methodenwissen erwerben und ausbauen konnte und in Forschungs- und Lehrprojekten Verantwortung übernehmen durfte. Mein Dank gilt insbesondere auch dafür, mein Interesse an Forschung geweckt und mich nachhaltig für sie begeistert zu haben.

Meinen Kollegen Alexander Freund, Jörg-Tobias Kuhn, Nina Zeuch und Sergej Davidov danke ich für die Unterstützung bei Konzeption und Durchführung der hier beschriebenen Studien.

Den zahlreichen engagierten studentischen Hilfskräften und Diplomanden, insbesondere Robert Wunderlich, Kai Knipping, Evelyn Alex und Boris Forthmann, danke ich herzlich für die wichtigen Beiträge zur Itementwicklung sowie für die tatkräftige Unterstützung bei Datenerhebungen.

Ganz besonders danke ich Masha Bertling, die mich nicht nur an entscheidenden Stellen bei der Realisierung dieser Arbeit unterstützt hat, sondern auch verstanden hat, mich in den richtigen Momenten zu motivieren und dabei immer an mich geglaubt hat. Mein persönlicher Dank gilt ebenso Monika, Baldur und Gerd Bertling für das bedingungslose Vertrauen, die kontinuierliche Unterstützung und die vielen guten Gespräche, deren Inhalte in die Gestaltung der Arbeit eingeflossen sind.

Zusammenfassung [Summary]

Die Fähigkeit schlussfolgernd zu denken ist eine zentrale Voraussetzung für Lernen und Problemlösen. Sie wird als wichtige Komponente menschlicher Intelligenz aufgefasst. Seit Beginn des 20. Jahrhunderts wurde eine Vielzahl von Messinstrumenten entwickelt, mit denen Ausprägungen der beiden Komponenten schlussfolgernden Denkens (Induktion und Deduktion) erfasst werden können.

Während in der psychologischen Grundlagenforschung große Fortschritte bei der Beschreibung und Erklärung von menschlicher Wahrnehmung und Informationsverarbeitung gemacht wurden, sind diese Erkenntnisse bisher nur kaum in die Konstruktion von diagnostischen Tests zur Erfassung der Intelligenz im Allgemeinen und des schlussfolgernden Denkens im Speziellen eingeflossen. Häufig werden in Anwendungskontexten veraltete Messinstrumente eingesetzt, die nur unzureichend die Prozesse menschlicher Intelligenz abbilden, und über deren Konstruktvalidität – abgesehen davon, dass sich in der Regel recht gut externe Erfolgsgrößen wie Studien- oder Berufserfolg vorhersagen lassen – wenig bekannt ist.

Methoden der regelgeleiteten automatischen Aufgabengenerierung (AIG) ermöglichen die Berücksichtigung von kognitiven Theorien und Modellen bereits während der Neukonstruktion von Testverfahren. Ziel dieser Methoden ist es, Schwierigkeiten von Testaufgaben aufgrund der Teilschwierigkeiten der kognitiven Prozesse vorherzusagen, die für die Aufgabenlösung erforderlich sind. Mithilfe von sogenannten erklärenden Item Response Modellen können Schwierigkeitsparameter geschätzt werden und die Übereinstimmung von vorhergesagten und tatsächlichen Schwierigkeiten, und somit die Eignung des jeweiligen kognitiven Modells zur Beschreibung der Testleistungen – oder umgekehrt: die Konstruktvalidität des Testverfahrens – überprüft werden. Erklärende IRT Modelle können somit einen Beitrag zu der Frage “was eigentlich gemessen wird” wenn ein Intelligenztest bearbeitet wird, leisten und Rückschlüsse auf die Gültigkeit von Theorien kognitiver Verarbeitung ermöglichen. Item-Cloning Ansätze beziehen sich auf die Frage, wie parallele Testversionen entwickelt werden können, deren Aufgaben strukturell und psychometrisch den Aufgaben einer Ausgangsversion entsprechen. Beim computergestützten adaptiven Testen werden beispielsweise große Mengen strukturgleiche Items benötigt, deren Konstruktion und Kalibrierung durch Item Cloning Techniken stark vereinfacht werden kann.

Die vorliegende Arbeit beschäftigt sich mit drei Anwendungen von AIG im Kontext der Erfassung schlussfolgernden Denkens.

Die erste Studie, “The Figural Analogy Test (FAT): Item Generation and Construct Validation”, beschreibt die Konstruktion und erste Validierung eines neuen figuralen Analogietests. Im Rahmen einer Studie mit $N=308$ Studierenden wird der Einfluss verschiedener Konstruktionsparameter auf die Aufgabenschwierigkeit geprüft. Ziel der Testentwicklung war es, einen rein figuralen, also vollkommen sprach- und numerikfreien, Analogietest zu konstruieren, dessen Schwierigkeiten sich auf Basis eines kognitiven Modells vorhersagen lassen. Bei der Bearbeitung des FAT müssen die Testpersonen unterschiedlich komplexe räumliche Relationen zwischen abstrakten Figuren erkennen und analog auf andere Figuren übertragen. Die Arbeit baut auf einer Arbeit von Beckmann (2008) auf, die einen figuralen Analogietest basierend auf alphanumerischen Symbolen vorgestellt hat. Zwei Forschungsfragen sind Gegenstand der empirischen Studie. Erstens wird überprüft, ob sich Itemschwierigkeiten des neuen Tests auf Basis eines Sets von vorher spezifizierten strukturellen Parametern erklären und vorhersagen lassen. Zweitens wird getestet, ob die geschätzten Modellparameter im Einklang mit

kognitiven Theorien des figural-räumlichen Denkens und analogen Schlussfolgerns stehen. Eine Reihe von spezifischen Hypothesen wird getestet. Verschiedene erklärende IRT Modelle werden verglichen, unter anderem Modelle mit Personen-mal-Item Interaktionen zur Prüfung von facetten-spezifischen Geschlechtseffekten. Alle Parameterwerte fallen im Einklang mit den Hypothesen aus. Außerdem wird gezeigt, dass Geschlechtsunterschiede, wie auf Basis von Theorien räumlicher Fähigkeiten erwartet, nur für bestimmte Itemmerkmale auftreten.

Die zweite Studie, "Development of the Number Series Test (NST): Item-generation and Investigation of Parallel Test Forms", beschreibt die Entwicklung eines Konstruktionsansatzes zur Generierung von Zahlenreihenaufgaben. Basierend auf Theorien des numerischen Schlussfolgerns und früherer Arbeiten zu Zahlenreihen, wird ein neuer Ansatz vorgeschlagen, auf Basis dessen sich neue Zahlenreihenaufgaben generieren lassen. Der Generierungsansatz wird anhand von Daten einer Studentenchichprobe (N=406) überprüft, die jeweils zwei strukturgleiche Testformen plus einen Aufwärmdurchlauf des neuen Aufgabentyps bearbeiteten. Dabei wird die Äquivalenz der Aufgaben geprüft, sowie der Einfluss von Oberflächenmerkmalen auf Lösungsprozesse und die Attraktivität von bestimmten Falschlösungen betrachtet. Erklärende IRT Modelle verschiedener Komplexitätsstufen werden verglichen. Die Ergebnisse der Studie zeigen, dass parallele Testformen auf Basis des neuen Regelsets konstruiert werden können wenn Quellen für unerwünschte Schwierigkeitsvariation kontrolliert werden. Itemschwierigkeiten können zu großen Teilen durch die relationale Komplexität von jeweils zwei aufeinanderfolgenden Zahlen erklärt werden. Zudem deuten die Ergebnisse auf hinreichende Robustheit bezüglich von Itemoberflächenmerkmalen hin. Korrelationen mit schlussfolgerndem Denken und Schulnoten bestätigen die Validität des NST.

Die dritte Studie, "A cross-cultural Investigation of the Latin Square Task", beschreibt die Überprüfung der interkulturellen Validität eines figuralen Tests vom Typ Lateinische Quadrate zur Erfassung des schlussfolgernden Denkens anhand Stichproben russischer (N=201) und deutscher (N=452) Studierender. Zum Aufgabentyp Lateinische Quadrate, welcher Ähnlichkeiten mit den beliebten SUDOKU-Rätseln aufweist, existieren bereits eine Vielzahl von Befunden, unter anderem zur automatischen Aufgabengenerierung und zu Retest-Effekten. Die vorliegende Studie ergänzt diese Anwendungen um eine kulturvergleichende Perspektive. Dabei wird überprüft, inwieweit die der Aufgabenlösung zugrunde liegenden Prozesse unabhängig von der Kulturzugehörigkeit der Testperson sind, oder ob – wie bei vielen kognitiven Tests der Fall – Verzerrungen im Sinne von Differential Item Functioning (DIF) zugunsten einer kulturellen Gruppierung auftreten. Ergänzend werden Ergebnisse von Differential Facet Functioning (DFF) Analysen berichtet um zu überprüfen, ob DIF Effekte auf Facettenebene abgebildet und vorhergesagt werden können. Während die Ergebnisse der Studie keine Anzeichen auf bedeutende Unterschiede hinsichtlich der Konstruktvalidität der Testaufgaben zeigen, gibt es deutliche Hinweise auf Differential Item Functioning für die beiden Studentengruppen. Qualitative Analysen der Testitems deuten darauf hin, dass DIF möglicherweise eher durch bestimmte Oberflächenmerkmale, und nicht nur das Auftreten bestimmter, kognitiv komplexer struktureller Parameter zustande kommt.

Insgesamt liefert diese Arbeit aus inhaltlicher Sicht einen Beitrag zu einem besseren Verständnis der Konstruktvalidität von Testverfahren des schlussfolgernden Denkens. Mit engem Bezug zu spezifischen Theorien kognitiver Verarbeitung werden neue Konstruktionsansätze für zwei Testverfahren vorgestellt, die für Anwendungen im Bereich des computergestützten adaptiven Testens Verwendung finden könnten. Die hier beschriebenen Studien haben den Charakter von Pilotstudien für die Entwicklung vollends automatischer Itemkonstruktionsprogramme. Eine Computersoftware zur automatischen

Generierung und Vorgabe von Testaufgaben auf Basis der Ergebnisse dieser Dissertation ist momentan in Entwicklung. Aus methodischer Sicht illustriert diese Arbeit die Anwendung von erklärenden IRT Modellen zur Vorhersage von Itemschwierigkeiten und Erstellung von Paralleltests, sowie mögliche Anwendungen zur Prüfung der Robustheit von Itemgenerierungsansätzen psychometrischer Testverfahren.

Contents

1. General introduction	1
1.1. Research goals	3
1.2. Outline	4
2. Theoretical background	8
2.1. Reasoning ability as a construct	8
2.2. Cross-cultural validity of reasoning tests	12
2.2.1. The term “culture” in cross-cultural studies	12
2.2.2. The challenge of “culture-fair” testing	14
2.2.3. Cognitive and cultural complexity as possible factors for bias	18
2.3. Item difficulty modeling and rule-based automatic item generation	20
2.4. Explanatory item response modeling	27
2.4.1. Models with item predictors	28
2.4.2. Models with person predictors	37
2.5. Differential item and facet functioning	38
3. The Figural Analogy Test (FAT): Item generation and construct validation	46
3.1. Introduction	47
3.1.1. Figural-spatial analogies as indicators of fluid intelligence	47
3.1.2. Rule-based generation of figural analogy items	50
3.1.3. Research questions	55
3.2. Method	56
3.2.1. Development of the new item-generative framework	56
3.2.2. Specific hypotheses	68
3.2.3. Sample	69
3.2.4. Instruments and procedure	69
3.3. Results	71
3.3.1. Prediction of item difficulty parameters	72
3.3.2. Construct validity	75
3.4. Discussion	80
3.4.1. Conclusions regarding the research questions	81
3.4.2. Limitations and future prospects	86

4. The Number Series Test (NST): Item generation of parallel forms	90
4.1. Introduction	91
4.1.1. Number series items as indicators of reasoning ability	91
4.1.2. An information processing model for number series	95
4.1.3. Item difficulty modeling of number series items: Previous attempts and problems	97
4.1.4. Research questions	105
4.2. Method	106
4.2.1. Development of the new item-generative framework	106
4.2.2. Sample	114
4.2.3. Instruments and procedure	115
4.3. Results	119
4.3.1. Equivalence of parallel test forms	120
4.3.2. Item difficulty modeling	126
4.3.3. Predictive power of the model	133
4.4. Discussion	134
4.4.1. Conclusions regarding the research questions	135
4.4.2. Limitations and future prospects	139
5. A cross-cultural investigation of the Latin Square Task	142
5.1. Introduction	143
5.1.1. The Latin Square Task	144
5.1.2. Similarities to the popular number placement game SUDOKU	146
5.1.3. Cross-cultural validity of the LST	148
5.1.4. Research questions	150
5.2. Method	151
5.2.1. Sample	151
5.2.2. Instruments and procedure	152
5.3. Results	154
5.3.1. Psychometric properties of the LST for the two samples	155
5.3.2. Item difficulty modeling for the two samples	158
5.3.3. Differential item functioning analyses	161
5.3.4. Differential facet functioning analyses	166
5.3.5. Qualitative analyses of DIF in LSTs	172
5.4. Discussion	177
5.4.1. Conclusions regarding the research questions	179
5.4.2. Limitations and future prospects	181
6. Epilogue	186
6.1. Prediction of item difficulties by means of explanatory IRT models	187
6.2. Understanding and enhancing construct validity	189

6.3. Limitations	191
References	193
A. Study 1 – Additional Materials	212
B. Study 2 – Additional Materials	230
C. Study 3 – Additional Materials	235

List of Tables

1.1.	List of abbreviations frequently used in this thesis	7
2.1.	Reasoning in prominent structural models of human intelligence	10
2.2.	Prominent content areas for rule-based automatic item generation	21
2.3.	Explanatory and descriptive IRT models	28
2.4.	Differences in model foci between explanatory IRT models with different types of design matrices	34
2.5.	Overview of methods for detecting differential item functioning for two groups	41
2.6.	Structure of a contingency table for non-IRT DIF methods	41
3.1.	Spatial rules and cognitive processes in figural-spatial tasks (Study 1)	49
3.2.	Item-generative framework behind Beckmann’s analogy test: Radicals	53
3.3.	Possible rule-combinations in the FAT (Study1)	64
3.4.	Descriptives for the FAT and other measures used (Study 1)	71
3.5.	Classical item statistics, Rasch parameters, and item fit statistics (Study 1)	73
3.6.	Standardized absolute errors for the alignment of rescaled LLTM and Rasch parameters (Study 1)	76
3.7.	Explanatory IRT modeling for the FAT: item difficulty modeling (Study 1)	77
3.8.	Explanatory IRT modeling for the FAT: gender effects (Study 1)	78
3.9.	Correlations between FAT scores and other variables (Study 1)	79
3.10.	Prediction of FAT performance by other tests and gender (Study 1)	80
3.11.	Prediction of math grades by FAT scores and other tests (Study 1)	81
4.1.	Examples of typical number series tasks	91
4.2.	LLTM parameter estimates for item facets manipulated in Porsch’s study . .	98
4.3.	Design matrix for item types of the NST (Study 2)	116
4.4.	Detailed description of all item types of the NST (Study 2)	117
4.5.	Summary statistics for all instruments (Study 2)	120
4.6.	Correlations of the three parallel item sets with math grade and g (Study 2)	120
4.7.	Three parallel NST sets, answer frequencies and item statistics	122
4.8.	“Virtual item model” results, comparison of set A and set B (Study 2) . . .	124

4.9. “Virtual item model” results, comparison of warm-up items and set A (Study 2)	125
4.10. LR model comparison tests for the the three different explanatory IRT models (Study 2)	126
4.11. Rescaled item difficulty parameters for two different LLTM models for the NST (Study 2)	127
4.12. LLTM modeling for the NST (Study 2)	131
4.13. Frequent wrong answers for the NST and possible explanations (Study 2)	133
4.14. Predictive power and sparseness of different explanatory IRT models for the NST (Study 2)	134
5.1. Means and standard deviations for the German and Russian sample (Study 3)	154
5.2. Rasch parameters and item fit statistics for the LST in both samples (Study 3)	156
5.3. Correlations between the LST and other variables (Study 3)	157
5.4. LLTM with basic design matrix for each sample (Study 3)	159
5.5. LLTM with extended design matrix for each sample (Study 3)	160
5.6. Results for country-dependent DIF in the LST (Study 3)	162
5.7. Results for pre-knowledge dependent DIF in the LST (Study 3)	163
5.8. Surface characteristics and applicability of two simple solution heuristics for LST items flagged as DIF	168
5.9. LLTM & DFF modeling for the total sample (Study 3)	184
5.10. Model fit indices for the different DFF-models (Study 3)	185
A.1. Design matrix for the 40-item FAT investigated in this study	227
B.1. Correlations of responses on WT items with responses of non-WT items, general cognitive ability, and scholastic performance (Study 2)	233
B.2. Results for separate LLTM models for Russian and German test takers (Study 2)	234
C.1. Sum-normed RM item difficulty parameters for full Russian sample and two subsamples (Study 3)	250
C.2. Frequencies of A, B, C DIF for items that (not) allow for a reduction of considerable response alternatives (Study 3)	251
C.3. Frequencies of A, B, C DIF for items that (not) allow for the application of an easy falsification strategy (Study 3)	251

List of Figures

2.1. Process model of reasoning based on stages distinguished by Wilhelm (2005)	11
2.2. Four essential steps of the automatic item generation process	23
2.3. Illustration of different possible design matrices on the continuum of explanatory IRT models: Number of radicals	35
2.4. Illustration of different possible design matrices on the continuum of explanatory IRT models: RM, LLTM, and Item Cloning	36
2.5. Uniform and nonuniform-DIF in terms of the ICC of an item	39
3.1. Process model of analogical reasoning (Study 1)	48
3.2. Example Item with two rules from Beckmann’s analogy test (Study 1) . . .	52
3.3. Sample FAT item with 9 response alternatives (Study 1)	57
3.4. Combination of main shapes and features into figural objects in the FAT (Study 1)	58
3.5. Exemplary illustration of all rules that apply to the main shape in the FAT (Study 1)	58
3.6. Exemplary illustration of all rules that apply to the features in the FAT (Study 1)	59
3.7. Illustration of the complexity parameter “Type of Form” in the FAT (Study 1)	60
3.8. Exemplary illustration of the complexity parameter “Type of Form” in two FAT items (Study 1)	61
3.9. Illustration of the complexity parameter “Additional Feature” in the FAT (Study 1)	62
3.10. Exemplary illustration of the complexity parameter “Random Change of Feature Characteristics” in two FAT items (Study 1)	63
3.11. Examples for incidentals in the FAT (Study 1)	65
3.12. Overview of all possible main shapes in the current AIG framework for the FAT (Study 1)	66
3.13. RM and rescaled LLTM parameters for the FAT (Study 1)	72
3.14. Distribution of standardized absolute errors for LLTM 1 and LLTM 2 (Study 1)	74

4.1.	Overview of different possible attributes of number series tasks (Study 2)	93
4.2.	Process model for number series (Study 2)	96
4.3.	Four example NST items generated based on the new AIG framework (Study 2)	107
4.4.	Illustration of surface differences in structurally identical items caused by variation of item incidentals in the NST (Study 2)	112
4.5.	Test design and testing time for the NST (Study 2)	118
4.6.	Item parameters across parallel sets (Study 2)	123
4.7.	Alignment of RM item difficulties and rescaled LLTM difficulties for the two LLTM models for the NST (Study 2)	128
4.8.	Relationship of RM and rescaled item parameters for the two LLTM models for the NST (Study 2)	129
4.9.	Parameter Differences in logits between RM and rescaled LLTM item difficulties (Study 2)	130
4.10.	Possible alternative future versions of the NST (Study 2)	138
5.1.	Cognitive complexity determinants in the LST: Binary, ternary and quaternary processing (Study 3)	145
5.2.	Example SUDOKU puzzle (Study 3)	147
5.3.	Incidental item surface characteristics in the LST: example of two items based on the same item radicals (Study 3)	149
5.4.	Overview of country-dependent DIF in the LST	164
5.5.	Overview of pre-knowledge dependent DIF in the LST	165
5.6.	Number of items flagged as cDIF and sDIF by the five methods used (Study 3)	167
5.7.	Cross-cultural DFF in the LST (Study 3)	171
5.8.	LST items that allow or not allow for the application of a quick exclusion of response alternatives strategy (Study 3)	174
5.9.	LST items that allow for application of a quick falsification heuristic (Study 3)	175
5.10.	LST items that allow both for a quick reduction of response alternatives and for application of a falsification heuristic (Study 3)	176
5.11.	LST items that allow neither for a quick reduction of response alternatives nor for application of a falsification heuristic	178
A.1.	Figural Analogy Test: Items 1-3	213
A.2.	Figural Analogy Test: Items 4-6	214
A.3.	Figural Analogy Test: Items 7-9	215
A.4.	Figural Analogy Test: Items 10-12	216
A.5.	Figural Analogy Test: Items 13-15	217
A.6.	Figural Analogy Test: Items 16-18	218
A.7.	Figural Analogy Test: Items 19-21	219

A.8. Figural Analogy Test: Items 22-24	220
A.9. Figural Analogy Test: Items 25-27	221
A.10. Figural Analogy Test: Items 28-30	222
A.11. Figural Analogy Test: Items 31-33	223
A.12. Figural Analogy Test: Items 34-36	224
A.13. Figural Analogy Test: Items 37-39	225
A.14. Figural Analogy Test: Item 40	226
A.15. Optimal design SAS input file (Syntax)	226
A.16. Optimal design SAS output file	228
A.17. Item characteristic curves for all 40 FAT items (Study 1)	229
B.1. Item characteristic curves for all 33 NST items (Study 2)	231
B.2. “Wrong-track” items (Study 2)	232
C.1. LST: Items 1-6	236
C.2. LST: Items 7-8	237
C.3. LST: Items 13-18	238
C.4. LST: Items 19-24	239
C.5. LST: Items 25-30	240
C.6. Category frequencies for all LST: Items, comparison of Russian and German samples (Study 3)	241
C.7. Category frequencies for all LST: Items, comparison of Russian and German samples (cont’d)	242
C.8. Frequencies for “not solvable” choices among Russian and German test-takers (Study 3)	243
C.9. Item characteristic curves for uniform country-DIF based on the logistic regression model (Study 3)	244
C.10. Item characteristic curves for non-uniform country-DIF based on the logistic regression model (Study 3)	245
C.11. Item characteristic curves for uniform country-DIF based on Lord’s approach (Study 3)	246
C.12. Item characteristic curves for uniform pre-knowledge-DIF based on the logistic regression model (Study 3)	247
C.13. Item characteristic curves for non-uniform pre-knowledge-DIF based on the logistic regression model (Study 3)	248
C.14. Item characteristic curves for uniform pre-knowledge-DIF based on Lord’s approach (Study 3)	249
C.15. Alignment of sum-normed RM item difficulty parameters for full Russian sample and two subsamples (Study 3)	251

1

General introduction

Reasoning ability is an important requirement for learning and problems solving and at the core of what is typically labelled “Intelligence” or g (Spearman, 1923). Since the beginning of the 20th century, numerous psychological tests have been developed that measure reasoning ability, and practical applications have demonstrated the importance of reasoning for the prediction of important outcome variables (see e.g., Ones, Viswesvaran, & Dilchert, 2005; Schmidt & Hunter, 1998). While the practical importance of assessments has grown rapidly during the last decades, the question of construct validity, that is whether and how test results actually represent the intended underlying psychological abilities and skills, is widely ignored in testing practice. Test development has not kept pace with recent developments in cognitive psychology as well as neuroscience and related disciplines. While general cognitive ability can be measured with some precision, the “construct of g is poorly understood” (Carlstedt, Gustafsson, & Ullstadius, 2000, p. 145). Moreover, as Gierl and Lai (2012) stated, “there are currently no published studies describing either the principles or practices required to develop item models” (p. 27).

Methods of rule-based Automatic Item Generation (AIG; Irvine & Kyllonen, 2002) term the typically computer-based generation of test items based on a-priori defined algorithms. Item generation attempts are driven by (at least) two forces. On the one hand, research in cognitive sciences and the investigation of reasoning processes and resources is an important theoretical basis for AIG. As Wilhelm (2005) pointed out, generative item production can be “a side product of such efforts [efforts in the theoretically driven investigation of reasoning processes]” (p.388). On the other hand, the need for large calibrated item pools in international testing settings and enhanced test-efficiency through AIG generates new

insights into cognitive processing as well. From that perspective, improved construct validity and a better knowledge of cognitive processes and resources underlying reasoning performance can be seen as a side product as well. These two “faces” of AIG are reflected in Drasgow, Luecht, and Bennett’s (2006) distinction of “strong” theory and “weak” theory approaches. An example of a weak theory approach is given in Gierl and Lai (2012): the authors mention item cloning where item clones are created based on features of a parent or family item and “the determinants of item difficulty for the manipulated elements in the model must be discerned through the guidelines, judgments, and experiences of the content specialist” (p. 36). While weak models might be sufficient to model items based on characteristics of such parent items, a core drawback is, according to the authors, that “relatively few elements can be manipulated in the model because their effect on the psychometric characteristics or the generated items is poorly anticipated” (Gierl & Lai, 2012; p. 26). Further, “clones are believed to be easy to detect by coaching and test preparation companies and, therefore, of limited use in operational testing programs” (p. 27). Drasgow et al. (2006) recommended adopting weak theory approaches only when little theoretical knowledge or limited theoretical models on the cognitive processes underlying item responses are available. On the other hand, item generation based on “strong” theory is given when a cognitive model is used to define so-called radicals and incidentals of the item model for predicting the psychometric characteristics of the generated items (Drasgow et al., 2006; Gierl & Lai, 2012). Speaking with Embretson and Gorin (2001), “the most important potential for cognitive theory is test design.” (p. 364). Rule-based item generation approaches allow the integration of psychometric test models with psychological theories and test-development principles.

It is important to note that, regardless of whether strong or weak item models are used, there is no guarantee that item models capture the true cognitive processes underlying test performance. As Box and Draper (1987) stated, “all models are wrong, but some are useful” (p. 424). This is the case here as well. Item generation models address the concern that “since Spearman (...) the development of good reasoning tests has been almost an art form, owing more to empirical trial-and-error than to systematic delineation of the requirements such tests must satisfy” (Kyllonen & Christal, 1990, p. 426) by providing an empirical base for evaluating construct validity on the item level. However, despite their potential usefulness in explaining variation in item difficulties the use of item models *per se* does not guarantee high validities or capture of true underlying cognitive processes. Still, if explanatory models can be considered already during the development of new reasoning measures (instead of applying such models post-hoc to tests that were developed in a non-systematic idiosyncratic way), this provides an important basis for practical and high-stakes applications of generating items automatically and predicting item difficulties “on-the-fly” as part of computer-adaptive test systems. The investigation of construct validity can then be extended to the item and item-facet level. As Bejar (2012) stated, “there is a symbiotic relationship between theory and test development based on item generation (...) when items in a test become a psychological experiment, which in turn may

lead to the improvement of both theories and tests” (p. 45). If the intend is not only predicting item difficulties based on models that are *consistent* with theoretical assumptions, but capturing actual underlying cognitive processes, other methods in addition to modelling item difficulties based on explanatory rules, such as think-aloud studies, response time analysis, eye-tracking or neurophysiological methods should be applied. Establishing and validating strong theory item models for cognitive tests used in testing practice is only a first, but important, step towards a full understanding of the cognitive processes underlying response processes for reasoning tests.

1.1. Research goals

The focus of this thesis is on modeling the cognitive task structure of reasoning items and developing rule-based item generation models that link item generation with underlying theories of reasoning ability, that is, a “strong” theory approach (Drasgow et al., 2006) is taken where cognitive models are used to define item generation models. Two new rule-based item generation frameworks for a figural-spatial and a numerical reasoning test are presented and tested empirically. In addition, the validity of an existing figural reasoning measure in a cross-cultural context and the generalizability of the underlying item-difficulty model across heterogeneous test-taker populations are tested. Two general research goals link the three studies:

1. First, the *usefulness of item-generation models in predicting item difficulties* is investigated. An accurate prediction of item difficulties by the set of underlying pre-specified task parameters is a necessary condition for most applications of AIG in practical testing contexts, especially when AIG is combined with computerized-adaptive testing technology. The goal here is investigating the usefulness of existing psychometric item difficulty models for item difficulties for different types of reasoning items under realistic conditions. For instance, how robust are item generation frameworks against variation of item surface characteristics? Are item-difficulty models generalizable for test-takers from different cultural backgrounds? Are structurally parallel test forms also parallel in a psychometric sense?
2. Second, the *value of item-generation models for a deeper understanding and improvement of construct validity* of reasoning measures is investigated. For instance, to what extent can item-generation models improve the understanding of what a test measures and how test scores relate to underlying abilities and skills? How can item facets be defined and item features be designed to guarantee definite solutions? Are parameter estimates for item facets truly in line with theoretical assumptions? Can facet-level analyses contribute to an identification of drivers for group performance differences?

From a theoretical point of view, the three studies presented here contribute to the clarification of what reasoning tests measure and how tests can be designed to be consistent with cognitive theories about information processing. From an applied point of view, the studies are pilot studies for the development of fully computerized automatic item generators that are suitable to design large numbers of new test items with sufficiently well predicted item difficulties “on-the-fly” during testing in high-stakes large-scale settings.

1.2. Outline

This thesis is structured as follows. First, a general theoretical background is given, then three studies are presented that address the two general research goals. Study 1 focuses on the construct validation of a strong-theory item-generation framework, study 2 focuses on the question whether structurally parallel test forms are also psychometrically parallel, and study 3 focuses on the cross-cultural validity of an established AIG framework. The thesis closes with an Epilogue that discusses the contribution of the three studies to the two general research goals. Findings of all three studies are discussed with respect to the overall objectives. This includes a discussion of the limitations of the studies presented here and possible future directions.

Study 1 Chapter 3 describes the development of a new reasoning measure, the *Figural Analogy Test* (FAT) that extends earlier research by Beckmann (2008) who developed an analogy measure based on alphanumerical symbols and figural-spatial rules. The current study aims at the development and validation of a purely figural measure that requires no mathematical or verbal abilities. The generative framework is exclusively based on theories of analogical reasoning, specifically research on geometric analogies and spatial ability. The validity of the new item-generative framework is tested in an empirical study with $N = 308$ university students. Two main research questions are addressed, first, the appropriateness of the set of pre-specified item radicals to model item difficulties in terms of a reliable prediction of difficulty parameters. Second, it is tested whether the parameter estimates of the item-difficulty model are in line with assumptions about figural-spatial processing and analogical reasoning. A set of specific hypotheses related to the impact of each of the item radicals manipulated are tested. Several explanatory IRT models are compared, including models with item-predictors only and models with person-by-item interactions. Results show that item difficulties can be predicted based on the new AIG framework. Absolute parameter differences between true and predicted difficulty parameters are, however, considerable and constitute a threat to potential operational application of on-the-fly item generation and estimation. All parameters are in line with theories of figural-spatial reasoning. Gender differences are driven by specific item features. Furthermore, scores on the new test correlate with other established measures of

fluid reasoning and spatial ability and demonstrate incremental validity for the prediction of school grades. Future studies should investigate the generalizability of these results to fully automatically generated FAT items and the feasibility of the item difficulty modeling approach in computerized adaptive testings scenarios.

Study 2 Chapter 4 describes the development of a new item-generative framework for generating number series items. The focus of this study is on the question whether item-generation models can facilitate the construction of structurally and psychometrically parallel test forms. Two main research questions are addressed. First, the appropriateness of the new item-generative framework for the construction of parallel tests; second, whether estimates of the item-difficulty model are in line with findings from cognitive psychology on mathematical processing and numerical reasoning. The validity of the framework, especially for the generation of parallel test forms, is investigated in a study with $N = 406$ university students. Virtual item models are applied to test the stability of item parameters across parallel item sets. Warm-up effects are distinguished from true parallel-test effects. Results demonstrate that parallel forms can be constructed based on a generative framework if sources for heterogeneity in item difficulties are carefully controlled. Item difficulty is predominantly determined by the relational complexity of two consecutive numbers. Complexity levels could be manipulated considerably by combining a set of relatively simple arithmetic rules requiring only addition and subtraction. LLTM modeling results show that item difficulties could be well explained by underlying radicals when both arithmetic rules and their combination principles were included as item predictor variables. The item-generative framework was shown to be relatively robust against irrelevant surface patterns in the number of a series caused by random incidentals. After a warm-up run, item difficulties could be predicted very reliably for two parallel test forms. Correlations with a general reasoning measure and maths grades further confirmed the criterion-related validity of the new instrument.

Study 3 Chapter 5 investigates the cross-cultural validity of the Latin Square Task (LST), a figural reasoning measure that can be generated based on a set of item-generative rules. Performance differences on reasoning measures in cross-cultural settings are a well documented finding, but still only little is known about the bias-generating processes on the item-level. Two research questions are addressed. First, it is asked whether relational complexity theory is a cross-culturally valid framework to generate figural reasoning items. Second, it is investigated whether item difficulties are comparable across countries or whether bias on the item level (i.e., Differential Item Functioning) and on the facet level (i.e., Differential Facet Functioning; see chapter 2.5 for a definition of DIF) is present. Qualitative analyses of DIF versus non-DIF items were conducted to achieve a better understanding of the generating processes for DIF in the LST. Cultural background was investigated in a broad sense by comparing students from two countries representing tra-

ditionally individualistic (Germany, $N = 452$) versus collectivistic (Russia, $N = 201$) cultures. Countries of medium cultural distance and moderate differences in school systems and educational expenditures per child were chosen. Additionally, performance on the LST dependent on the (non-)existence of test-specific pre-knowledge was investigated. Knowledge of the number-placement game SUDOKU was assessed as a proxy of relevant pre-knowledge that might facilitate LST performance. Results confirm the cross-cultural validity of the LST in a broad sense but also point to problems with the functioning of individual items in a cross-cultural context. Item surface characteristics could be identified that contribute to the emergence of DIF and should be controlled in future studies or applications of the LST.

Common themes and differences between the three studies presented While studies 1 and 2 present the development of new rule-based item generation frameworks, study 3 represents an application of a previously developed and validated item type to a cross-cultural context. The three studies all address the two overall research goals, but their focus is different. Study one is designed to develop and validate an item generation model for figural reasoning items, study two investigates the feasibility of generating parallel test forms of number series items based on a newly proposed generation model. Study three investigates differential item functioning and possible factors for the emergence of DIF in a cross-cultural setting for an established figural reasoning measure with an existing underlying cognitive item model.

The common theme of the three studies is the use of explanatory IRT models to investigate students' performance on non-verbal reasoning tests. In all three studies, the population of interest is university students. Test-takers received comprehensive instructions of all item-generative rules at the beginning of the testing session and a set of warm-up items was administered prior to actual test items. Explanations of rules and warm-up items were included based on positive impact on test validity reported in the literature. For instance, Anastasi (1981) recommended to implement short orientation and warm-up sessions to establish comparable testing conditions for all subjects. Warm-up items allow subjects to learn the correct solution strategies and then utilize them on subsequent problems (Verguts & De Boeck, 2002). Beckmann (2008) demonstrated that an explanation of all rules to the test-takers can actually increase the validity of abstract reasoning items. While all studies involve warm-up runs their length differs across studies with study 2 containing a much longer warm-up run than studies one and three. The fact that only few warm-up items were used in studies 1 and 3 constitutes a major limitation for these studies.

DIF across culturally heterogeneous populations is only investigated in study 3. The first two studies aim at the development of two new item types and a first validation of the internal cognitive structure of these items. Integrating psychometric test models with psychological theories and item-generation frameworks is a necessary step towards a

Table 1.1.

List of abbreviations frequently used in this thesis

Abbreviation	Spelled-out form
AIC	Akaike Information Criterion
AIG	Automatic item generation
BD	Beslow-Day
BIC	Bayesian Information Criterion
CFT	Culture Fair Test
DFE	Differential Facet Functioning
DIF	Differential Item Functioning
FAT	Figural Analogy Test
ICC	Item Characteristic Curve
ICM	Item Cloning Model
IDM	Item Difficulty Modeling
IRT	Item Response Theory
LLTM	Linear Logistic Test Model
LR-LLTM	Latent Regression LLTM
LST	Latin Square Task
MH	Mantel-Haenszel
NST	Number Series Test
RCT	Relational Complexity Theory
RM	Rasch Model
3DW	Three-dimensional cube test

better understanding of the construct validity of the new item types. While it is necessary to test the cross-cultural fairness of both the FAT and the NST before items could be administered operationally in cross-national settings, this goes beyond the goals of this thesis. Study 3, though, illustrates based on an established reasoning measure how facet level analyses might benefit the detection and explanation of item-by-country effects (i.e., DIF). Future studies are needed to cross-validate these findings and to investigate the cross-cultural validity of the two new items types described in studies one and two.

Table 1.1 provides a list of abbreviations frequently used throughout the following chapters.

2

Theoretical background

This chapter provides the reader with the general background that constitutes the theoretical basis necessary for understanding the three studies that are presented in the consecutive chapters.

2.1. Reasoning ability as a construct

The ability to make inferences based on the processes of inductive and deductive reasoning constitutes a core part of thinking. It has been a main theme of philosophical inquiry ever since the beginning of scientific endeavor. Thinking and underlying processes have been studied by philosophers, by cognitive and experimental psychologists as well as biologists and neuroscientists yielding complex theories of human information processing. Carroll (1993) has defined cognitive abilities very broadly as abilities “that concern some class of cognitive tasks” (p. 10). Gottfredson (1997) defined intelligence as “the very general mental capacity that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience.” (p. 13). Guilford (1985) defined intelligence as “a systematic collection of abilities or functions for processing information of different kinds in various form” (p. 231). The identification and application of rules through induction and deduction is viewed as a central component of almost all models of human intelligence. *Reasoning* is emphasized as a prerequisite for learning and problem solving (see e.g., Sternberg, 1984; Snow, Federico, & Montague, 1980). Since Binet’s (1903) first introduction of the concept

of intelligence mainly as an ability to adapt to novel situations, reasoning ability has a central place in all prominent theories of the structure of intelligence (see e.g., Wilhelm, 2005, for an overview).

Different models and different researchers have used slightly different labels to refer to a largely equivalent ability construct. For instance, Carpenter, Just, and Shell (1990) used the term *analytic intelligence* to refer to the “ability to reason and solve problems involving new information, without relying extensively on an explicit base of declarative knowledge derived from either schooling or previous experience” (p. 404). This definition is almost completely equivalent to what others have labelled *fluid intelligence* or g_f (Cattell, 1971). Fluid intelligence has been shown to be the best predictor of general intelligence g , which has been defined as the ability of the “education of relations and correlates” (Spearman, 1927, p. 165). Also, the term relational reasoning (see e.g., Crone et al., 2009) has been used extensively, especially cognitive and neurophysiological researchers have used this label for “the ability to consider relationships between multiple mental representations” (Crone et al., 2009, p. 55) instead of referring to fluid intelligence. Relational reasoning is thought to be instrumental in the learning of tasks requiring complex spatial, numerical, or conceptual relations.

Table 2.1 gives an overview of the most prominent structural models of human intelligence and the role of reasoning in these models. A state-of-the-art overview of intelligence research can be found in Wilhelm and Engle (2005). The models in Table 2.1 can be categorized into two groups, that is into single factor (g) theories and variants of multiple-factor theories. The latter account for the fact that, as Stankov (2005) termed it, human minds are “far too complex, and individual differences cannot be adequately accounted for by an overly parsimonious construct” (p. 290). At the same time, most of them do not deny the existence of an overarching general ability component. For instance, the Gf-Gc theory distinguishes components of intelligence while at the same time assuming a higher-order factor g of general reasoning.

Carpenter et al. (1990) stated regarding their analyses of the cognitive processes during working on complex reasoning tasks that “the processes that distinguish among individuals are primarily the ability to induce abstract relations and the ability to dynamically manage a large set of problem-solving goals in working memory.” (p. 404). Working memory as “a system for the temporary holding and manipulation of information during the performance of a range of cognitive tasks” (Baddely, 1986, p. 34) was shown to be very closely related (although not identical) to Reasoning (e.g., Gustafsson & Undheim, 1996; Kyllonen & Christal, 1990).

The reasoning process can be divided into four stages, out of which three stages involve the same processes for inductive and deductive reasoning (Wilhelm, 2005). This process is illustrated in Figure 2.1.

Table 2.1.
Reasoning in prominent structural models of human intelligence

	Author(s)	Intelligence Model	Reasoning in this model
SF	Spearman (1927)	<i>Theory of general intelligence (g)</i>	Reasoning as the best single indicator of g ; measures are highly g -loaded and demonstrate low proportions of specific variance
MF	Thurstone (1938); Thurstone and Thurstone (1941)	<i>Primary Mental Abilities</i>	Reasoning as one of 7 primary factors, no distinction between deductive and inductive reasoning
MF	Cattell (1971); Horn and Cattell (1967)	<i>Gf-Gc Theory</i>	Distinction of reasoning along two dimensions (inductive and deductive, verbal and figural-spatial)
MF	Carroll (1993)	<i>Three-stratum theory</i>	Reasoning constitutes the fluid intelligence (g_f) second-stratum factor; distinction of three components (Sequential/deductive reasoning, Induction, Quantitative Reasoning)
MF	Guilford (1967)	<i>Structure of Intellect (SOI) model</i>	distinction of four reasoning factors (General Reasoning, Thurstone's Induction, Commonalities, Deduction)
MF	Jäger (1982)	<i>Berlin Intelligence Structure Model (BIS)</i>	Reasoning as one of four operation facets; distinction between verbal, quantitative, and spatial reasoning

Note. SF: single-factor model; MF: multiple-factor model

Since the last century, reasoning measures are routinely administered in both educational and employment settings. Reasoning is one of the most relevant psychological construct in the prediction of professional work performance (e.g., Ones et al., 2005; Schmidt & Hunter, 1998). Typical reasoning tasks are figural matrices (e.g., Freund, Hofer, & Holling, 2008), analogies (e.g., Whitely & Schneider, 1981), latin squares (e.g., Birney, Halford, & Andrews, 2006) or number series (e.g., LeFevre & Bisanz, 1986; Quereshi & Seitz, 1993). Test scores can be used to predict important, real-world criteria at a relatively low test administration cost (Domino & Domino, 2006; Jensen, 1998; Kuncel, Hezlett, & Ones, 2001; Schmidt & Hunter, 1998).

Cognitive scientists and neuroscientists have gained a sound description and understanding of rather isolated cognitive processes underlying human perception and action (e.g., Goldman-Rakic, 1995; Gray, Chabris, & Braver, 2003; Kane & Engle, 2002). In contrast, most intelligence tests used in applied settings rely on rather old principles and task types. Many reasoning tests were developed in a very idiosyncratic way. It is hard to tell, if not impossible, what they truly measure. Carpenter et al. (1990), for instance,

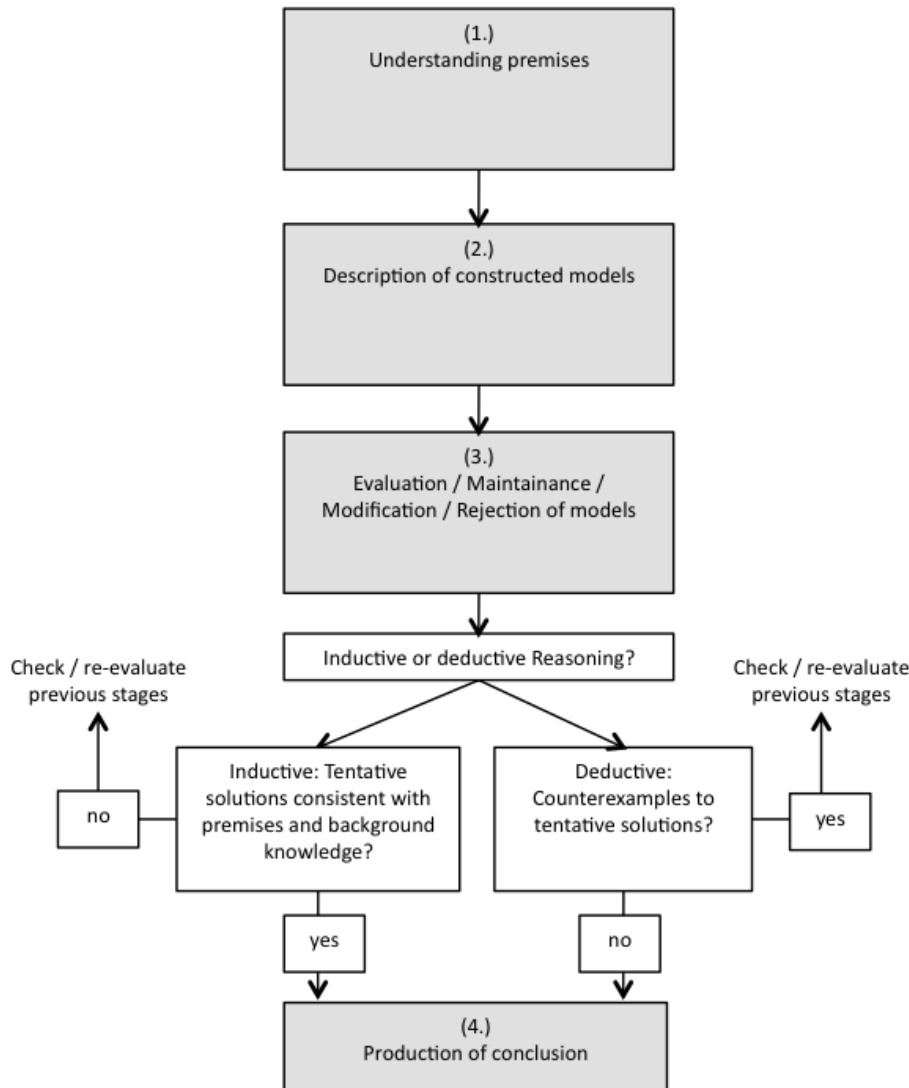


Figure 2.1.

Process model of reasoning based on stages distinguished by Wilhelm (2005)

analyzed Raven’s Advanced Progressive Matrices (APM; Raven, 1962), up to now one of the most used instruments to assess abstract cognitive abilities. They critically investigated the item generation principles applied by Raven, concluding that Raven “used his intuition and clinical experience to rank order the difficulty of the six problem types (...) and the descriptions of the abilities that Raven intended to measure are primarily characteristics of the problems, not specifications of the requisite cognitive processes.” (p.408). Kyllonen and Christal (1990) came to a similar conclusion that, “since Spearman (...) the development of good reasoning tests has been almost an art form, owing more

to empirical trial-and-error than to systematic delineation of the requirements such tests much satisfy” (p. 426). Wilhelm (2005) elaborated further on the distinction between reasoning tasks applied in cognitive experiments and reasoning tasks typically used a parts of assessment batteries, concluding that “there is an enormous gap between theoretically established models of intelligence research and widely used tests of cognitive abilities” (p. ix). Clearly, more research on the underlying cognitive processes of reasoning instruments and test-takers’ performance is needed.

2.2. Cross-cultural validity of reasoning tests

While the practical importance of assessment has grown rapidly during the last decades, the question of construct validity, that is whether and how test results actually represent the intended underlying psychological abilities and skills, is widely ignored in testing practice. This refers especially also to the understanding of the cross-cultural validity or reasoning measures.

Psychological studies that collect data from different countries are often referred to as “cross-cultural” studies. The number of such studies investigating data from culturally different populations has grown rapidly during the last decades and cross-cultural studies have become an integral part of psychological and educational research (e.g., Matsumoto & Yoo, 2006; Organisation for Economic Co-Operation and Development, 2004).

Paralleling the economic processes of globalization and the increasing blurring of national borders in business and educational settings, researches have taken more and more interest in cross-cultural phenomena, including cultural effects in psychological assessment settings. One of the most important questions in the field of testing has been whether test scores obtained in different cultural populations are invariant across cultural borders. That is, whether test scores can be interpreted in the same way for test-takers with different cultural backgrounds.

2.2.1. The term “culture” in cross-cultural studies

For many years, tests developed in Western societies have been applied in developing and emerging countries assuming that the measurement properties of the instruments are identical across the countries or cultures (see Misra, Sahoo, & Puhan, 1997). Different terms have been introduced for abstract, language-free tests that suppose to be equally valid for different cultural groups, with the three most common ones, “culture-free” (Cattell, 1940) “culture-reduced” (Jensen, 1980), and “culture-fair” (Cattell, 1949).

Many of these studies give no explicit definitions of what they conceptualize as culture. It is implicitly assumed that participants from different countries differ also in terms of their “cultural backgrounds”. When samples from different countries are compared,

culture is confounded with society. True experiments are impossible because individuals simply cannot be assigned randomly to different societies and different cultural groups. Still, culture and society are not equivalent (Berry, Poortinga, Segall, & Dasen, 2002a). A contemporary definition of society is that society describes “people who interact in a defined space and share culture” (Macionis & Plummer, 1998, p. 66).

An important difference between culture and society is that culture can be defined in a psychological way, for instance., in terms of attitudes or values. Society is primarily a description of a group of people living close to each other and interacting in their daily lives. Culture emerges from adaptive interactions between humans and environments (Leung & Van de Vijver, 2008).

The first scientific definition of the term culture was given in the 19th century by anthropologist Tylor who defined culture as “that complex whole which includes knowledge, belief, art, morals, laws, customs, and any other capabilities and habits acquired by man as a member of society” (Tylor, 1871, p. 1) Since then, definitions of culture have not changed substantially, though researchers have focused on different aspects of culture and further extended Tylor’s definition. While some definitions focused mostly on behavioral, objective manifestations of culture (e.g., Herkovits, 1948; Kroeber & Kluckhohn, 1952), others have stressed more its psychological, subjective aspects (e.g., Rohner, 1984; Triandis, Bontempo, Villareal, Asai, & Lucca, 1988). In general, definitions can be based on physical(objective) culture (in terms of the human-made part of the environment, e.g. streets, houses, infrastructure, etc.) or psychological (subjective) culture (in terms of shared experiences, social norms, roles, beliefs, and values). Kroeber and Kluckhohn (1952) defined that, “culture consists of patterns, explicit and implicit, of and for behavior acquired and transmitted by symbols, constituting the distinctive achievements of human groups, including their embodiment in artifacts” (p. 181). Triandis (1972) distinguished between physical elements of culture, such as buildings and transportation networks, and subjective elements, such as values and norms. Smith and Bond (1998) gave a broad definition of culture as “a relatively organized system of shared meanings” (p. 39). Fiske (2002) defined culture as “a socially transmitted or socially constructed constellation consisting of such things as practices, competencies, ideas, schemas, symbols, values, norms, institutions, goals, constitutive rules, artifacts, and modifications of the physical environment” (p. 85).

A difference that has been the focus of many cross-cultural studies is the distinction between individualistic and collectivistic cultures. Individualism–Collectivism has been used as a predictor for group differences on many psychological constructs (e.g. Triandis, 1972). Individualistic cultures, traditionally represented by Western European and North-American societies, foster a unique sense of self and autonomy. Clear boundaries between an individual and others are drawn, encouraging the individual to value one’s needs, wishes, and desires over collective concerns. On the contrary, collectivistic cultures, traditionally represented by Eastern societies, teach individuals to value needs, wishes, and

desires of the collective over personal interests and motives. Harmony, cooperation, group cohesion, and conformity are values that play an important role in collectivistic cultures. The process of globalization has made societies increasingly multicultural. Two examples are Asian people living in the U.S. or people from eastern Europe living in Germany. While there is no doubt that these individuals are part of the American, respective German, society, it is also clear that their ways of life still represent in many ways the culture of their home countries. Culture and society are not independent as cultural habits and beliefs change with changes in society (e.g., Leung & Van de Vijver, 2008). For instance, the second or third generation of immigrants in Germany have adopted some of the cultural beliefs and attitudes of their German neighbors, while still holding on to many other parts of their “own” culture. Acculturation research (e.g., Van de Vijver, Helms-Lorenz, & Feltzer, 1999) investigates this process.

Different samples from different countries reflect different cultures and different societies to varying degrees as the cultural heterogeneity and the strength of cultural specifics varies. It is therefore important to be aware of the specific frame of reference of each study when referring to culture. It has also been suggested to replace the global, abstract concept of “culture” with more specific “context variables” (e.g., Matsumoto, 2001). The replacement of the term culture by specific variables can lead to a better specification of measurement approaches and allow to actually test the degree to which cultural differences are related to such context variables. Study 3 incorporates this thinking into the selection of grouping variables. It includes country as a proxy of broad culture and a few further background variables on the individual level to relate cognitive performance to specific context factors.

2.2.2. The challenge of “culture-fair” testing

“Culture-fair” intelligence tests were received very positively, but unfortunately, they evoked similar problems in multi-country assessment settings as other tests not specifically designed for culture-fair assessment. Empirical studies often showed, for instance, that migrant pupils score consistently lower on these tests than native pupils (e.g., Van de Vijver, 1997). In general, meta-analytic findings demonstrate that the largest performance differences appear for tasks that were developed in Western societies based on the values, beliefs and shared knowledge of cultural groups represented by these countries (Van de Vijver, 1997). Whenever multi-cultural samples are investigated, researchers and practitioners have to deal with bias. That is, they have to face the situation that individuals with the same latent ability might be evaluated differently based on the instruments used because the latter favor one or several specific cultural groups over others. On the one hand the factorial structure of human abilities, not just the g factor, is known to be relatively invariant across cultures (Irvine & Berry, 1988); on the other hand, there is multiple evidence that tests that are expected to be culture-fair demonstrate bias in favor

of certain culture-specific groups. Despite its abstract and largely language-free character, the largest cross-cultural differences have been reported in fluid reasoning measures (e.g., Brouwers, Van de Vijver, & van Hemert, 2009; Carroll, 1993; see also Jensen, 1998 or Hartmann, Kruuse, & Nyborg, 2007; Lynn & Owen, 1994; Te Nijenhuis & van der Flier, 2001).

Berry et al. (2002a) distinguished three overarching approaches concerning the relationship between cognitive performance and cultural variables. The three approaches offer different explanations for the observed cross-cultural differences and the interpretation of the manifest score differences.

1. *Absolutistic approaches* assume that test scores can be directly compared between people from different cultures because they capture cognitive processes in an absolute way that is not dependent on the cultural background. Any manifest differences in test scores between different cultural groups reflect true differences in the underlying latent abilities under this approach.
2. *Universalistic approaches* assume that cognitive processes are universal, but that their manifestation is shaped by context-factors. That is, performance on cognitive tests can be seen as a culturally shaped behavior as well (e.g., Lonner, 1980; Segall, Lonner, & Berry, 1998). From this perspective, the distinction between cognitive abilities and cognitive performance (or between intelligence and intelligence scores; cf. Vernon, 1979) is important. Debilitating or facilitating context-factors might cause disparities between true abilities and measured test scores. Some evolutionary psychologists have argued that many cultural practices are environmentally evoked and context-dependent (e.g., Kenrick et al., 2002; Schmitt, 2006; Tooby & Cosmides, 1995). Any manifest differences in test scores between different cultural groups reflect the way cultures shape these universal properties in their own way, and not necessarily true differences in the underlying latent abilities. Cultural variables have, therefore, to be taken into account when cognitive abilities should be assessed in cross-cultural settings. Most modern cross-cultural researchers have adopted the assumptions of the universalistic model (e.g., Hakstian & Vandenberg, 1979; Hennessy & Merrifield, 1976; Irvine, 1969; Irvine & Berry, 1988; Naglieri & Jensen, 1987; Ree & Carretta, 1995; Sung & Dawis, 1981).
3. *Relativistic models* conceptualize all psychological findings as linked to a specific cultural context. Direct comparisons of manifest test scores between cultures are therefore, under this approach, not possible. For instance, Frijda and Jahoda (1966) argued that both the definition of intelligence itself as well as its expression are cultural. Therefore, all cross-cultural comparisons would, by definition, be confounded with cultural influences.

Under the assumption of universal cognitive processes, observed differences in intelligence scores are due to other factors than actual cognitive processing. Proponents of so-called

Bias Models have argued that it is not reasonable to interpret country-related score differences (only) as a manifestation of differences in the values of the underlying construct (see Van de Vijver, 1997). That is, performance differences are not necessarily indicators for differences in basic cognitive processes, but more often shortcomings of the measurement instruments. Differences of country scores on reasoning tasks can be (partly) due to construct-irrelevant factors.

Bias can be defined as the effect of a multitude of factors that can threaten the validity of comparisons between groups with different cultural backgrounds (Van de Vijver & Hambleton, 1996). As such, it is a manifestation of a test's cultural loading in terms of the extent to which the test implicitly or explicitly refers to a particular cultural context. Bias is a sign for country or group differences on variables that influence test performance but are not related to the latent trait that is supposed to be measured. It has been argued that the magnitude of cross-cultural differences is dependent on the nature of the task, specifically on the complexity of the item. Three major types of bias have been distinguished in the literature (see e.g., Van Hemert, Van de Vijver, & Poortinga, 2004; Van de Vijver & Tanzer, 2004).

1. *Construct bias*: An instrument does not measure the same psychological construct in culturally different samples. One example is that the factor structure of a measure is not the same across samples. While a measure might allow for the differentiation of several sub-factors or facets of a construct in one culture, this might be not true for another culture. Also, based on the educational background, a test measuring cognitive abilities in one sample might be merely a reflection of practice and educational training in another sample. Cross-cultural comparisons are seriously limited when construct bias is present. That is why the non-existence of construct bias is a fundamental requirement for any quantitative cross-cultural comparison. It has been shown that many cognitive test batteries fulfill this requirement at least to a sufficient degree (see Berry, Poortinga, Segall, & Dasen, 2002b).
2. *Method bias*: The methodology of a study produces performance differences that do not reflect differences in the underlying latent abilities (Van de Vijver & Leung, 1997). There are two types of method bias, bias related to the instrument and bias related to the test administration procedure. The latter is not directly related to test development and item generation; carefully planned test administration procedures can largely rule out method bias due to test administration. Bias due to the instrument is a more severe problem. For instance, stimulus familiarity can differ across cultural backgrounds (representing the cultural complexity of the tasks; cf. previous section).
3. *Item bias*: Item-specific problems cause performance differences between culturally diverse groups. Typical examples of item bias in verbal items are effects due to inadequate translations or the use of words that are not equally well known in the two cultures. That is, the cultural complexity of the item is high. For figural items,

item-bias can be caused by the usage of specific shapes and forms that represent a certain cultural background. However, the bias-generating processes are less clear for figural than for verbal or numerical items. Item bias directly corresponds to what has been called “Differential Item Functioning” (DIF; see e.g., Holland & Thayer, 1985). Items are said to demonstrate DIF when subjects from different groups but with the same ability level have different probabilities of answering an item correctly. The statistical models that can be used to test DIF were summarized in detail in Chapter 2.5 of this thesis.

Method bias is related to the whole instrument whereas item bias is related to specific test items part of an instrument. The two are related to each other in that method bias is the specific case of item bias where functioning of all items is affected (uniformly) by cultural variables. For instance, test-takers from a specific cultural group could be more familiar with a multiple choice (MC) test format than other test-takers. The MC format is the same for all items and the response format will therefore influence performance on all items. Here, method bias manifests itself on the item level; the major difference is that item bias is related to specific features of individual items. Explanatory models of cross-cultural bias form two categories, correlational studies that relate bias findings to context variables on the group or country level (*post-hoc* approaches; Van de Vijver & Leung, 2000), and models that include context-variables on the individual level to explain cross-cultural performance differences.

Many researchers have investigated the relation between country characteristics and bias on individual test scores (e.g., Blaira, Gamsonb, Thornec, & Baker, 2005; Brouwers et al., 2009; Ceci, 1991; Flynn, 1987; Luria, 1976; Lynn & Vanhanen, 2002; Rindermann, 2007; see Van de Vijver & Tanzer, 2004, for an overview). Indicators tied to the educational systems in different countries have been shown to play an important role for the emergence of cross-cultural differences: several educational variables (e.g., expenditure per capita, teacher qualifications, enrollment into primary, secondary, and tertiary education) were identified as robust predictors of country-level scores on cognitive instruments (e.g., Georgas, Van de Vijver, & Berry, 2004; Van de Vijver, 1997). These indicators quantify the degree to which formal education has shaped society in a given country. For instance, years of schooling and educational expenditure are positively related to performance differences (Ceci, 1999; Gustafsson, 2001; Herrnstein, Nickerson, de Sanchez, & Swets, 1986; Van de Vijver, 1997; Winship & Korenman, 1997). Even for simple mental tasks, national affluence is a successful predictor of cross-cultural performance differences (Van de Vijver, 1997). Schooling broadens the domains in which cognitive skills can be successfully applied. Consequentially, it facilitates cognitive tasks because of training and by exposure to psychological and educational tests. Similar effects were reported for general wealth indicators, such as GDP, the availability of certain technologies and products, or socio-economic status (e.g., Turkheimer, Haley, Waldron, D’Onofrio, & Gottesman,

2003). Cross-cultural performance differences increase with age (cf. *cumulative differences model*, Jensen, 1977) and years of schooling (Van de Vijver, 1997).

2.2.3. Cognitive and cultural complexity as possible factors for bias

The *cognitive complexity model* is based on Spearman's (1923) work. He hypothesized that tasks with a higher cognitive complexity reveal larger cross-cultural score differences. *Cognitive complexity* refers to the complexity of stimulus transformations that are required to arrive at a solution. Scientific support for the cognitive complexity model comes from several studies. For instance, tests that require simple information processing steps typically show smaller cross-cultural differences than tests addressing complex information processing (Vock & Holling, 2008). Despite the idea of culture-fair testing especially of the fluid component of intelligence, the largest cross-cultural differences have been reported in fluid reasoning measures (e.g., Brouwers et al., 2009; Carroll, 1993; see also Jensen, 1998 or Hartmann et al., 2007; Lynn & Owen, 1994; Te Nijenhuis & van der Flier, 2001). Cumulative research findings from studies published between the years 1973 and 1994 support that cross-cultural differences are related to the cognitive complexity of the tasks (Van de Vijver, 1997). Cross-cultural performance differences were positively related to stimulus complexity but not to response complexity in this meta-analysis.

Several theories of cognitive complexity of reasoning items have been suggested. Some approaches are purely empirical and define complexity in a post-hoc way based on empirical data. Other approaches are more closely linked to cognitive theories.

- *Purely empirical approaches:* Cognitive complexity of a given task can be defined by correlating performance on the task with general intelligence g . The larger the correlation between these two variables, the higher the cognitive complexity of the task. This approach requires no prior theory about complexity generating factors. Examples for studies using this strategy to define bias are the works by Marshalek, Lohman, and Snow (1983), or Spilisbury, Stankov, and Roberts (1990). Others (e.g., Vernon & Jensen, 1984) defined complexity based on the response time required to solve a task. From this perspective, tasks that pose higher demands in terms of response time are more complex. Both approaches are purely empirical in that they establish complexity in a post-hoc fashion without a prior cognitive theory.
- *Approaches based on cognitive theories:* Two directions of approaches can be distinguished, approaches that focus purely on the memory load of a task and approaches that focus on the relational complexity of a task. Both approaches are related to working memory theories and assume that reasoning ability is limited by working memory (WM, e.g., Just & Carpenter, 1992); their focus, though, is different. The former focuses on the sheer number of distinct elements or element relations in a

task (Carpenter et al., 1990; Holzman, Pellegrino, & Glaser, 1983; Primi, 2001): more difficult items in a psychometric test require more WM capacity because more elements have to be stored and manipulated simultaneously. A drawback of this definition is that the sheer number of elements might be not sufficient to determine cognitive complexity. Elements and element relations themselves can vary in their complexity. Chunking and other processing strategies can further influence the complexity of distinct elements of a task. The depth of processing required to solve a task is not considered. Simply increasing storage demands of an otherwise simple task will produce higher task difficulty, but not necessarily higher cognitive complexity. Halford, Wilson, and Phillips (1998) argued that it is not the amount of information per se, but the complexity between the pieces of information that have to be processed which is subject to capacity limitations. This complexity is called relational complexity. Halford et al.'s approach is known as *Relational Complexity Theory* (RC). It defines cognitive complexity in cognitive tasks independent of the domain of the task. The validity of the RC approach could be demonstrated in multiple studies, for instance, for the prediction of the difficulty of deductive reasoning tasks (Birney et al., 2006; Lee, Goodwin, & Johnson-Laird, 2008). RC plays a role in cognitive development (Andrews & Halford, 2002), logical reasoning in adults (Birney & Halford, 2002), and applied areas such as mathematics literacy (English & Halford, 1995). The RC-Theory approach is especially valuable for the investigation of cross-cultural bias because it provides a basis for an empirical test of the Cognitive Complexity assumption for rule-based generated cognitive items.

More complex tasks are more prone to cultural influences because they rely more strongly on (culturally) acquired knowledge and skills (e.g., Jensen, 1998). Cultural knowledge that is required to master a test can be declarative as well as procedural. Tests are highly *culturally complex* if the variation in familiarity with the type and content of the test between different cultural groups is high (e.g., if specific item types are used as training material in schools in one culture, or if a specific cognitive games are popular especially in one cultural group). An example was given by Demetriou et al. (2005) who report large differences in complex visuo-spatial tasks between Greek and Chinese children. In their study, Chinese children clearly outperformed Greek children. The finding can be related to the massive visuo-spatial practice Chinese children receive when they learn to write Chinese. Meta-analytic results support the importance of cultural task complexity for performance in international studies: Tasks characteristics (e.g. cognitive complexity) were the most powerful predictor for performance in national studies (i.e., studies with samples from only one country). On the contrary, performance differences in international studies could be better predicted by individual and group characteristics (Van de Vijver, 1997). Helms-Lorenz, Van de Vijver, and Poortinga (2003) showed that performance differences between majority-group members and migrant students and students without migration background were better predicted by cultural complexity (c) of a test than by the cognitive complexity (g) of the measure. This would mean that the

relational complexity of the cognitive operations required to solve a reasoning item should be a good predictor for item difficulties in culturally homogeneous, but not in culturally diverse samples.

2.3. Item difficulty modeling and rule-based automatic item generation

Towards the end of the 20th century researchers have begun to try to relate psychometric intelligence factors to ability constructs identified by tasks from experimental cognitive psychology (e.g., Engle, Tuholski, Laughlin, & Conway, 1999; Freund et al., 2008; Hambrick & Engle, 2002; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). In some areas, these approaches have yielded very useful results, demonstrating, for instance, strong associations between working memory and fluid intelligence (e.g., Ackerman, Beier, & Boyle, 2005; Kyllonen & Christal, 1990; Süß et al., 2002). But these models are only rarely used in assessment practice. One reason for this might be that these models are confirmatory in the sense that they require strong theory in item generation and model specification (cf. Embretson & Schmidt-McCollam, 2000). Many tests at use lack strong theories that enable an empirical test of the functioning of underlying item properties. Speaking with Deary (2001), “linking mental test scores to cognitive variables is only really productive when the cognitive variables are themselves theoretically traceable. Otherwise one has merely linked an unknown to another unknown.” (p. 167). If test construction is not strictly theory-driven based on a cognitive model of thought processes, the number of mental models needed to solve a specific test item can hardly be determined (Yang & Johnson-Laird, 2001; see also Wilhelm, 2005). Likewise, it is quite possible to find two reasoning tests that suppose to measure the same construct but differ considerably in their features, attributes, and requirements (cf. Wilhelm, 2005). In the same way, reasoning tests that share comparable features, attributes and requirements are not necessarily equivalent in a psychometric sense (see e.g., Porsch, 2007); on the contrary, attempts to construct truly parallel tests have been rather inconclusive. In most testing batteries that are used in practice, parallel test forms are simply identical test forms with a changed item order (e.g., Amthauer, Brocke, Liepmann, & Beauducel, 2001; Weiß, 2007).

Cronbach (1957)’s claim that more research is needed that combines the experimental and differential traditions to explore the relation between basic cognitive mechanisms and intelligence is, unfortunately, still valid today. More theoretically driven work in the development and validation of reasoning measures is one essential stepping stone on the “fruitful avenue to future research on measuring and understanding reasoning ability” (Wilhelm, 2005, p. 388). So far, the literature is still lacking studies describing clear guidelines for generating and criteria for evaluating item models (Gierl & Lai, 2012).

Table 2.2.
Prominent content areas for rule-based automatic item generation

Content Area	Item type	Applications (examples)
General cognitive abilities	Mental rotation	Bejar, 1990
	Spatial sense	Gittler, 1990
	Abstract reasoning	Embretson, 1999
	Numerical flexibility	Arendasy, Sommer, & Hergovich, 2007
	Figural matrices	Freund et al., 2008; Arendasy, 2005
	Figural analogies	Beckmann, 2008
	Object Assembly	Embretson & Gorin, 2001
Quantitative Skills	Mathematical word problems	Enright, Morley, & Sheehan, 2002; Holling, Bertling, & Zeuch, 2009; Zeuch, Geerlings, Holling, Van der Linden, & Bertling, 2010
	Quantitative comparison problems	Bejar et al., 2002
	Mathematics tasks	Singley & Bennett, 2002
Verbal Skills	Reading comprehension	Sonnleitner, 2008
	Paragraph Comprehension	Embretson & Gorin, 2001
Other innovative constructs	Traffic risk behavior	Arendasy, Hergovich, Sommer, & Bogner, 2005

Integrating psychometric test models with psychological theories is a first important step towards a better understanding of from where item responses to the various kind of reasoning items truly originate. As such, it is an important step to approach Hunt (1976)’s claim that “the psychology of intelligence must be a part of the psychology of cognition” (p. 257). In other words, knowing which item properties trigger specific cognitive processes as specified in an underlying theory can contribute significantly to the establishment of construct validity.

Besides (and probably more important from a practical point of view), a better understanding of important thought processes underlying specific types of reasoning tasks opens the avenue for a more efficient generation of test items of anticipated difficulties. Approaches of *Rule-based Automatic Item Generation* (AIG; see Irvine and Kyllonen (2002) for an early and Gierl and Haladyna (2012) for a recent overview of the field) in combination with the use of *explanatory Item Response Theory* (explanatory IRT, De Boeck & Wilson, 2004a) put cognitive theories on a testable fundament. Actual performance on reasoning measures can be related to specific underlying cognitive processes.

AIG terms the typically computer based generation of test items based on a-priori defined algorithms. There is a rapidly growing research tradition that created numerous applications with regard to the measurement of cognitive (e.g., Arendasy et al., 2007; Embretson, 1999; Freund et al., 2008) as well as non-cognitive and educational (e.g., Arendasy et al., 2005; Holling et al., 2009) constructs (see Table 2.2 for an overview). Further, AIG becomes more important as computerized and web-based item delivery create new challenges for exposure control (Bejar, 2012).

Figure 2.2 illustrates the four main steps of the general AIG process. Defining a new item-generative framework starts with the definition of the latent construct to be measured: the cognitive processes, solution strategies and knowledge structures that characterize the latent construct have to be identified. Based on knowledge of cognitive psychology principles, a cognitive model has to be developed. Components of items that influence item complexity and difficulty are analyzed. Then, these features can be combined to generate items. Components which are crucial for the solution process predominantly influence item difficulty. They should be well-defined for item design, item generation and item application. These item features are called “*radicals*” (Irvine & Kyllonen, 2002). Radicals systematically affect the difficulty of an item; they determine the cognitive processes needed to solve the items. “*Incidentals*” (Irvine & Kyllonen, 2002) do not influence the difficulty of a task. Incidentals are item characteristics that cause only surface differences in the appearance of the items. They make psychometrically equivalent items *look* different. Once items have been assembled into tests, hypotheses with regard to the pre-specified item facets can be tested empirically. Embretson and Gorin (2001) described principles of evaluating cognitive models for rule-based generated item sets by predicting item performance. Item difficulties or other response variables such as response times can be regressed on the item structures and stimulus features that were chosen to operationalize the relevant cognitive processes. A detailed summary of the statistical models that can be used here will be given in chapter 2.4 of this thesis. If the empirical test of the underlying generation model has been successful, new items of pre-specified complexity levels can be generated by the combination of item facets. Note, the internal procedure of testing the construct-validity of tests through the modeling of cognitive item structures does not render the validation of new measures by investigating external correlates of test scores unnecessary (Embretson & Gorin, 2001).

With regard to a theory-driven development of new reasoning measures in specific, Wilhelm (2005) has described four key aspects that need to be considered during the construction and evaluation of new instruments. These are also important aspects for the development of new item-generation models.

1. *Operation*: First, the operational requirements of the new tests need to be defined. It has to be answered what the cognitive operations that must be mastered in order to solve items of the new test are. For reasoning tests, possible requirements mentioned by Wilhelm (2005) are the inductive creation of semantic information,

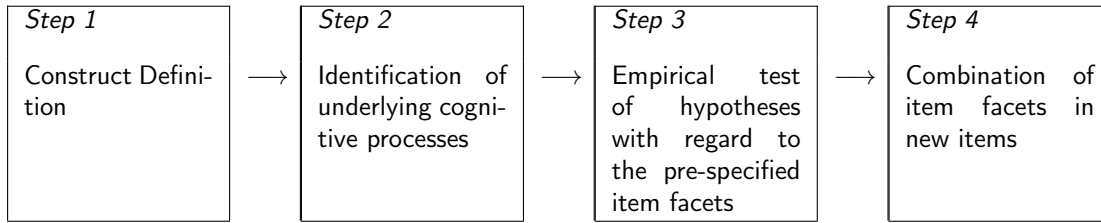


Figure 2.2.

Four essential steps of the automatic item generation process

deductive maintenance of semantic information, derivation of inferences, judgment, decision making, and planning.

2. *Content*: Second, the actual item content needs to be specified. This involves a decision about the inclusion of figural, numeric and/or verbal material. Depending on the type of material chosen, a new test will focus on different aspects of the reasoning construct. Experimental manipulations of the item content can also add to the understanding of the structure of reasoning ability.
3. *Instantiation and non-reasoning requirements*: Third, a decision has to be made about how the formal reasoning process should be initiated for the test-taker. Wilhelm (2005) described this decision as going “through a decision tree” (p. 380). A first decision is whether to use concrete or abstract material. For example, a figural reasoning test can make use of concrete shapes relating to numeric and/or alphabetical knowledge as in the analogy test presented by Beckmann (2008) or apply abstract shapes and forms without that semantic component as, for instance, in the Culture-Fair Test (CFT-20R; Weiß, 2007). When abstract material is used, “nonsense” and “variable” instantiations can be distinguished. Nonsense instantiations present a logical connection between abstract (“nonsense”) stimuli (e.g., phantasy words in a verbal task or abstract shapes in a figural task) whereas variable instantiations induce the formal reasoning process by referring to variables (e.g., “X” and “Y”). When concrete material is used, instantiations of reasoning problems can be in accordance with prior knowledge (e.g., factual or possible instantiations; see Wilhelm, 2005, for examples) or in contradiction to it (e.g., counterfactual or impossible instantiations; see Wilhelm, 2005, for examples). The explicit specification of the instantiation of the reasoning process is of high importance for the development of new AIG frameworks. Structurally identical reasoning items can differ in their cognitive demand and consequentially in their psychometric properties when different forms of instantiations are applied (e.g., Beckmann, 2008; Gilinsky & Judd, 1993; Klauer, Musch, & Naumer, 2000). In the terminology of Irvine and Kyllonen (2002) the form of instantiation would take the role of an item radical here. This means that instantiations should either be held constant across a set of items (in order to minimize the amount of variance in item difficulties not explained by

the underlying task parameters) or included explicitly in the rule-based generation framework. In the latter case, the cognitive model would explicitly predict differences in item difficulties for different types of instantiations. The same is true for the use of abstract versus concrete material.

4. *Vulnerability to reasoning strategies*: Forth, the degree to which reasoning items are vulnerable to the use of reasoning strategies can diminish the equivalence of structurally identical items. The use of most existing reasoning measures is based on the assumption that all individuals approach the problems in the same way. If this is not the case, the diagnostic value of the test can be diminished. It might be the case that some test-takers are more successful not because their reasoning ability exceeds that of other test-takers, but because they are more familiar with the test material or had more extensive practice to prepare for the testing situation. Also, test-takers with different cultural backgrounds might approach identical reasoning problems using different, and possibly differentially effective, strategies. The consequence would be that the test is measuring different abilities for different subgroups which is equivalent to “Differential item functioning” (DIF; see e.g., Holland & Thayer, 1985). If such interactions of person and item or test characteristics are not modeled explicitly, resulting biases can seriously diminish the construct and predictive validity of a test.

The situations where AIG is beneficial for test developers and administrators are manifold. Mostly, these are situations where several comparable test forms are needed. For instance, when tests are administered in high stakes settings or test-takers are allowed to take tests multiple times, parallel forms are needed to assure test security. One common strategy to deal with the problem of item exposure (i.e., the familiarity of subjects with item content) is to use item subsets from a large pool of calibrated items. Each test taker then gets a different subset of items. It is assumed that the pool of items is large enough to control for item exposure effects. However, the creation of such an item pool is both costly and time intensive. The processes involved, such as the writing, reviewing, pretesting and calibration of individual items are practical constraints that can interfere with the development of large item pools. After all, there is no guarantee that large item pools prevent item exposure problems; they just make it harder, but not impossible, for test takers to get to know item content from a test. Three important arguments why AIG procedures should be used for the generation of new measures can be outlined.

AIG can makes item-generation processes more efficient. By its necessity to formulate item design principles in an algorithmic way, AIG builds the basis for the construction of items of equal structure and quality. Computers can generate new items within milliseconds. Furthermore algorithms can control for possible alternative solutions and guarantee that every item has only one single right solution. The results of such algorithms exceed the power of manual inspection and control by human item developers. Freund et al.

(2008) demonstrated for the classical type of matrix items that even experienced item writers cannot overcome their idiosyncratic styles during the process of item design. For instance the selection of figural elements and their arrangement is traditionally based on an item developer's personal taste to make items look smooth and aesthetic. Independent item developers with the same level of expertise might be able to construct similar looking items without too much effort. Yet, it is possible that these items have completely different statistical properties if the underlying cognitive processes needed to solve an item are not identical (Freund et al., 2008).

Proponents of manual item construction might argue that AIG makes test items look technical and boring for the individuals taking the test. Yet, this can be true for any instrument; the difference between traditional item generation and AIG is that the creative process of item development is moved to another phase during the development of new instruments. Traditional item generation involves creative input on the item level. AIG builds upon an item generation framework that relies on human item writers and subject matter experts. Expertise and creative ideas are needed to develop such frameworks to the same extent as during the development of individual test items. Algorithms and automatization come in only once the framework is established. AIG makes the process of item generation more effective. New items can be generated faster, sparser, and with fewer construction errors. This can improve both objectivity and reliability.

This reasoning represents a technical perspective on rule-based item generation that has been taken by many researchers and test-developers. For instance, Lai, Alves, and Gierl (2009) described cost benefits, enhanced test security and decreased item exposure, and a more accurate estimation of examinee ability as the three main reasons why AIG should be implemented. This perspective focuses primarily on enhanced efficiency and increased security of the test design and assessment process, not on questions related to the construct validity of the generated items.

AIG can help establishing construct validity. Traditional item descriptions are mostly related to item content or overall characteristics. However, they contribute only marginally to an understanding of what test performance says about the underlying cognitive processes and the problem solving capabilities and knowledge of the test-taker.

Rule-based item generation makes it possible to carry over classical experimental approaches to the generation of diagnostic instruments. These instruments cannot only be used in applied settings for the measurement of specific traits. They can also build up a basis for the test of cognitive theories. Embretson (1983, 1998) has described this perspective in her "*Cognitive Design Systems*" (CDS) approach. She describes the sequence of steps that are necessary to link cognitive theories with adequate test and measurement models. Her framework constitutes the basis for item design and is related primarily to construct validity. Both classical procedures (i.e., a-posteriori estimated, correlative relationships between one instrument and others or external criteria; labelled "Construct

representation”, cf. also Cronbach & Meehl, 1955) as well as Item Response Theory (IRT) approaches to test construct validity based on item characteristics underlying test performance (labelled “Nomothetic span”; Embretson, 1983) are part of Embretson’s framework. Item characteristics and components identified based on findings from cognitive science can be integrated into a theory-driven rule-based item-generative framework. Test results can be used to test hypotheses with regard to the construction rationale and the underlying cognitive model. Rule-based AIG might then serve “to build construct representation directly into the item construction process by combining research in cognitive psychology, individual differences and applied psychometrics.” (Arendasy et al., 2007, p. 567). If such a model is lacking, rule-based item generation can, at most, be of heuristic value. This is because item explanatory models remain, ultimately, arbitrary in the way they define item difficulty as a combination of underlying component difficulties. Different variants of item explanatory models might explain difficulties equally well with no proof of which model is the “true” model. That said, AIG models can provide a useful basis for evaluating the plausibility of competing cognitive models for a given test or item type. AIG models provide a basis for testing whether empirical data for a given test is compatible with a certain hypothesized model of underlying cognitive processes. They cannot, however, provide unambiguous information on what the “true” cognitive processes that underlie test performance are. As Bejar (2012) notes in a recent chapter on the implications of AIG for test validity, “merely referring to a theory is not sufficient to establish construct representation. We can choose a theory and claim that we have developed a test based on that theory, and further assert that the scores from such a test have specific theoretical attributes as a result. However, it is also necessary to demonstrate that the theoretically expected results are actually observed.” (p. 45).

While acknowledging the practical benefits of such item-generation principles, core of this reasoning is the goal to improve the construct validity of cognitive tests, not primarily the enhanced efficiency and increased security of the test design and assessment process. A sufficient set of theoretical principles to generate items can provide strong support for construct validity of an ability test. As such, “AIG offers a framework for item writing that draws it closer to the scientific approach of experimental stimulus design than does an artistic process” (Gorin & Embretson, 2012; p. 136).

AIG can increase transparency and test-fairness. While the first two reasons presented here have been described in the item-generation literature in large detail, the chances to increase test-fairness through enhanced test transparency have not been discussed as intensively. While the lack of test-security and item exposure are severe threats to traditional tests, exposure to test items takes a less critical role when rule-based item generation principles are used. Item construction based on generative rules provides the opportunity to explain *all* relevant *solution principles* to the test-taker *before* the assessment without making the right solutions to an item item obvious. Individuals can

make themselves familiar with the rules and practice on the kind of items applied before taking the test. The knowledge of all relevant rules (i.e., the distinctive definition of a solution space) guarantees that all solutions are truly unique (cf. Freund et al., 2008; Preckel, 2003). Especially when it comes to assessment of giftedness, explaining the rules beforehand can be of great advantage. Gifted students occasionally tend to produce creative solutions to existing problems that differ from the solutions intended by the test developer. This might lead to coding answers as wrong, and consequentially to errors in the estimation of true abilities. One might argue that explaining rules changes the test from a purely inductive reasoning measure to a measure of more specific and more narrow processes. But Beckmann (2008) demonstrated that an explanation of all rules beforehand actually enhanced also criterion-related validities. More studies should focus on the influence of prior item exposure and practice on the internal cognitive structure of automatically generated test items in order to evaluate the true potential of rule-based item generation for applied settings.

Taken together, AIG can help understand item response processes better by forcing the test developer to explicitly identify item radicals. Compared to manual item construction based on idiosyncratic principles and ideas of the test developer, AIG can speed-up the item generation process and establish a basis for mass-generation of structurally equivalent items. Yet, important questions remain that have to be answered by future research, including questions such as: Do structurally equivalent items actually have the same statistical properties? To what degree can item difficulties be, ultimately, predicted by knowledge of the most important underlying radicals? How many radicals are needed? To what degree are item difficulties influenced by “irrelevant” item features that are supposed to be incidental? Do empirical facet difficulty estimates align with theoretically expected patterns of facet difficulties? Are item generation frameworks valid for cross-cultural assessment as well, or does the cognitive structure of a test vary between culturally diverse populations?

Two classes of statistical models will be described in the following. First, explanatory item response models that will be used in all three studies to predict item difficulties based on the underlying item structures, will be summarized. Second, Differential Item Functioning (DIF) and Differential Facet Functioning (DFF) models that will be used in study three to test the equivalence of item characteristics across samples defined by cultural background or other criteria, will be described.

2.4. Explanatory item response modeling

Traditionally, cognitive psychology did not play a prominent role in construct validation as the meaning of a construct could only be established after a test was developed (Embretson & Gorin, 2001). This situation has changed with the development of rule-

Table 2.3.
Explanatory and descriptive IRT models

“Item-side”	“Person-side”	
	Absence of predictors (<i>descriptive</i>)	Inclusion of person variables (<i>explanatory</i>)
Absence of predictors (<i>descriptive</i>)	RM	LR-RM
Inclusion of item facets (<i>explanatory</i>)	LLTM	LR-LLTM

Note. This classification is based on De Boeck and Wilson (2004a).

based item generation models and related item response models. The term *Explanatory Item Response Modeling* was introduced by De Boeck and Wilson (2004a) to describe how “the domain of item response models (...) can be broadened to emphasize their explanatory uses beyond their standard descriptive uses” (p. vii). De Boeck and Wilson (2004a) use the attribute “explanatory” because their models allow modeling item responses as a function of predictors of various kinds. Responses to individual items can be *explained* by characteristics of the items, of persons, or of combinations of persons and items. Predictors can be either observed or latent, and can be continuous or categorical. Explanatory IRT models are therefore ideally suited to explain difficulties of automatically generated items based on known item radicals: item difficulties can be predicted based on a set of predefined task parameters.

Table 2.3 shows the four types of possible descriptive and explanatory models. Item response models can be exclusively descriptive (“double descriptive”) when neither item- nor person-predictors are included. An example is the Rasch model (RM; Rasch, 1860). The opposite case is that models can be “double explanatory” when both item- and person-predictors are included. That is, variables that explain why different items have different solution probabilities (“item-predictors”) and variables that explain why persons differ in terms of their ability to solve items correctly (“person-predictors”) are included. Models are half explanatory and half descriptive when either item- or person-predictors only are included.

2.4.1. Models with item predictors

The *Linear Logistic Test Model* (LLTM; Fischer, 1973) is descriptive on the person side and explanatory on the item-side. It can be used to predict item difficulties of the Rasch Model (Rasch, 1860) as a linear-combination of a vector of basic task parameters:

$$\eta_{pi} = \theta_p - \sigma_i = \theta_p - \sum_{k=0}^K \beta_k X_{ik}. \quad (2.1)$$

η_{pi} is the so-called *linear predictor* (De Boeck & Wilson, 2004b) of the binary random variable Y_{pi} representing the item response of person p on item i . The Rasch model uses a logistic link function, i.e. $\eta_{pi} = f_{\text{logit}}(\pi_{pi}) = \log(\pi_{pi}/(1 - \pi_{pi}))$. θ and σ denote person abilities and item difficulties, respectively. X_{ik} is a $k \times i$ design matrix that indicates the cognitive task parameters/radicals for every item, plus a constant to make the model identifiable. β denotes the vector of weights for each of these parameters. Item difficulties are determined by is the *number* and the *nature* of the cognitive demands involved. In this regard, the LLTM can be “viewed as formalizations of structural hypotheses regarding the psychological complexity of test items” (Fischer & Formann, 1982, p.397).

Equation 2.1 implies the strong assumption that item difficulties can be predicted perfectly by the underlying item facets. This constraint may be relaxed by including item-related random effects. Adding random effects also to the item side of the model leads to a crossed random-effects model with both person and item random effects (RE-LLTM; R. Janssen, Schepers, & Peres, 2004; Van den Noortgate, De Boeck, & Meulders, 2003):

$$\eta_{pi} = \theta_p - \sum_{k=0}^K \beta_k X_{ik} + \varepsilon_i, \quad (2.2)$$

with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Here, η_{pi} is conditional on θ_p and ε_i . That is, items with an identical configuration have an expected value of σ_i , but random variation is possible, and captured by the variance σ_ε^2 . σ_ε^2 represents the residual variance when regressing item difficulties in the Rasch model on the linear predictors specified in X_{ik} . Variability in item difficulties can be ascribed to surface characteristics of the items, with effects being random across items belonging to the same item group (Rijmen & De Boeck, 2002). The RE-LLTM is especially helpful in assessing the explanatory power of the cognitive model under investigation (De Boeck, 2008). By modeling random effects on the item side of the model that account for the variance in item difficulties not explained by the set of basic parameters, different explanatory models can be compared. The amount of random item variation that can be accounted for by the predictor variables included in the model indicates the explanatory power of the model. In the simple Rasch model, this random effect variance equals zero because one parameter for every item is estimated.

In most tests, the processes needed to solve reasoning items are not compensatory. Items cannot be solved if a test-taker does not possess minimum values of *all* abilities needed. Strictly speaking, this violates the additivity assumption of the LLTM. However, LLTM models have brought forward many applications, especially in the assessment of cognitive constructs (e.g., Arendasy et al., 2007; Embretson, 1999; Enright et al., 2002; Freund

et al., 2008;Holling et al., 2009). LLTMs have been considered theoretical useful to link test development with cognitive theories and models of human information processing (cf. Embretson, 1998). Yet, accuracies of item parameter prediction have so far mostly not met the demands of high-stakes assessment settings. Typical values of proportional error reduction in the prediction of item difficulties based on explanatory IRT models reported in previous studies lie in a range from $R^2 = .50$ to $R^2 = .80$ (e.g., Freund et al., 2008; Preckel, 2003). Arendasy (2005) suggested that sufficient construct representation is only given, when the underlying item facets explain at least $R^2 = .80$ of the variation in Rasch item difficulties. Zeuch (2011) demonstrated that even such high values of proportional error reduction should only with great caution be interpreted as sufficient construct representation. She analyzed standardized absolute differences between rescaled LLTM and Rasch parameters for the Latin Square Task, a rule-based generated figural instrument. Even though roughly 87 percent of the whole variance in item difficulties were explained by the basic design parameters in a sample of $N = 850$ examinees, the average standardized absolute difference denoted to 3.12 standard error units (Zeuch, 2011, p. 51). Bejar's (1993) idea of fully functional item generators that can generate item isomorphs of predicted difficulties "on-the-fly" during test administration is still a vision. On the other hand, it is without doubt that applications of the LLTM to cognitive instruments can contribute to a better understanding of the construct validity of these instruments. It is important to distinguish between the different purposes of statistical item difficulty modeling: one purpose is a better understanding of item response processes and a clarification of the questions of construct validity. This has been the focus of most of the aforementioned LLTM applications. A second purpose is an enhanced efficiency of testing by using item-generative models and predicting item difficulties instead of calibrating individual items. The limitations of the LLTM described in this paragraph relate mostly to the second purpose. As shown by Zeuch (2011) most applications so far have not been successful in predicting item difficulties sufficiently well. On the other hand, others (e.g. Freund et al., 2008) have shown that deviations in the prediction of true item difficulties do *not* influence the estimation of person parameters dramatically.

Item cloning (e.g., Bejar, 1993; Glas & Van der Linden, 2003) terms the idea that certain item types are duplicated to produce a theoretically unlimited number of equally difficult item clones. Item "families" are defined by particular demands on information processing. Item "clones" (also denoted as item "siblings", see e.g., Sinharay, Johnson, & Williamson, 2003) result from changing surface characteristics for items belonging to one particular family. If the demands on information processing are captured well by the item-generative framework (i.e., by the specified radicals) all items from one family should have the same psychometric characteristics. A test taker's reaction should not depend on the incidentals of items with the same cognitive demands. Instead of modeling item difficulties as a linear combination of item radicals (Irvine & Kyllonen, 2002), item cloning models estimate means and variances for item families each consisting of a number of item clones all sharing the same characteristics. Within-family variances indicate the

variation among items that should, theoretically, be interchangeable exemplars of the same item type (“The more efficient the item-cloning techniques are, the smaller the amount of within-family item variability is and the better the test adapts to the examinee’s ability level.”; Glas & Van der Linden, 2003, p. 260).

The term item cloning is strongly connected to statistical item cloning models (ICMs) that were developed by Wim Van der Linden and his group (e.g., Glas & Van der Linden, 2003). However, first item cloning approaches were already presented a few decades ago. Overviews of item cloning techniques can be found in Roid and Haladyna (1981) and Bejar (1993). Sinharay and colleagues (e.g., Sinharay et al., 2003; M. S. Johnson & Sinharay, 2003) distinguished three different item cloning models.

1. In the *Unrelated Siblings Model* (USM) a separate, unrelated item response function for all items is assumed. Family membership is not accounted for statistically. M. S. Johnson and Sinharay (2003) introduced this model as the “gold standard approach for modeling item response functions” (p. 3). Each item is assumed to be independent of all other items, regardless of whether they stem from the same family or not. A notable disadvantage of this model is, however, that each item has to be individually calibrated. The relationship between clones (i.e., the inherent multilevel structure of the data) is ignored. This can enlarge standard errors of item parameters and require large sample sizes to reach sufficient calibration precision (M. S. Johnson & Sinharay, 2003).
2. In the *Identical Siblings Model* (ISM) the same item response function is assumed for all items that belong to the same family. Here, the statistical model incorporates the family structure of shared radicals between items while ignoring possible within-family variation. If this variation is larger than zero, the ISM provides biased estimates of the item parameters, the amount of item information is usually overestimated (M. S. Johnson & Sinharay, 2003). The assumptions of the ISM are very similar to the classical LLTM without random effects: in the LLTM, it is assumed that the configuration of item radicals as specified in the design matrix fully determines an item’s difficulty. Variation in item difficulties across items that share the same vector of item radicals is not accounted for in the model. Only when random effects are defined on the item-level is this assumption relaxed (R. Janssen et al., 2004; see also Equation 2.2 in this thesis).
3. The *Related Siblings Model* (RSM) is a hierarchical model that assumes both separate response functions for each item or item clone and, on a higher level, a hierarchical component for each family. The *Family Expected Response Function* (FERF, Sinharay et al., 2003) for an item family describes the probability of correct response for a randomly selected item from this family for an individual with an ability parameter of $\theta = 0$. In this model, both variation within item families and variation between item families is accounted for. Item families can be calibrated without the

unrealistic assumption that difficulties for all items from one family are the same. The USM and ISM and restricted cases of the RSM.

Glas and Van der Linden (2003) and Geerlings, Glas, and Van der Linden (2011) presented item cloning models that are able to capture the special features of the RSM approach. The two-level item cloning model (ICM) presented by Glas and Van der Linden (2003) is the most general model. On level 1, parameters are defined based on a three-parameter logistic (3PL) model. On level 2, variation within item families is captured by a distribution. Both persons and items are random with the common assumption that person parameters stem from a standard normal distribution, and the vector of the item parameters being drawn from a multivariate normal distribution. As pointed out by Zeuch (2011), cognitive demands on information processing can, in a special case, also be represented by basic parameters or radicals and make “the validation process more straightforward as item difficulties within item families can then be ascribed to certain basic parameters and their combinations” (p. 11). Rule-based generated items with the same combination of basic parameters can be considered item clones belonging to the same family (Zeuch, 2011). Geerlings et al. (2011) presented an extended item cloning model, the item cloning linear model (ICLM). The ICLM includes a design matrix with item radicals. The entries of this design matrix determine the cognitive complexity of each item based on an additive function of stimulus features. That is, the ICLM combines the LLTM and the ICM by conceptualizing not only item families but by referring to the underlying rules (i.e., item radicals specified in a design matrix) that define these families.

Glas and Van der Linden illustrated the accuracy of a Bayesian ICM procedure by means of a large simulation study using real item difficulties from a large item pool. Unfortunately, comprehensive software packages to apply ICMs and ICLMs are still missing.

Zeuch (2011) showed that also the LLTM and RM are restricted cases of a general ICM. LLTMs model item difficulties as a linear combination of indicator variables specified in a design matrix. Typically, each column in the design matrix represents one item radical. Items sharing identical configurations of basic parameters differ only with regard to the values of their incidentals and can be considered item clones (Zeuch, 2011). In line with this a number of LLTM models will be applied to investigate item-generation and item-cloning principles for two new and one existing instrument. The classical LLTM estimates parameters on the level of item-facets. If the interest is not in predicting item difficulties based on radicals (i.e., item facets) but based on the difficulties of certain item types (i.e., item families), the design matrix can be changed to include one column for each item family or item type. As long as there are more item types than radicals (which should be the case in most, if not all, applications), such a model will contain a larger number of parameters. The additivity of basic parameters assumption of the classical LLTM is not needed here. Explanatory IRT models of that type estimate parameters on the level of item families, respectively item types:

$$\eta_{pi} = \theta_p - \sum_{f=1}^F \beta_f X_{if}. \quad (2.3)$$

Equation 2.3 is identical to Equation 2.1 besides the index f that replaced the index k . Here, f refers to the item family underlying the respective item.

Changes in item difficulties across two or more sets of (structurally) identical items can easily be modeled by adding specific fixed effects to the model. If k items are presented at at least two different times, Fischer (1995) suggested to specify all item \times time point combinations as $k \times t$ “virtual items” (p. 158), that is as interaction effects in X_{ik} . With only two measurement points ($s = 1, 2$), item difficulties for one “pair” of virtual items will be σ_{1i} and σ_{2i} . For the first measurement point $\sigma_{1i} = \sigma_i$, whereas for the second $\sigma_{2i} = \sigma_i + \tau_s$. In this linear combination the σ_{si} -parameters are composed additively by means of an initial item parameter σ_i and fixed temporal effects τ_s . In Fischer’s virtual item model, only one temporal effect is modeled for all items. Modeling such an effect would mean that all item difficulties from a second item set differ in the same way from the first item set. As Mair and Hatzinger (2007) illustrate, this concept extends to an arbitrary number of time points or testing occasions. Instead of modeling one general temporal effect, item specific effects can be estimated when temporal effects are modeled. These effects can be easily defined on the item-level (τ_{si}), or on the facet-level (τ_{sk}). The additional index i (or k) indicates that each item (or each item facet) has its own temporal effect. These model extensions are shown for two types of LLTM-type models by the following equations:

$$\eta_{pi} = \theta_p - \sum_{k=0}^K \beta_k X_{iK} + \tau_{sk}. \quad (2.4)$$

$$\eta_{pi} = \theta_p - \sum_{f=1}^F \beta_f X_{if} + \tau_{sf}. \quad (2.5)$$

Model 2.4 is an LLTM with fixed temporal effects on the facet level, model 2.5 is an explanatory model with fixed temporal effects on the level of item families. If several items sharing the same construction principles are administered to the same person, models including a parameter for each item type and models including a parameter for each basic item only can be compared. The application of such model extensions is not limited to the investigation of generation procedures for structurally parallel tests. Zeuch (2011) demonstrated how the modeling of fixed temporal effects on the facet level can be applied to investigate practice or training effects for cognitive items as well. Similar approaches have been presented by Fischer (1989), Formann and Spiel (1989), Glück and Spiel (2007), or Hohensinn et al. (2008).

Table 2.4.

Differences in model foci between explanatory IRT models with different types of design matrices

Columns of Design Matrix indicate	Describe	Generate	Understand	Predict
Items (diagonal matrix, RM)	✓	–	–	–
Item basic parameters / facets (LLTM)	–	✓	✓	(✓)
Item types / families (Parallel tests, Cloning)	–	✓	–	✓

Differences between explanatory IRT models with different types of design matrices are illustrated in Table 2.4.

- The *RM* is useful to describe the difficulty of a given item. It cannot facilitate the generation of new items and provides no information for a better understanding of item difficulties. It cannot be used to predict difficulties or new test items. It is a purely (“double”) descriptive model.
- The *LLTM* cannot describe individual item difficulties as accurately as the RM, but can be used to *predict* item difficulties of uncalibrated new items that were designed based on the same principles as an existing instrument. The power of the prediction depends on how well item difficulties can be explained by an additive combination of facet difficulty parameters (i.e., on the strength of the explanatory model). The LLTM helps to generate new items and can, if the explanatory model is valid, lead to a better understanding of item response processes. Classical LLTM models are explanatory on the item-side, but descriptive on the person-side of the model.
- Explanatory IRT models with design matrices containing item type or item family indicator variables are less useful for the understanding of item response processes than models based on item basic parameters because their focus is not on the underlying cognitive processes. When family indicator variables are specified as item explanatory variables, the existence of certain item types is taken as a given, regardless of how these item types are defined. With models of this type, the psychometric equivalence of structurally identical items can be investigated. As the LLTM, such models are explanatory on the item-side, but descriptive on the person-side of the model as well.

At first glance, this distinction and the depiction of models in Table 2.3 and Table 2.4 suggests that the described models are discrete classes of models. However, this is not the case. Conceptually, the only difference between the doubly descriptive RM and explanatory models with a focus on item facets or item types is a different layout of the design matrix \mathbf{X} . The most precise model is the Rasch Model with one difficulty parameter for each item. It provides a perfect description of item difficulties. When the design matrix is reduced to contain less column variables than items, a trade-off between

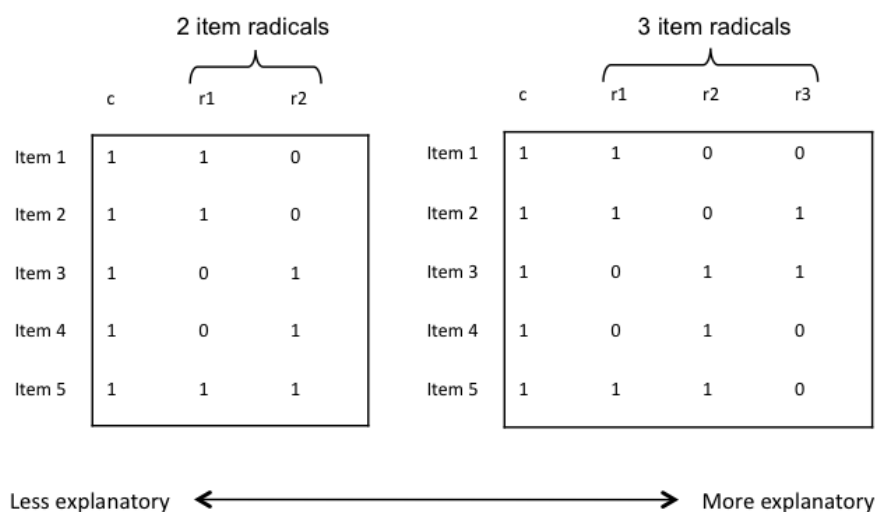


Figure 2.3.

Illustration of different possible design matrices on the continuum of explanatory IRT models: Number of radicals

predictive power and sparseness of the model is made. In the most extreme case, an empty model with an intercept only can be modeled (cf. De Boeck, 2008). This means that models can be more or less explanatory without constituting a discrete alternative class of models. Moreover, the labels “descriptive” and “explanatory” highlight the main focus of the model, rather than terming a different class of models. This is illustrated in Figures 2.3 and 2.4. Here, the columns of the design matrices represent item explanatory variables, i.e. variables that are applied to predict difficulty parameters of the items of a test. The two design matrices in Figure 2.3 represent two different complex LLTM models, one containing two item radicals and the other one containing 3 radicals. That is, the right model is more explanatory because it includes more item predictors; the amount of unexplained variation in item difficulties is at minimum as large as for the sparser model.

Figure 2.4 depicts four different design matrices, the leftmost representing a model without any item predictor; here the same item difficulties are predicted for each of the five items. Clearly this model has no explanatory value. The second model is equivalent to the first model in the previous illustration, it represents a LLTM with two item-explanatory variables, i.e. two item radicals. Applying the terminology of De Boeck and Wilson (2004a) this model is explanatory because it attempts to explain item difficulties by underlying task parameters instead of merely describing item difficulties. The third model is a model with three item explanatory variables representing three different item radical combinations, i.e. three item families. The rightmost model, then, is a model with five item explanatory variables, exactly one indicator for one item. This model is equivalent to a Rasch model, which is, following De Boeck and Wilson’s classification less explanatory

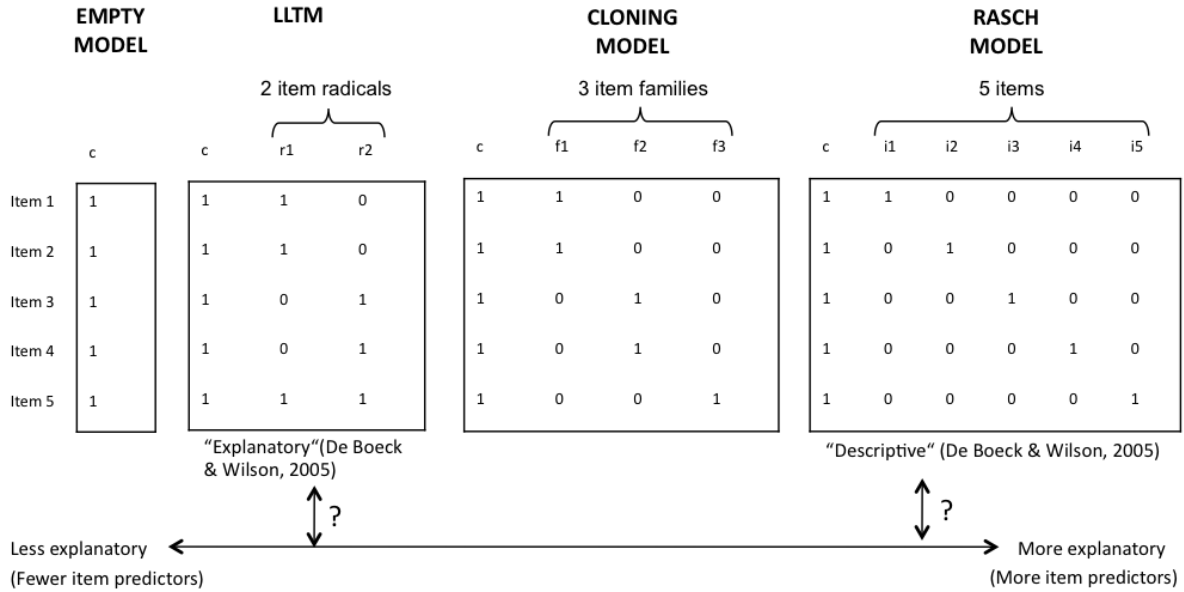


Figure 2.4.

Illustration of different possible design matrices on the continuum of explanatory IRT models: RM, LLTM, and Item Cloning

than the LLTM. This illustration shows the conceptual differences in the distinction between descriptive and explanatory IRT models proposed by De Boeck and Wilson (2004a) and the distinction of models according to their explanatory power in terms of predicting item difficulties. The term “explanatory” in De Boeck and Wilson (2004a) is not related to the explanatory power of the model in a statistical sense; moreover whether a model is defined as explanatory or descriptive depends on the *meaning* of the indicator variables used. If these variables are indicators of cognitive processes underlying item performance in the sense of radicals, the model is labelled explanatory; if the variables are indicators of specific types of items or individual items, the model is labeled descriptive. Due to this inconsistency in the definition of explanatory and descriptive models, I will not use the terms to distinguish between models in this thesis. Rather in the following chapters I will refer to explanatory IRT models as the broad class of models that allow to model item responses as a function of predictor variables of various kinds.

Also, it should be noted that the term “explanatory” cannot be interpreted literally in the way that models from the LLTM family truly explain item difficulties in a psychological, information-processing sense. Moreover, explanation of item difficulties in this context refers to the prediction of item difficulties based on a set of underlying variables as specified in the design matrix \mathbf{X} (also referred to as “Q” in the literature). As noted above, only in case that these variables characterize information processing steps derived from cognitive theory (i.e, if a strong theory approach is chosen) can resulting parameter

estimates be interpreted in a meaningful way. On the contrary if no theoretical model of item difficulties is available, parameters for the variables in a design matrix have, at most, heuristic value. Any design matrix is only unique up to allowed linear transformations that maintain full column rank. Ultimately, any design matrix that is derived as a linear transformation of the original matrix maintaining full rank will allow for an equally good prediction of item difficulties as the original design matrix (see e.g., Bechger, Verstralen, & Verhelst, 2002). That said, it is important not to mistake basic parameter estimates for LLTM components as true difficulty generating parameters. Item explanatory models from the LLTM family can provide a useful basis for evaluating the plausibility of competing cognitive models for a given test or item type. They provide a basis for testing whether empirical data for a given test is compatible with a certain hypothesized model of underlying cognitive processes. They cannot, however, provide unambiguous information on what the “true” cognitive processes that underlie test performance are. Further, design matrix mis-specifications can lead to biases during item-parameter estimation, such as over- or underestimation if too few or too many attributes are specified and the proportion of no-zero elements in the matrix is low (e.g., Rupp & Templin, 2008; Baker, 1993; Kunina-Habenicht, Rupp, & Wilhelm, 2012).

2.4.2. Models with person predictors

Doubly explanatory models include predictors for items difficulties as well as predictors for person abilities. Extending the LLTM by including additional predictors on the person side yields the so-called *latent regression LLTM* (LR-LLTM; e.g., Wilson & De Boeck, 2004). It takes person-level covariates into account in order to explain differences at the level of the individual:

$$\eta_{pi} = \sum_{j=0}^J \vartheta_j Z_{pj} + \varepsilon_p - \sum_{k=0}^K \beta_k X_{ik}, \quad (2.6)$$

Here, Z_{pj} is the value of person p on person property j ($j = 1, \dots, J$), ϑ_j is the fixed regression weight of person property j , and ε_p is a person random effect representing the remaining person effect after the effect of the person properties is accounted for. One can think of the latent person variable θ_p as being regressed on external person variables (Adams, Wilson, & Wu, 1997). The “virtual item” models described above can be conceptualized as latent-regression LLTMs as well: the fixed item-set effects can be interpreted as person covariates indicating that the probability of correct response for a given item changes for a second item set based on experiences with a first item set. Alternatively, these effects can as well be interpreted as item covariates, indicating that item difficulties change over time. Within the metric of the RM and the LLTM, any change in solution probabilities between testing occasions can be described without loss

of generality as either a change in terms of the person parameter or as a change of the item parameters (Mair & Hatzinger, 2007).

The LR-LLTM allows to move the investigation of construct validity one step further. The internal structure of the respective task can be analysed both with regard to the constituting task parameters and the contribution of broader ability constructs to solution probabilities. That is, instead of correlating test scores with scores from other (related or non-related) scores, scores on other tests or other person variables (e.g., gender, ethnicity) can be incorporated directly into the explanatory model.

Meulders and Xie (2004) have described a third class of models, that is models that include additional interaction effects between item- and person-predictor variables. This class of so-called “Differential Facet Functioning” (DFF) models will be described as part of the following section.

2.5. Differential item and facet functioning

Items are said to demonstrate “Differential Item Functioning” (DIF; see e.g., Holland & Thayer, 1985) or “Bias” (see e.g. Jensen, 1980) when subjects from different groups but with the same ability level have different probabilities of answering the item correctly. While DIF is the term usually used in psychometrics to characterize the phenomenon of group-specific item response functions, Bias is a widely used term in the field of cross-cultural Psychology or other more applied, test-fairness related fields of study. DIF research is especially important for all applications of AIG. Only if difficulty parameters for automatically generated items are truly identical for all groups of potential test-takers, or if differences in item parameters between different groups of potential test-takers can be accounted for by inclusion of appropriate interaction terms in the model can the true potential of AIG be used in practical testing situations.

DIF-research has a long tradition and there are many different models, non-IRT and IRT based approaches to determine whether items of a test are subject to DIF. Mathematically, DIF-effects are quantifiable differences in measurement properties of an item for two or more groups. When responses of two groups of test-takers are compared, one group is called the *reference group*, the other the *focal group*. According to an IRT model, an item displays DIF if the shape of the Item Characteristic Curve (ICC) varies across studied groups, given equivalent levels of the underlying construct. Equation 2.7 gives a mathematical definition:

$$P(Y_{pi} = y_{pi} | \sigma_i, \theta_p, z_p) \neq P(Y_{pi} = y_{pi} | \sigma_i, \theta_p), \quad (2.7)$$

with z_p indicating membership of person p to the focal or reference group, respectively. Here, the probability of a specific response is not only dependent on the ability param-

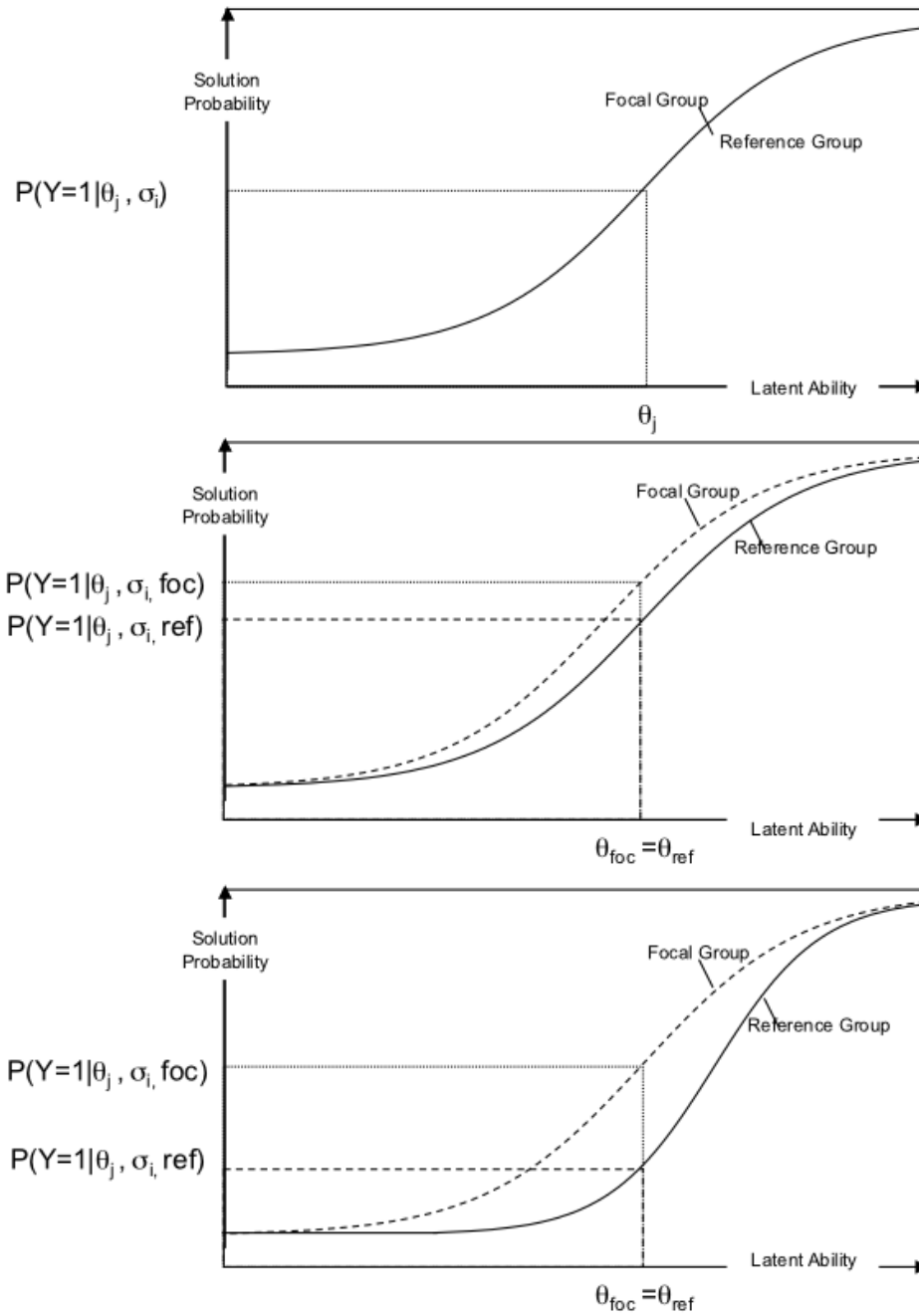


Figure 2.5. *Uniform and nonuniform-DIF in terms of the ICC of an item; top = no DIF, middle = uniform DIF, bottom = non-uniform DIF*

eter and the item difficulty parameter, but also on group membership. The conditional

probability of a response y_{pi} given the item difficulty parameter σ_i , the person ability parameter θ_j , and the group membership indicator z_p differs from the conditional probability of a response y_{pi} given the item difficulty parameter σ_i , and the person ability parameter θ_j only. This diminishes the validity of a test because cultural background or group characteristics are measured unintentionally instead of a pure measurement of the latent ability. On the contrary, if items are *not* subject to DIF, the location of items along the measurement scale is the same across all different subgroups. Given a certain level on the latent ability, only the difficulty of the item, and not group membership is predictive for the solution probability:

$$P(Y_{pi} = y_{pi} | \sigma_i, \theta_p, z_p) = P(Y_{pi} = y_{pi} | \sigma_i, \theta_p). \quad (2.8)$$

DIF methods enable the test developer to judge whether a test functions the same manner in various groups of examinees or across several testing occasions. In broad terms, this is a matter of measurement invariance (cf. Zumbo, 2007). In a globalized world with multicultural societies, DIF constitutes a severe threat to test-fairness and standardized assessments in educational and selection settings. Especially when test results are used to justify selection decisions, the equivalence of item characteristics and the related statistics across different cultural groups is a key variable to test fairness. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) indicate that DIF in a test also diminishes the practical value of the assessment. In selection contexts, DIF research can reveal the degree to which a test administrator has been successful in establishing fair “starting conditions” for all test-takers or whether specific group-characteristics are influential for the cognitive processing on certain items.

Magis, Beland, Tuerlinckx, and De Boeck (2010) recently presented a “General Framework for DIF Analysis” that allows for modeling most of the existing classical DIF approaches for dichotomously scored items. They divide DIF methods in (a.) methods based on Item Response Theory (IRT) vs. methods not based on IRT (i.e. the methodological approach), (b.) uniform vs. nonuniform DIF (i.e. the type of DIF effect), (c.) methods involving single vs. multiple focal groups, and (d.) iterative vs. non-iterative elimination of DIF items (i.e., methods with or without item purification). Table 2.5 lists these approaches. The differences between uniform and nonuniform DIF is depicted in Figure 2.5. ICCs for items showing uniform DIF are only different in terms of their location on the horizontal axis. ICCs for items showing non-uniform DIF differ also with regard to their discrimination parameter.

In the following selected methods for each of framework/DIF-effect constellations will be described in more detail. DIF methods that do not rely on item response theory usually detect DIF based on statistical methods for categorical data. Here, the total test score

Table 2.5.

Overview of methods for detecting differential item functioning for two groups

Method	Reference	Framework	DIF effect
Mantel-Haenszel (MH)	Mantel and Haenszel (1959)	Non-IRT	U
Standardization (Stand)	Dorans and Kulick (1986)	Non-IRT	U
Breslow-Day (BD)	Breslow and Day (1980)	Non-IRT	NU
Logistic Regression (Log)*	Swaminathan and Rogers (1990)	Non-IRT	U, NU
LRT	Thissen, Steinberg, and Wainer (1988)	IRT	U, NU
Lord	Lord (1980)	IRT	U, NU
Raju	Raju (1988)	IRT	U, NU

Note. U: uniform, NU: non-uniform; *Logistic regression methods do not require the estimation of a specific IRT model. However, they share some similarities with IRT methods. Therefore, they represent a “bridging method” between IRT and non-IRT methods (cf. Camilli & Shepard, 1994).

Table 2.6.

Structure of a contingency table for non-IRT DIF methods

	$Y_{pi} = 1$	$Y_{pi} = 0$
Reference Group	A_s	B_s
Focal Group	C_s	D_s

Note. $Y_{pi} = 1$ denotes a correct and $Y_{pi} = 0$ an incorrect response; s denotes the total score.

is used as a matching criterion. In DIF methods that rely on item response theory, the estimation of an IRT model is required for DIF testing.

Contingency-table based methods

The *Mantel-Haenszel* (MH) method and the *Breslow-Day* (BD) method are both based on contingency tables. The difference between the two methods is that MH allows to detect uniform DIF whereas BD allows for the detection of nonuniform DIF. For any tested item i , all examinees with a given total test score are cross-classified into a 2×2 (group \times correctness of response) contingency table (see Table 2.6 for an example).

A_i and B_i refer to the frequencies of correct and incorrect responses among all test-takers with total score s ($s = 1, \dots, k$) in the reference group, C_s and D_s denote these frequencies for the focal group. One contingency table for every possible total score s is investigated. The MH method conditions on the sum score and tests whether there is a relationship between group membership and item responses. The MH statistic is given as:

$$\text{MH} = \frac{(|\sum_i A_i - \sum_i E(A_i)| - 0.5)^2}{\sum_i \text{Var}(A_i)}. \quad (2.9)$$

with $E(A_i)$ and $\text{Var}(A_i)$ denoting the expected value (under the assumption of no DIF) and variance of A_i (see e.g., Magis et al., 2010 for the formulas for the derivation of these statistics). Under the null hypothesis of no DIF, the MH statistic is asymptotically $\chi^2(1)$ -distributed. Items are classified as DIF when a critical value of this distribution is exceeded. An alternative test statistic for the MH approach is λ_{MH} , which is derived as the logarithm of the odds ratio for the frequencies in Table 2.6:

$$\lambda_{\text{MH}} = \log \left(\frac{\sum_i A_i D_i / n_i}{\sum_i B_i C_i / n_i} \right). \quad (2.10)$$

Here, n_i gives the number of test-takers with test score i in the total sample. λ_{MH} is used as a common effect size for DIF, with the *ETS Delta scale* (Holland & Thayer, 1985) providing a widely accepted classification for this effect size. Three categories of DIF effects are proposed, i.e. negligible effects (ETS Delta = A), moderate effects (ETS Delta = B) and large effects (ETS Delta = C).

The BD method (Breslow & Day, 1980) is a widely used non-IRT method to detect non-uniform DIF. As the MH statistic, the BD statistic is also based on a contingency table of correct and incorrect item responses. The BD test tests whether the association between item response and group membership is homogeneous across different values of total test scores. The BD statistic is given as:

$$\text{BD} = \sum_i \left(\frac{[A_i - E(A_i)]^2}{\text{Var}(A_i)} \right), \quad (2.11)$$

with $E(A_i)$ and $\text{Var}(A_i)$ denoting the expected value (under the assumption of no DIF) and variance of A_i (see e.g., Magis et al., 2010 for the formulas for the derivation of these statistics). Under the null hypothesis of no DIF, the BD is asymptotically χ^2 distributed with the number of degrees of freedom matching the number of different total test scores investigated (Aguerri, Galibert, Attorresi, & Maranon, 2009).

Logistic Regression based DIF detection

The *Logistic Regression* method (Swaminathan & Rogers, 1990) is based on a logistic model for the probability of correct response. Uniform DIF effects are tested by investigation of main effects, interaction effects can be investigated to test nonuniform DIF. The Logistic Regression method does not strictly fall into the category of IRT-based methods

but was called a “bridging method” between non-IRT and IRT approaches (Camilli & Shepard, 1994). The logistic regression model can be written as:

$$\text{logit}(\pi_{pi}) = \beta_0 + \beta_1 s_p + \beta_2 z_p + \beta_3 s z_i, \quad (2.12)$$

with $s_p i$ the total test score for person p , z_p the group membership for this person, and $s z_p$ the interaction of the two factors. Both uniform and nonuniform DIF effects can be tested using this model and the usual statistical test procedures (e.g., Wald or LR test), depending on whether interaction effects are specified or not. A model-fit statistic ΔR^2 (Zumbo & Thomas, 1997) can be calculated based on Nagelkerke’s R^2 (Nagelkerke, 1991) for nested models. Two different categorizations for the ΔR^2 effect size were proposed (Zumbo & Thomas, 1997; Jodoin & Gierl, 2001). The categorization proposed by Jodoin and Gierl (2001) (JG) is less conservative than the earlier classification by Zumbo and Thomas (1997) (ZT).

DIF frameworks based on specific IRT models

In DIF methods that rely on IRT, the estimation of a specific IRT model is required for DIF testing. For dichotomously scored items, IRT-methods can, for instance, rely on the 1PL, 2PL or 3PL model. 2PL and 1PL can be derived from the most general formula of the 3PL when item guessing parameters and (in the 1PL also) item thresholds are fixed. The 3PL model is defined as¹:

$$P(Y_{pi} = 1 | \theta_p, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{\exp [a_i(\theta_p - b_i)]}{1 + \exp [a_i(\theta_p - b_i)]}. \quad (2.13)$$

As in the models defined in the previous sections, Y_{pi} is the binary response of person p to item i . θ_p is the position of person p on the latent trait. a_i is the item discrimination parameter. b_i is the item difficulty parameter. c_i is the guessing parameter of item i . Besides the Likelihood-Ratio-Test method and the methods by Raju, Lord’s approach is one of the most important IRT-based DIF methods. *Lord’s* method (Lord, 1980) tests the equality of item parameters in focal and reference group based on a χ^2 -Statistic Q_i . It is based on the assumption that the maximum likelihood item parameter estimates are asymptotically normally distributed. The method is very flexible and allows to test both uniform and nonuniform DIF based on any type of fitted IRT model. Under the assumption that the item difficulties of a test can be accurately described by a 1PL (Rasch) model, Lord’s test statistic is given as:

¹Here, the typical 3PL notation with a_i , b_i , c_i as the three item parameters is used. This deviates from the Rasch-model notation used generally in this thesis where the item difficulty parameter is denoted σ_i and not b_i

$$Q_i = \frac{(b_{iR} - b_{iF})^2}{\text{SE}(b_{iR})^2 + \text{SE}(b_{iF})^2}. \quad (2.14)$$

Whether uniform or nonuniform DIF is investigated depends on the choice of the specific IRT model. The Rasch model provided a framework for uniform DIF (all item discrimination parameters are set to one), whereas the 2PL model allows to investigate non-uniform DIF (here, the item discrimination parameters are freely estimated).

These “classical” approaches to DIF are useful to determine which items are prone to bias in cross-cultural applications. However, these methods do not allow to identify sources of bias in characteristics of the respective items. In terms of Zumbo (2007)’s (2007) general framework of DIF methods, these approaches represent the “first and second generation of DIF”. They remain largely technical and show how DIF in individual items affects the distribution of the test scores in different groups. In contrast, approaches of the “third generation of DIF” are suitable for the analyses of construct bias on the level of facets underlying each item, and therefore the identification of factors for DIF (Zumbo, 2007).

Person-by-Facet Interactions: Differential Facet Functioning (DFF)

Meulders and Xie (2004) have described a third class of explanatory IRT models, that is models that include additional interaction effects between item- and person-predictor variables. By the inclusion of such predictor variables models can be build that account for person- or group-specific differences in item parameters. *Differential Facet Functioning* (DFF; Engelhard, 1992; Meulders & Xie, 2004) falls in the category of “third-generation approaches” (Zumbo, 2007; p. 229) to DIF. In a similar way like the LLTM attempts to explain item difficulties based on a set of underlying item facets, DFF attempts to explain differences in item difficulties between groups by the inclusion of interaction effects, specifically person*facet interactions. Engelhard suggested that “studies of differential facet functioning can be conducted by a variety of procedures that are conceptually similar to current approaches for studying differential item functioning” (p. 175). In comparison with DIF, in DFF a more explanatory investigation of component difficulties due to specific person properties (e.g., WM capacity) is feasible. As shown by Meulders and Xie (2004) a DFF model can be conceptualized as an extension of the models described in 2.1 and 2.6 as

$$\eta_{pi} = \theta_p - \left(\sum_{k=0}^K \beta_k X_{ik} + \sum_{k=0}^K \gamma_k X_{ik} Z_p \right) \quad (2.15)$$

where γ_k denote the weights for the interactions of item facet difficulties with qualitative person predictors indicating group membership (i.e. $X_{ik}Z_p$), respectively. This model, then, allows to assess whether difficulties of item components (“facets”) vary with

person properties, indicating possible differences in cognitive processing while working on the items. Significant interaction effects γ_k imply that the effect of item facets on item difficulty depends on the group. As Meulders and Xie (2004) point out, modeling DFF can be used to investigate both main and interaction effects. Whereas main effects represent mean differences in item difficulties between groups that are not indications for LLTM parameter invariance, interaction effects capture facet bias. Facet bias diminishes the construct validity of an item because the item measures the set of abilities needed to solve an item to different degrees in the two groups. That is, if the relative contributions of the item facets to global difficulties are not the same across groups, the constructs measured in each of the groups are, at least to some degree, not the same. DFF has been proposed as an addition to classical DIF methods to add an explanatory component to the pure detection and quantification of DIF: “The LLTM is useful in testing how the response data conform to the structure of the test design. The DFF model helps to explain the DIF effects more substantively.” (Xie & Wilson, 2008, p. 414).

The DIF and DFF methods described in the previous sections will be applied in Study 3 of this thesis to investigate the cross-cultural comparability of test-scores on a rule-based generated figural reasoning measure.

3

The Figural Analogy Test (FAT): Item generation and construct validation

This study describes the development of a new reasoning measure, the *Figural Analogy Test* (FAT), that extends earlier works by Beckmann (2008) who developed an analogy measure based on alphanumeric symbols and figural-spatial rules. The current study aims at the development and validation of a purely figural measure that requires no mathematical or verbal abilities. The generative framework is exclusively based on theories of analogical reasoning, specifically focusing on geometric analogies and spatial ability. The validity of the new item-generative framework is tested in an empirical study with $N = 308$ university students. Two main research questions are addressed, first, the appropriateness of the set of pre-specified item radicals to model item difficulties in terms of a reliable prediction of difficulty parameters. Second, it is tested whether the parameter estimates of the item-difficulty model are in line with assumptions about figural-spatial processing and analogical reasoning. A set of specific hypotheses related to the impact of the item radicals manipulated are tested. Several explanatory IRT models are compared, including models with item-predictors only and models with person-by-item interactions. Results show that item difficulties can be predicted based on the new AIG framework. Absolute Parameter differences are, however, considerable and constitute a threat to potential operational applications. All parameters are in line with theories of figural-spatial reasoning. Gender differences are driven by specific item features. Furthermore, scores on the new test correlate with other established measures of fluid reasoning and spatial ability and demonstrate incremental validity for the prediction of school grades. Future studies should investigate the generalizability of these results to fully automatically generated FAT items and the feasibility of the item difficulty modeling approach for computerized adaptive testings scenarios.

Keywords. Figural-spatial reasoning, Analogical reasoning, g , Automatic Item Generation, Explanatory IRT, LLTM

3.1. Introduction

This introduction is structured into three parts. First, figural-spatial analogies and their value as indicators of fluid intelligence will be described. Second, previous attempts to construct figural analogies based on item-generative rules and predict item difficulties will be reviewed. Third, the research questions of the current study will be derived.

3.1.1. Figural-spatial analogies as indicators of fluid intelligence

The ability to make inferences based on the processes of inductive and deductive reasoning has been a main theme of philosophical enquiry ever since the beginning of scientific endeavor. It is a prerequisite for learning and problem solving (e.g., Glaser, 1982; Snow et al., 1980; Sternberg, 1984) and a vital part of most important theories on human intelligence (e.g., Carroll, 1993; Cattell, 1971; Jäger, 1982; Spearman, 1904; Thurstone, 1938). Abstract Analogical reasoning items of the form

$$\mathbf{A} : \mathbf{B} :: \mathbf{C} : ?$$

combine the two core processes induction and deduction. Analogical reasoning technically terms the process of extrapolating a function from a pair of source objects ($\mathbf{A} : \mathbf{B}$) and the application of the function to a pair of target objects ($\mathbf{C} : ?$). The target pair is incomplete (i.e., \mathbf{D} is replaced by a question mark) and has to be completed using the function induced from the pair of source objects.

A number of different information processing theories have been proposed, describing the component processes of analogical reasoning (Spearman, 1923; Sternberg, 1977b; Sternberg, 1977a; see Beckmann, 2008 for a comprehensive review of these theories). *Analogical mapping* has been considered the core process of analogical reasoning. Analogical mapping is the process of establishing a structural alignment between and projecting inferences based on two represented situations. The problem solving process entails the establishment of an explicit set of rules that describe the correspondences between the elements of the two situations. It can be divided into two main stages (see also Figure 3.1):

1. The first phase is called *preparation period*. The relation between the A term and the B term has to be deduced by induction. The sequence of processes during this period is: (1.) inspecting the $\mathbf{A} : \mathbf{B}$ -relation, (2.) investigating the relational similarity, and (3.) inducing of formal logical rules (i.e., the function that has to be applied to create the missing target object).

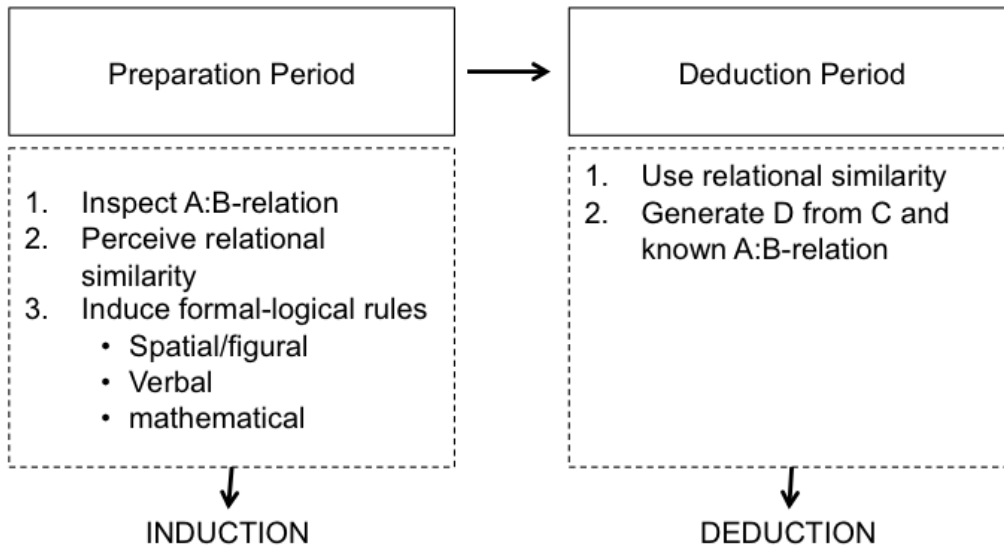


Figure 3.1.
Process model of analogical reasoning (Study 1)

2. Deduction is executed when the inferred relation is applied onto the C term to generate the D term (*deduction period*). Here, the rules induced during the preparation phase are applied to the first target object, **C**.

According to D. M. Johnson (1962), item difficulty can be led back to either induction or deduction, depending on the degree of familiarity with the stimuli. The solution of analogy items requires, therefore, exactly the two processes that build the core of analogical thinking, i.e. the abilities to perceive and use relational similarity. This process is very general and applies to a multitude of possible analogy problems. Analogy problems can be verbal, asking the test-taker to detect semantic similarities between words. Analogies can be complex realistic problem solving tasks, asking respondents to induce general rules or principles from complex text or simulations, and apply them in new situations. For the current study, only abstract figural analogies are investigated. This is due to the goal to generate test items of logical reasoning that are as free as possible from language, mathematical rules, and references to cultural-specific knowledge.

The solution of figural-geometric analogy items requires the usage of spatial abilities. Spatial abilities and reasoning are closely related, though they are still two distinct dimensions (Gittler, 1999). The importance of image rotation in human intellectual function has been repeatedly highlighted (e.g., W. Johnson & Bouchard, 2005). At the same time spatial abilities are often assessed only poorly in cognitive test batteries. As shown in Table 3.1, spatial tasks can be classified according to the type of spatial rules as well as according to the types of cognitive processes associated with them.

Table 3.1.

Spatial rules and cognitive processes in figural-spatial tasks (Study 1)

	Visualization	Orientation
Spatial distortion rules	✓	–
Spatial displacement rules	✓	✓

Factor analytic studies of spatial ability tasks have provided strong support for the existence of two distinct spatial abilities, usually termed *visualization* and *orientation* (e.g., Hegarty & Waller, 2004): *Visualization* refers to the ability to mentally rotate, manipulate, and twist two- and three-dimensional objects in tasks. *Orientation* refers to comprehension of the arrangement of elements within a visual pattern and the ability to retain spatial orientation in changing conditions.

During the solution process of figural-geometric analogies, two types of mental transformations have to be applied (cf. Linn & Petersen, 1985; Shepard & Metzler, 1971; Whitely & Schneider, 1981): Transformations referring to disorientation of elements from *A* to *B* such as rotation, reflection and exchanges are labelled *spatial displacements*. Transformations that refer to the size, shade, shape, and number of geometric elements are labelled *spatial distortions*. Only spatial displacements truly require the two spatial abilities of visualization and orientation. Spatial distortions refer to changes in geometric elements as well, but these changes are not directly linked to their orientation and relation in space.

While the large majority of research findings indicate that there are no gender differences in general intelligence (e.g., Colom, Juan-Espinosa, Abad, & Garcia, 2000; see Halpern & La May, 2000 for a review), gender differences on specific cognitive abilities have been reported consistently (e.g., Birenbaum, Kelly, & Levi-Keren, 1994; Dykiert, Gale, & Deary, 2009; Irwing & Lynn, 2005; Irwing & Lynn, 2006; W. Johnson & Bouchard, 2007). Specifically, gender differences on tasks involving spatial abilities are well-documented (Cooke-Simpson & Voyer, 2007; Jordan, Wüstenberg, Heinze, Peters, & Jäncke, 2002; Linn & Petersen, 1985; Masters & Sanders, 1993; Terlecki, Newcombe, & Little, 2008; Voyer, Voyer, & Bryden, 1995). Meta-analyses show an average standardized mean difference of $d = 0.73$ for mental rotation tasks (Birenbaum et al., 1994; Linn & Petersen, 1985) and faster mental rotation for males (e.g., Heil & Jansen-Osmann, 2008). The largest effect sizes are reported for the Mental Rotation Test (MRT; Linn & Petersen, 1985; Voyer et al., 1995). Effects increase with the complexity of the stimuli to be rotated (Heil & Jansen-Osmann, 2008). Scores in the MRT increase with age but this increase is stronger for males (Geiser, Lehmann, & Eid, 2008). Males invest more time in activities that require spatial cognition and therefore train their spatial ability over lifespan more than women do (Baenninger & Newcombe, 1989; (1995)).

Mental rotation can be performed analytically or holistically. Many studies have shown that females tend to focus more strongly on analytic strategies while solving spatial tasks whereas men favor holistic strategies (Heil & Jansen-Osmann, 2008; Hugdahl, Thomsen, & Erslund, 2006; A. B. Janssen & Geiser, 2010; A. B. Janssen & Geiser, in press; Moe, Meneghetti, & Cadinu, 2009).

- *Analytic Processing:* When *analytic strategies* are used to perform mental rotation, stimuli are rotated in a piecemeal fashion, i.e., every part of the stimulus is processed individually. In this case, the speed of mental rotation is a function of stimulus complexity. That is, the more complex a stimulus is, the longer it takes to mentally rotate it. For instance, polygons with higher degrees of angular disparity, more corners or less distinct shapes will be more complex than simpler polygons. All individual features of the stimulus are represented and rotated separately when this strategy is followed.
- *Holistic Processing:* Test-takers using *holistic strategies* to perform mental rotation represent the stimulus as a whole. In this case, mental rotation performance is expected to be less affected by stimulus complexity. Only one operation is needed no matter whether the stimuli has many or only few features.

In line with these general assumptions, longer reaction times for mental rotation have been reported for men vs. women (e.g., Geiser, Lehmann, & Eid, 2006; Heil & Jansen-Osmann, 2008). Large proportions of the robust gender-effect in mental rotation performance can, indeed, be explained by the use of different strategies and the extent of previous experiences with spatial problems. Feng, Spence, and Pratt (2007) trained men and women with an action video game. They report substantial gains in spatial attention in general and mental rotation in specific. Most notably, gain effects were significantly larger for women. The gender effect could be considerably reduced by the intervention. Gender differences on spatial tasks can be, at least to some degree, explained by prior task experiences and preferences for different processing strategies. It is important to consider these findings when new measures based on figural-spatial rules are constructed and validated.

3.1.2. Rule-based generation of figural analogy items

The impact of the components of figural-geometric analogy items on item difficulties has been investigated in several studies. When the contributions of pre-specified task parameters to item difficulty are investigated, a cognitive theory of the item solving process can be tested. Statements on the construct validity of the items can be made. The existence of a sound cognitive model describing task performance is one prerequisite for the application of rule-based AIG in practical assessment settings: knowing the contribution of item radical difficulties to item difficulty allows to create items of designated difficulties (see Chapter 2.3 of this thesis for a discussion of rule-based AIG).

Whitely and Schneider (1981) explored the information structure for geometric analogies, using three spatial displacement transformations (rotation, reflection, and spatial exchanges) and six transformations from the category of spatial distortions (adding, removing or dividing elements, shade, size, shape). Only spatial displacements led to an increase in item difficulty whereas spatial distortion rules caused, in fact, an opposite effect (the “number of spatial displacement transformations was positively related to item difficulty, while number of spatial distortions was negatively related”, p. 395). In a similar experiment, Novick and Tversky (1987) analyzed the impact of eight different types of transformations on item difficulties. Their conclusions were almost identical with the results reported by Whitely and Schneider (1981). Novick and Tversky’s findings confirmed increased item difficulties for tasks containing spatial displacement in contrast to distortion tasks. In order to construct tests of higher difficulty level, tasks should, therefore, obtain more spatial displacements.

Murray (1997) reported shorter reaction times for flipping of objects (i.e., mirroring) versus spinning of objects (i.e. rotation). It is assumed that flipping is performed without formation of intermediate representations and is therefore more efficient (Kanamori & Yagi, 2002).

Mulholland, Pellegrino, and Glaser (1980) analyzed the factors contributing to the difficulty of geometric analogy items and concluded that the number of transformations yielded a significant effect on error rates in true analogy tasks. Further, the number of transformations and the number of elements interacted and significantly affected the amount of errors made. Mulholland et al. (1980) concluded that “the largest single source of errors was multiple transformations of single elements” (p. 282). The number and complexity of the transformation to be performed represents the relational complexity of the tasks (cf. RC theory; Halford et al., 1998 or Chapter 5 in this thesis). Working memory load is assumed to account for an increase in error rate due to an increase in the complexity of the spatial relations between elements and thus an increased amount of information needs to be stored and processed in working memory.

While geometric analogies have been studied intensively since the early 1980s, these studies did not link the investigation of key factors of item difficulty with the rule-based generation of new instruments. Beckmann (2008) closed this gap between intensive research on analogical processing on the one hand and the lack of sound psychometric instruments making use of the principles of rule-based AIG on the other. She proposed a new item-generative framework and tested the applicability of this framework empirically in a sequence of three pre-studies and one larger main study. The results from her studies can be summarized with regard to two main research goals, (1.) the identification of key drivers of item difficulty for figural analogy items and their usefulness in modeling item difficulties by explanatory IRT models, and (2.) validity findings for her new measure with regard to general cognitive ability as well as scholastic performance, and the impact of a

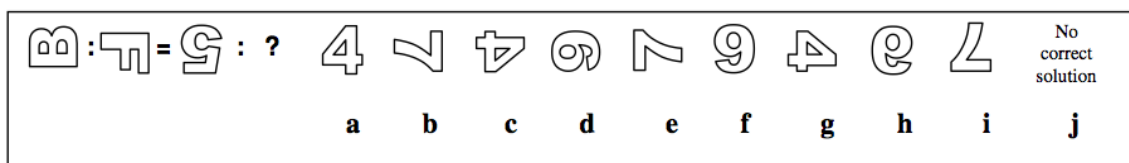


Figure 3.2.

Example Item with two rules from Beckmann’s analogy test; Two rules are applied here, i.e., rotation by 180 degrees, and sequence plus; figure from Beckmann (2008)

comprehensive instruction materials on test validity. In the following, I will summarize her findings and describe why an extension of Beckmann’s study was necessary.

Beckmann (2008) identified important factors of item difficulties for the type of figural analogies by applying a number of explanatory IRT models. The item facets constituting her item-generative framework are summarized in Table 3.2, together with the facet difficulty parameters estimated by means of LLTMs. An example item of Beckmann’s instrument is shown in Figure 3.2. She combined two spatial distortion with 5 spatial displacement rules and two additional alphanumerical rules in a set of $k = 44$ items. An individual item contained between 1 and 3 rules with a maximum of 1 rule from each type. Subjects worked on the test in four blocks with time limits given for each block. Item solution probabilities ranged from $p = .11$ to $p = .93$ with a mean difficulty of $p = .47$. Out of 44 items, 32 items showed good fit to the Rasch model. Confirmatory factor analyses showed misfit for a one-factor model for all items, but acceptable fit statistics for the reduced set of Rasch-homogenous items. The amount of heterogeneity among the test items was reflected also in the internal consistency of the test: With $\alpha = .76$, the 44 item-version shortly missed the desired value of $\alpha > .80$ (Anastasi, 1981).

Regarding Beckmann’s item difficulty modeling approach, all item facets expected to function as item radicals reached statistical significance in the LLTM. 71% of variation in item difficulties could be explained by the linear combination of LLTM item facet parameters. Signs and relative heights of all facet parameters were mostly in line with previous research (see e.g., Whitely & Schneider, 1981): While changes in the printed size of letters and numbers showed facilitating effect on item difficulties, presence of all other facets led to increased item difficulty. Item difficulties did not differ significantly depending on the direction of rotation (clockwise vs. counter clockwise); a reduced model with only two rotation parameters (instead of 3) did not fit considerably worse. It turned out sufficient to model rotation by 90 degrees when predicting item difficulties without distinguishing between clockwise and counter-clockwise rotation. Contrary to earlier studies, rotation by 180 degrees had lower impact on item difficulty than rotations by only 90 degrees. This finding is inconsistent with research on mental rotation. One of the most robust findings in mental rotation research is that difficulties of mental rotation tasks increase with increasing rotation angles (Shepard & Metzler, 1971).

Table 3.2.

Item-generative framework behind Beckmann’s analogy test: Radicals

Type of cognitive rule	Item Facet	LLTM weight
Spatial Distortion	Increase in Size	0.29**
	Decrease in Size	0.29**
Spatial Displacement	Rotation by 90 degrees clockwise	-1.19**
	Rotation by 90 degrees counter-clockwise	-1.16**
	Rotation by 180 degrees	-0.54**
	Mirroring at the horizontal axis	-1.91**
	Mirroring at the vertical axis	-1.42**
Numerical/alphabetical	Sequence plus	-0.96**
	Sequence minus	-0.60**

Note. LLTM weights are values on logit scale, smaller values represent higher difficulties; table displays results from Beckmann (2008)

Beckmann’s findings with regard to the two new “sequence” rules were only partly satisfactory. Application of these rules did not require any figural or spatial abilities, but their application relies only on numerical and alphabetical knowledge (“The rules sequence plus and sequence minus required the participant to go forward or backward in the alphabet when letters were presented and to add or subtract when digits were presented”, Beckmann, 2008, p. 90). Both letters and digits cannot be considered *abstract* figural elements. They represent objects with a semantic meaning which is more or less easy to detect depending on their orientation in space. This is reflected in Beckmann’s results: considerable influences of surface characteristics related to these rules were found: for instance, the orientation of the A elements played a significant role for the difficulty of preparation period processing. Item difficulties differed considerably dependent on the initial position of the letter or digit in the **A**-element: an item containing exactly the same cognitive rules was considerably more difficult when the letter was presented 90 degrees clockwise rotated instead of in the upright “normal” position. The same was true for the orientations of letters and digits in the **C**-element. While initially conceptualized as incidentals to be chosen randomly, stimuli positions and orientations turned out to function as item radicals here, albeit with inconsistent magnitudes. This finding complicates reliable prediction of item difficulties of uncalibrated items based on the underlying item facets. It is an open question whether the unexpected result regarding 180 degree rotations is also connected to the special types of geometric forms used.

Mixed results were found as well when item solution probabilities were regressed on the type of combination of letters and digits in a given item: Items with digit-digit combinations (i.e. item where were **A** : **B** elements and **C** : **D** elements were both digits) were significantly easier than items with letter-digit, letter-letter or digit-digit combinations.

Here, empirical findings on the drivers of item difficulties are not in line with the previously assumed pattern of item radicals and incidentals as well. Research could analyze the impact of specific digits or numbers on the ease of spatial displacements. However, such analyses could not solve the problem that the alphanumerical rules do “not correspond to the characteristic of pure culture-fair tests, since participants of different backgrounds might not meet these requirements or do not provide such knowledge” (Beckmann, 2008, p. 170). An alternative way to deal with these problems could be to avoid letters and digits and to build an analogy test that is purely based on abstract figural elements. This procedure is followed by the current study.

In addition to the investigation of the goodness of fit of the new item-generative framework, Beckmann (2008) investigated the effects of two different types of instruction. Specifically, Anastasi’s claim was tested, that short orientation and practice sessions preceding the actual assessment can establish comparable testing conditions for all subjects. The so-called *Test Sophistication Hypothesis* (Anastasi, 1981) suggests that brief practice increases the construct validity of ability tests by reducing confusion and test anxiety. The measurement of the construct becomes less contaminated with other factors and the error variance is reduced, thereby enhancing its measurement properties. Two randomly assigned groups were compared in a between subject design. While one group received a comprehensive instruction comprising all rules and example items before the assessment, another group of test-takers received no instruction. Mean scores of the instruction vs. non-instruction groups differed significantly with members of the instruction group solving on average 2 items more, representing an effects size of $d = 0.36$.

Scores on Beckmann’s instrument correlated substantially with the revised German version of Cattell’s *Culture Fair Test* (German adaption CFT 20r; Weiß, 2007) only in the instruction-group ($r = .54$ vs. $r = .24$ in the no-instruction group). Test performance in the total sample correlated significantly with GPA ($r = .18$), with a higher correlation for a math & science composite score compared to an arts & music composite ($r = .22$ vs. $r = .16$).

Beckmann’s (2008) work builds a strong starting point for automatic rule-based generation of new test items. Beckmann identified core drivers of item difficulty. By specifying what item facets contribute to item difficulty and what facets are negligible when item difficulties should be modeled, her results provide a good basis for further developments of rule-based analogy items. However, parameter estimates reported by Beckmann are not completely in line with theoretical assumptions about the underlying construct. That is, while the items developed based on her framework showed, overall, good psychometric properties, the construct validity of the new measure could only be partly established. While spatial displacement and distortion rules functioned mostly as expected, the usage of letters and digits and the two new “sequence” rules produced inconsistent results. Surface characteristics related to specific digits and numbers and their orientation in space turned out to function as item radicals rather than as incidentals. Also, the usage of

letters and digits conflicted with the goal to develop a culture-fair reasoning measure that is valid across cultural borders and independent of language or mathematical knowledge. Unwanted multidimensionality due to the heterogeneity of cognitive rules might be one reason for the large amount of 25% of Rasch-misfitting items and the rather low internal consistencies and criterion-related validities. Carlstedt et al. (2000) showed that an intelligence test made up of homogenous items loaded higher on a general intelligence factor than did a similar test made up of heterogeneous items. Furthermore, Bethell-Fox, Lohman, and Snow (1984) showed that analogy items containing spatial displacements evoked other cognitive processes compared to items without such spatial rules. Due to constraints with regard to the combination of rules, at maximum only one spatial displacement rule could be used in one item, yielding items of mostly medium difficulty level. Beckmann (2008) suggests that, in order to construct tests of higher difficulty level, tasks should obtain more spatial displacement rules.

3.1.3. Research questions

The overarching goal of the current study is to extend Beckmann's work in order to develop an item-generative framework that is suitable for computer-based AIG. It will be investigated how a rule-set for the generation of a new, purely figural, analogy measure can be developed based on cognitive theories of analogical reasoning and figural-spatial processing. Items of the instrument should be purely figural, highly *g*-loaded and focus more strongly on spatial displacement rules. The set of item-generative rules should allow to construct items suitable for assessment of the whole range of the ability continuum. Two main research questions are investigated:

1. First, the appropriateness of the set of pre-specified item radicals to model item difficulties in terms of a reliable prediction of difficulty parameters will be critically investigated. This refers to the mostly technical benefits of rule-based item generation, i.e. an accurate prediction of item difficulties based on pre-calibrated item facet parameters instead of individual item calibrations. How well can item difficulties be predicted by the set of underlying item facets? How large is the deviation of true and predicted item difficulties?
2. Second, it is tested whether the parameter estimates of the item-difficulty model are in line with findings from cognitive psychology on figural-spatial processing and analogical reasoning. That is, are items generated based on the new framework construct valid both in terms of their construct representation and in terms of nomothetic span? Do item facet parameters go in line with their theoretically assumed direction and magnitude? Can the pattern of gender effects found for figural-spatial tasks be replicated regarding the set of new item radicals? Is the new test capturing the intended construct based on its relationships with other measures? Specific hypotheses tested regarding the item radicals and incidentals

of the new measure will be lined out after the description of the item-generative framework in the next section.

Regarding the AIG process illustrated in Figure 2.2 the research questions of this study focus on the first three steps, i.e. construct definition, identification of underlying cognitive processes and item radicals, and the empirical test of hypotheses regarding the item radicals. Step 4, the combination of item facets in new items, goes beyond the scope of this study. It is only indirectly addressed here. As a result of the first three steps, a set of item radicals is established that provides a basis for the generation and empirical investigation of new items. The development of a computer-software for the generation of new items, calculation of difficulty parameters and estimation of person abilities is currently under development.

3.2. Method

There are two parts to the method section. First, the derivation of radicals and incidentals for the new item-generation framework is described. Second, the design of the empirical study is described that addresses the research questions outlined above. Specific hypotheses tested regarding the construct representation of the new measure will be described at the end of the description of the framework.

3.2.1. Development of the new item-generative framework

Based on the findings summarized above, a number of changes to and extensions of Beckmann’s item generative framework were made. These changes are summarized in the following:

1. Item facets that did not have considerable impact or inconsistent effects on item difficulties were removed from Beckmann’s item-generative framework.
2. Instead, a more homogeneous set of rules was applied to enhance the validity of the item generative framework and the explanatory value of the item facets (cf. Carlstedt et al., 2000). All rules were chosen to be spatial displacements. In contrast to the previous item facets, no verbal or mathematical rules were used.
3. In total, six displacement rules were applied, with up to 3 rules used simultaneously in one item. Others have shown that “the largest single source of errors was multiple transformations of single elements” (Mulholland et al., 1980, p. 282). The combination of more than one spatial displacement rule in a given item should allow especially for the generation of difficult items suited for discrimination in high ability ranges.

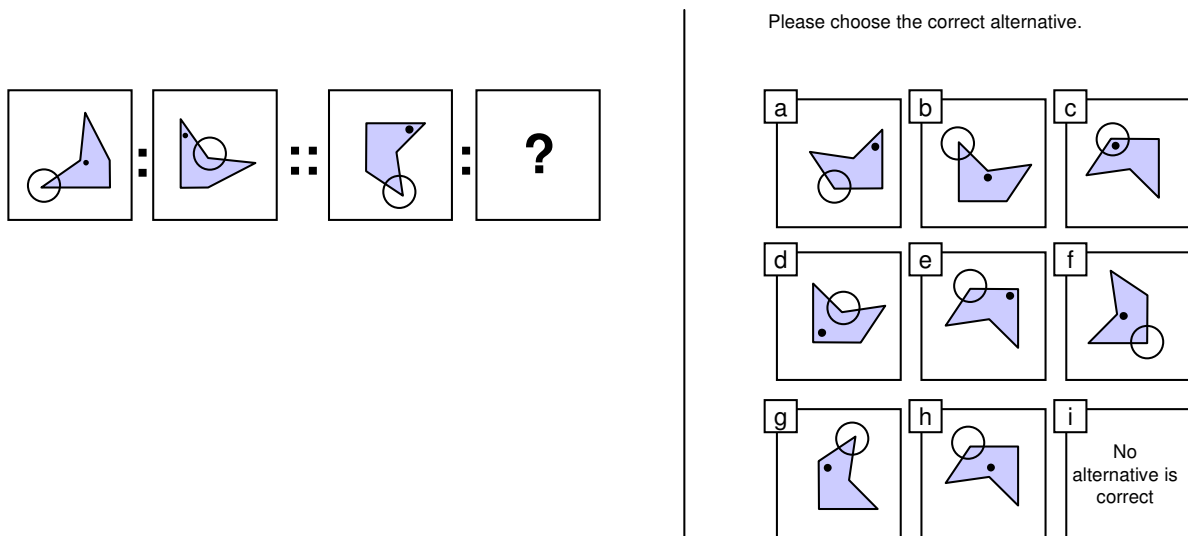


Figure 3.3.

Sample Figural Analogy Test item with 9 response alternatives

4. Purely abstract geometric figures were used instead of letters and digits. All references to specific cultural and educational variables were excluded. Previous findings indicate that reasoning measures with figural content are the best measures of fluid intelligence (e.g., Undheim & Gustafsson, 1987). An increased coverage of spatial abilities was one consequence of the explicit focus on language- and knowledge-free cognitive rules. This was expected turn out beneficial from a practical point of view as well as performance on spatial tasks, particularly those involving mental rotation, predicts success in many professional settings (e.g. aviation, engineering, visual arts) better than general intelligence or verbal abilities.

Figural stimuli Instead of the letters and digits used in Beckmann’s test, each element of an analogy item (i.e., **A**, **B**, **C**, and **D**) is composed of one grey pentagonal main shape and a number of additional figural features (see Figure 3.3 for an example). In order to prevent problems with initial stimuli positions as described above, 8 exchangeable pentagonal main shapes were designed sharing a set of general characteristics while still being visually distinguishable (see section about item incidentals in the following).

Each main shape could be combined with up to four additional features that show a specific spatial relation to the main form. The idea to lay the focus on the spatial relation of figural objects to each other was successfully applied by other tests before. For instance, the subtest “Topologies” from the CFT (Weiß, 2007) uses a similar principle. In the Topologies test respondents have to detect the spatial relations of geometrical figures,

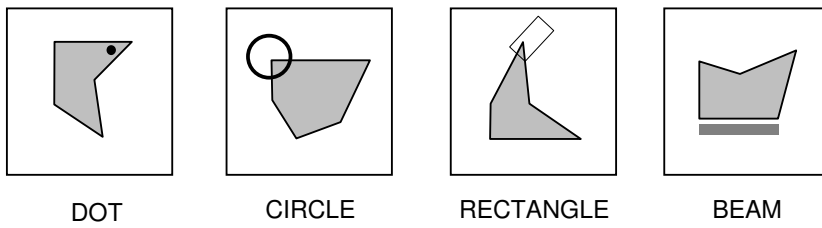


Figure 3.4.
Combination of main shapes and features into figural objects in the new item-generative framework of the FAT

lines and dots to each other in one configuration, and then pick another configuration that pertains the same relations of elements by building an analogy between the two sets. In case of the new item-generative framework presented here, four distinct features were chosen that can be positioned clearly either on the edges or corners of each pentagonal shape. The introduction of these features allowed for the generation of more complex stimuli that comprise a higher number of spatial displacements in one item. Figure 3.4 exemplarily shows how main shapes and features could be combined to build figural objects. Circle and rectangle were always positioned at the corners of the main shape; the point was always positioned in one corner of the main shape, and the beam parallel to an edge of the respective main shape. All possible positions of each feature were thereby unequivocally defined.

Item Radicals The choice of item radicals was guided by Beckmann’s findings as well as the objective to allow for more spatial displacement rules and a combination of 2 or

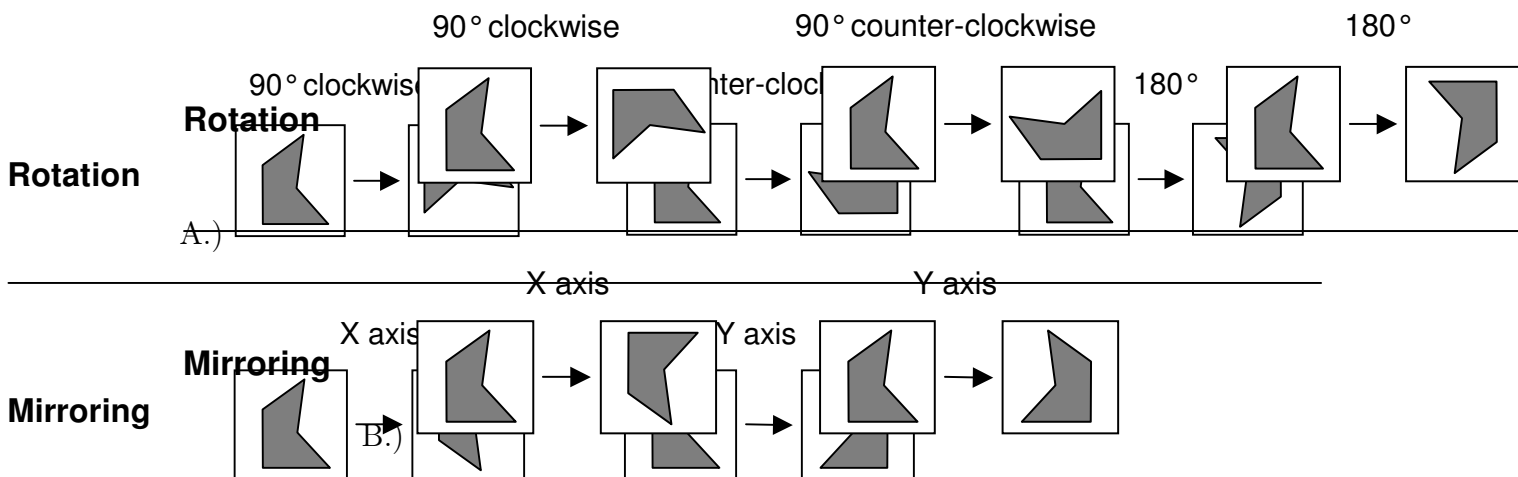


Figure 3.5.
Exemplary illustration of all rules that apply to the main shape, (A.) rotation rules, (B.) mirroring rules

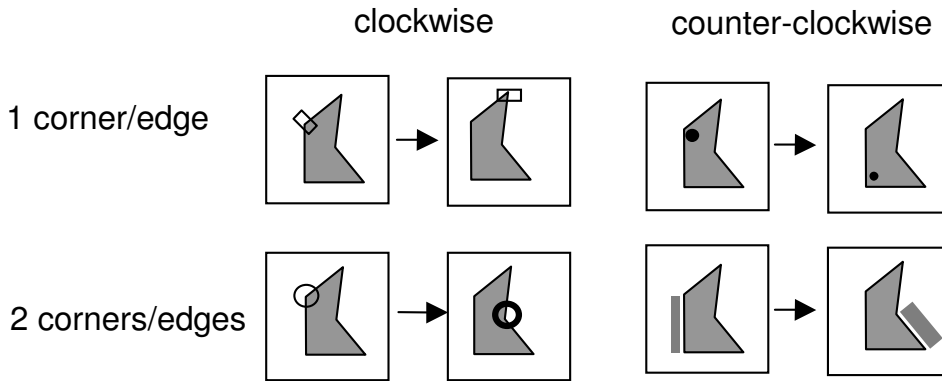


Figure 3.6.

Exemplary illustration of all rules in the FAT that apply to the features

more displacement rules in one item. The spatial displacement rules used here can be divided into two groups, rules that apply to the main form (1-4, see Figure 3.5), and rules that apply to the figural features (5-6, see Figure 3.6); in the new framework all logical connections between the elements **A** and **B** as well as **C** and **D** are exclusively defined by this set of spatial displacement rules:

- R1: *Mirroring at the X-axis (M-X)*: When this rule is applied the main form is reflected at the horizontal axis. This rule separately is illustrated in Figure 3.5. Figure 3.9 demonstrates this rule for an actual item: Here the pentagonal main shape is reflected at the horizontal axis from its orientation in **A** to its orientation in **B**.
- R2: *Mirroring at the Y-axis (M-Y)*: When this rule is applied the main form is reflected at the vertical axis. This rule separately is illustrated in Figure 3.5. The two variants of mirroring were modeled as separate item facets because vertical and horizontal mirroring requires not the same cognitive processes. Due to the item direction from left to right mirroring at the vertical axis could be facilitated by use of a simple flipping strategy (Murray, 1997) whereas this strategy is not applicable as easily for mirroring at the horizontal axis.
- R3: *Rotation of the main shape $\pm 90^\circ$ (R90)*: When this rule is applied a rotation by 90 degrees is applied to the main shape. Rotation could be either clockwise or counterclockwise. This rule separately is illustrated in Figure 3.5. Figure 3.8 (upper part) demonstrates this rule for an actual item: Here the pentagonal main shape is rotated by 90 degrees counter-clockwise from its orientation in **A** to its orientation in **B**.
- R4: *Rotation of the main shape by 180° (R180)*: When this rule is applied the main shape is rotated by 180 degrees. This rule separately is illustrated in Figure 3.5 as well. Figure 3.8 (lower part) demonstrates this rule for an actual item: Here the

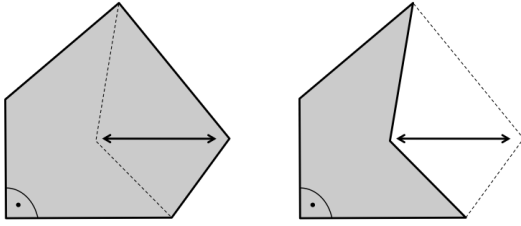


Figure 3.7.

Illustration of the additional complexity parameter “Type of Form” in the FAT; Left: convex shape, without landmark; right: concave shape with landmark; both shapes are identical besides for the position of one corner; the lower left angle is rectangular in both shapes

pentagonal main shape is rotated by 180 degrees from its orientation in **A** to its orientation in **B**.

- R5: *Change of feature-position ± 1 corner/edge (Cp1)*: When this rule is applied, one of the features in the arrangement changes its relative position in space to the main form. The feature affected by this rule is moved one position (i.e., one corner or one edge) clockwise or counterclockwise. This rule separately is illustrated in Figure 3.6. Figure 3.9 demonstrates this rule for an actual item: Here the dot is moved one corner clockwise from its position in **A** to its position in **B**.
- R6: *Change of feature-position ± 2 corners/edges (Cp2)*: This rule is similar to the previous rule: One of the features in the arrangement changes its relative position in space to the main form. The feature affected by this rule is moved two positions (i.e., two corners or two edges) clockwise or counterclockwise. This rule separately is illustrated in Figure 3.6. Figure 3.9 demonstrates this rule for an actual item: Here the beam is moved two edges counter-clockwise from its position in **A** to its position in **B**.

The logical structure of the items is only based on the first category of radicals, spatial displacements. However, three further complexity factors were introduced as radicals in the new item-generation framework. These complexity factors are related to the actual appearance of item features and the number of stimuli used in each figural-spatial configuration. They do not represent item radicals in a narrower sense as they are not associated with the cognitive rules defining the analogy. They rather define the general complexity level of an item independent of the complexity of the analogy (or nuisance factors that make it harder for the individual to detect, recognize and apply the analogy-defining rules).

- R7: *Type of Form (ToF)*: The shape of the pentagonal main shapes used was either convex or concave. Concave main forms were identical with the convex main forms except for one corner that was “moved inwards” (see Figure 3.7 for an illustration).

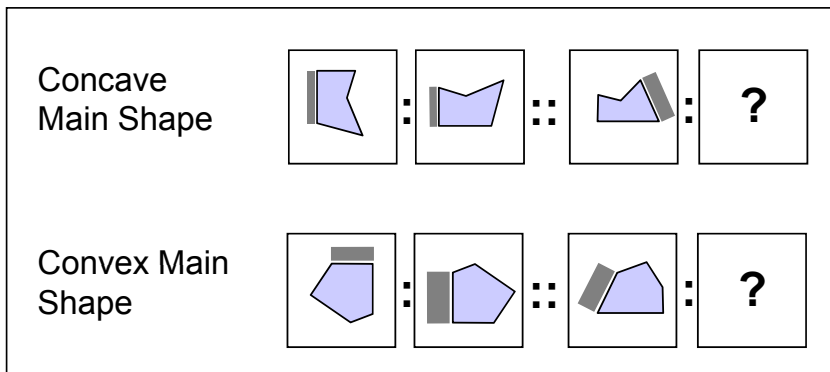


Figure 3.8.

Exemplary illustration of the complexity parameter “Type of Form” of the FAT in two items

Figure 3.7 depicts an item example. In the example shown, two spatial displacement rules are present in both items. Main shapes in the upper example are concave, main shapes in the lower example are convex.

Concave shapes are more distinct and therefore might serve as a “landmark” and thus help to rotate or to mirror the main form. Hochberg and Gellman (1977) showed that “landmarks” in polygons can facilitate mental rotation. Landmarks allow especially for the application of holistic processing strategies, i.e. strategies that are based on a holistic perception of the complete object instead of strategies focusing on a decomposition of the objects into its constituting elements. A higher amount of analytic processing is needed to detect changes in polygons without a distinctive landmark. Figure 3.8 demonstrates this rule for two items: The item in the upper part of the figure was generated based on concave main shapes, the item in the lower part was generated from convex main shapes. In order to reduce the amount of item difficulties not explained by the set of rules, the type of the main shape can only differ between items, not within items here. That is, all shapes in **A**, **B**, **C**, and **D** must fall into the same category of either a concave or a convex shape.

R8: *Additional Feature (AF)*: Every figural-spatial arrangement had one or two figural features. Items comprising the “additional feature”-rule contained one extra feature, i.e. in total three figural features. That is, these items comprised a more complex spatial configuration than the remaining items. This is illustrated in Figure 3.9. The analogy is defined by exactly the same spatial displacement rules in the two examples. However, a third feature (here: rectangle) is added to the second item. This feature does *not* change its position or relative orientation to the main shape but makes the whole figural configuration visually more complex. Note that,

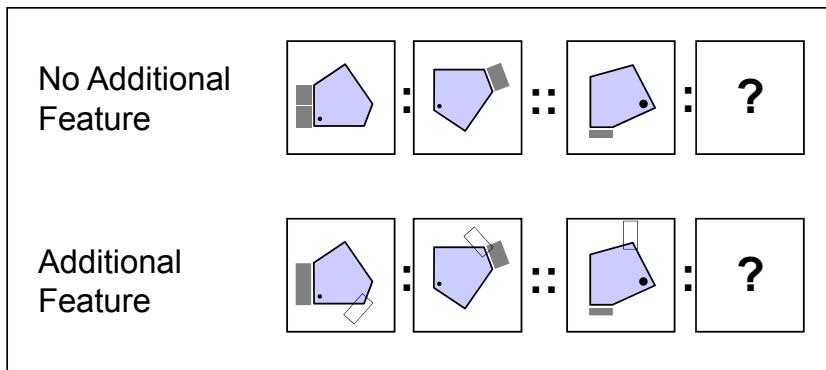


Figure 3.9.

Illustration of the additional complexity parameter “Additional Feature” in the FAT

the item-generative framework allows, in principle, to add as many features to the configuration as perceptually distinguishable. The current parameter value of a maximum of three features in one item was based on a number of smaller pilot studies with university students. A result of these studies was that students tended to be confused and used considerably more time for their answers if more features were combined in one item. It was decided to rather allow for an increase of the number of items that can be presented in a given amount of time than for an increase of item complexity to a maximum. Future studies could, however, investigate this issue further by systematically analyzing items with larger numbers of features.

- R9: *Random Change of Features Characteristics (RCF)*: If this rule was used in an item the surface characteristics of every feature were allowed to vary between **A** and **B** as well as **C** and **D**: The circle always had the same size, but the thickness of the line was allowed to vary. The dot could vary in its diameter. While the length of the beam was always determined by the length of the respective edge of the main shape, its wideness could vary from **A** to **B** or **C** to **D**. For the rectangle, random change was implemented by allowing its relative orientation in space to the main shape to vary.

The complexity parameter RCF was introduced to make sure that individuals working on the test were truly “forced” to represent the spatial relations of the elements to each other instead of trying to solve the items based on surface characteristics. The detection of Random Change of Features required test-takers to follow an analytic processing strategy. An example for the application of this rule is shown in Figure 3.10. Two item examples are shown along with two sets of response alternatives. The spatial displacement rules applied in both items are exactly the same. The main shape is rotated by 90 degrees clockwise, the beam changes its relative position by two edges counter-clockwise, and the circle changes its position by one

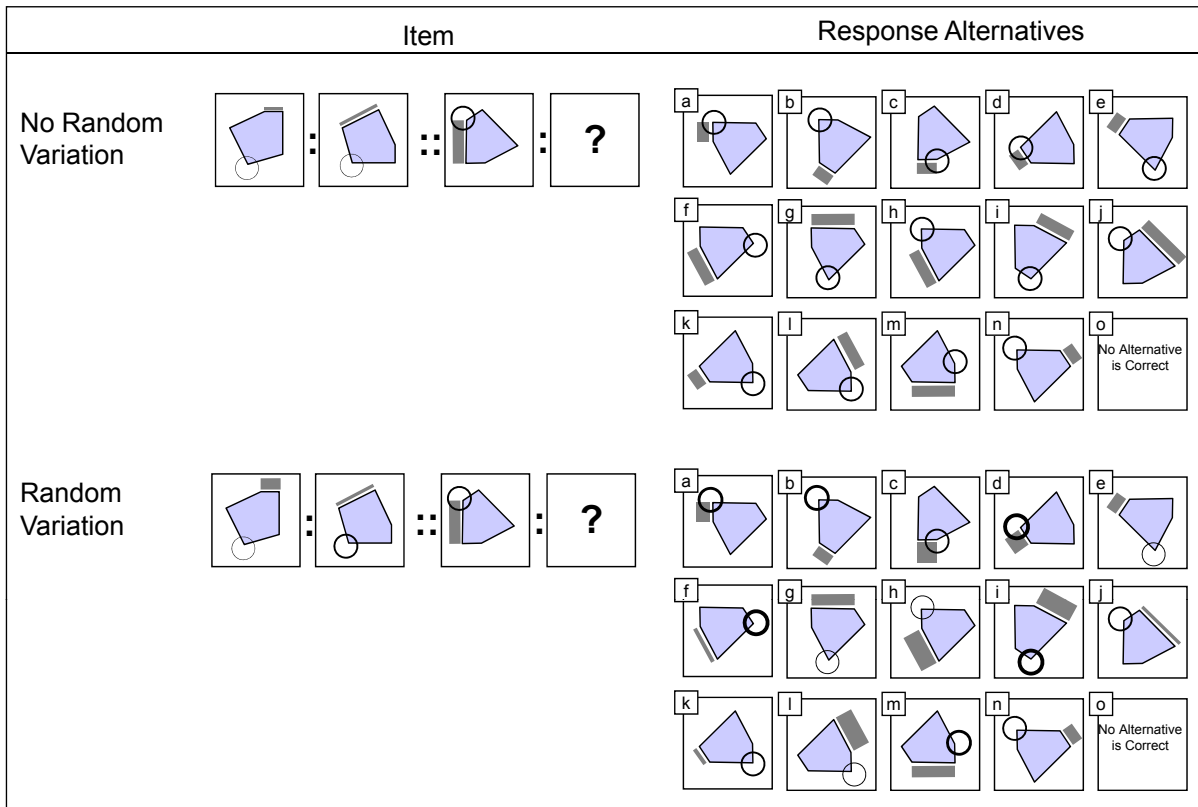


Figure 3.10.

Exemplary illustration of the complexity parameter “Random Change of Feature Characteristics” in two items (Study 1)

corner counter-clockwise. The difference between the two examples is that surface characteristics of the item features are constant in the upper example and allowed to vary in the lower example. This makes especially the distractor set much more heterogeneous. Test-takers need to abstract from these irrelevant changes and focus on the deep structure of the spatial displacements in order to find the correct response alternative. This rule separately is illustrated in Figure 3.6. Figure 3.10 demonstrates this rule for an actual item: Here an item with exactly the same logical structure and the same values on all other parameters is displayed once without RCF (top) and with RCF (bottom).

One problem that can arise in LLTM applications is that not all rules can be freely combined. Table 3.3 shows the possible rule-combinations for the FAT. Three different cases can be distinguished:

1. “✓” indicates that a rule-combination is possible and can be clearly interpreted. This is the case for instance for the combination of a change in a feature position and

Table 3.3.
Possible rule-combinations in the Figural Analogy Test

Rule	MX	MY	R90	R180	Cp1	Cp2	ToF	AF	RCF
MX		= R180	(✓)	= MY	✓	✓	✓	✓	✓
MY	= R180		(✓)	= MX	✓	✓	✓	✓	✓
R90	(✓)	(✓)		= R90	✓	✓	✓	✓	✓
R180	= MY	= MX	= R90		✓	✓	✓	✓	✓
Cp1	✓	✓	✓	✓		✓	✓	✓	✓
Cp2	✓	✓	✓	✓	✓		✓	✓	✓
ToF	✓	✓	✓	✓	✓	✓		✓	✓
AF	✓	✓	✓	✓	✓	✓	✓		✓
RCF	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. ✓: possible combination, (✓): possible but ambiguous combination, “=” indicates that a rule-combination leads to another rule in the set of radicals

rotation of the main shape by 180 degrees. In general, all feature-rules can be freely combined with all main shape-rules. All combinations that fall into this category are reasonable rule-combinations in a new item-generative framework. Here, the order of rule-application does not play a role and the resulting figural configuration can unequivocally be tracked back to the underlying logical processes.

- “(✓)” indicates that a rule-combination is possible but leads to ambiguous results. For instance, when the main shape is reflected at the horizontal axis (MX) and then rotated by 90 degrees clockwise, the resulting main shape orientation is the same as when the main shape is reflected at the vertical axis (MY) and then rotated by 90 degrees counter-clockwise. A correct response of the test-taker can, in this case be not unequivocally tracked back to certain cognitive processes. Especially when it is assumed that the cognitive processes while mirroring an object at the horizontal axis are not identical with those needed to reflect an object at the vertical axis (because of the (non-)availability of simple flipping strategies (cf. Kanamori & Yagi, 2002 or Murray, 1997) such rule-combinations should be avoided when designing new test items.
- The third case arises when the combination of two rules yields a result that is identical to the result from the application of just one rule. That is, one rule is superfluous because the figural transformation can be explained in a simpler way. In the item-generative framework of the FAT this is the case for a combination of two mirroring or two rotation rules. For instance, when the main shape is first rotated by 180 degrees and after that rotated by 90 degrees, the result is, again, a 90 degree rotation. If the main shape is rotated by 180 degrees and then reflected at the horizontal axis, the resulting orientation is identical to a reflection at the

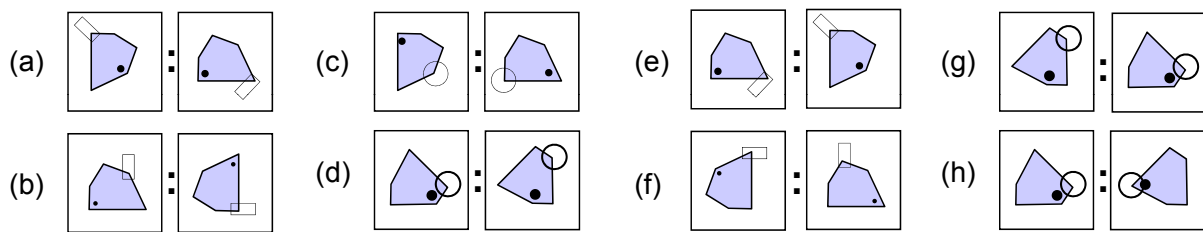


Figure 3.11.

Examples for Incidentals in Study 1; (a) - (h) are all based on the same radicals

vertical axis. In the design matrix for the FAT, it was decided to code always the most simple explanation for a given stimulus transformation.

Item incidentals Figure 3.11 shows eight sets of stimulus combinations **A** : **B** that are constructed from exactly the same item radicals, R90, CP1, and CP2. The different “look” of the stimulus pairs is due to variation of incidentals. Given that the radicals capture the main sources of variation in item difficulty there should not be any systematic differences in difficulties for all eight items. In the extreme case of a within-family variance of zero (i.e. a perfect explanation of item-difficulties by the set of radicals) the difficulty parameters for the eight items should be exactly the same. In the following, all item incidentals are described in detail. In total, 7 incidentals can be distinguished.

- i1 *Individual main shape*: First, the exact variant of the main shape was defined as incidental. 4 figural shapes were designed in each main shape category (see Figure 3.12). That is, four concave main shapes and four convex main shapes were designed. All shapes share a set of characteristics. They are all pentagons and all shapes have exactly one right angle corner. Also, the sizes of the shapes are very similar. All shapes are abstract and purely figural. There is no semantic meaning in any of the shapes and it was taken care that none of them resembles symbols that are used in language or mathematics. Several pretests with university students did not indicate any considerable differences in terms of the distinctiveness of these stimuli. In the current item-generative framework, each main shape had a chance of $\frac{1}{4}$ to be chosen. Extensions to a larger number of shapes is possible.
- i2 *Starting orientation of the main shape*: Second, the starting orientation of the main shape chosen was set as incidental as well. Beckmann (2008) showed that the starting orientation of alphanumeric stimuli actually influenced the difficulty of an analogy. This is because letters and numbers might be recognized more easily when they are presented in their “canonical” orientation. The speed of object recognition varies as a function of the depiction of an object in its normal, upright, “canonical” orientation or in a deviating orientation (see e.g., Jolicoeur, 1985; Lawson, Humphreys, & Jolicoeur, 2000). For such abstract stimuli as the pentagons used as

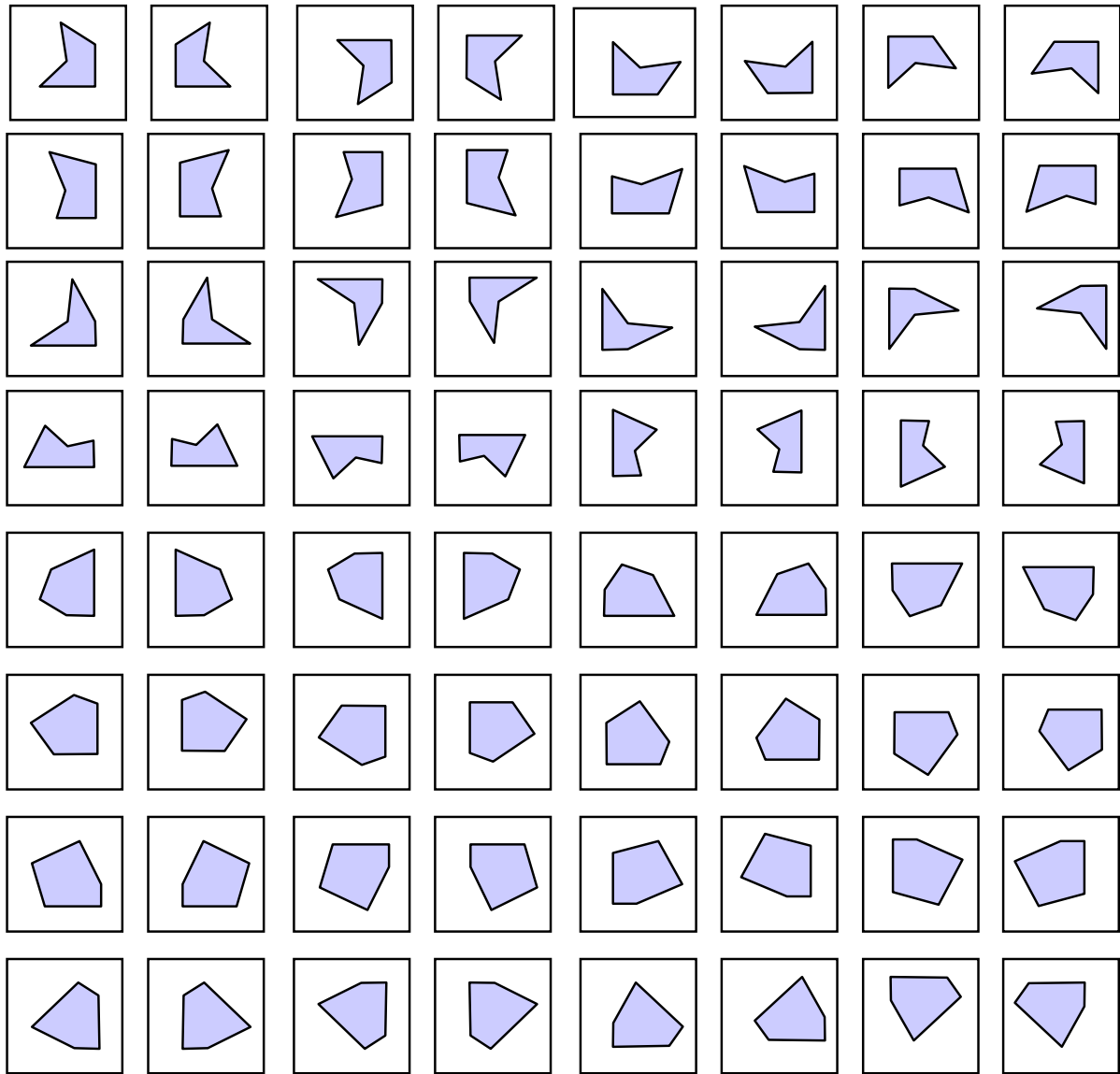


Figure 3.12.

Overview of all possible main shapes in the current AIG framework for the FAT; Each of the 8 main shapes is depicted in all four possible orientations; upper half: concave; lower half: convex

main shapes, there is no such one canonical orientation. However, due to the one right angle corner it is a reasonable assumption that recognition and manipulation of a shape is facilitated when the right angle corner corresponds to the axes of a coordination system. To account for this the random choice of a starting orientation was constrained to be a choice of only a limited number of possible orientations. The

different orientations differed from each other exactly by 90 degree rotations. That is, whatever orientation was chosen, the right angle corner corresponded exactly to the horizontal and vertical axes. In addition, all shapes were allowed to be reflected on the vertical axis, resulting in twice the number of possible stimuli. Figure 3.12 depicts all possible shapes and possible starting orientations. In total, $8 \times 8 = 64$ different main shapes can be distinguished, 32 variants if $ToF = 0$ and 32 variants if $ToF = 1$. In the current item-generative framework, each specific combination of main shape and starting orientation had a chance of $\frac{1}{32}$ to be chosen after having chosen the category of shape in the first step. The choice of a shape for **C** was constrained by the rules that (a.) the same category of shape had to be chosen and (b.) the actual form of the shape in **A** and **C** were not allowed to be the same. That is, after having determined a specific shape for **A** for each item there were 3×8 possible shapes (3 shapes in 8 different orientations each) to be chosen for **C**.

Figure 3.11 provides examples for the manipulation of these incidentals. For instance, the initial orientation of the main shape differs between (a) and (b). In (a) the right angle corner is in the upper left while in (b) this corner is in the lower left.

- i3 *Figural features*: Third, the specific features used in an item are chosen randomly with the constraint that each feature can appear only once in an item. For instance, for an item with two features, there are 6 possible combinations: {beam, point}, {beam, circle}, {beam, rectangle}, {circle, point}, {circle, rectangle}, {rectangle, point}. This is demonstrated in Figure 3.11 as well. (a) and (b), for instance, use rectangle and point, (c) and (d) use circle and point.
- i4 *Surface characteristics of figural features*: As with the main shapes, the actual features was chosen randomly. To allow for considerable variation in design across items and in order to allow for the variation of surface characteristics (see radical 9), the specific layout of each feature could vary across different items. The circle always had the same size, but the thickness of the line was allowed to vary between 1/4 and 2 pt. The diameter of the dot varied between 1 and 3 pt. While the length of the line/beam was always determined by the length of the respective edge of the main shape, its wideness could vary between 2 and 6 pt. The orientation of the rectangle varied freely in 30 degree steps. These ranges of parameter values were chosen based on a few pretests with university students. It was attempted to choose values that were clearly distinguishable but not too extreme to draw attention away from the underlying structural differences in the analogies.
- i5 *Starting position of figural features*: The starting position for all features relatively to the main shape was chosen randomly. For example, in Figure 3.11, a different main shape was chosen in (d) compared to (a).
- i6 *Direction of rotation*: When radical 3 (rotation by 90 degrees) was applied to an item, the direction of the rotation was set as incidental. That is, it was randomly

determined whether a main shape was rotated clock- or counterclockwise. The decision for defining the direction of rotation as incidental was made based on findings reported by Beckmann (2008). She demonstrated that difficulties did not differ between clockwise and counterclockwise rotations.

- i7 *Direction of change of feature positions*: Again, it is not distinguished between the direction of change of feature positions. That is, based on Beckmann's (2008) findings change of a feature position by one corner clockwise is assumed to be as difficult as a change in position one corner counterclockwise. This is depicted in Figure 3.11.

The same item facet parameters can be used during the generation of distractor stimuli among the multiple choice answer alternatives. A detailed description of the technical steps of item generation is given by Bertling and Holling (2009).

3.2.2. Specific hypotheses

Based on the item-generative rules described in the previous section, a number of specific hypotheses regarding the construct representation of the new item type were tested in addition to the investigation of the research questions lined out in section 3.1.3. These hypotheses are related to the third research question, that is the question whether the parameter estimates of the item-difficulty model are in line with theories of figural-spatial processing and analogical reasoning.

The first four hypotheses refer to the general item radical functioning and their consistency with the cognitive model.

- *Hypothesis 1*: All spatial displacement rules increase item difficulty.
- *Hypothesis 2*: Usage of convex (i.e., less complex) polygons instead of concave (i.e., more complex) polygons decreases item difficulty.
- *Hypothesis 3*: Not the number of elements but the number of relations between the elements of each configuration determines the complexity of an item. The addition of a third additional feature to the figural configuration therefore does not influence item difficulty.
- *Hypothesis 4*: Random change of features will make encoding of the relation between the "A" and "C" term, i.e. the preparation period of analogical reasoning harder and therefore increase item difficulty.

Hypotheses 5 to 7 address the typical gender differences in figural-spatial processing that were described in the introduction. As not all item radicals are directly related to processes of mental rotation and spatial reasoning, differential effects for some item radicals are hypothesized for female and male test-takers. It is hypothesized that the impact on item difficulty of the following item radicals will be moderated by gender:

- *Hypothesis 5*: Rules requiring mental rotation skills are easier for men.
- *Hypothesis 6*: Rules that require the application of analytic processing strategies are easier for women.
- *Hypothesis 7*: Use of convex (i.e., less complex) polygons instead of concave (i.e., more complex) polygons decreases item difficulty for men more strongly than for women. Convex polygons can be processed easier following a holistic processing strategy that is based on the landmarks in such shapes.

3.2.3. Sample

Participants were recruited at a large German university and received feedback of their results as an incentive. The total sample consists of $N = 308$ individuals (76.3 % female). The mean age was 22.51 years ($SD = 4.19$). 113 participants (36.8%) reported prior experience with IQ tests. This percentage of prior experience is representative (cf. meta-analytic findings by Hausknecht, Halpert, Di Paolo, & Gerrard, 2007). All participants gave consent that their data be used for scientific purposes.

3.2.4. Instruments and procedure

Figural Analogy Test (FAT) A new 40-item version of the FAT was constructed manually by the author. All items were checked by several student assistants. All 40 items are given in the Appendix of this thesis. The principles of rule-based AIG were used to generate items as well as wrong answer options. Wrong answers were generated as partly correct stimuli showing different degrees of discrepancy to the right answer alternative. Thereby typical problems with regard to distractors that have been reported previously (e.g., Mittring & Rost, 2008) were prevented in the new instrument. In order to meet the demands of Mittring and Rost (2008) that—across all items of a test version—the test-taker should be unable to make inferences regarding the right answer by looking only at the set of response alternatives, frequencies of features and main forms were combined in a way that did not allow for counting strategies. In each distractor set one false feature position or main form orientation appeared more often than the correct one. This way it was guaranteed that a counting strategy would not lead to the correct solution.

For the version of the FAT administered here, four items (i.e. 10 percent) without a correct solution among the distractor set were administered. Here, the option “No correct answer” had to be checked to answer the item correctly. Consistent with Gittler (1990) this answer option was added to ensure that participants could not rely on a falsification strategy by excluding distractors (see also Preckel, 2003). With this option being possible for every item the correct solution could only be identified by means of a verification strategy. Participants could also choose the answer option “I don’t know the answer”,

which was added to minimize guessing effects (cf. Gittler, 1990). In total, each item had 15 response alternatives. A large number of response alternatives was chosen for the first empirical evaluation of the test to minimize the possibility of guessing effects.

The individual items featured between 2 and 6 item facets. Items were grouped in four blocks of 10 items each. The design matrix was constructed based on optimal design analyses using the OPTEX procedure in SAS. D-, A-, and G-efficiency criteria were calculated. The design matrix for all 40 items is given in the Appendix along with the input and output files for the optimal design derivation.

Based on several smaller pilot studies, a time-limit of 45 seconds per item was chosen. Five additional seconds were given to mark responses on an answering sheet, yielding a total maximum response time per item of 50 seconds. Compared to average response times allowed for widely used reasoning tests, timing was allocated rather amply here (e. g., in the CFT-20R (Weiß, 2007), 56 items have to be completed in a total time of 14 min, yielding an average response time per item of 15 seconds). After each block of items there was a short break to give the participants some relief. Including the time for the instruction the total test time for the FAT added up to a maximum of 50 minutes per test-taker.

Preceding the actual assessment, a detailed description of the item type was provided. All rules were explained in detail including examples for each rule. Participants were informed that several rules could be combined in one item. The multiple possible combinations of rules were *not* instructed. After this instruction, all participants were asked whether they had understood the rules. Only data from participants who had understood all rules (which was, actually, true for all individuals in the sample) was used for consecutive analyses. Test-takers worked on 4 warm-up items before starting to work on the the actual 40 test items. Anastasi (1981) recommended to implement such short orientation and warm-up sessions to establish comparable testing conditions for all subjects. In this case, subjects can learn the correct solution strategies and then utilize them on subsequent problems (Verguts & De Boeck, 2002). It is the ability to learn and implement strategies across problems that is, according to this view, important for reasoning performance. Beckmann (2008) demonstrated that an explanation of all rules beforehand can actually amplify the validity of a reasoning test.

General intelligence As a measure of general cognitive ability g , the four subtests from the revised german version of the Culture-Fair Test (CFT 20-R; Weiß, 2007), were utilized. The CFT 20-R is a paper-and-pencil test which provides high loadings on fluid intelligence and has good psychometric properties. It consists of four different subtests: Series completion, Classifications, Matrices and Topologies. The Topology subtest shows the largest similarity with the FAT. It requires to represent the spatial relations between several figural elements and map stimulus configurations to each other according to the structure of these spatial relations.

Table 3.4.
Descriptives for the FAT and other measures used (Study 1)

Instrument	<i>k</i>	<i>M</i>	<i>SD</i>	Min	Max	Skewness	Kurtosis
Figural Analogy Test (FAT)	40	14.93	6.29	3.00	36.00	0.60	0.13
CFT 20-R Test 1	15	13.14	1.61	7.00	15.00	−0.89	0.47
CFT 20-R Test 2	15	11.57	2.00	4.00	15.00	−0.64	0.14
CFT 20-R Test 3	15	12.21	1.88	3.00	15.00	−1.12	2.93
CFT 20-R Test 4	11	7.68	1.64	2.00	11.00	−0.50	0.54
CFT 20-R Total	56	44.60	4.91	28.00	56.00	−0.57	0.42
3DW	15	7.05	3.74	1.00	15.00	0.05	−0.73
Grade compound (GPA)	1	11.73	1.78	6.80	14.67	−0.86	−0.01
Math Grade	1	11.61	3.01	1.00	15.00	−1.24	1.17

Spatial ability As a measure capturing spatial abilities, Gittler’s (1990) threedimensional cube test (“Dreidimensionaler Würfetest”, 3DW) was administered. The 3DW is a Rasch-scaled paper-and-pencil test that measures spatial ability as one of the primary factors of human intelligence. Every item contains one drawing of a three-dimensional cube with different symbols on each side. The test-taker has to make the decision which out of 6 possible cubes is the same cube in a different spatial orientation. The specific usefulness of three-dimensional cube items in the diagnostics of mental rotation ability has been demonstrated as well as its strong relationship with general cognitive ability *g* (e.g., Gittler, 1999). Subjects worked on 15 items under power-conditions in accordance with the guidelines provided by Gittler.

High school grades Participants were asked to report their most recent grades in math, science, languages (German, English), and arts. A compound score similar to GPA was computed as the arithmetic average of these grades.

3.3. Results

Table 3.4 shows the most important descriptive statistics for all instruments administered. On average, subjects were able to solve 14.93 items ($SD = 6.285$) correctly, with considerable variation among test scores. The minimum score achieved was 3 items whereas the best test-taker was able to solve 36 items correctly. Probabilities for a correct response cover the whole range from 6% (Item 40) to 94% (Item 9) with a mean probability of correct response of 37.3%. No difference between female and male subjects with regard to general cognitive ability as measured by the CFT-20r was found ($t(284) = -0.834$; $p = .405$), whereas test performance differed significantly both on the

3DW ($t(306) = -3.697$; $p < .001$) and the FAT ($t(306) = -3.051$; $p = .002$), in both cases favoring men. Effect sizes for these gender differences were $d_{3DW} = 0.49$ and $d_{FAT} = 0.39$, respectively. Internal consistencies (Cronbach's α) for the three cognitive measures were $\alpha_{3DW} = .811$, $\alpha_{CFT} = .732$, and $\alpha_{FAT} = .827$.

3.3.1. Prediction of item difficulty parameters

Rasch model parameters and fit statistics for all 40 items are summarized in Table 3.5, along with classical item statistics. For the assessment of item fit, z -transformed Q -indices (Rost & Davier, 1994) as well as Infit and Outfit statistics (see Linacre, 2010) were computed. The estimated item difficulty parameters σ_i range from -2.47 to 3.72 . The combination of spatial displacement rules yields items that cover a wide range of the ability continuum and are well-suited to assess reasoning ability across the whole scale with a slight focus on the upper half of the ability scale (total information $I = 33.26$, 34.53% in $(-5, 0)$ and 58.68% in $(0, 5)$, respectively). None of the 40 items shows considerable misfit to the RM. 3 items (items 23, 24, and 35) show slight misfit based on the Q index. However, these items fit well on both the outfit and the Infit statistic. The item characteristic curves (ICCs) for all items are given in the Appendix.

Two LLTM models of different complexity were estimated to test the appropriateness of the set of pre-specified item radicals to model item difficulties. The first Model is a LLTM model that includes only spatial displacement rules (in the following called LLTM 1). Model 2 includes the three additional complexity parameters as well (in the following

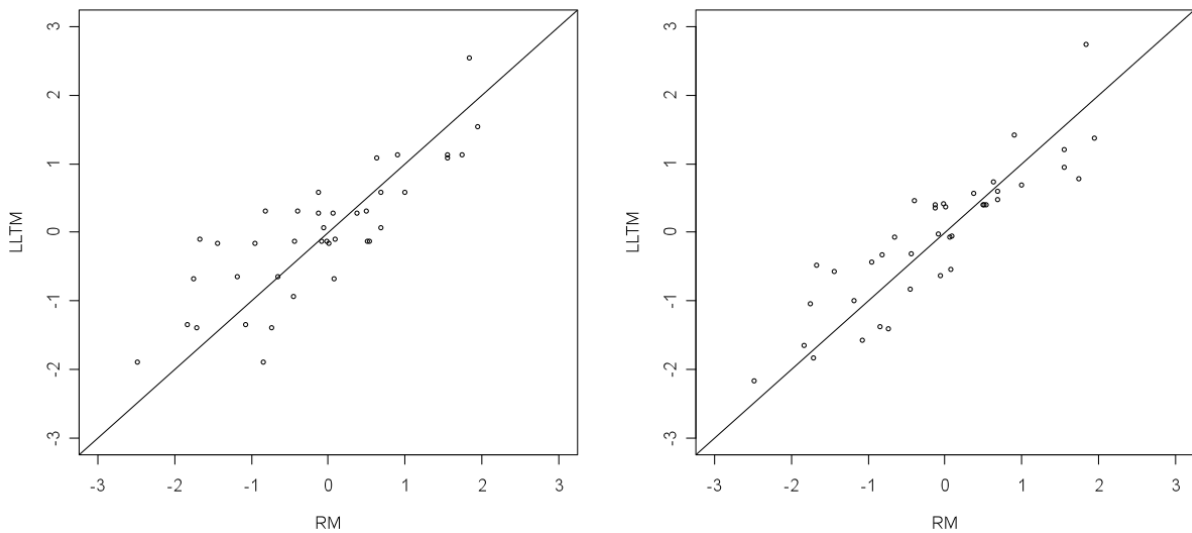


Figure 3.13.

RM and rescaled LLTM parameters; left side: LLTM 1; right side: LLTM 2

Table 3.5.

Classical item statistics, Rasch parameters, and item fit statistics (Study 1)

<i>i</i>	Classic. Stat.		Descriptive IRT					Explanatory IRT	
	<i>p</i>	<i>r_{it}</i>	RM σ_i	SE(σ_i)	<i>p</i> (<i>Q</i>)	Outfit	Infit	LLTM 1	LLTM 2
01	.734	.253	1.85	0.14	.45	1.00	0.99	1.68	2.18
02	.468	.284	0.55	0.12	.28	1.06	1.03	-0.16	0.20
03	.237	.282	-0.64	0.14	.32	1.04	1.01	-0.68	-0.26
04	.279	.418	-0.39	0.13	.94	0.86	0.91	0.42	0.29
05	.373	.220	0.10	0.13	.06	1.11	1.09	-0.14	-0.43
06	.461	.220	0.52	0.12	.07	1.13	1.09	-0.16	0.20
07	.162	.332	-1.17	0.16	.79	0.85	0.94	-0.68	-1.37
08	.110	.248	-1.66	0.19	.50	1.00	0.95	-0.14	-0.91
09	.942	.126	3.72	0.24	.47	1.03	0.93	1.68	2.66
10	.682	.273	1.57	0.13	.42	1.04	1.00	1.13	1.46
11	.568	.300	1.01	0.12	.41	1.00	1.01	0.59	1.00
12	.104	.254	-1.74	0.19	.56	1.02	0.93	-0.71	-1.41
13	.351	.323	-0.01	0.13	.48	1.00	1.00	-0.17	0.18
14	.370	.340	0.09	0.13	.57	0.99	0.98	-0.71	-0.77
15	.367	.427	0.07	0.13	.93	0.86	0.91	0.39	-0.38
16	.357	.332	0.02	0.13	.53	0.99	0.99	-0.19	0.17
17	.500	.370	0.70	0.12	.73	0.94	0.96	0.00	0.77
18	.545	.342	0.91	0.12	.66	0.95	0.97	1.13	1.44
19	.211	.344	-0.81	0.15	.72	0.92	0.96	0.42	-0.35
20	.753	.204	1.96	0.14	.27	1.15	1.02	1.68	1.36
21	.718	.377	1.76	0.13	.89	0.86	0.88	1.13	0.98
22	.487	.315	0.64	0.12	.48	1.00	1.00	1.11	0.97
23	.331	.478	-0.11	0.13	.99	0.78	0.86	0.39	0.10
24	.133	.094	-1.43	0.17	.01	1.27	1.12	-0.19	-0.95
25	.458	.295	0.51	0.12	.33	1.07	1.02	0.42	0.79
26	.682	.200	1.57	0.13	.18	1.17	1.04	1.11	0.96
27	.500	.396	0.70	0.12	.86	0.89	0.93	0.59	0.50
28	.331	.367	-0.11	0.13	.72	0.96	0.95	0.59	0.34
29	.338	.399	-0.08	0.13	.87	0.89	0.93	-0.17	0.35
30	.273	.394	-0.43	0.14	.87	0.91	0.92	-0.16	-0.28
31	.208	.326	-0.83	0.15	.67	0.94	0.96	-1.97	-0.87
32	.107	.147	-1.70	0.19	.05	1.12	1.09	-1.45	-1.53
33	.269	.370	-0.45	0.14	.82	0.88	0.95	-0.88	-0.32
34	.224	.251	-0.72	0.14	.16	1.17	1.02	-1.45	-1.05
35	.432	.176	0.38	0.12	.01	1.23	1.13	0.39	0.73
36	.192	.281	-0.95	0.15	.44	0.95	1.00	-0.19	-0.47
37	.344	.295	-0.04	0.13	.33	1.03	1.02	0.00	-1.01
38	.097	.248	-1.82	0.20	.50	1.28	0.93	-1.43	-1.70
39	.175	.236	-1.07	0.16	.23	0.98	1.03	-1.43	-1.54
40	.055	.268	-2.47	0.25	.86	0.87	0.89	-1.97	-2.16

Note. *p* = average proportion of correct solution across all test-takers; *r_{it}* = Item-total correlation; *p*(*Q*) = significance of *Q* index; LLTM 1= rescaled item difficulty parameters based on LLTM with spatial displacement parameters only; LLTM 2= rescaled item difficulty parameters based on LLTM with extended design matrix including both spatial displacement and additional complexity parameters.

Basic LLTM			Extended LLTM		
Frequency	Stem &	Leaf	Frequency	Stem &	Leaf
6.00	0 .	023678	7.00	0 .	1255789
10.00	1 .	1226788999	8.00	1 .	01224566
3.00	2 .	249	10.00	2 .	1245677789
6.00	3 .	134477	6.00	3 .	002459
3.00	4 .	689	6.00	4 .	122269
6.00	5 .	224668	1.00	5 .	6
2.00	6 .	29	1.00	6 .	6
2.00	7 .	59	1.00	7 .	3
2.00	8 .	12			
Stem width:	1.00		Stem width:	1.00	
Each leaf:	1 case(s)		Each leaf:	1 case(s)	

Figure 3.14.

Distribution of standardized absolute errors for LLTM 1 and LLTM 2 (Study 1)

called LLTM 2). The two rightmost columns in Table 3.5 show rescaled item difficulty parameters based on the basic and extended LLTM. Altogether, the 6 spatial displacement rules in LLTM 1 explain $R^2 = 69\%$ of variation in item difficulties. Item random effect variance is reduced from $s^2 = 1.456$ in a model without any item predictor to 0.456 in a model with the 6 item covariates. When the three additional complexity parameters are included as item predictors in LLTM 2, R^2 for the explanation of variation in item difficulties is increased to $R^2_{\text{Model B}} = .86$ ($\Delta\chi^2(3) = 18.59, p < .001$). Item random effect variance is reduced from $s_e^2(I) = 0.46$ in LLTM 1 by nearly 40 percent to $s_e^2(I) = 0.21$ in LLTM 2.

Figure 3.13 displays the accuracy of Rasch difficulty parameters by the combination of item facets in LLTM 1 and LLTM 2. Rasch item difficulties are plotted on the horizontal axis against reconstructed LLTM item difficulties on the vertical axis. In both models no systematic bias is visible, i.e. some rescaled item parameters overestimate difficulty and others underestimate the difficulty of an item. As can be seen in Figure 3.13, rescaled item difficulties lie closer to actual item difficulties in LLTM 2 compared to LLTM 1. This is also reflected by the model fit indices, both favoring the model with the extended design matrix that does not only comprise spatial displacement rules but also takes other complexity factors into account ($\text{AIC}_{\text{basic}} = 13179.668$ versus $\text{AIC}_{\text{ext}} = 13156.468$; $\text{BIC}_{\text{LLTM1}} = 13213.239$ versus $\text{BIC}_{\text{LLTM2}} = 13201.229$). The change in model fit is significant (LR χ^2 for a comparison of the LLTM with the basic and extended design matrix: $\chi^2(3) = 29.201, p < .001$).

In addition to these classical model comparisons of the LLTM and Rasch Model, standardized absolute errors were analyzed as proposed by Zeuch, 2011: Absolute differences

between Rasch and rescaled LLTM item difficulty parameters were computed and standardized by dividing them by the standard errors of the Rasch parameters. Table 3.6 summarizes the results for the two alternative LLTM models. Figure 3.14 shows the distribution of standardized absolute differences. On average, standardized absolute errors denote to $M_{LLTM1} = 3.521$ ($SD = 2.390$) in LLTM 1 and $M_{LLTM2} = 2.688$ ($SD = 1.720$) in LLTM 2. Unstandardized absolute differences between Rasch and predicted difficulty parameters denote to $M_{LLTM1(us)} = 0.54$ logits in LLTM 1 and $M_{LLTM2(us)} = 0.39$ logits in LLTM 2. In the model with the basic design matrix, 21 differences are larger than 3 standard error units and 15 exceed 4 standard error units. The number of such extreme differences is reduced to 15 and 9 in the model with an extended design matrix. The absolute errors are reduced by 23.05% from LLTM 1 to LLTM 2. These values are in about the range as values reported by Zeuch (2011) for another figural reasoning measure. The implications of such extreme errors in the prediction of true item parameters by means of the LLTM will be discussed in section 3.4.2.

3.3.2. Construct validity

Hypotheses 1 to 4 were tested by investigating the LLTM facet parameters for both the LLTM with spatial displacement parameters only and the LLTM with an extended design matrix. Results for both models are shown in Table 3.7. All item facets representing spatial displacement rules contribute significantly to item difficulties, confirming Hypothesis 1. Rotation by 180 degrees is the most difficult item facet, while mirroring at the vertical axis is the easiest rule. When the main form is rotated by 180 degrees, the logit for a correct response to the respective item is decreased by nearly 2 points on the logit scale ($\beta = -1.973$, 95% CI $[-2.905, -1.041]$) compared to an item where no rotation has to be performed. Mirroring at the vertical axis is associated with considerably less increase in difficulty, i.e. the logit is changed by -0.876 (95% CI $[-1.806, 0.054]$).

All three additional complexity parameters are statistically significant:

1. Changes in the type of shape of the main form (convex vs. concave) lead to changes in the logit by -0.64 (95% CI $[-0.948, -0.332]$). This change is statistically significant, confirming Hypothesis 2.
2. When an additional feature is added to the figural configurations in “A”, “B”, “C”, and “D”, the logit increases by 0.66 (95% CI $[0.324, 0.996]$). This result is not in line with the expectations formulated by Hypothesis 3.
3. When the appearance of item features is allowed to vary (i.e., random variation of surface characteristics is allowed), difficulty is significantly increased, confirming also Hypothesis 4 ($\beta = 0.48$, 95% CI $[-0.787, -0.171]$).

Hypotheses 5 to 7 were tested based on explanatory IRT models including additional facet*gender interactions. Estimation results for these two models are given in Table

Table 3.6.

Standardized absolute errors for the alignment of rescaled LLTM and Rasch parameters (Study 1)

i	Standardized Absolute Error	
	LLTM 2	LLTM 1
FAT01	2.415	1.167
FAT02	2.810	5.678
FAT03	2.653	0.253
FAT04	4.921	5.885
FAT05	4.160	1.877
FAT06	2.567	5.432
FAT07	1.224	2.996
FAT08	3.938	7.956
FAT09	4.287	8.207
FAT10	0.783	3.318
FAT11	0.130	3.404
FAT12	1.690	5.230
FAT13	1.423	1.228
FAT14	6.662	6.203
FAT15	3.531	2.458
FAT16	1.112	1.644
FAT17	0.551	5.632
FAT18	4.290	1.765
FAT19	3.060	8.189
FAT20	4.223	1.941
FAT21	5.683	4.620
FAT22	2.701	3.765
FAT23	1.556	3.776
FAT24	2.700	6.967
FAT25	2.257	0.677
FAT26	4.627	3.474
FAT27	1.614	0.899
FAT28	3.405	5.290
FAT29	3.281	0.714
FAT30	1.065	1.943
FAT31	0.226	7.561
FAT32	0.901	1.271
FAT33	0.890	3.102
FAT34	2.197	4.960
FAT35	2.793	0.037
FAT36	3.065	4.875
FAT37	7.394	0.319
FAT38	0.596	1.896
FAT39	2.925	2.278
FAT40	1.215	1.941

Note. LLTM 1: standardized absolute difference between rescaled item difficulty parameters based on model with spatial displacement parameters only and original Rasch parameters; LLTM 2: standardized absolute difference between rescaled item difficulty parameters based on extended LLTM with spatial displacement and additional complexity parameters and original Rasch parameters.

Table 3.7.

Explanatory IRT modeling for the FAT: comparison of a LLTM including spatial displacement rules only (LLTM 1) and an extended LLTM with three additional complexity parameters (LLTM 2).

<i>Fixed Effects</i>	Empty Model		LLTM 1		LLTM 2	
	Est	SE	Est	SE	Est	SE
Constant	-0.687**	0.198	1.8542**	0.4522	1.9084**	0.3585
<i>(Spatial Displacements)</i>						
mx			-1.4324**	0.4648	-1.1680**	0.3326
my			-0.8756*	0.4643	-0.6100*	0.3325
r90			-1.4529**	0.4565	-1.1764**	0.3263
r180			-1.9726**	0.4655	-1.6330**	0.3372
cp1			-1.2642**	0.2209	-1.2338**	0.1544
cp2			-1.2961**	0.2208	-1.2662**	0.1543
<i>(Additional Complexity Parameters)</i>						
tof					-0.6375**	0.1540
fp					0.6573**	0.1678
rcf					-0.4785**	0.1539
<i>Random Effects</i>	VAR	SE	VAR	SE	VAR	SE
$s_e^2(\text{Item})$	1.456	0.334	0.4562	0.1089	0.2108127	0.0524
$\Delta s_e^2(\text{Item})$	0.00%		-68.66%		-85.52%	
$s_e^2(\text{Person})$	0.685	0.071	0.6854	0.0705	0.6851032	0.0704
$\Delta s_e^2(\text{Person})$	0.00%		0.00%		-0.04%	
<i>Model Fit</i>						
<i>ll</i>	-6603.22		-6580.8342		-6566.2339	
<i>df</i>	3		9		12	
<i>AIC</i>	13212.44		13179.6684		13156.4678	
<i>BIC</i>	13223.63		13213.2393		13201.229	

Note. Delta parameters $\Delta s_e^2(\text{Item})$ and $\Delta s_e^2(\text{Person})$ quantify the reduction in random effect variance in comparison to the empty model (RE-RM; cf. De Boeck, 2008) containing only a constant and neither person nor item predictors; information criteria for best model fit in boldface.

Table 3.8.
Explanatory IRT modeling for the FAT: gender-effects

<i>Fixed Effects</i>	LR-LLTM gender		LR-LLTM Gender*Facet	
	Est	SE	Est	SE
Constant	1.8234**	0.3594	1.8982**	0.3650
mx	-1.1681**	0.3326	-1.1297**	0.3375
my	-0.6100**	0.3325	-0.6602**	0.3374
r90	-1.1764**	0.3262	-1.2763**	0.3314
r180	-1.6330**	0.3371	-1.7455**	0.3427
cp1	-1.2338**	0.1544	-1.2481**	0.1572
cp2	-1.2661**	0.1543	-1.2719**	0.1571
tof	-0.6377**	0.1540	-0.6843**	0.1567
fp	0.6572**	0.1678	0.6183**	0.1706
rcf	-0.4785**	0.1539	-0.4360**	0.1565
gender	0.3604**	0.1206	0.0745	0.2615
gender*mx			-0.1619	0.2209
gender*my			0.2140	0.2173
gender*r90			0.3933*	0.2136
gender*r180			0.4369*	0.2232
gender*cp1			0.0460	0.1070
gender*cp2			-0.0021	0.1067
gender*tof			0.1826*	0.1049
gender*fp			0.1536	0.1139
gender*rcf			-0.1746*	0.1040
<i>Random Effects</i>	VAR	SE	VAR	SE
$s_e^2(\text{Item})$	0.2108	0.0524	0.2121	0.0527
$\Delta s_e^2(\text{Item})$	-85.52%		-85.43%	
$s_e^2(\text{Person})$	0.6613	0.0684	0.6644	0.0687
$\Delta s_e^2(\text{Person})$	-3.52%		-3.06%	
<i>Model Fit</i>				
<i>ll</i>	-6561.8321		-6548.3568	
<i>df</i>	13		22	
<i>AIC</i>	13149.6642		13140.7136	
<i>BIC</i>	13198.1555		13222.7758	

Note. $\Delta s_e^2(\text{Item})$ and $\Delta s_e^2(\text{Person})$ quantify the reduction in random effect variance in comparison to the empty model (RE-RM; cf. De Boeck, 2008) containing only a constant and neither person nor item predictors; information criteria for best model fit in boldface; * $p < .05$ and ** $p < .01$ (one-sided)

Table 3.9.
Correlations between FAT scores and other variables (Study 1)

	FAT	CFT1	CFT2	CFT3	CFT4	CFT total	3DW	GPA	Math
FAT	1	.371**	.367**	.410**	.423**	.569**	.597**	.199**	.299**
CFT1	.371**	1	.307**	.249**	.227**	.624**	.313**	.085	.095
CFT2	.367**	.307**	1	.321**	.308**	.734**	.345**	.215**	.218**
CFT3	.410**	.249**	.321**	1	.355**	.713**	.319**	.183	.197
CFT4	.423**	.227**	.308**	.355**	1	.670**	.338**	.114	.126 *
CFT	.569**	.624**	.734**	.713**	.670**	1	.478**	.224**	.238**
3DW	.597**	.313**	.345**	.319**	.338**	.478**	1	.224**	.303**
GPA	.199**	.085	.215**	.183**	.114	.224**	.224**	1	.758**
Math	.299**	.095	.218**	.197**	.126*	.238**	.303**	.758**	1

Note. * $p < .05$. ** $p < .01$. GPA: grade compound score, similar to grade-point-average

3.8. 2 models are compared. Model A includes gender as a person predictor without additional gender*facet-interactions. This model estimates the global change in the logit of the probability of a correct response across all items between males and females. This represents the general gender effect that is also captured by the effects size d . Model B includes additional gender*facet-interactions, i.e., one gender effect per facet. This model allows to make specific conclusions on gender effects on the level of item facets. It can be analyzed whether the gender difference is truly global across all items or whether the performance differences are driven, as hypothesized, by very specific differences of subprocesses needed to solve an item. If there is no interaction, all facets are affected by gender in the same way. Significant interaction effects show that some facets produce larger gender effects than others.

When gender is included as a predictor variable on the person-side of the model (Model A), random person variance is reduced by 3.52%. The model fits significantly better than a respective model without the person parameter ($\Delta\chi^2(1) = 8.80$, $p = .003$). The difference in the logit between female and male participants is 0.36. Results are in line with Hypotheses 5, 6, and 7.

The inclusion of gender*facet-interaction parameters shows that rotation rules are facilitated in men ($p < .05$, Hypothesis 5). The facet difficulty for rotations by 90 degrees is 0.393 lower for males (95% CI: [0.042, $+\infty$]); for rotations by 180 degrees the logit increases by 0.437 (95% CI: [0.071, $+\infty$]) when male instead of female test-takers are examined.

Dealing with random feature changes is more difficult for male test-takers than for female participants ($p < .05$, Hypothesis 6). When feature surface characteristics are allowed to vary within items, the logit of a correct response increases by 0.175 (95% CI: [0.004, $+\infty$]) for female test-takers relative to male test-takers.

Table 3.10.
Prediction of FAT performance by other tests and gender (Study 1)

Predictor	Unstand. Coeff.		β	Correlations	
	b	SE		zero-order	partial
Intercept	-10.666	2.595			
CFT	0.463	0.063	.362**	.569	.402
3DW	0.708	0.083	.425**	.610	.452
Gender	0.916	0.652	.062	.172	.083

Note. R^2 for all predictors is .475; ΔR^2 for Model including gender vs. a model including CFT and 3DW only is $\Delta R^2 = .004$ ($p = .161$, n.s.).

Changing the shape of the main form to include a more clear “landmark”, again, facilitated performance for men ($p < .05$, Hypothesis 7). When main shapes are concave and do not pertain clear landmark features, item difficulties increase by 0.183 more (95% CI: [0.011, $+\infty$]) for men relative to the increase in item difficulty for women.

In line with the expectations, all other gender*facet interaction effects did not reach statistical significance, and the effect sizes for these interaction effects are low. Model fit statistics did not clearly favor one of these models. AIC indicated superior fit for Model B, BIC indicates better fit for Model A.

Correlations between the FAT and all other instruments are summarized in Table 3.9. Performance on the new measure was substantially related to performance on both other cognitive tests, the 3DW ($r = .597$, $p < .001$) and the CFT20 ($r = .569$, $p < .001$). Both measures together explain almost 50% of variation ($R^2 = .472$) in FAT scores in a multiple regression analysis. The partial correlations displayed in Table 3.10 shows that the two measures explain almost mutually exclusive parts of variation in FAT scores. Among the four subtests of the CFT, the highest correlation was found for subtest 4, Topologies ($r = .423$, $p < .001$). The FAT also correlates significantly with the grade compound score ($r = .199$, $p < .001$). Among the individual school grades, the highest correlation was found for maths ($r = .299$, $p < .001$). The stepwise regression results summarized in Table 3.11 reveal that performance on the FAT can explain additional variation in maths grades beyond the other two cognitive measures ($\Delta R^2 = .016$; $p = .025$).

3.4. Discussion

Ever since the development of intelligence tests, establishing construct validity has been a central goal of all test developers. There have been numerous suggestions how to link psychological theories of information processing with the design of new instruments. The

Table 3.11.
Prediction of math grades by FAT scores and other tests (Study 1)

Predictor	Unstand. Coeff.		Correlations		
	b	SE	β	zero-order	partial
Intercept	7.819	1.672			
CFT	0.035	0.043	.056	.238	.048
3DW	0.138	0.058	.173*	.306	.142
FAT	0.083	0.037	.174*	.312	.134

Note. R^2 for all predictors is .121; ΔR^2 for Model including FAT vs. a model including CFT and 3DW only is $\Delta R^2 = .016$ ($p = .025$).

use of explanatory IRT models in combination with approaches of rule-based AIG is a promising way to put cognitive theories on a testable fundament. With the current study, a new item-generative framework was presented that is purely figural, thereby providing a basis for truly language-free and culture-fair testing. Thereby, this study addresses Gierl and Lai’s (2012) claim that “the theory and practices that underlie item model development must be studied” (p. 37). The validity of the new framework was tested based on data from $n = 308$ university students. Results of this study will be discussed along the research questions and specific hypotheses.

3.4.1. Conclusions regarding the research questions

Two main research questions were tested. First, can a thorough item design based on psychological and cognitive theories provide a basis for reliable prediction of item difficulties? Second, are item facet difficulties in line with theoretical assumptions? Here, a set of more specific hypotheses was tested. Correlations with other instruments were investigated to test the “nomothetic span” of the new item generation framework.

Based on the investigation of existing tests and theories of figural-spatial reasoning and analogical processing, a set of item generative rules was developed that allows the generation of a large range of analogy items of almost all difficulty levels. The generation of a new measure with good psychometric properties based on exclusively figural material without any reference to verbal or numerical content, was successful: Items of the new FAT are Rasch-scalable, and the test is informative along the whole person parameter continuum. As expected, items with only 2 rules were most informative in the lower ability range, pertaining solution probabilities higher than 90% while items with a maximum combination of 6 item facets were most informative in the upper ability range. Solution probabilities for these items were lower than 10%. In comparison with the analogy test described in Beckmann (2008), test information decreased less with extreme ability levels

making the test also applicable for the assessment of giftedness. They alone could reduce random item variation by nearly 70 percent. Compared to other item-generation studies (e.g., Freund et al., 2008), this is already a very satisfying explanatory power of the underlying LLTM model. The inclusion of the three additional complexity rules could contribute an additional 15 percent to the explanation of item difficulties, resulting in a total explanatory power of more than 85 percent.

However, analyses of absolute errors in the prediction of true item difficulties showed that even such high values of explained variation bear large prediction errors. Based on the height of prediction errors, the application of the LLTM for the prediction of item difficulties based on estimates of facet parameters cannot be regarded successful. While LLTM parameters allow insights into the structure of item difficulties and the contribution of certain task parameters to item difficulties, a reliable prediction of item difficulties was not successful. If wrongly predicted item difficulties would be used in operational testing settings, potentially involving computer-adaptive testing, the error in person parameter estimates might be tremendous. This is a severe threat to the validity of automatically generated items that are not calibrated individually. The concrete extent of error in person parameter estimates should be investigated by means of systematic simulation studies. Cloning approaches along with more complex psychometric modeling strategies might be more helpful as well when a stronger alignment of predicted and true item parameters is required, for instance during computerized-adaptive testing.

From a construct-validity standpoint, all changes to the item-generative framework presented by Beckmann proved successful. None of the spatial displacement and stimulus complexity parameters has effects that are inconsistent with theoretical models of figural-spatial reasoning. The increased homogeneity of the rules applied proved appropriate to model item difficulties across a wide range on the ability-difficulty continuum. Compared to Beckmann's findings, the data fits the RM better with less items showing misfit to the model. The internal consistency of the test is higher as well.

The internal cognitive structure of the FAT was investigated to test the construct validity in terms of what cognitive processes are needed for successful completion of the test. Four hypotheses were formulated. Three of these four hypotheses (H1, H2, and H4) could be confirmed based on the data from the empirical study. One hypothesis (H3) could be only partly confirmed.

First, it was hypothesized that all spatial displacement rules increased item difficulty. The FAT was developed with a strong focus on theories of mental rotation and figural-spatial reasoning. That is, the intended construct of the test should be represented in the empirical results from explanatory IRT modeling. All spatial displacement rules have facet parameters that are significantly larger than zero. Rotation by 180 degrees was the most difficult operation while mirroring at the vertical axis was the easiest operation. This finding makes sense as a simple flipping strategy can be applied for these items. The two displacement rules that apply to the four possible features contribute to item difficulties in

the hypothesized way as well; in the student sample investigated changes in the positions of a feature by 2 edges contribute slightly more to item difficulty than changes by only one edge. However, this difference was not statistically significant. In general, all parameter estimates were consistent with findings reported by other researchers in earlier studies.

Second, it was hypothesized that the type of figural shape used influenced the ease of cognitive processing. In line with theories of spatial processing, convex shapes with clear landmark features were assumed to facilitate analogical reasoning in the FAT relative to concave shapes that do not have such distinct landmarks. The variation of the type of shape was included in the new item-generative framework to allow for a generation of items that cover the whole range of difficulties. Therefore, hypothesis 2 reflects not only an important theoretical assumption from cognitive theory, but also an important technical question for the development of a new measure. As predicted did landmarks in convex polygons facilitate representation in working memory and mental rotation of the respective form. This findings emphasizes the need to control for as many generation principles as possible when items are generated automatically. It cannot be assumed that item features that are not primarily linked to cognitive rules needed for correct solution of the task necessarily take the role of incidentals. While there are multiple sources for complexity (and difficulty) changes when letters and digits are used (as in Beckmann's instrument), stimulus complexity could be controlled sufficiently well in the new FAT. Whether a shape is concave or convex can be manipulated easily here, and changes in stimulus complexity are in line with theories of spatial reasoning.

Third, it was hypothesized that not the number of elements but the number of relations between the elements of each configuration determined the complexity of an item. This hypothesis was developed from RC theory that assumes that the relational complexity between stimuli is more influential for the difficulty of cognitive processing than the sheer number of elements that have to be held active in working memory. According to this hypothesis, the addition of a third additional feature to the figural configuration without the simultaneous addition of a third rule associated with this feature should not influence item difficulty. Opposed to this hypothesis did the addition of a third feature to the figural configuration of an item actually lead to a decrease in item difficulties. Items comprising three features were, on average, easier than items containing only 2 features. This finding cannot be explained by the assumptions of RC theory and seems, at first sight, rather contra-intuitive: an additional feature makes the figural configuration more complex and does not reduce complexity. However, a viable explanation for this finding might be that the figural configuration was more distinct when a third feature was added. The additional feature might have served as a "landmark" with regard to the main form. This explanation is especially reasonable with regard to the specific constraints posed by the FAT items used in this study: even when a third feature was included, at maximum two features changed their position in order to make test items not too hard for the participants. That means that the third feature stayed constant across the whole analogy. In order to

clarify on the effects of overall stimulus complexity future studies are needed. A satisfying answer to this research question can be found if the number of features and the number of feature-rules applied are manipulated independently in an experimental design. Again, the findings here show that supposedly irrelevant item features might influence item difficulty in unforeseen ways. Only a careful control and empirical investigation of facet difficulties for all item facets (as done in the current study) can guarantee that a sufficient degree of construct representation is reached.

Forth, it was hypothesized that random change of the surface characteristics of the features used in the FAT items would turn out another viable means to manipulate item difficulties. This hypothesis was based on findings that test-takers use either analytic or holistic processing strategies, that is, strategies that focus on an analysis of each figural element separately versus that try to process the complete stimulus with all its features as a whole. Holistic strategies facilitate cognitive processing in a similar way as chunking techniques because they reduce the complexity of the cognitive process. When surface characteristics vary in a random way, test-takers are encouraged to apply analytic strategies in order not to run into mistakes due to distracting surface similarities or differences between the stimuli. Specifically, random variation should make the encoding of the relation between the “A” and “C” term, i.e. the preparation period of analogical reasoning harder and thusly increase item difficulty. The prediction that random change of the surface characteristics of main form features would place more cognitive demands during the encoding of the relation between the “A” and “C” was supported by the empirical data. The character of the rcf rule is similar to that of the tof rule. Both rules are not from the class of spatial displacement rules. Both rules do not determine the internal structure of the tasks. But both rules must be considered item radicals when item difficulties should be predicted based on task characteristics. It is important that these rules are incorporated into an automatic item generator and any subsequent estimation procedures.

Image rotation abilities have repeatedly shown the most robust sex differences among cognitive abilities (see Voyer et al., 1995). At the same time, most tests of fluid intelligence are based on figural item types that often include mental rotation principles. That is, score differences between female and male test-takers on such instruments might be partly due to differences in spatial abilities. The explanatory IRT models applied in the current study allowed for a deeper analysis of the factors for gender effects on the level of task characteristics. I formulated three gender-related hypotheses that were tested by including gender*facet interaction as well as gender-main effects in the LLTM model.

The first gender-related hypothesis, Hypothesis 5, stated that rules requiring mental rotation skills would be easier for men. This hypothesis was based on existing research showing robust gender effects for mental rotation tasks. The expected direction of gender effects was found on the manifest level on both FAT and 3DW scores, favoring men. The effect size for this gender effect was in the lower range of gender differences reported for other measures that involve mental rotation abilities. Exploratory IRT modeling results

demonstrated that these manifest gender effects on the FAT were driven, as hypothesized, by very specific differences of subprocesses needed to solve an item. Not all cognitive operations were easier for men, but it's specifically the mental rotation part in the items that were processed easier for men. All other spatial displacement rules that did not directly involve mental rotation functioned the same for male and female test-takers.

The two other gender-related hypotheses, Hypotheses 6 and 7, emanated from findings that males and females prefer different types of processing strategies when solving abstract cognitive tasks. Women tend to use analytic processing strategies, i.e., strategies that decompose a stimulus into its constituting parts and then focus on each of these parts separately. Men tend to use holistic processing strategies, i.e., strategies that treat a stimulus as one entity, regardless of how many features a stimulus has or how complex it is. In the FAT, two item facets were designed to trigger specific processing strategies. The facet "Type of Form" (tof) and the facet "Random change of feature characteristics" (rcf). As hypothesized rules that required the application of analytic processing strategies were easier for women compared to men. For women, solution probabilities differed less when the complexity of the main shapes was reduced and, thereby, holistic processing was encouraged. Women also dealt more easily with random variations in the surface characteristics of item features. That is, women were less distracted by random variations in the feature appearance that make holistic processing strategies less applicable. These findings add an important facet to research on gender differences on cognitive tests. They demonstrate that gender differences can have multiple complex factors and that, even on spatial tasks, males do not outperform women in all regards but depending on the nature of the item facets, it can be female test takers as well who outperform the other gender. These results are also consistent with previously reported findings on spatial visualizers versus object visualizers that showed that the latter tend to encode images holistically as a single perceptual unit while the former encode and process images part by part in an analytical, sequential way (Kozhevnikov, Kosslyn, & Shephard, 2005). A shortcoming of the current study is the relatively small sample size especially for the male subsample. Consequentially, standard errors for the gender effects reported here are relatively large. However, the effects reach statistical significance — a result that is promising for more extensive future research on gender effects.

All gender effects are in line with theories on gender differences on cognitive tasks and provide strong prove for the construct validity of the FAT. Construct representation is given not only in terms of an accurate prediction of item difficulties by facet parameters, but also in terms of a theory-consistent patterning of gender differences. Item facet parameters did actually cover the cognitive processes that they were intended to cover. The results from additional multiple regression analyses were in line with these findings as well. Gender did not predict performance on the FAT when controlled for general cognitive ability and spatial abilities. That is, given a certain level of general intelligence

and spatial abilities, the FAT turned out to be gender fair: males and females with the same ability level did not perform differently on the FAT.

Criterion-related validities were also promising. The FAT correlated both with g as well as mental rotation capabilities. Most notably, multiple regression analysis showed that the FAT predicted school grades better than a combination of CFT and 3DW scores. The high correlation with math is in line with typical correlations of intelligence and scholastic performance. The substantial correlation with the Topologies subtest from the CFT further supports the construct-validity of the FAT. The Topology subtest in the CFT also requires the representation of spatial relations of figural elements to one another and shows the highest similarity with the items of the FAT. The increased coverage of spatial abilities proved useful in terms of the criterion-related validity of the new measure. The FAT added incremental validity to the prediction of math grades above an already good prediction based on a reasoning and a mental rotation test.

(Gierl & Lai, 2012) suggested that item models should be evaluated not only in terms of their statistical properties but also in terms of two additional principles, their generative capacity and their generative veracity. The item model presented in this study has high generative capacity, that is, a large number of items can be generated based on the manipulation of radicals and incidentals. Further, the item model has high generative veracity, that is, items can be clearly interpreted regarding the underlying cognitive processes and, as the illustrative facet-by-gender interaction analyses showed, allow for investigation of additional hypotheses regarding the cognitive processing applied by the test-takers.

3.4.2. Limitations and future prospects

While the development of a new item-generative framework was successful, and first empirical results demonstrated a strong relationship between true and rescaled item difficulties, many challenges remain for future research, especially the question of how a better alignment in absolute parameters between true and predicted item difficulties can be achieved to enable implementation of the FAT in a fully computerized adaptive testing module. Many of the in the following outlined future prospects are actually being followed at this moment. The results of this study provided the basis for the development of a computerized generative and adaptive test system. The FAT will be one component among other item types that will be implemented in this system.

The test items applied in the current study were generated manually based on the new item-generative framework. A comparison of manually and automatically generated test items is still pending. The same is true for a comparison of paper-pencil and computerized test administrations. It should be tested how robust the computational algorithms to generate items are and if items generated based on the same set of radicals are truly parallel. The modeling of explanatory IRT with random effects is a first step to predicted item parameters. Item-Cloning models could be applied to test the suitability of

FAT items for computer-based mass-generation in computerized adaptive testing settings. When multiple item clones should be generated during testing without a calibration of each item clone, an accurate prediction of item difficulty is one core requirement. More extensive studies with several item clones for a specific radical-combination could further clarify on the robustness of the item generative framework. Item Cloning models could help to provide accurate estimates for the within-family variance of item difficulties for specific item clones. As mentioned above, the magnitude of absolute errors in parameter estimates between sum-normalized Rasch and sum-normalized rescaled LLTM parameters was not satisfactory. Future studies need to investigate how the alignment of parameters can be improved. Such attempts should include both further studies of the cognitive task structure and the systematic experimental or quasi-experimental investigation of item radicals and incidentals, as well as the use of more advanced statistical item cloning models (e.g., Geerlings et al., 2011). For instance, future studies should systematically investigate structurally parallel item sets to achieve a better understanding of the factors for within-family variation in item difficulties.

An alternative, though closely connected, angle on the same problem could be to further investigate the effects of item mis-calibrations on person parameter estimates. Under the assumption that even modified item-generation models will always leave certain proportions of variation in item difficulty unexplained, an important question is what degree of uncertainty in item parameters constitutes an acceptable level to still be able to estimate person abilities with the necessary accuracy. The current work mainly contributes to the clarification of difficulty generating processes and the establishment of construct validity by analyzing contributions of different task parameters to global item difficulties. When item-generative frameworks are used in operational high-stakes assessment settings, it must be known whether person parameters estimated based on rescaled LLTM item difficulty parameters are accurate or whether ability estimates are systematically biased when rescaled LLTM parameters are used instead of “original” Rasch parameters. Future studies should systematically investigate possible biases in person parameter estimation due to imprecisely calibrated item parameters. Simulation studies (cf. e.g., Bertling, 2007) could be one viable strategy here. Also, item misfit should be further investigated in future studies. Although overall the items in this study fitted the RM well when multiple fit indices were considered, individual items showed misfit as measured by specific fit indices. For instance, items 23, 24, and 35 showed misfit on the Q-statistic. Inspection of these items did not reveal any specific item features that have caused model misfit. For future applications it is important to further investigate whether certain item attributes might cause misfit. Demonstrating that no systematic bias is introduced by lack of fit for certain items is an important requirement for the operational use of the AIG model presented here.

Future applications should also use the potential of AIG to a more fully degree than the current test development pilot study. For instance, a more thorough analysis of distractor

stimuli and respondents' distractor choice behavior might be promising research areas. In the current study, distractor stimuli were generated based on the same rule-based framework as regular test items. However, no specific analyses were performed with regard to these distractor stimuli. Future studies could, for instance, analyze the process of distractor choice by the test-taker more thoroughly. In principle the application of polytomous IRT models to FAT data should be possible as well. Partly-correct answers could be identified based on the knowledge of what distractors share what features with the correct solution. This could also comprise an analysis of *Differential Distractor Functioning* (Green, Crone, & Folk, 1989) for different groups of test-takers.

Replications with larger samples and more diverse samples are necessary to cross-validate the findings with regard to the functioning of each of the item facets and the results on gender effects. Gender effects on the FAT and the underlying item facets should be investigated in larger and more equally sized samples. In the current study, the proportions of male and female test-takers are not optimal. Gender-effects were analyzed, but they were not the focus of the empirical study. Due to the small samples, standard errors for the gender-effects estimated were rather large. Future studies should investigate gender differences on the FAT in more detail. This could also include qualitative approaches such as cognitive labs and think-aloud studies. Furthermore, differences in test-performance do not only depend on the underlying ability but also on motivational and pre-knowledge based factors. It would be interesting to use specific methods (e.g. think-aloud protocols) to take a deeper look into the strategies, that test-takers use explicitly during test completion in order to further validate the item-generative framework and test the underlying psychological theories. Lee et al. (2008) showed that performance on complex Sudoku items strongly depends on the familiarity with the relevant strategies how to solve that type of item. Sudoku puzzles are one of the most popular, yet not the only available cognitive puzzle. Numerous cognitive games and "brain training" tools have gained worldwide popularity during the last years. Training effects can be a threat to the validity of cognitive instruments (see e.g. Freund and Holling (2011)). That is, investigating links between performance on cognitive ability tests in general, the engagement in cognitive puzzles, as well as specifically designed training or coaching programs should be one field of future research. The automatic design of tests with sufficient robustness against context effects and training will be one major challenge for test development in the twenty-first century. This should also comprise the investigation of the cross-cultural fairness of the item-generative framework.

Finally, the current study provided first prove for the criterion-related validity of the FAT by investigating correlations with other cognitive measures and with school grades. Future studies should focus on a more comprehensive validation of the FAT using internal as well as external validity criteria. Especially the validity of the FAT in high-stakes settings, for instance as part of workforce readiness or personnel selection assessments,

needs to be investigated to come to conclusions about the feasibility of the FAT for such applications.

4

The Number Series Test (NST): Item-generation and investigation of parallel test forms

This study describes the development of a new item-generative framework to generate number series items. The focus of this study is on the question whether item-generation models can facilitate the construction of truly (i.e., structurally and psychometrically) parallel test forms. Two main research questions are addressed. First, the appropriateness of the new item-generative framework for the construction of parallel tests is investigated. Second, it is asked whether estimates of the item-difficulty model are in line with findings from cognitive psychology on mathematical processing and numerical reasoning. The validity of the framework, especially for the generation of parallel test forms, is investigated in an study with $N = 406$ university students. Virtual item models were applied to test the stability of item parameters across parallel item sets. Warm-up effects are distinguished from true parallel-test effects. Results demonstrate that parallel forms can be constructed based on a generative framework if sources for heterogeneity in item difficulties are carefully controlled. Item difficulty is predominantly determined by the relational complexity of two consecutive numbers. Complexity levels could be manipulated considerably by combination of a set of relatively simple arithmetic rules requiring only addition and subtraction. LLTM modeling results showed that item difficulties could be well explained by underlying radicals when both arithmetic rules and their combination principles were included as item predictor variables. The item-generative framework was shown to be relatively robust against irrelevant surface patterns in the number of a series caused by random incidentals. After a warm-up run, item difficulties could be predicted very reliably for two parallel test forms. Correlations with a general reasoning measure and maths grades further confirmed the criterion-related validity of the new instrument.

Keywords. Series completion tasks, Numerical Reasoning, g , Automatic Item Generation, Explanatory IRT, LLTM, Virtual Item Model, Parallel Tests

4.1. Introduction

This introduction consists of three parts. First, characteristics of number series completion tasks as one of the most common indicators for numerical reasoning are described. Second, information processing theories and previous attempts to model item difficulties will be reviewed. Findings regarding cognitive task parameters as well as strategy-related factors will be summarized. Third, the research questions and hypotheses of the current study will be derived.

4.1.1. Number series items as indicators of reasoning ability

Number series are one of the most frequently used numerical item types in psychometric tests of intelligence. Almost all aptitude test batteries include this task type (e.g., Amthauer et al., 2001, Jäger et al., 2006, Weiß, 2007). Number series are indicators of general reasoning ability and popular instruments in personnel selection settings. Number series are part of school books in mathematics, and the internet offers numerous online number series tests and training tools. Number series are not only used in psychometric tests, but frequently applied as cognitive measures in experimental psychology as well (e.g., Hackett, Betz, O'Halloran, & Romac, 1990; Verguts & De Boeck, 2002).

Table 4.1 gives examples of typical number series items. In a typical number series test, a sequence of about 4 to 10 natural numbers is presented and the test-taker has to induce the arithmetic rules that define the number sequence. Most number series must be solved by continuing the series by one or two correct subsequent elements. The problem solver either has to write down the answer in a constructed response (CR) format, or he has to choose the correct solution among a selection of multiple choice (MC) alternatives. Alternative

Table 4.1.
Examples of typical number series tasks

Authors	Example Item								Solution
1.) Amthauer et al., 2001	13	15	18	14	19	25	18	?	26
2.) Amthauer et al., 2001	9	6	18	21	7	4	12	?	15
3.) Heller et al., 1976	2	2	3	3	5	5	8	?	8
4.) Holzman et al., 1983	22	22	21	21	20	20	?		19
5.) Holzman et al., 1983	45	36	44	36	43	36	?		42
6.) Holzman et al., 1983	64	36	24	32	12	24	16	4 ?	24
7.) Verguts & De Boeck, 2002	2	4	6	10	16	26	?	?	42
8.) Porsch, 2007	20	9	20	32	15	32	56	27 ?	56
9.) Porsch, 2007	-216	-102	-56	-24	-6	-8	0	?	6
10.) Weiß, 2007	8	11	10	15	12	19	?	?	14
11.) Weiß, 2007	2	3	5	9	17	33	?	?	65

types of number series must be solved by identifying a rule-discrepant element or by simply naming the rules (e.g., Verguts & De Boeck, 2002). In any case, solving number series items requires the ability to discover one or several general rules or relations among numeric elements and to apply these rules to new elements. Apart from the differences in the specific abilities required, number series items require similar *general* cognitive processes as the analogical reasoning items presented in Chapter 3 of this thesis: in a first step, the rules have to be identified by means of inductive reasoning. The second step involves application of the rule to a new element (i.e. a new “situation”).

Number series have a number of special characteristics that seem beneficial from a test development point of view. Only knowledge about natural numbers and some elementary arithmetic operations is required, making number series also adequate task types for language-free assessments as well. Number series items can be administered using a CR format. This helps to avoid problems concerning distractor generation and a reduction of guessing parameters. Number series are easy to understand given test-takers are already familiar with numbers. Compared to other task types, such as matrices or analogies, people usually deal with numbers in everyday life. Number series have a high face validity because numeracy and an understanding of mathematical concepts and operations is considered a key competence necessary for all kinds of activities in education and in the workforce (e.g., Organisation for Economic Co-Operation and Development, 2010; Mullis, Martin, Ruddock, O’Sullivan, & Preuschoff, 2009; Lemke & Gonzales, 2006). This might also be a reason why number series are judged as more interesting than other cognitive tasks by adult test takers (Quereshi & Smith, 1998).

Probably the most important advantage of number series items for rule-based AIG is that the algorithmic nature makes it easy to formalize their structure and create a universe of items (e.g., Quereshi & Smith, 1998; Korossy, 1998). Number series can be *constructed* much easier than other reasoning tasks. Compared to figural matrices or analogy items, as far as a set of item-generative rules is given, the actual construction of a number series item is possible even using simply paper and a pencil or by means of basic software, such as Microsoft Excel. Computerized test administrations can also be realized easily because items comprise no complex graphical configurations and could, for instance, be administered even on small handheld devices without high demands to graphical processing, storage capacity, or display size. However, there is no consensus on *how* number series items should be generated to make sure that they truly represent the intended cognitive processes.

Number series tasks can be classified with regard to a number of properties. Figure 4.1 gives an overview of different possible attributes of number series tasks. With regard to surface characteristics visible at first glance when inspecting the number series in Table 4.1, number series can differ with regard to their length (i.e. the number of elements assigned in a row), and the magnitude of the numbers applied and the rank ordering of elements. While some series include only small positive two-digit numbers, others include

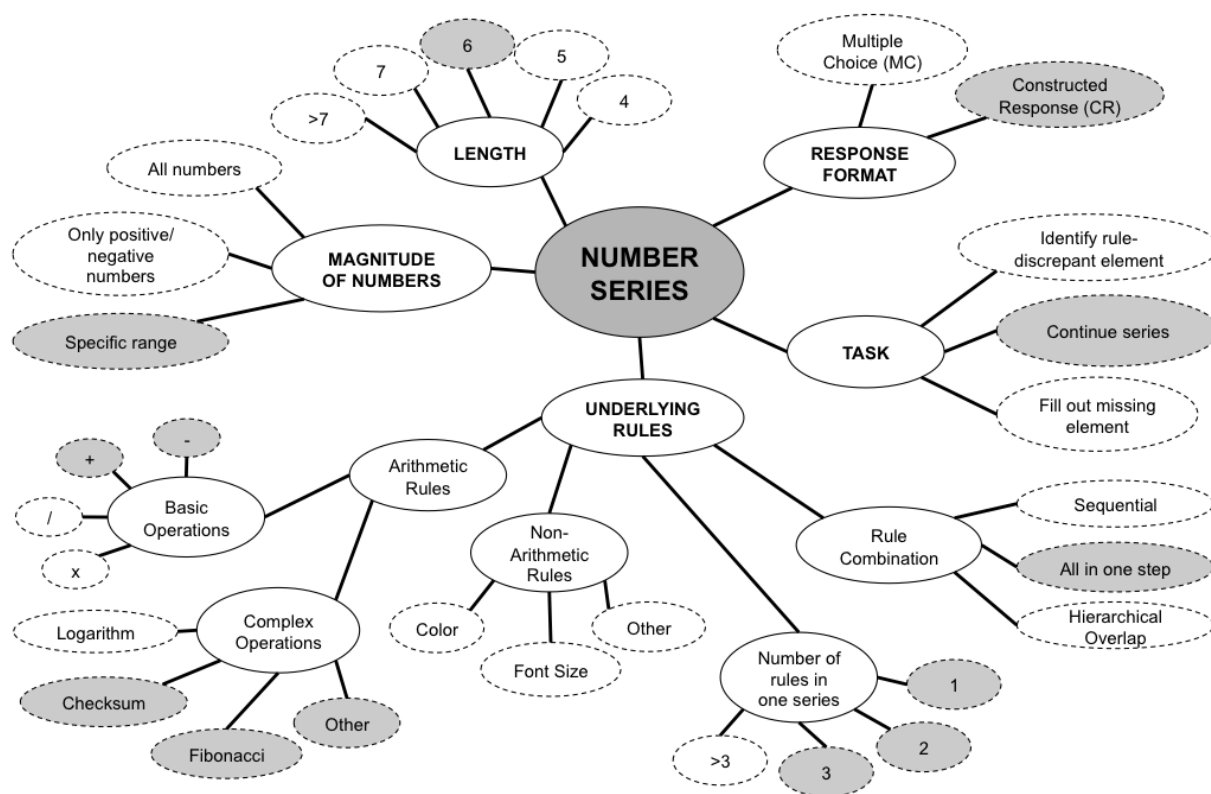


Figure 4.1.

Overview of different possible attributes of number series tasks; Grey ovals mark attributes of the number series test in this study.

larger numbers and negative numbers as well. Some series are monotone increasing or decreasing; others are alternating with no clear monotone rank order in the elements in the series). With regard to the arithmetic rules applied in each series, items can be distinguished according to (a.) the nature of rules, (b.) the number of rules combined in one series, (c.) the way of combination of rules in one series.

The most frequently rules are simple counting rules and the basic arithmetic operations of addition, subtraction, multiplication and division. Some items also make use of more advanced rules such as computation of the logarithm. A rule that is also found in many series is a simple identity rule, i.e. an element of the series remains unchanged from one position to another. Rules can require a combination of two or more elements of the series by basic arithmetics and the combination of one element with a constant not part of the series. For instance, the Fibonacci-rule in series 7 given in Table 4.1 requires the test-taker to calculate a new element of the series recursively as the sum of the two previous numbers. In contrast, series 1 requires the test-taker to calculate a new element

of the series by adding a constant to the current element. Both series include only one arithmetic operation (i.e. addition), but the nature of this rule is different.

Number series comprising only one rule (see example 1 in Table 4.1) are very easy and, therefore, in most cases of limited diagnostic value. In example 1, the next number can be found by simply counting in steps of two without the need to represent several rules or intermediate results in working memory. Still, number series with one rule are often used as instruction or warm-up items in cognitive test batteries. In order to measure inductive reasoning and not only the knowledge of basic arithmetic operations, typically items with at least two rules are used. Test batteries such as the Intelligence Structure Test (IST-2000; Amthauer et al., 2001) include items with up to four rules combined in one series. For example, item 11 in Table 4.1 combines 2 rules (multiplication and subtraction) that have to be applied simultaneously to each element. Item 2 is build from 4 arithmetic rules (subtraction, multiplication, addition, division) that have to be applied sequentially. After execution of all three rules, the series starts again with the first rule again, and so forth. In general, three principles of rule-combination can be distinguished.

1. The first principle is a *combination of rules in one step*. That is, multiple rules determine the difference between two consecutive numbers. This is the case in series 8 or 9.
2. An alternative principle is the *sequencing of rules* one after another. This is the case in series 2. Here, period length is defined as the number of rules applied one after another to continue a given number series.
3. A third method to combine rules is *hierarchical overlap of rules*, also termed *hierarchy* (Porsch, 2007) or *interpolation* (Verguts & De Boeck, 2002). In this case, there is no mathematical connection of two directly neighboring elements of the series. When two rules overlap, the first rule can, for instance, apply to all odd elements and the second rule to all even elements: in series 10 the third element is computed by applying the rule $+2$ to the first element, while the fourth element is computed by applying the rule -2 to the second element. The two rules carry on alternately across the whole series.

When rules are combined one after another, the number of rules applied defines the so-called period length of the series (Holzman et al., 1983). Depending on the nature of the rules that are combined, eye-catching *breaking points* might appear in a given series. Breaking points can make changes in rules visible by abrupt changes in the magnitude of the numbers or systematic patterns across a series of numbers. This is the case in items 4 and 8. In the IST 2000 (Amthauer et al., 2001) or in Porsch’s instrument, items comprise sequences of up to three operations. A drawback of this technique to create relational complexity is that more complex number series always contain more elements. One period has to be repeated at least once in order to allow for a unique solution. Only if the rules contains “breaking points”, i.e. points where the rules start to repeat, a test-

taker can induce the period length unequivocally. In order to construct items of sufficient complexity, some tests use two or more combination principles together in one item. For instance, in series 8 in Table 4.1, 2 arithmetic rules are combined in one step and there is a sequence of rules. In the example, the three 2-rule-combinations that have to be applied are $/2 - 1$, $\times 2 + 2$, and $\times 2 - 8$. After the application of the third rule, the series starts again with the application of the first rule-combination. One problem, however, is that a way must be found to guarantee for the uniqueness of solutions.

It has been shown that the sequence of rules is an important factor for differences in item difficulties of number series (Ebert & Tack, 1974; Porsch, 2007). While other task types (matrices, analogies) allow, in most cases, for a sequential processing of all rules present in one item, number series are different. Figural matrices and analogies usually consist of a configuration of several figural elements, and each rule is applied to some of these elements. That is, while solving a complex matrix item, the test taker has to induce from the complete figural configuration the rule that applies to element A, hold the rule in mind, then induce the rule applied to element B, hold it in mind, and so forth. Numbers do not have this composite character. As soon as two operations are applied to one number in order to calculate the next number, it is not possible to induce one rule first by inspecting the number sequence, hold it in mind and then induce the next rule. As a consequence, the test-taker has to hold active in mind several possible rule-combinations while storing intermediate result(s) in working memory as well. This does not only make the solution of such a number series very hard; it also allows for different strategic approaches to reduce complexity. This can cause large differences in solution probabilities Porsch (2007). Key findings regarding this problem and other difficulties regarding the prediction of item difficulties will be reviewed after a summary of a widely used information processing model of number series.

4.1.2. An information processing model for number series

Kotovsky and Simon (1973) investigated the information processing steps individuals engaged in while solving letter series items. Based on their results, Holzman et al. (1983) identified the cognitive processes involved in series-completion problems and specified a framework for solving number series. Even though their model is almost 30 years old, it is still considered an important theoretical starting point for more recent applications (e.g., Verguts & De Boeck, 2002; Porsch, 2007). Their theoretical framework makes use of four key factors (see Figure 4.2 for an illustration):

1. *Relations detection*: In a first step, the individual has to detect relations between elements. This requires the test-taker to scan the series and to formulate hypotheses on how one element of the series is related to another. (LeFevre & Bisanz, 1986). Relations between elements can be distinguished based on the mathematical rules and their combination as described in the previous section. Furthermore, arithmetic

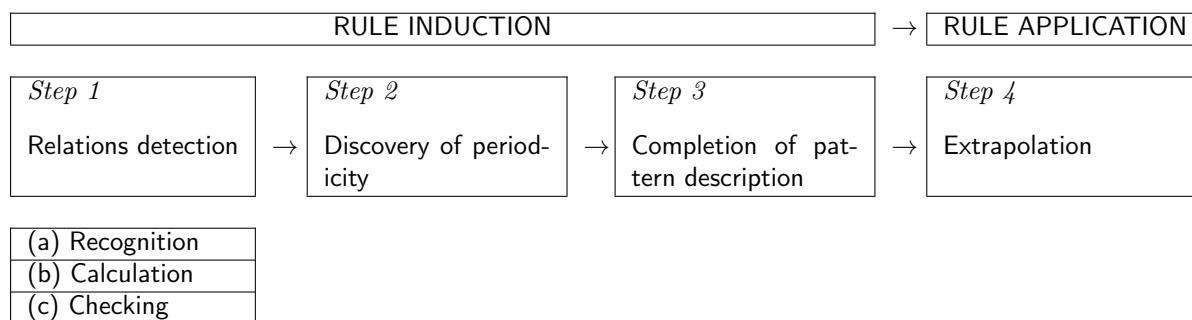


Figure 4.2.

Process model for number series (based on Holzmann et al., 1983, and LeFevre & Bisanz, 1986)

operations pose different cognitive demands based on the number of cognitive steps and their working memory load: solving calculations requires more working memory resources than counting tasks; complex arithmetic tasks require more working memory resources than simple tasks. In general, the relational complexity of a cognitive problem can be defined by the number of relationships between elements that define the right solution (cf. RC-Theory, e.g., Halford et al., 1998). Three important factors that determine the difficulty of this processing step, are type of operation, magnitude of numbers, and memory load.

In the context of relations detection, the *Problem-size Effect* (e.g., Ashcraft, 1992; LeFevre, Sadesky, & Bisanz, 1996) is one of the most robust findings in research on mental arithmetic. It describes that mental calculations become slower and more error prone with larger numbers (e.g., $7 + 8 = 15$ versus $2 + 3 = 5$). Discriminations between small numbers can also be processed systematically faster than discriminations between larger numbers. For smaller problems, answers are more frequently retrieved from long-term memory, whereas larger problems require the use of procedural strategies (Imbo & Vandierendonck, 2008). Holzman et al. (1983) and Kotovsky and Simon (1973) made no distinction between sequences that are already stored as units in semantic memory, e.g. sequences such as $(2, 4, 6, 8, \dots)$, $(3, 6, 9, 12, \dots)$, or $(25, 50, 100, 200, 400, \dots)$, and sequences that are not yet represented, for instance $(24, 48, 96, 192, 384)$. LeFevre and Bisanz (1986) proposed to extend the process model by defining three core processes of relations detection: recognition of memorized sequences, calculation, and checking. This is shown in Figure 4.2. Only when retrieval strategies fail are test-takers assumed to engage in actual calculations.

2. *Discovery of periodicity*: The periodicity of a given series can be determined by investigating systematic “breaking points” between single elements or rows of elements. Periodicity equals one if the same rule is applied to determine the relations

between *all* numbers. Series 4 in Table 4.1 has a periodicity of 2, i. e., the rule -1 is repeated only every two elements.

According to Holzman et al. (1983), a test-taker has to search for a new relation whenever a breaking point in a previously induced rule is observed. Together with complexity, period length determines the amount of working memory place-keepers required to come to a solution (Holzman et al., 1983). WM place-keepers (WMPK) can be defined as the rules and intermediate results that have to be held in working memory in order to solve a number series item. For instance, if three simple rules are combined sequentially (i.e. the period length equals 3), the item will require to hold at least 3 intermediate results (i.e., 3 WMPKs) in working memory. The number of place-keepers was found to be one of the best predictors of item difficulty with correlations between the number of place-keepers and item difficulty exceeding $r = .70$ (Holzman et al., 1983).

3. *Completion of pattern description:* In the third step, according to Holzman et al., rules which integrate the remaining elements of the series into already discovered relations and period lengths have to be identified. This involves also the detection of higher hierarchy principles, for instance the combination of two or three overlaying rules. Knowing the correct rules and the period length is a prerequisite for this step.
4. *Extrapolation:* The final step is to apply the arithmetic rules identified to continue the series or fill out a missing element. Extrapolation describes the application of the induced rule(s) to identify the position of the period occupied by the missing element of the series. This includes the processes of isolating the part of the rule governing that position, as well as the subsequent application of that part of the rule (Holzman et al., 1983).

Taken together, the first three steps comprise the process of rule-induction while the last step involves the application of the induced rules to a new number. The benefit of Holzman et al.'s model is its clear structure and the explicit description of the cognitive processes involved. However, the serial structure of the model with rather separated steps following each other is very restricted. Especially the assumption that a complete pattern description is necessary to solve an item is a very strict assumption for many number series items. When items involve periodicity, and clear breaking points are visible for the test-taker, the correct solution can sometimes be found by inspecting the general pattern of numbers without inducing and representing every single rule in working memory.

4.1.3. Item difficulty modeling of number series items: Previous attempts and problems

Porsch (2007) constructed a new number series test using the rules proposed by Holzman et al. (1983). He generated number series by means of rule-based item generation. Porsch's

Table 4.2.

LLTM parameter estimates for item facets manipulated in Porsch's study

Item facet	Set A	Set B	Consistent across parallel versions?
Addition	0.43**	-0.31**	no
Subtraction	0.17*	-0.07**	no
Multiplication	-0.34**	0.85**	no
Division	-0.52**	-0.03**	no
Hierarchy	2.55**	3.29**	yes
Periodicity	0.65**	0.87**	yes

Note. Parameters are logits; the model intercept is not displayed; * < .05, ** < .01; parameters in this table are reprinted from Porsch

work builds one starting point for the current study. It was the first study aiming at systematic construction of number series items ready for mass-construction and item cloning. The rules used by Porsch were: Addition, subtraction, multiplication, division, hierarchy, and periodicity. In line with Holzman et al.'s information processing theory, the two rules hierarchy and periodicity were chosen to manipulate the cognitive complexity of the series completion items. Two complexity levels and two levels of periodicity (2 vs. 3) were investigated. In low hierarchy items, one mathematical operation explained the difference between two successive elements, whereas in high hierarchy items, two mathematical operations were combined simultaneously. This is illustrated in the following example item (item 8, see also Table 4.1):

$$20 \quad 9 \quad 20 \quad 32 \quad 15 \quad 32 \quad 56 \quad 27 \quad ? \quad (56).$$

Here, hierarchy is 2, i.e. two mathematical operations are combined. Periodicity equals 2, i.e. the rules repeat only after 3 elements of the series. The first relation is defined by $/2 - 1$, the second by $\times 2 + 2$, and the third by $\times 2 - 8$ respectively. Here, Porsch defined the actual numerical realizations of each rule as incidentals, assuming no influence on item difficulties due to variation in numerical values within a specified range and changes in the order of the operations. Operators were allowed to vary freely between -9 and 9 . No limitations were specified for the number range of the whole series. The lengths of the series varied between 5 and 9 elements.

Porsch generated 26 items based on 13 different design vectors. Two clones were generated for each design vector. The two item sets were administered to two student samples ($n_1 = 235$, $n_2 = 233$). Porsch tested the contribution of each of the six rules to item difficulty by means of both LLTM models and regression analyses. While both tests were received positively by the test takers and showed good psychometric characteristics in terms of internal consistencies, split half reliabilities, and correlations with a figural

reasoning measure, the results with regard to the underlying cognitive model and the procedure of parallel test generation were not satisfactory. The goal to generate items covering the whole ability continuum by the combination of the 6 item-generative rules was not reached. Table 4.2 shows the facet parameter estimates from Porsch (2007). The only two item radicals showing consistent functioning across the two samples were hierarchy and periodicity. Results for the contributions of the four different arithmetic rules to item difficulties were inconsistent. Addition and subtraction had positive weights for the first test version and negative weights for the parallel version. Multiplication had negative impact on item logits in the first version, and facilitating effect in the parallel version. Division was not significant in the second version. At least two different explanations for these findings seem reasonable: First, a viable explanation for these findings is that almost all variation in item difficulties was captured by the rule hierarchy. A closer inspection of solution probabilities for low hierarchy and high hierarchy items shows that there is almost no overlap between these items. Low hierarchy items were, on average, solved by $p = .79$ (set 2: $p = .88$) of the test-takers ($SD = .08$, set 2 $SD = .10$), while the average solution probability for high hierarchy is only $p = .35$ (set 2: $p = .41$) with a standard deviation of $SD = .16$ (set 2: $SD = .22$). Variation of hierarchy produced items that were either very easy or very difficult. Only 4 out of 26 items in each set had solution probabilities between $p = .4$ and $p = .6$. In Porsch's study, the correlation of item difficulties with item complexity manipulated through the usage of either only one rule at each point of time or two rules simultaneously was $r > .80$. Second, an alternative explanation for the low agreement in item parameters for the parallel test versions could be that test-takers didn't use the cognitive processes expected based on the theoretical assumptions, but made use of other strategies that caused unexpected differences in solution probabilities between the two sets. This might not be the case for all items, but could be a factor for parameter invariance across several number series. For instance, item type 8 has a solution probability of $p = .31$ in the first set, and $p = .91$ in the second set. In theory, both items should have the same difficulty parameters. If item difficulties should be predicted based on a set of item facets, it is necessary to know the contribution of each facet parameter to global item difficulties. Also, the contribution of each facet parameter must be (at least to some degree) the same across the set of all possible items generated by means of the item-generative framework. A closer inspection of the two parallel items generated based on item type 8 shows why the solution probabilities might be so different:

Set A : 9, 16, 12, 18, 34, 21, 36, 70, ? *Set B* : 20, 9, 20, 32, 15, 32, 56, 27, ?

Due to the variation of incidentals, it is possible to detect a specific pattern of numbers in clone B (see underlined numbers) but not in clone A: in the given series, the same number always appears twice, with a — from the perspective of the test-taker — possibly random number in between. That is, instead of trying to induce a rule how several mathematic operations are to be combined, many test-takers might have just chosen 56 as an answer to

this item. This in, in fact, the correct solution, explaining the huge difference in solution probabilities between the two sets. Whereas solving clone A truly requires processing of all rules, clone B can be solved following a simple heuristic. When item generation models should be used to generate large numbers of items with sufficiently known parameters, it must be assured that incidentals (in this case the specific operators chosen in the item) do not cause such solution-facilitating “patterns” in some items, but not in others.

The finding that number series often allow for multiple solution strategies and sometimes for multiple possible solutions as well is known as the *non-uniqueness problem of number series*. The problem is that “aside from the implemented rule of the sequence and the keyed answer considered correct, many differing answers basing on other unintended (say accidental) regularities of the number sequence seem possible and might be judged to be correct” (Korossy, 1998, p. 44). However, this problem was widely ignored in testing practice. Criticism with regard to the uniqueness of number series problems has been rejected as “utterly trivial (...) because the other correct solutions are usually possible only for a mathematician” (Jensen, 1980, p. 153) or as practically irrelevant: For instance, Verguts and De Boeck (2002) constructed number series consisting of 6 numbers each based on a framework comprising 4 item-generative rules: addition, fibonacci, interpolation, and multiplication. The way that they combine these rules led to non-unique solutions and most of their items could be solved also by applying completely different strategies. However, they report that “it turned out that none of [the] participants ever used a valid rule other than the ones mentioned.” (Verguts & De Boeck, 2002, p.47). The non-uniqueness problem of number series (i.e. the problem that many number series often can be explained (and solved) by different sets of rules) might be negligible when items are used as indicators of cognitive variables in experimental settings and only easy items with rather obvious rule-combinations are administered (cf. Verguts & De Boeck, 2002). Yet, the prediction of item difficulties for automatically generated items will be less accurate when items, themselves, already contain ambiguity.

One effective means of reducing the universe of possible “correct” answers to a number series problem is a more precise explanation of rules to the test-takers in advance of the assessment. Wilhelm (2005) gave an example of the extent of rule-explanation necessary to guarantee unequivocal solutions for a series item:

“If the premises of such a number-series task are explicitly stated — for example in ‘Continue the number series 1, 3, 5, 7, 9, 11 by one more number, the operations you can use are +, -, / and * and all results are positive integers, rules are indicating regularities in proceeding through the number series, and these regularities can include rule-based changes to the rule’ there might be just one option that meaningfully continues the sequence: 13.” (p. 376)

Many existing tests do not explain rules in a test to that extent in order not to diminish the test’s validity by making solution principles obvious (and therefore, easy) for the test-taker. On the other hand, it has been shown that — when tests are constructed based

on a set of pre-specified item radicals — a comprehensive explanation of the abstract principles underlying items of a test did not diminish the validity of test scores, but could even improve criterion-related validities (e.g., Freund et al., 2008; Beckmann, 2008). Also, if test-takers have a chance to get familiar with the rules of a test while working on homogeneous item sets rather than test batteries with many different item types, higher g -loads have been reported (Carlstedt et al., 2000). Findings that criterion-related validities are reduced in retesting mainly rely on studies, where individuals worked on identical, or at most pseudo-parallel versions of a test (i.e. parallel test versions with the same items administered in a different order; cf. e.g., Amthauer et al., 2001), and can be largely attributed to memory effects (e.g., Lievens, Reeve, & Heggestad, 2007; Amthauer et al., 2001). Comprehensive retest-studies with truly parallel tests generated based on cognitive item difficulty models are still missing.

Irle focused less on specific item-generative rules and the cognitive processes during number series completion. Instead, the main emphasis of his work was on strategies used by test-takers when confronted with number series items. Irle (1969) hypothesized that knowledge of specific solution strategies might be relevant for the ease of solution of number series items. This is also a possible explanation for the theory-divergent findings reported by Porsch. The two analytic strategies identified by Irle are *calculating differences* and *calculating ratios*.

1. *Calculating Differences*: The first strategy is to calculate differences between each pair of successive elements in the series. An operation which transforms a number series $[x_1, x_2, x_3, x_4]$ into $[x_2 - x_1, x_3 - x_2, x_4 - x_3]$ is used.

If a series contains two alternating rules, -1 and $+3$, for instance, the successive element of the series can be found easily. Suppose a series

$$(12, 14, 13, 16, 15, ?).$$

Here, the calculated differences are:

$$(-1, +3, -1, +3, -1).$$

Simple inspection of the pattern of differences shows that the rule $+3$ must be applied to find the next element. The result of the second rule, -1 , does not have to be represented in working memory to find this solution. Only if the series should be continued by two elements would the second rule have to be processed. The same principle applies to series 2 from Table 4.1:

$$(9, 6, 18, 21, 7, 4, 12, ?).$$

Here, four rules are used, but the solution can be found by simply inspecting the general pattern of differences as well:

$$(-\underline{3}, +12, +\underline{3}, -14, -\underline{3}, +8).$$

By inspecting this pattern, it can be induced that every second difference denotes to 3, alternating with positive and negative signs. If the other differences (+12, -14, +8) are totally ignored, a test-taker can still find the correct successive element, 15. Only one arithmetic operation (+3) has to be applied. The problem in these cases is *not* that the number series can be solved by looking at the pattern of differences. The problem is that the number series can be solved *as well* by looking at the pattern of differences *as* by inducing all four rules. If a test-taker solves this item correctly, no conclusion about the cognitive steps the individual followed during the solution process can be made. If item difficulties should be predicted as a linear combination of item radical difficulties, this ambiguity causes a severe problem.

2. *Calculating Ratios:* The second strategy identified by Irle is the calculation of ratios. A given series (x_1, x_2, x_3, x_4) is transformed into $(x_2/x_1, x_3/x_2, x_4/x_3)$. That is, a ratio between all consecutive elements is calculated. Irle hypothesized that this strategy is used if the first strategy (i.e., calculating differences) does not lead to to detect a complete pattern detection. If the series is

$$(10, 5, 25, 20, 100, 95, ?),$$

the vector of differences is

$$(-\underline{5}, +20, -\underline{5}, +80, -\underline{5}).$$

Inspecting this vector shows that every second difference denotes to -5 , but this knowledge cannot be used to find the successive number in this case. An additional inspection of the vector of ratios

$$(0.5, \underline{5}, 0.8, \underline{5}, .95)$$

shows that a subtraction rule (-5) and a multiplication rule ($\times 5$) are applied alternately. A “bibliography series” (Irle, 1969) can be formed, which makes the systematic structure of the number series problem visible.

Irle (1969) postulated that the differences strategy is usually applied before the ratio strategy. Only when the increase in numbers of the series is very fast, like for instance in the example series with a progression of +20 from the second to the third element

and +80 from the fourth to the fifth element, individuals will carry out a ratio-strategy first. The availability of these solution strategies is not a threat to the validity of number series tasks *per se*. The problem is that multiple processing strategies come to the same solution and item response processes cannot be linked to specific cognitive processes of the test-taker.

Ebert and Tack (1974) investigated a set of number series items differing with regards to the combination and sequence of difference and ratio rules. Subjects needed more time for calculating ratios than differences, and the second operation in the items had a larger impact on processing speed. Number series with a difference-ratio (DR) structure were more difficult and more time-consuming than number series with ratio-difference (RD) structures. That is, a cognitive model defining item difficulties based on the two item-generative rules (here: addition and multiplication) would make wrong predictions about item difficulties when the ordering of operations is not parameterized in the statistical model.

Findings from Verguts and De Boeck (2002) also support the hypothesis that strategy knowledge is an important factor for solution probabilities in number series. They could show that strategies can be induced simply by means of a comprehensive instruction in the beginning of test administration. Strategy explanations as part of the instruction had significant impact on the effectivity of solution strategies. In each of two experimental groups, only one of 4 rules in a number series test was instructed to the participants. Results showed that individuals yielded higher scores on those items requiring the strategy learned in the instruction phase compared to the remaining items. That is, item difficulties depended not only on item characteristics manipulated but also on the activation of specific memory contents relevant or not relevant for rule-induction.

Klahr and Wallace (1970) investigated letter series and described further variables that highlight the importance of test-taker behavior and strategies for the processing of series completion tasks. In their model test-takers are assumed to apply a matching procedure to complete the processing step of pattern description. Here, the directionality of the series, and the presence of specific patterns in the beginning or end of the series come into play. According to Klahr and Wallace (1970), the first item in the series is tested against the second in order to identify a legitimate relation. If a relation is found, the test-taker will try to apply the model to the entire series. That is, the elements in the beginning of a series play an important role in cognitive processing. However, the process models by Holzman et al. (1983) and Kotovsky and Simon (1973) do not account for differences in item difficulties caused by specific patterns of elements in the beginning of a series. As illustrated by Hersh (1974), irrelevant relations in the beginning of a series can influence the cognitive process of the test-taker considerably:

“For example, Simon and Kotovsky’s model would judge the two series MMMNM0 and AMANA0 to be of equal difficulty, since each can be generated by the same pattern description (a repeated letter alternating with a progression of

the alphabet). However, even to the casual observer, the second series appears much easier to continue than the first.” (Hersh, 1974, p. 771)

Hersh’s analyses of series completion tasks showed that test-takers tend to make more errors when irrelevant relations occur at the beginning or at the end of a series (Hersh, 1974). Besides, irrelevant relations at the beginning or end of a sequence increase solution times. Increases of solution times because of the retesting of false hypotheses have been reported for mathematical tasks as well (e.g., Huesmann & Cheng, 1973). Hersh (1974) argued that irrelevant relations affected only the induction phase and not the extrapolation or production phase of the information processing process. Therefore, given that a test-taker has induced the correct pattern of rules present in a given series, errors occurring during the production phase should be independent of the existence of irrelevant relations in the series. Since Hersh’s findings, the influence of irrelevant relations on the difficulty and the cognitive structure of series completion tasks has not been investigated systematically.

Number series have a special character because of their reliance on purely numerical elements. They allow, on the one hand, for a multitude of operations to manipulate relational complexity. On the other hand, they are more affected by test-takers’ strategies towards arithmetic manipulations and the detection of patterns from numerical stimuli.

Despite the huge amount of number series tests and the enormous popularity of this item type, attempts to generate truly parallel items based on a set of item-generative rules have not been successful so far. One conclusion from findings covering the last 30 years is that difficulties of number series seem, notwithstanding their high suitability for algorithmic generation, harder to predict than one would expect based on promising findings on other reasoning measures (cf. e.g., Freund et al., 2008; Embretson, 1999; Holling, Bertling, Zeuch, & Kuhn, 2010). It has been shown, for instance, that the sequence of the rules is an important factor for differences in item difficulties of number series (Ebert & Tack, 1974; Porsch, 2007), that item incidentals (i.e. features of an item that are theoretically expected not to influence item difficulty) can determine item difficulties to enormous extents (e.g., Porsch, 2007), and that the complexity parameters introduced by Holzman et al. (1983) are not sufficient to determine item difficulty and generate truly parallel test forms. If the cognitive complexity of a number series item could be manipulated in a way that several cognitive operations have to be processed simultaneously and that these operations stay the same across a whole number series, this would be a great benefit for rule-based generation of number series. Also, the prediction of item difficulties for automatically generated items will be less accurate when items, themselves, already contain ambiguity. That is, findings solutions for the non-uniqueness problem of number series is an important necessary condition for establishing item difficulty models that can explain substantial amounts of variation in item difficulties.

4.1.4. Research questions

The goal of the current study is to derive a new set of item-generative rules that allows for the generation of structurally and psychometrically parallel number series tests, and to test this new rule-set in a first empirical study. Based on previously reported problems regarding the prediction of item difficulties based on explanatory IRT models, and the psychometric equivalence of “parallel” forms, a revised set of rules and their combination into new items is proposed. The new framework addresses typical problems of previous number series tests, specifically the item difficulty modeling by explanatory IRT models, and the parallelism of structurally identical form. Two main research questions related to the validity of the new AIG framework are addressed:

1. *Are structurally identical items also psychometrically equivalent?* Previous studies have demonstrated the difficulties to generate parallel test forms. It has been shown that items designed based on the same structural principles are not necessarily equally difficult (e.g., Porsch, 2007). It is hypothesized that structurally equal items generated based on the revised rule-set are also equally difficult, i.e., it is expected that the new item-generative framework can model item difficulties across parallel forms. This question corresponds to the first major research goal outlined in the general introduction of this thesis, the investigation of the technical feasibility of item generation approaches.
2. *Can item difficulties be predicted by the hypothesized underlying cognitive processes?* It is tested whether the parameter estimates of the item-difficulty model are in line with the cognitive model that guided the item construction process. Solution probabilities should be determined by the complexity of a series as defined by the item radicals, and not by the variation of surface characteristics. That is, can item difficulties be modeled based on the set of pre-specified item radicals? This research question extends the first research question. If parallelism of structurally identical test forms is given, this demonstrates the technical feasibility of an item generation framework based on cloning structurally identical items. However, the construct validity of the generated items is dependent on the functioning of item radicals and incidentals in line with theoretical expectations about the measured construct. It is hypothesized that the numerical reasoning construct is represented by the pre-specified item radicals representing hypothesized underlying cognitive processes, and not by item incidentals defining surface differences between items. Two drivers for relational complexity are distinguished, the complexity of each individual mathematical rule, and the principles of combining rules in one item. Both rule complexity itself, and the combination of rules are expected to increase the relational complexity, and thereby the difficulty, of a given number series. This corresponds to the second major research question outlined in the general introduc-

tion, the investigation whether item generation approaches can enhance construct validity.

In addition to these two research questions, the constructed response format of the new item type will be investigated in more detail. Because of the constructed-response format of the new measure, patterns of wrong answers need to be checked for plausibility as well in addition to the computation of the typical item statistics. It is expected that the new generation approach produces items with clear solutions. Frequent wrong solutions should represent partly correct solutions in the sense that test-takers will apply only some of the necessary rules or make mistakes while applying more complex rules. For validation purposes, relationships of performance on the new measure with other ability constructs will be also investigated.

4.2. Method

There are two parts to the method section. First, the derivation of radicals and incidentals for the new item-generation framework is described. Second, the design of an empirical study is described that addresses the research questions and hypotheses summarized in the previous paragraph.

4.2.1. Development of the new item-generative framework

Analogous to other item types for which item difficulty models have been successfully implemented (cf. e.g., Figural Analogy items in study 1 of this thesis), a guiding principle for the new AIG framework is the premise that the difficulty of a number series can only be clearly defined if exactly the same cognitive operations are needed to find any missing element of the series. That is, difficulty should not depend on whether element 5 or 6, or 8 or 9 is missing, but on the set of rules and their combination accounting for the complete pattern of the number sequence. That said, all rules that were applied to generate a given number series should also be needed to solve that series, i.e., if 3 rules were used to generate an item, cognitive processing of all 3 rules should be necessary to solve this item). Several changes to previously presented generative rules for number-series were made in the item-generative framework applied here.

1. The arithmetic rules needed to continue a number series are *exactly* the same in each part of the series instead of combining several rules serially.
2. Based on Oberauer, Süß, Wilhelm, and Wittmann's (2008) research on cognitive complexity in reasoning items, item difficulties of the number series used here were defined with regard to the cognitive complexity of the relational representations between two consecutive elements. Rules were limited to simple arithmetic operations.

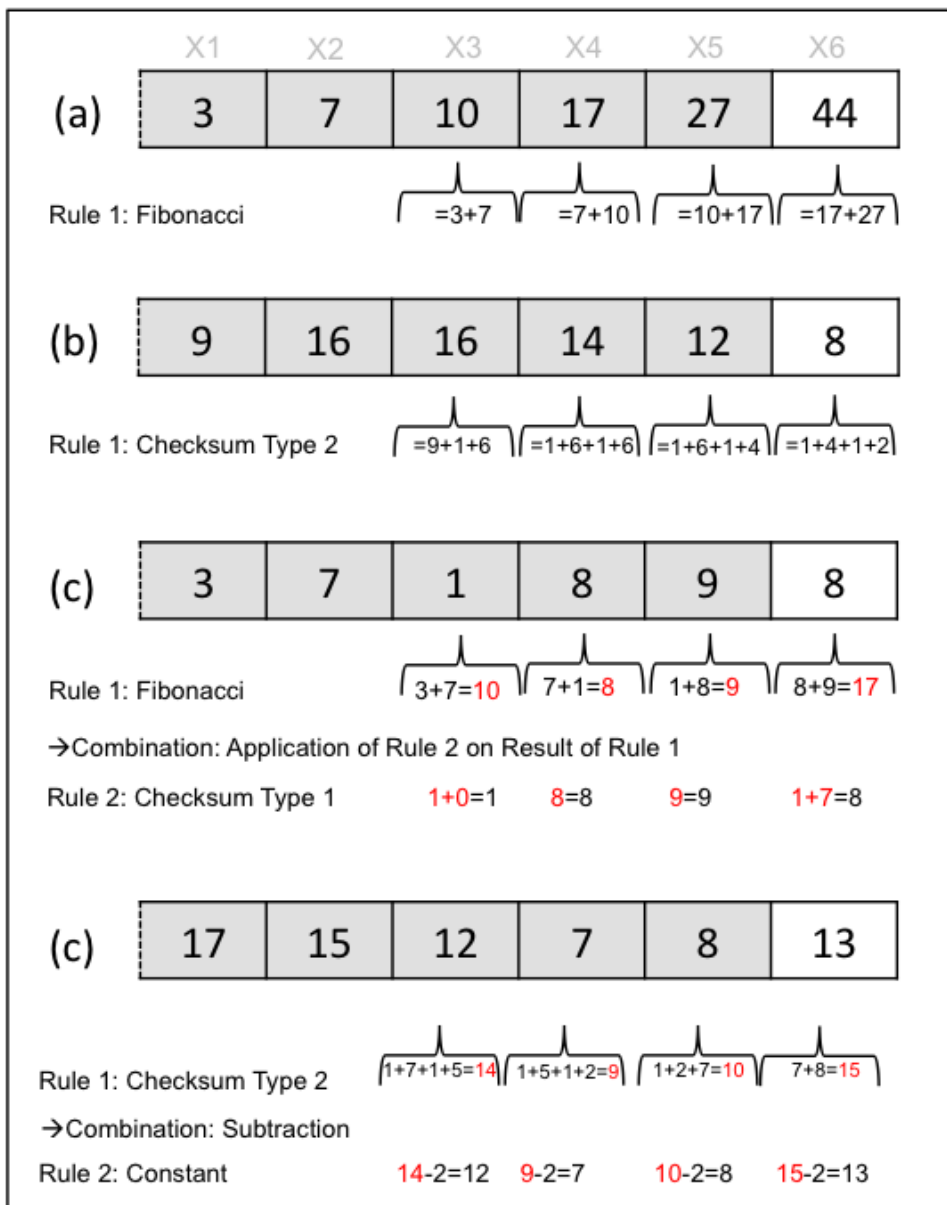


Figure 4.3.

Four example NST items generated based on the new AIG framework; under each item, the generating rules and combination principles are given; intermediate results are printed in red.

Difficulties are not defined by the mathematical complexity but by relational complexity (cf. also Halford et al., 1998) and working memory load of the items. This

decision was made because the goal of test development was to develop reasoning items, and not items assessing mathematics proficiency.

3. Holzman et al.'s factors *Periodicity* and *Hierarchy* were held constant across all number series: All items have a periodicity of 1. This ensures that the transitions between *all* elements of the series can be explained exactly by the *same* sets of mathematical operations (i.e., the rules that determine the transition from element three to element four are the same as from element four to element five. Application of the Rasch model and thereupon based explanatory models require that all items measure the same latent ability, and the existence of clear solution principles with unequivocal solutions.
4. No higher hierarchy rules are used.
5. The complexity of the numbers used was held at a minimum. All number series of the current test comprise only numbers from 0 to 99. Also only constants in the range of 1 to 3 are used when rules introduce values that are not derived from the current or previous elements of the series. These constraints were introduced to minimize unwanted effects on item difficulty such as an attenuation by the problem-size effect (e.g., Ashcraft, 1992).
6. By means of a comprehensive instruction prior to testing, ambiguity in solutions and processing strategies should be minimized.

Item Structure Each item of the new NST consists of a series of five numbers with the sixths number missing. All series have exactly the same length. In order to minimize problem size effects, the range of numbers used was constrained. All number series only contain positive whole numbers the range from 0 to 99. As mentioned above, the rules determining the relation between consecutive elements in the series are based only on simple arithmetic operations. The operations multiplication and division are not needed for any of the items generated based on the NST framework. In order to assure that the cognitive operations needed to fill in any element for the number series were exactly the same, no higher hierarchy rules were used and the periodicity was set to 1 for all items. Figure 4.3 shows four example items generated based on the new AIG framework along with the underlying cognitive rules. Items (a) and (b) are generated from one rule each, items (c) and (d) comprise two rules. In order to solve a NST item, in a first step, test-takers have to identify the relational representations between two consecutive elements, i.e. identify the rules used and the principle how they are combined. This step represents Holzman's *Relations detection* step. In a second step, the corresponding relation has to be applied to continue the series. This corresponds to the *Extrapolation* step.

In total, the new AIG framework consists of 4 basic rules and three combination principles for these rules, i.e., a total of 7 item radicals. In order to allow for the generation of structurally identical but phenotypically different items, several item characteristics are

defined as incidental. All radicals and incidentals will be explained in more detail in the following paragraphs.

Item Radicals The choice of item radicals was guided by previous findings (Holzman et al., 1983; Porsch, 2007) and the changes summarized above. As mentioned in the previous paragraph, two categories of radicals are used in the new AIG framework, first arithmetic principles that were explained to the test-takers and define operations independent of the actual combination in number series. Second, combination principles that do not represent stand-alone rules but define the mathematic operations that provide the basis for combining rules into actual items.

Four simple rules requiring only the operations addition and subtraction were identified based on inspection of existing number series tests. The set of rules was pretested in several smaller groups of university students willing to contribute to the development of a new measure. All four rules did not pose high demands on mathematical knowledge or numeracy, were easy to understand and to apply.

R1: *Constant (Const)*: As a first rule, a constant number c could be introduced as an element of the formula needed to continue a given number series.

$$\text{Const} = c, \text{ with } c \in \mathbb{N} \quad (4.1)$$

.

This rule simply means that a constant is introduced, not matter what operation is actually applied to the constant. In some existing number series tests, the introduction of a constant was defined not separately but directly combined with addition or subtraction of a constant. In the current framework, the use of a constant in a series is distinguished from what operation is actually performed with the constant, defined as the combination principle (see below).

R2: *Checksum type 1 (CS1)*: As a second rule, the checksum of one element y of the number series could be calculated, such as:

$$\text{CS1} = \text{checksum}(y_i). \quad (4.2)$$

This rule only defines the type of operation needed to solve a number series, not necessarily the complete logical pattern of the series itself. Depending on what combination principle is used, the result of this checksum rule could directly give the next element of the series, or an intermediate result to be combined with another rule.

R3: *Checksum type 2 (CS2)*: As a third rule, the checksum across two consecutive elements of the number series could be calculated, such as:

$$\text{CS2} = \text{checksum}(y_{i-1}) + \text{checksum}(y_i). \quad (4.3)$$

Again, this rule only defines the type of operation needed to solve a number series, not necessarily the actual logical pattern of the series itself. Depending on what combination principle is used, the result of this checksum rule could directly give the next element of the series, or an intermediate result to be combined with another rule.

R4: *Fibonacci (Fib)*: As a fourth rule, the sum of the last two elements could be calculated:

$$\text{Fib} = y_i + y_{i-1} \quad (4.4)$$

This rule uses the same principle as the well-known *Fibonacci sequence* based on the works of the Italian mathematician Leonardo Fibonacci in the middle ages. In the Fibonacci sequence of numbers, each number is the sum of the previous two numbers, starting with 0 and 1. Depending on what combination principle is used, the result of this rule could directly give the next element of the series (see example (a) in Figure 4.3), or an intermediate result to be combined with another rule (see example (c) in Figure 4.3).

The four rules used differ in their cognitive complexity. The application of some rules requires only 1 operation while others require several cognitive steps. Previous research (e.g., Porsch, 2007) showed that one driver of item difficulty in number series tasks besides the complexity of the actual arithmetic rules was the combination of several rules in one step. For the new AIG framework, it was decided to follow this rationale, i.e. generating item difficulty by combining several rules in one step. Three different logical principles (in the following denoted as *combination principles*, CP) were used, first an additive combination, second a subtractive combination, and third a combination by using the result of one rule as an argument in another rule.

CP1 *Addition (Add)*: When this CP was used in a series, the mathematic operation addition had to be applied by the test-taker to derive a consecutive element of the number series. For instance, the result of the CS1 rule or a specific constant c could be added to the current element of the series:

$$y_{i+1} = y_i \pm \text{checksum}(y_i), \quad (4.5)$$

$$y_{i+1} = y_i \pm c. \quad (4.6)$$

Another example could be the additive combination of the results of two individual rules, such as:

$$y_{i+1} = \text{checksum}(y_i) \underline{+} c. \quad (4.7)$$

Here, each element of the number sequence is derived as the checksum of the previous element plus a constant. Note that, Addition refers to the combination of rules in actual test items, not to the mathematical operation Addition when part of a rule itself (such as the addition part of the Fib rule). This distinction is made here to account for possible differences in cognitive processing when (a) a previously instructed rule involves the mathematical operation addition, or (b) the test-taker has to infer from the sequence of numbers that the addition operation is used to combine or manipulate results of these known rules. In order to make clear what additions are referred to here, the operations defining the CP are underlined in the equations. Note that, while the result of one rule is added to the current element of the series, y_i in equations 4.2.1 and 4.2.1, equation 4.2.1 involves the addition of two (intermediate) results. In the current framework, this difference is formalized by specifying one parameter less in the design matrix for the former case (i.e., no separate basic parameter is specified for the inclusion of y_i). A number of small pilot testings with more complex items showed that using both variants in one item (such as $y_{i+1} = y_i \underline{+} \text{checksum}(y_i) \underline{+} c$) were problematic both in terms of their difficulty level, and of complying to the constraint not to exceed the maximum value of 99 in a given series.

CP2 *Subtraction (Sub)*: This CP is equivalent to CP1 with the only difference that the mathematical operation subtraction is used instead of addition. An example for the application of this CP could be:

$$y_{i+1} = [y_i + y_{i-1}] \underline{-} [\text{checksum}(y_i)]. \quad (4.8)$$

Here, each element of the number sequence is derived as the difference of the result of the Fibonacci rule and the Checksum Type 1 rule. The rectangular brackets indicate the two rules.

CP3 *Sequence (Sequ)*: When this CP is used in a series, the result of a first rule was used as an argument in a second rule, i.e. a certain sequence of rule application has to be followed. For instance, the CS1 rule could be applied to the result of the Fib rule:

$$y_{i+1} = \text{checksum}(y_i + y_{i-1}).$$

Number Series							X0	X1	c	
(a)	X0	X1								
	2	6	5	8	10	15	22	2	6	3
			-1	+3	+2	+5	+7			
(b)	2	6	6	10	14	22	34	2	6	2
		+0	+4	+4	+8	+12				
(c)	2	6	7	14	20	33	52	2	6	1
		+1	+7	+6	+13	+19				
(d)	8	7	11	14	21	31	48	8	7	4
		+4	+3	+7	+10	+17				
(e)	8	7	12	16	25	33	60	8	7	3
		+5	+4	+9	+8	+27				

Figure 4.4.

Illustration of surface differences in structurally identical items caused by variation of item incidentals; all items are generated based on the same underlying logical structure, $y_{i+1} = y_i + y_{i-1} - c$.

Here, the Fibonacci rule has to be applied first, and then the checksum rule is needed in a second step. Figure 4.3 (d) gives an example for this rule-combination. While CP1 and CP2 allow the separate application of the two rules before combining the results afterwards, CP3 requires a specific sequence of applying the two rules.

Item incidentals Figure 4.4 shows 5 NST items that were constructed from exactly the same item radicals, *Fib*, and *Const*, using CP2 (subtraction) to combine these two rules. The different “look” of the items is due to variation of item incidentals. Given that the radicals capture the main sources of variation in item difficulty there should not be any systematic differences in difficulties for all 5 items. In the extreme case of within-family variance of zero the difficulty parameters for these items should be exactly the same.

In the following, all item incidentals are described in detail. In total, two incidentals are distinguished.

1. *Starting numbers:* First, the specific numerical values for the starting numbers are set incidental. As for the application of several rules, two previous elements of the series had to be used in the calculation, two initial elements, X_0 and X_1 , were randomly chosen for each item. The first element, X_0 , was, then, not displayed as part of the actual item. This procedure was chosen to guarantee inherent logical number series. Given the “random component” in the first numbers, test-takers might detect inconsistencies (e.g., negative numbers) when tracing back the series to the element prior to the first element shown. By generating one element more, it could be guaranteed that all numbers shown, also the first, were consistent with the logical rules and fell into the range of allowed numerical values. X_0 and X_2 both had to be non-negative numbers within the range from 0 to 99. Also, numbers had to be chosen in a way that the complete series would not reach values higher than 99 or lower than 0 until the seventh element, X_6 . Figure 4.4 illustrates the choice of different numbers for X_0 and X_1 . In series (a)-(c), the numbers “2” and “6” were used, in series (d) and (e), the numbers “8” and “7” were used. All other things identical, the surface layout of these two series shows already considerable differences making it hard for the test-taker to see that the underlying structure is identical without actually analyzing the pattern of numbers. Also, as demonstrated in Figure 4.4, varying only these two incidentals leads to completely different patterns of differences between consecutive numbers. Strategies as those analyzed by Irle (1969) could not lead to successful solutions here. The non-applicability of shortcut strategies and heuristics such as the *differences strategy* was considered an important requirement for the assumption of unequivocal solution processes for each series.
2. *Constant:* Second, the numerical value of the constant was set incidental for all items that involved a constant. Based on the findings on the *Problem Size Effect*, however, values were not chosen completely at random. Only values between one and four were considered. The number four poses a natural limit to human working memory (Cowan, 2010). Figure 4.4 illustrates the effect on the surface appearance of the items caused by this incidental. Items (a)-(c) use exactly the same starting numbers but were generated using different values for c . Apart from the first number, X_1 , none of the elements or differences between consecutive elements is the same across the three variants of the same structural item template.

As shown in Figure 4.4, the definition of incidentals applied in the NST framework produces number series that differ considerably with regard to the surface pattern of numbers. This can have both beneficial and potentially distracting effects. Surface patterns, such as the repetition of the same number in two consecutive elements (see Figure 4.4 (b)), or a pattern across a subset of all numbers suggesting the presence of a rule which is in fact, when encoding the whole series, not valid are two examples of such potentially distracting

effects. For instance, the following two items were generated based on the same set of rules.

$$A \quad 6, \underline{4}, \underline{1}, \underline{5}, \underline{6}, ?$$

$$B \quad 4, 11, 6, 8, 5, ?$$

In both items, the Fibonacci and the Checksum type 1 rule are combined. In a first step, the sum of the two last elements of the series have to be computed (Fib). Then, the checksum rule is applied to the results of this rule. Only by inspecting the whole series can this rule-combination be identified. When only the last 4 elements are inspected (and the first element is ignored; see underlined elements) one can wrongly induce the Fibonacci rule for the item in set A, yielding the response “17” instead of the correct solution, “8”; for the item in set B, no patterning of numbers is as attractive when all but the first elements of the series are inspected. Even though the two items are, in theory, identical in terms of their cognitive structure, they look different. While item B has no distinctive patterning of numbers, the surface pattern in item A might suggest a wrong rule when only elements 2-5 are inspected. Based on the findings on the role of irrelevant relations summarized in the Introduction (Hersh, 1974; Huesmann & Cheng, 1973) it seems reasonable that test-takers might be influenced by the existence of certain irrelevant patterns in the new number series items. If, in contrast, item difficulties are not significantly affected by such surface patterns, this would be a good result for the robustness of the item-generation approach. It was decided to investigate the susceptibility of the NST item difficulty model to such surface patterns empirically instead of directly constraining the AIG model not to allow items to demonstrate such patterns.

4.2.2. Sample

Participants were recruited at two universities in Russia and Germany (B.Sc. students of psychology) and received feedback of their results as an incentive. The initial sample consisted of $N = 406$ persons (80 % female). 2 participants were excluded from data-analyses because they didn't understand the rules of the test, yielding a final sample for the analyses of $N = 404$ (229 Russian and 175 German students). The mean age was 21.45 years ($SD = 4.18$). 154 participants (38.0 %) reported prior experience with IQ tests, and 151 participants (37.3%) reported prior experience with number series tasks. This percentage of prior experience is representative (cf. meta-analytic findings by Hausknecht et al., 2007). All participants gave consent that their data be used for scientific purposes. The fact that data was collected in 2 countries makes the sample more heterogenous than typical samples for test development studies. Cross-cultural fairness problems have been reported especially for tests that were developed in one cultural setting and then, lat-

eron, carried over to test-takers not sharing the same cultural background as test-takers of the piloting and calibration samples. The investigation of a less homogeneous sample as early as during test development and pilot testing is supposed to produce results of higher generalizability. The hypotheses of this study refer not to cross-cultural performance differences on the new number series items. Due to the relatively small sample size, no explicit comparison between the samples and the item parameters in the two samples is made. The focus is on the parallelity of item sets, not on the comparison of person parameter estimates for test-takers with different cultural backgrounds. However, in order to make sure that conclusions regarding the research questions asked are not biased because of potential cross-cultural effects, analyses that are most central to evaluating the hypotheses were conducted separately for the two subsamples as well. These analyses are summarized in the Appendix.

4.2.3. Instruments and procedure

Number series A new set of number series items based on the item-generation framework described in the previous sections was administered in this study. Three items each for the 11 item types presented in Table 4.3, each defined by a specific item radical combination, were investigated¹. Every item type was represented by three exemplars all sharing the same structural characteristics, while differing in their incidentals. Changing the incidentals resulted in phenotypically different item isomorphs (see Irvine & Kyllonen, 2002).

Table 4.3 gives the design matrix for all 11 item types. In addition to the item radicals (left side of table), a formal description of each item type is given in the table (middle). Also, at the right hand side, the number of cognitive steps needed to follow to solve an item is given as an additional complexity indicator. Whereas the item radicals present in an item defines rules as chunks that are processed as a unit by the test-takers (e.g., the CS2 rule contains multiple steps itself but is conceptualized as one rule) while the number of steps describes the actual number of cognitive steps that need to be performed to arrive at the right solution. Miller (1956) introduced the concept of chunking to explain how people can overcome limitations in WM storage capacity. Chunking can reduce the cognitive demand of processing complex relations by recoding a high-dimensional relation into a lower-dimensional one. Table 4.4 gives a more detailed description of the 11 item types used.

A number of quality checks were performed for all items. For all items, it was checked whether it was possible to find the right solution without application of *all* rules present in the item and whether there were items allowing for multiple solutions. All items had

¹Note that initially 13 item types were designed. However, due to an error during manual item construction, one item type had to be dropped. Another item type was dropped because it was too difficult for the test-takers in the sample that was investigated.

Table 4.3.
Design matrix for item types of the NST (Study 2)

	Design Matrix						Formal description of item	Steps	
	Rules			Combination Principles					
	Const	CS1	CS2	Fib	CP1	CP2			CP3
1	1	0	0	0	1	0	0	$y_{i+1} = y_i + c$	1
2	0	0	0	1	0	0	0	$y_{i+1} = y_i + y_{i-1}$	1
3	1	0	0	0	0	1	0	$y_{i+1} = y_i - c$	1
4	0	0	1	0	0	0	0	$y_{i+1} = \text{checksum}(y_{i-1}) + \text{checksum}(y_i)$	3
5	0	1	0	1	0	0	1	$y_{i+1} = \text{checksum}(y_i + y_{i-1})$	2
6	1	0	1	0	1	0	0	$y_{i+1} = \text{checksum}(y_{i-1}) + \text{checksum}(y_i) - c$	4
7	1	0	0	1	0	1	0	$y_{i+1} = y_i + y_{i-1} - c$	2
8	0	1	0	0	1	0	0	$y_{i+1} = y_i + \text{checksum}(y_i)$	2
9	0	0	1	0	1	0	0	$y_{i+1} = y_i + \text{checksum}(y_{i-1}) + \text{checksum}(y_i)$	4
10	0	1	0	1	1	0	0	$y_{i+1} = y_i + y_{i-1} + \text{checksum}(y_i)$	3
11	0	1	0	1	0	1	0	$y_{i+1} = y_i + y_{i-1} - \text{checksum}(y_i)$	3

Note. Steps = number of mathematical operations needed to apply rules

Table 4.4.

Detailed description of all item types of the NST (Study 2)

Type	Description
1	For <i>item type 1</i> , $y_{i+1} = y_i + c$, any subsequent element of the series, y_{i+1} is given as the current element, y_i , plus a constant c . Only one mathematical operation (+) needs to be applied.
2	<i>Item type 2</i> , $y_{i+1} = y_i - c$, corresponds to type one, but the mathematical operation addition is exchanged for a subtraction. Any subsequent element of the series, y_{i+1} is given as the current element, y_i , minus a constant c .
3	<i>Item type 3</i> , $y_{i+1} = y_i + y_{i-1}$, is constructed from only one rule, Fib. Any subsequent element of the series, y_{i+1} is given as the sum of the current element, y_i , and the preceding element, y_{i-1} . Again, the number of steps equals one as only one mathematical operation (+) needs to be applied. The difference compared to item types 1 and 2 is that the preceding element is added instead of a constant. The test-taker can solve the number series without representing any number or intermediate result in working memory while applying the rules.
4	<i>Item type 4</i> , $y_{i+1} = \text{checksum}(y_{i-1}) + \text{checksum}(y_i)$, is constructed from only one rule as well, CS2. However, more steps are needed to apply this rule. Any subsequent element of the series, y_{i+1} is given as the sum of the two checksums $\text{checksum}(y_{i-1})$ and $\text{checksum}(y_i)$. This combination of steps was introduced as one rule to the participants in the instruction and can, therefore, be represented as one chunk by the test-taker. The number of steps denotes to three as test-takers have to perform three mathematical operations (compute the first checksum, compute the second checksum, add the two checksums to each other).
5	<i>Item type 5</i> , $y_{i+1} = \text{checksum}(y_i + y_{i-1})$, is constructed from two rules, namely CS1 and Fib, that are combined using CP3. First, the test-taker has to calculate the sum of the current element, y_i , and the preceding element, y_{i-1} . Second, the test-taker has to calculate the checksum of this intermediate result. The result from this calculation is the subsequent element of the series, y_{i+1} .
6	<i>Item type 6</i> , $y_{i+1} = \text{checksum}(y_{i-1}) + \text{checksum}(y_i) - c$, combines CS2 and C using the subtraction. This item type is similar to type 4, but with the additional step of subtracting a constant number, c . Therefore, the number of cognitive steps is four (compute the first checksum, compute the second checksum, add the two checksums to each other, subtract a constant).
7	<i>Item type 7</i> , $y_{i+1} = y_i + y_{i-1} - c$, combines Fib and C, again using subtraction as the combination principle. This type of number series extends item type 3 by adding the step of subtracting a constant. The number of cognitive steps is two (apply the fibonacci rule, subtract a constant).
8	<i>Item type 8</i> , $y_{i+1} = y_i + \text{checksum}(y_i)$, combines CS1 with CP1 by adding the result of the CS1 rule to the current element, y_i . This item type is similar to type 1, but with the additional step of calculating the checksum. Whereas in type 1, the same number is added to the current element in each part of the series, here the number that is added has to be calculated for each element of the series.
9	<i>Item type 9</i> , $y_{i+1} = y_i + \text{checksum}(y_{i-1}) + \text{checksum}(y_i)$ combines CS1 with CP1 by adding the result of the CS2 rule to the current element, y_i . This rule combination is very similar to type 8 with the only difference that CS2 is applied instead of CS1.
10	<i>Item type 10</i> , $y_{i+1} = y_i + y_{i-1} + \text{checksum}(y_i)$, combines Fib and CS1 using the first combination principle, addition. First, the sum of the current element, y_i , and the preceding element, y_{i-1} , has to be computed, and the result has to be represented in working memory. Second, CS1 has to be applied to the current element, y_i , and the result has to be represented in working memory as well. Third, the two intermediate results have to be added to each other, resulting in the subsequent element of the series.
11	<i>Item type 11</i> , $y_{i+1} = y_i + y_{i-1} - \text{checksum}(y_i)$ corresponds to item type 10, with the only difference that the subtraction rule is applied instead of addition.

	15 min	15 min	15 min
Instruction & Examples	Warm-Up Set 11 items initial order	Set A 11 items order changed	Set B 11 items order changed

Figure 4.5.

Test design and testing time for the NST (Study 2)

only one possible solution. However, some items differed from other items in that they “offered” a certain attractive wrong solution if individuals did not look at the pattern of the whole series. For instance, in the checksum type 2 item in the warmup-set, one could wrongly induce the rule “minus 2” if only the last three elements were considered. Only by looking at the whole series, one could identify the correct rule. It was decided to explicitly include these items in the test because this represents a realistic case for applied testing settings. Even though it might be technically possible to control for such surface patterns, a new item-generative framework would benefit tremendously from a sufficient robustness against such structurally irrelevant item characteristics.

The three item sets were administered one after another with short breaks in between. The order of the items was iterated between the sets. This is shown in Figure 4.5. The first set was considered a warm-up run to make subjects get used to the nature of the task and the rules to be applied. While only a few items were included as warm-up items in the other two studies presented in this thesis, an extended warm-up set was included here because of the specific research question regarding the equivalence of parallel item sets. Based on the literature review of other findings on number series it seems reasonable that processing strategies play an even more important role for this task type. Anastasi (1981) recommends to implement short orientation and practice sessions to establish equal or at least comparable testing conditions for all subjects. By including a warm-up item set, variation on one potentially disturbing third variable, i.e. test familiarity, should be controlled at least to a certain extent. In specific, it was decided to administer the complete set of all item types as a warm-up set to make sure that test-takers were equally familiar with all item types investigated here. The two parallel forms that are investigated in order to answer the main research questions are set A and set B. Analyses for the warm-up set will be reported to gain additional insights into the role of practice runs and test familiarity.

Participants were provided with comprehensive instruction material: The arithmetic rules were explained, together with example number series on several pages. Participants were informed that up to three rules could be combined in one item, and this was

also shown in a sample item. Participants were *not* instructed about the specific rule-combination principles. That is, they had to discover during test completion how the rules were combined. This procedure corresponds exactly to the procedure used in Study 1 of this thesis. Making participants familiar with the item content and the rules by providing them with comprehensive instruction materials helps to make sure that all participants have equal conditions and do not need to invent any new rules not considered for item construction. Beckmann (2008) demonstrated that an explanation of all rules beforehand actually amplifies the validity of the instrument. Only 2 subjects indicated that they did not understand at least one rule. Data from these 2 subjects was excluded from all consecutive data analyses.

Based on several pretests with a few subjects, a testing time of 15 minutes for each set (i.e., a total testing time of 45 minutes for 33 items) was given. Subjects were instructed in the beginning of each test to try to finish all items of one set within 15 minutes. All items for one set were printed on one page. Test-takers did not have to turn pages; all items of one set were accessible for the test-taker while working on the respective set. Each item had an equal chance of being completed. After 15 minutes, test-takers were reminded to come to an end with the given set and proceed with the next set. Compared to average response times allowed for widely used reasoning power tests, timing is allocated amply here (e. g., in the CFT-20R (Weiß, 2007), 56 items have to be completed in a total time of 14 min, yielding an average response time per item of 15 seconds).

School grades Participants were asked to report their most recent math grades as an additional validity indicator for the number series items administered. Grades were transformed to a common scale from 1 (best grade) to 5 (worst grade) in advance of the analyses.

Culture Fair Test (CFT 20) A subsample of 175 German participants also completed the four subtests from the revised Grundintelligenztest Skala 2 (CFT 20-R; Weiß, 2007), a German adaptation of the Culture Fair Intelligence Test, Scale 2 (Cattell, 1973). The CFT 20R is a paper-and-pencil test which provides high loadings on fluid intelligence (Cattell, 1968) and has good psychometric properties. It consists of four different subtests: Series completion, Classifications, Matrices and Topologies.

4.3. Results

The results section is structured along the main research questions of this study. First, results regarding the equivalence of the parallel test forms are presented. Second, findings on the internal cognitive structure of the items based on LLTM analyses are summarized.

Table 4.5.
Summary statistics for all instruments (Study 2)

Instrument	k	M	SD	Min	Max	Skewness	Kurtosis	α	Time	SD
NST (warm-up)	11	6.19	2.19	0.00	13.00	0.25	0.01	.69	14.56	1.94
NST (A)	11	6.21	2.30	0.00	13.00	-0.03	-0.02	.72	13.95	2.73
NST (B)	11	6.56	2.40	0.00	12.00	-0.22	-0.25	.73	13.11	3.09
NST (total)	33	18.96	6.20	1.00	37.00	-0.01	-0.08	.86	41.68	5.61
CFT 20-R	56	44.72	5.04	29.00	54.00	-0.56	0.10	.73	--	--
Last math grade ²	1	1.79	0.85	1.00	5.00	0.95	0.67	--	--	--

Note. ² self-reported school grades were transformed to a common 5-point scale (low values represent better grades).

Third, results for additional analyses investigating the CR format and relationships with other measures will be presented.

4.3.1. Equivalence of parallel test forms

Table 4.5 shows the average scores and reliabilities, and the average time used for each of the three sets. On average, participants were able to solve about half of the test items correctly, with considerable variation among test scores.

Test performance increased significantly from the warm-up set to set B ($F(2, 598) = 7.451, p = .001$). With an effect size of $d = 0.16$ this gain effect lies in the lower range of retest-effects reported by meta-analyses (Hausknecht et al. (2007)). The average time needed to complete each item set decreases by about one minute from the warm-up items to set B ($F(2, 784) = 38.625, p < .001$) with the largest decrease after the warm-up items. Time was significantly related to test performance ($r_1 = .192, r_2 = .230, r_3 = .261$), indicating that high achieving subjects were spending more time to answer the items than subjects pertaining lower numerical reasoning ability.

Table 4.6.
Correlations of the three parallel item sets with math grade and g (Study 2)

	Math grade	CFT 20-R
Warm-Up	-.251**	.436**
Set A	-.192*	.435**
Set B	-.265**	.425**

Note. * : $p < .05$, ** : $p < .01$.

Cronbach’s α is lowest for the first set and considerably higher for the second and third set. The internal consistency doesn’t reach the desired value of .80 for any of the individual item sets. When all 33 items are included, Cronbach’s α denotes to $\alpha = .861$. The correlations displayed in Table 4.6 show that the new number series test correlates significantly with both scholastic performance and general reasoning ability. The values of the correlations are similar across the three subsets of the instrument.

Rasch item parameters and fit statistics for all 33 items are summarized in Table 4.7, along with classical item statistics. The table is organized according to item types, not according to the sequence of items in the actual test. For the assessment of item model fit, z -transformed Q -indices (Rost & Davier, 1994) as well as item Infit statistics (see Linacre, 2010) were computed (see Table 4.7). Aside from a single item in the warm-up set, all items across the three sets show at least good Rasch model fit. All Infit statistics lie in the range $[0.5, 1.5]$ and can, therefore, be considered productive for measurement (Linacre, 2010). The 33 items generated from 11 item types cover a wide range of the difficulty-ability continuum: estimated item difficulty parameters σ_i range from -4.609 to 4.126 .

Three different “virtual item models” (e.g., Fischer, 1995; Mair & Hatzinger, 2007) as described in Chapter 2.4 of this thesis were modeled to investigate the parallelism of structurally identical item sets. All models are estimated with random-effects in order to account for the fact that a random variance difficulty component adds to the predicted difficulty by the linear-combination of item predictor variables.

1. *Model 1* is a “virtual item” model (Fischer, 1995) based on a design matrix with one item predictor for each item type (i.e., for each item radical configuration). This reflects the case of perfect item cloning. This corresponds to Sinharay et al.’s (2003) identical sibling model (ISM), i.e., exactly the same item difficulty is estimated for two items representing the same item type.
2. *Model 2* includes an additional difference parameter to model parameter changes from one item set to the other. This effect can be understood as a general item set effect affecting all items of this set in an identical way.
3. *Model 3* is a “simple” Rasch model with one parameter per item, i.e. an LLTM with a diagonal design matrix. The way it is written here, though, is that one parameter for each item type is specified (analogue with Models 1 and 2), and item-specific change-parameter for the differences in difficulty between the two sets are estimated. This is the model that was described in Equation ?? in Chapter ?. This model corresponds to Sinharay et al.’s (2003) unrelated sibling model (USM), i.e., the difficulties of two structurally identical items are different for all item types and all item difficulties are only defined by the respective item.

Only two item sets each were compared, treating the item sets A and B as equivalent, and the warm-up set as potentially more different because of its warm-up character. That

Table 4.7.
Three parallel NST sets, answer frequencies and item statistics

Itemset	Type	Number Series							Item statistics		Rasch model			
		X1	X2	X3	X4	X5	?	p	r_{it}	σ	SE	pQ	Infit	
W	1		5	8	11	14	17	20	.978	.107	-4.336	0.353	.293	0.966
A	2	1	11	14	17	20	23	26	.936	.455	-3.104	0.224	.063	1.027
B	3		3	6	9	12	15	18	.943	.122	-3.255	0.235	.142	1.078
W	1		35	32	29	26	23	20	.983	.150	-4.609	0.396	.470	0.833
A	2	2	28	25	22	19	16	13	.963	.201	-3.763	0.282	.419	0.892
B	3		20	17	14	11	8	5	.938	.467	-3.152	0.227	.034	1.059
W	1		7	9	16	25	41	66	.973	.142	-4.115	0.323	.343	0.880
A	2	3	3	7	10	17	27	44	.916	.272	-2.760	0.201	.311	0.941
B	3		2	8	10	18	28	46	.906	.188	-2.612	0.192	.511	0.870
W	1		13	10	5	6	11	8	.691	.401	-0.698	0.130	.076	1.036
A	2	4	9	16	16	14	12	8	.809	.146	-1.570	0.149	.783	0.866
B	3		7	16	14	12	8	11	.829	.326	-1.746	0.155	.875	0.819
W	1		3	7	1	8	9	8	.319	.357	1.431	0.126	.022	1.095
A	2	5	6	4	1	5	6	2	.376	.464	1.098	0.122	.496	0.961
B	3		4	11	6	8	5	4	.515	.496	0.332	0.120	.017	1.084
W	1		17	15	12	7	8	13	.554	.601	0.112	0.121	.999	0.767
A	2	6	19	13	12	5	6	9	.636	.550	-0.360	0.125	.960	0.821
B	3		14	11	5	5	8	11	.557	.445	0.098	0.121	.834	0.938
W	1		14	8	19	24	40	61	.463	.556	0.617	0.120	.982	0.836
A	2	7	6	5	8	10	15	22	.673	.541	-0.588	0.128	.978	0.821
B	3		7	12	16	25	38	60	.646	.387	-0.420	0.126	.686	0.912
W	1		6	12	15	21	24	30	.480	.179	0.522	0.120	.000	1.389
A	2	8	14	19	29	40	44	52	.220	.390	2.093	0.140	.271	0.978
B	3		11	13	17	25	32	37	.319	.487	1.431	0.126	.417	0.946
W	1		2	7	16	30	40	47	.099	.383	3.305	0.192	.723	0.865
A	2	9	10	14	20	27	38	58	.161	.512	2.593	0.157	.413	0.941
B	3		5	11	18	29	49	73	.156	.421	2.640	0.159	.744	0.883
W	1		2	7	16	30	49	92	.054	.350	4.126	0.251	.747	0.837
A	2	10	2	5	12	20	34	61	.161	.381	2.593	0.157	.984	0.732
B	3		1	5	11	18	38	67	.139	.345	2.817	0.166	.565	0.889
W	1		5	12	14	21	32	48	.082	.345	3.574	0.209	.619	0.829
A	2	11	2	10	11	19	20	37	.097	.371	3.341	0.194	.683	0.874
B	3		9	17	18	26	36	53	.069	.394	3.800	0.225	.632	0.887

Note. W= warm-up set

is, the primary comparison was the comparison of set A and B. This is a comparison of two structurally equal sets of items after a phase of getting familiar with the item-type.

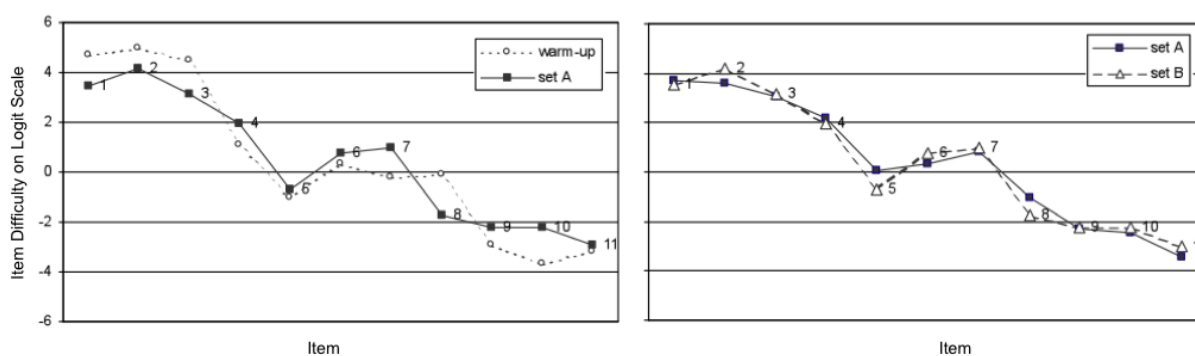


Figure 4.6.

Item parameters across parallel sets; left: warm-up set vs. set A; right: comparison of item parameters for set A vs. set B

An additional comparison of the warm-up set and set A served to quantify the effects of practice, or of getting familiar with the item types. Table 4.8 shows the parameters for the comparisons of item set A and item set B. These models are labelled Model 1a to Model 3a, with “a” denoting the primary comparison. Table 4.9 shows the parameters for the comparisons of the warm-up set and set A. These models are labelled Model 1b to Model 3b, with “b” denoting the secondary comparison. Figure 4.6 shows the congruence in item parameters for the 11 items constructed from the same vector of item radicals for the two comparisons. The logit for every item is plotted on the vertical axis against the item type on the horizontal axis. Smaller logits indicate more difficult items. Plots of the Item Characteristic Curves (ICCs) for all 11 item types and the three sets are given in the Appendix.

Parallelism of item sets A and B: From Table 4.8 and Figure 4.6 it can be seen that item difficulty parameters align mostly between sets A and B. Differences in item parameters are significant for items 2, 5, 6, 8 and 11, but these effects are not very large in size. However, they underline that, even though there is only very little overall variation in difficulty parameters between the two sets, there is some variation across parallel forms. This effect is not general in nature (as the nonsignificant overall difference parameter in Model 2a shows) but connected to specific items. AIC favors model 3a whereas BIC indicates best fit for Model 1a, i.e., the model comprising only one parameter for each item type. The LR statistics in Table 4.10 show that the model with one parameter for every actually administered item fits the data significantly better than the two sparser models. On the other hand, the random effects item variance is already close to zero in Model 1a ($s_e^2 = 0.0240$) indicating that almost all of the variation in item difficulties can be predicted based on the item type. Less than 1 percent of variation in item difficulties remains unexplained ($1 - R^2 = 1 - .996 = .004$) when the virtual item model is used instead of a Rasch model. In order to quantify the differences in parameters for parallel items,

Table 4.8.

Explanatory IRT modeling: “virtual item model” results, comparison of set A and set B

<i>Fixed Effects</i>	Model 1a		Model 2a		Model 3a	
	Est	SE	Est	SE	Est	SE
type1	6.7942**	0.28	6.7938**	0.28	7.1069**	0.33
type2	7.0471**	0.29	7.0460**	0.29	7.0028**	0.32
type3	6.2944**	0.26	6.2938**	0.26	6.4561**	0.30
type4	5.2566**	0.25	5.2562**	0.24	5.5828**	0.27
type5	2.8659**	0.23	2.8654**	0.23	3.4901**	0.25
type6	3.7212**	0.23	3.7207**	0.23	3.7277**	0.25
type7	4.0980**	0.23	4.0975**	0.23	4.2501**	0.25
type8	1.7988**	0.23	1.7981**	0.23	2.3584**	0.25
type9	0.9194**	0.24	0.9190**	0.24	1.1269**	0.27
type10	0.8336**	0.24	0.8333**	0.24	0.9509**	0.27
set B	---	---	0.0214	0.10	---	---
set B*type1	---	---	---	---	-0.1535	0.32
set B*type2	---	---	---	---	0.6178*	0.36
set B*type3	---	---	---	---	0.1496	0.27
set B*type4	---	---	---	---	-0.1772	0.21
set B*type5	---	---	---	---	-0.7871**	0.17
set B*type6	---	---	---	---	0.4623**	0.17
set B*type7	---	---	---	---	0.1690	0.18
set B*type8	---	---	---	---	-0.6796**	0.19
set B*type9	---	---	---	---	0.0476	0.22
set B*type10	---	---	---	---	0.2235	0.22
set B*type11	---	---	---	---	0.4392*	0.28
intercept (type 11)	-3.1835**	0.20	-3.1937**	0.20	-3.4173**	0.23
<i>Random Effects</i>	VAR	SE	VAR	SE	VAR	SE
$s_e^2(Item)$	0.0240	0.01	0.0236	0.01	0.0000	0.00
$\Delta s_e^2(Item)$	-99.59%		-99.60%		-100.00%	
$s_e^2(Person)$	2.1906	0.21	2.1905	0.21	2.2160	0.21
$\Delta s_e^2(Person)$	-0.76%		-0.75%		-1.93%	
Model Fit						
n	404		404		404	
ll	-3443.74		-3443.72		-3424.19	
df	13		14		24	
AIC	6913.48		6915.43		6896.38	
BIC	6965.50		6971.45		6992.41	

Notes. * $p < .05$ and ** $p < .01$ (2-sided); set B = general set effect; set B*type1-11 = item-specific effects (differences in item difficulty between set A and set B); information criteria for best fitting models in bold face.

Table 4.9.

Explanatory IRT modeling: “virtual item model” results, comparison of warm-up items and set A

<i>Fixed Effects</i>	Model 1b		Model 2b		Model 3b	
	Est	SE	Est	SE	Est	SE
type1	7.1030**	0.53	7.0993**	0.53	7.8844**	0.42
type2	7.6036**	0.55	7.6006**	0.55	8.1600**	0.45
type3	6.8158**	0.53	6.8123**	0.52	7.6605**	0.39
type4	4.5820**	0.50	4.5827**	0.50	4.2735**	0.24
type5	2.1929**	0.50	2.1932**	0.49	2.1301**	0.24
type6	3.5914**	0.50	3.5919**	0.49	3.4707**	0.24
type7	3.4487**	0.50	3.4492**	0.49	2.9623**	0.24
type8	2.1564**	0.50	2.1574**	0.49	3.0584**	0.24
type9	0.5040	0.50	0.5037	0.50	0.2583	0.27
type10	0.1477	0.51	0.1465	0.50	-0.5210	0.31
warm-up	---	---	0.0767	0.21	---	---
warm-up 1r	---	---	---	---	---	---
warm-up 2r	---	---	---	---	---	---
warm-up 3r	---	---	---	---	---	---
warm-up type1	---	---	---	---	-1.2389**	0.41
warm-up type2	---	---	---	---	-0.8540	0.48
warm-up type3	---	---	---	---	-1.3568**	0.38
warm-up type4	---	---	---	---	0.8567**	0.19
warm-up type5	---	---	---	---	0.3416*	0.17
warm-up type6	---	---	---	---	0.4690**	0.17
warm-up type7	---	---	---	---	1.2024**	0.17
warm-up type8	---	---	---	---	-1.6026**	0.18
warm-up type9	---	---	---	---	0.6968**	0.24
warm-up type10	---	---	---	---	1.4760**	0.28
warm-up type11	---	---	---	---	0.2238	0.27
intercept (type 11)	-3.0574	0.37	-3.0961	0.38	-3.1688	0.22
<i>Random Effects</i>	VAR	SE	VAR	SE	VAR	SE
$s_e^2(\text{Item})$	0.2178	0.08	0.2144	0.08	0.0000	0.00
$\Delta s_e^2(\text{Item})$	-97.06%		-97.10%		-100.00%	
$s_e^2(\text{Person})$	2.0596	0.20	2.0597	0.20	2.0965	0.20
$\Delta s_e^2(\text{Person})$	-0.31%		-0.32%		-2.11%	
Model Fit						
n	404		404		404	
ll	-3322.86		-3322.79		-3285.31	
df	13		14		24	
AIC	6671.71		6673.59		6618.61	
BIC	6723.73		6729.61		6714.64	

Notes. * $p < .05$ and ** $p < .01$ (2-sided); warm-up = general warm-up effect; warm-up type1-11 = item-type specific warm-up effects; information criteria for best fitting models in bold face.

Table 4.10.

LR model comparison tests for the the three different explanatory IRT models (Study 2)

Model comparison	Set A vs. Set B			Warm-up vs. Set A		
	LR χ^2	df	p	LR χ^2	df	p
Model 1 vs. 2	0.04	1	.8415	0.14	1	.7083
Model 1 vs. 3	39.10	11	.0001	75.10	11	< .0001
Model 2 vs. 3	39.06	10	< .0001	74.96	10	< .0001

absolute differences between logits for each item family across parallel sets were calculated. The average absolute difference in logits denoted to $M = 0.178$ logits ($SD = 0.124$) for set A and set B. That is, despite of the reduction of random effects variance to nearly zero in the explanatory model, parameter differences are present. Compared to the values reported by Zeuch (2011), the values here are small.

Warm-up effects: Substantial warm-up effects were found and are summarized in Table 4.9 and Figure 4.6. There is considerable variation in difficulty parameters for structurally identical items between the warm-up set and Set A. For 9 of 11 items, item-specific warm-up parameters (Model 3b) were significant, i.e. item difficulties changed considerably after the warm-up. There is no clear direction of an overall warm-up effect, i.e., participants did not perform better or worse on all items, but some item types were facilitated after warm-up while difficulties on others increased. In general, the range of item difficulties was reduced after warm-up. Both AIC and BIC show the best model fit for model 3b, that is, the RM is best suited for modeling item difficulties from the warm-up set and set A together. The virtual item model does not fit the data here. Random effects item variance can be reduced from 0.2178 in the virtual item model (Model 1b) to zero ($R^2 = 1$) in the Rasch model (Model 3b). 3 percent of the variation in item difficulties remains unexplained ($1 - R^2 = .03$) when the virtual item model is used instead of a Rasch model. The average absolute difference is $M = 0.469$ Logits ($SD = 0.232$) for the warm-up set and set A. These analyses show that differences in parameters are higher than for the two item sets A and B. Even though the virtual item model showed reasonable performance in terms of R^2 , the investigation of parameter differences showed that parameters in the warm-up set do not align with “true” parameters found after completion of the warm-up run.

4.3.2. Item difficulty modeling

Given the considerable warm-up effect, LLTM models were only run based on set A and B, excluding the 11 warm-up item from analyses. Two LLTM models of different complexity were estimated to address the second research question and test the appropriateness of

Table 4.11.

Rescaled item difficulty parameters for two different LLTM models for the NST (Study 2)

Item	RM	Rescaled Item Param.		Error in Prediction		Absolute Error	
		LLTM 1	LLTM 2	LLTM 1	LLTM 2	LLTM 1	LLTM 2
NST-A-01	3.2698	2.7711	3.1492	-0.4987	-0.1206	0.4987	0.1206
NST-A-02	3.1657	2.5254	3.3479	-0.6403	0.1822	0.6403	0.1822
NST-A-03	2.6189	2.7711	2.3726	0.1521	-0.2464	0.1521	0.2464
NST-A-04	1.7457	-0.0476	1.7506	-1.7933	0.0049	1.7933	0.0049
NST-A-05	-0.3470	-2.4831	-0.7174	-2.1361	-0.3704	2.1361	0.3704
NST-A-06	-0.1095	-0.8109	0.0675	-0.7015	0.1769	0.7015	0.1769
NST-A-07	0.4130	1.7620	0.8881	1.3490	0.4752	1.3490	0.4752
NST-A-08	-1.4788	-1.4741	-1.4823	0.0047	-0.0036	0.0047	0.0036
NST-A-09	-2.7103	-0.0476	-2.6659	2.6627	0.0443	2.6627	0.0443
NST-A-10	-2.8862	-2.4831	-2.9668	0.4031	-0.0806	0.4031	0.0806
NST-A-11	-3.8371	-2.4831	-3.7434	1.3540	0.0937	1.3540	0.0937
NST-B-01	3.1162	2.7711	3.1492	-0.3452	0.0330	0.3452	0.0330
NST-B-02	3.7835	2.5254	3.3479	-1.2581	-0.4356	1.2581	0.4356
NST-B-03	2.7685	2.7711	2.3726	0.0026	-0.3959	0.0026	0.3959
NST-B-04	1.5685	-0.0476	1.7506	-1.6161	0.1820	1.6161	0.1820
NST-B-05	-1.1341	-2.4831	-0.7174	-1.3490	0.4168	1.3490	0.4168
NST-B-06	0.3529	-0.8109	0.0675	-1.1638	-0.2854	1.1638	0.2854
NST-B-07	0.5820	1.7620	0.8881	1.1800	0.3061	1.1800	0.3061
NST-B-08	-2.1584	-1.4741	-1.4823	0.6843	0.6760	0.6843	0.6760
NST-B-09	-2.6627	-0.0476	-2.6659	2.6151	-0.0033	2.6151	0.0033
NST-B-10	-2.6627	-2.4831	-2.9668	0.1796	-0.3041	0.1796	0.3041
NST-B-11	-3.3980	-2.4831	-3.7434	0.9149	-0.3454	0.9149	0.3454

Note. Error in Prediction = Difference in sum-normed item difficulty parameters between RM and rescaled LLTM parameter; Absolute Error = Absolute value of Error in Prediction between Rasch and rescaled LLTM item difficulties.

the set of pre-specified item radicals to model item difficulties. The first model is an LLTM that includes only the four single explained rules, *Const*, *CS1*, *CS2*, and *Fib*. This model is used to estimate to what degree item difficulties can be explained by the set of individual rules alone, ignoring the combination principles. The second model is an LLTM with an extended design matrix that includes the same four radicals plus the three combination principles as additional radicals, i.e. in total 7 item-predictors.

Table 4.11 summarizes rescaled and sum-normalized LLTM item parameters for the 22 items of set A and B for the two different explanatory models, along with sum-normalized Rasch item difficulty parameters for a RM applied to the same 22 items. Differences and absolute differences between rescaled and actual item parameters are given at the right side of Table 4.11. Figure 4.7 gives a graphical summary of the alignment of parameters.

Several different criteria are considered to evaluate the fit of the LLTM model. First, the correlation of Rasch and rescaled LLTM parameters is investigated. This measure has

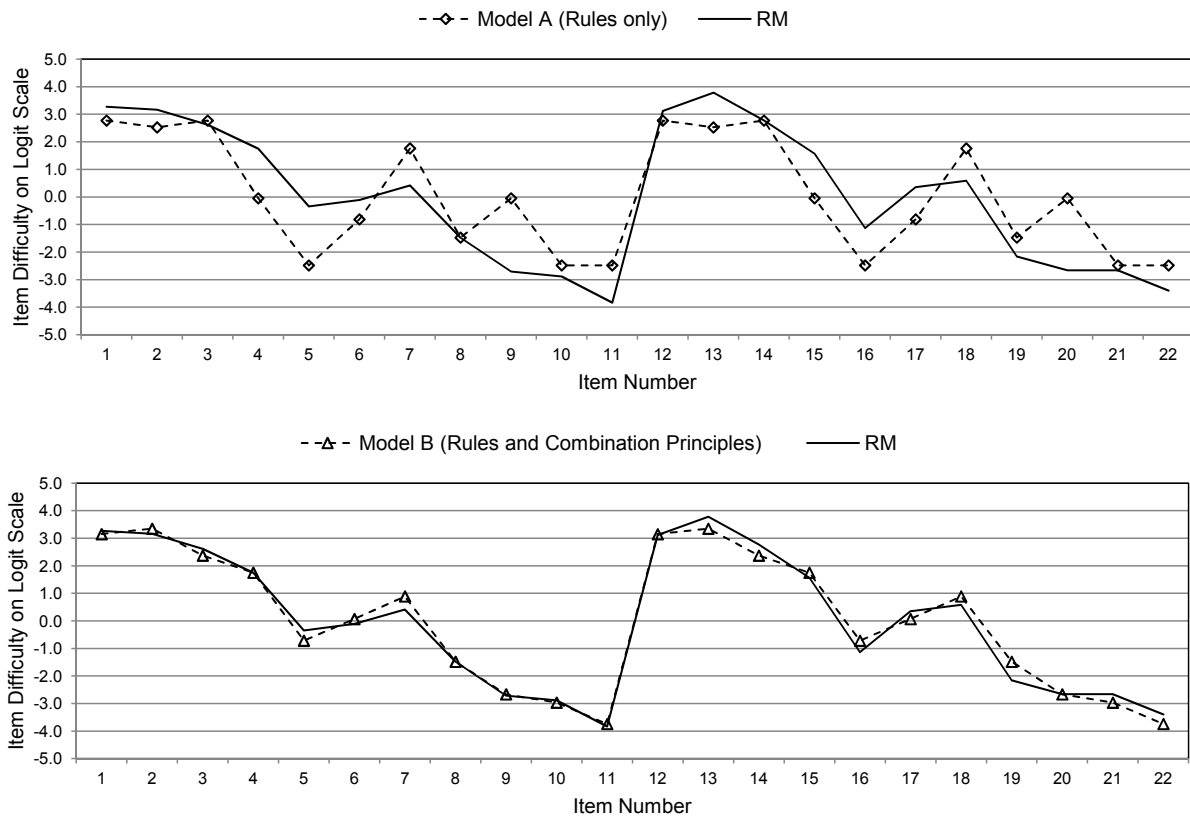


Figure 4.7.

Alignment of RM item difficulties and rescaled LLTM difficulties for the two LLTM models for the NST (Study 2)

been widely used in the literature to evaluate the construct representation of the LLTM. Correlations of $r > .80$ have been considered indicators of good construct representation (see e.g., Arendasy et al., 2007; Freund et al., 2008; Preckel, 2003). However, as already demonstrated in Study 1 in this thesis, the correlation alone might draw a much more favorable picture of the AIG model than a comparison of actual differences in LLTM and RM parameters suggests. When rescaled item parameters should be used without separate calibration in high-stakes settings, wrongly estimated difficulty levels can have severe consequences on ability estimates. For the Figural Analogy Test in Study 1, it was shown that substantial differences in parameters are found even for models with $R^2 > .80$. The picture for the NST looks somewhat similar.

In terms of the correlation with Rasch parameters results for both LLTM models are very promising: Parameters correlate $r = .8459$ ($R^2 = .716$) for Model 1 and $r = .9926$ ($R^2 = .985$) for the more elaborate Model 2. Around 70 percent of variation in item difficulties can be explained by the sparser LLTM model with only 4 parameters. This value

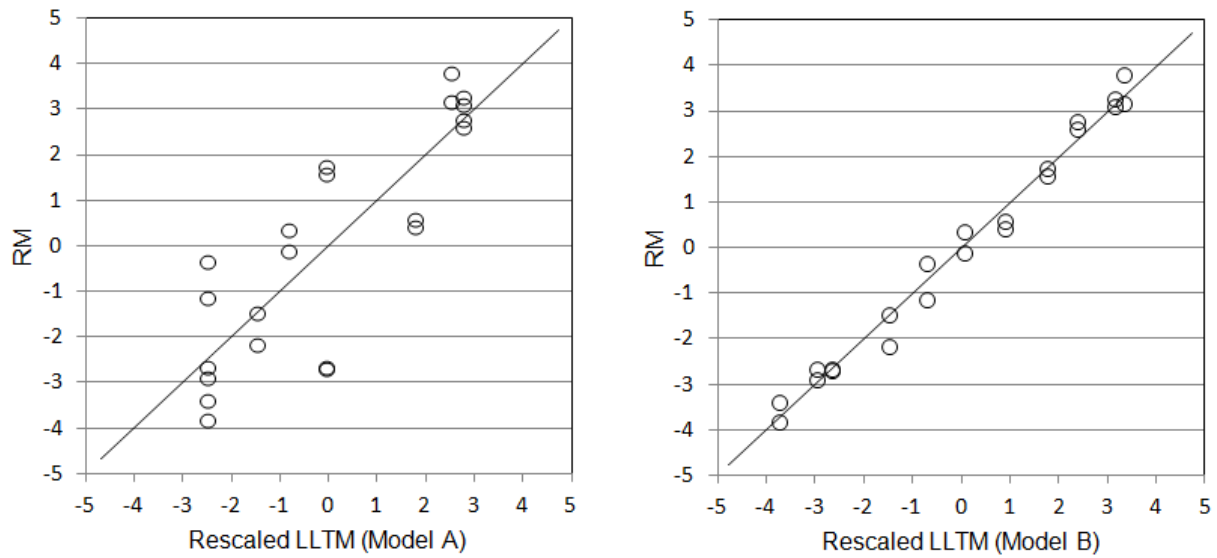


Figure 4.8.

Relationship of RM and rescaled item parameters for the two LLTM models for the NST (Study 2)

is very comparable to values reported for item-generation models presented in previous studies (e.g., Freund et al., 2008). The increase in model fit when the three additional combination principle radicals are included is tremendous. Only less than 2 percent of variation in item difficulties remains unexplained by the cognitive model with 7 parameters. This corresponds to a reduction in the number of item-predictors by 68.18%. Figure 4.8 illustrates the relationship of true and rescaled item parameters for the two models.

This figure gives already a more accurate picture of the explanatory power of the two models. Parameters lie very close to the diagonal for Model 2 but show substantial deviations for Model 1. Figure 4.9 gives a graphical summary of these deviations. While in Model 2, the average absolute error in prediction denotes to $M = 0.236$ Logits ($SD = 0.182$) with a maximum error of 0.676 logits for item 8 in set B, the average absolute error is $M = 1.046$ logits ($SD = 0.783$) for Model 1 with a maximum error of 2.66 logits for item 9 in set A. Facing these extreme errors in prediction, Model 1 seems not at all — despite for the relatively high correlation — a reasonable item difficulty model for the NST. Results for Model 2, though, are very promising. Item difficulties can be predicted reasonably well based on this item difficulty model.

Facet level parameter estimates for Model 2 were investigated more closely. Results for this model are shown in Table 4.12. Parameter estimates for all item-predictors reached statistical significance. Parameters for all rules reach statistical significance. The Fibonacci rule and the two checksum rules increase item difficulty. Items are especially more difficult if they contain the CS2 rule, here the logit decreases by around 3 points

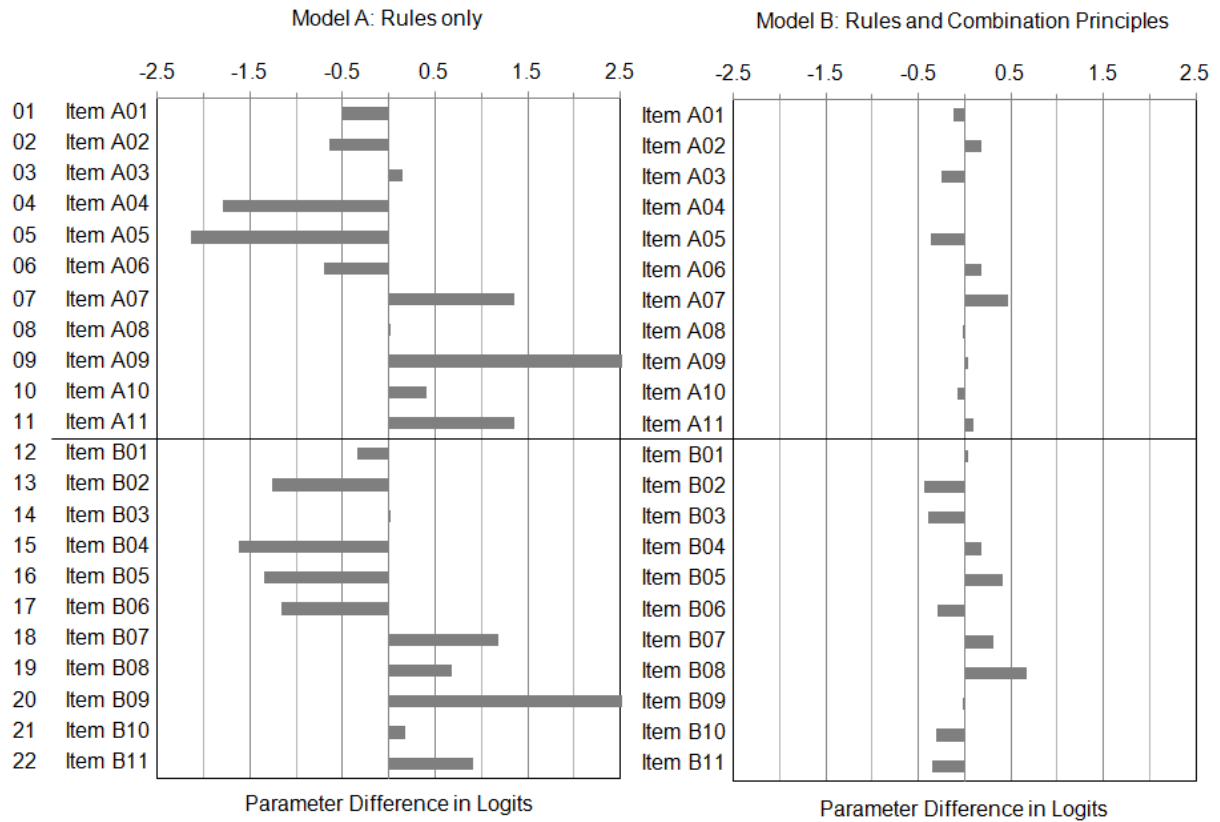


Figure 4.9. Parameter Differences in logits between RM and rescaled LLTM item difficulties; left: LLTM including explained rules only; right: LLTM including rules and combination principles.

Table 4.12.

Explanatory IRT modeling (LLTM) for the NST; Parameter estimates for LLTM with rules and combination principles.

<i>Fixed Effects</i>	Est	SE
Intercept	5.234**	0.252
Const	2.733**	0.270
CS1	-1.898**	0.303
CS2	-3.082**	0.222
Fib	-1.484**	0.183
Add	-4.417**	0.278
Sub	-5.193**	0.312
Comp	-2.167**	0.353
<i>Random Effects</i>	VAR	SE
$s_e^2(\text{Item})$	0.0584	0.0239
$s_e^2(\text{Person})$	2.1833	.20616
<i>Model Fit</i>		
ll	-3450.83	
df	10	
AIC	6921.653	
BIC	6992.578	

Note. * $p < .05$ and ** $p < .01$

($\beta_{CS2} = -3.082$) whereas the increase in item difficulty is lower for both the CS1 and Fib rules. The CS2 rule requires the test-taker to process more steps than the other two rules. The combination of multiple rules in one item represented an additional source of item complexity that would add to the individual contributions to item difficulty by each rule. The parameter estimates in Table 4.12 support this expectation. LLTM weights for all three combination principles are statistically significant and of considerable magnitude. Adding or subtracting a constant instead of a result of another rule reduces relational complexity in an item. According to the estimated item difficulty model, the probability of solving an item correctly increases by 2.733 Logits when a constant is used.

All analyses were conducted on a sample consisting student responses from a Russian and a German university. As mentioned in the introduction, the goal of this study was *not* to investigate cross-cultural differences or group differences. All main analyses were conducted on the full sample including Russian and German test-takers rather than incorporating the country as a grouping-variable into the models. However, to assure that facet-level results reported here are not distorted by potential cross-cultural bias, LLTM models were re-run for the two separate samples as well and compared for consistency. Results are summarized in the Appendix. There is a high consistency in parameters with no major differences in terms of the implications and conclusions. Only one parameter shows a larger difference, the complex combination principle of rules. It is important that this difference is investigated further in future studies to cross-validate this effect and understand why this radical might work differently across cultures. For the current study, there is no indication that conclusions regarding the research questions are distorted by the heterogeneity of the sample.

Inspection of constructed responses indicated that there were patterns of certain wrong responses that were chosen by a considerable number of test-takers. All wrong answers that were written down by at least 15 individuals are displayed in Table 4.13. A detailed qualitative analysis of these wrong answers showed that all of these solution attempts could, in fact, be explained by application of a wrong combination of rules. For some items, some test-takers seemed to be unable to induce a rule-combination and therefore used only one rule to compute the missing number. Also, some wrong responses reflect an application of rules that were not part of the set of allowed rules. A few takers applied, for instance, the rule +7 regardless of the instruction that stated that only the integers 1 – 4 could be used in connection with the addition rule. Further results for the investigation of items with distinct surface patterns are summarized in the Appendix. These analyses were done to assure the quality of the item-generation framework, but are not directly connected to the research questions of this study.

Table 4.13.
Frequent wrong answers for the NST and possible explanations (Study 2)

Set	F	WT	Number Series						Frequent wrong answers	
			X1	X2	X3	X4	X5	?	X	Possible Explanation
W	4	yes	13	10	5	6	11	8	17	Fib instead of CS2
A	4	yes	9	16	16	14	12	8	10	Sub instead of CS2
W	5	yes	3	7	1	8	9	8	17	Fib instead of FibCS1
A	5	yes	6	4	1	5	6	2	11	Fib instead of FibCS1
B	5	no	4	11	6	8	5	4	13	CS2 instead of FibCS1
W	6	no	17	15	12	7	8	13	15	Fib instead of QS2Sub
W	8	yes	6	12	15	21	24	30	27	Add instead of AddCS1
B	9	no	5	11	18	29	49	73	78	Fib instead of AddCS2
W	11	yes	5	12	14	21	32	48	39	Add (+7) instead of FibSubCS1
W	11	yes	5	12	14	21	32	48	50	FibSub (-3) instead of FibSubCS1
A	11	no	2	10	11	19	20	37	28	Add (+8) instead of FibSubCS1
B	11	no	9	17	18	26	36	53	44	Add (+8) instead of FibSubCS1

Note. Frequent wrong answers are answers with frequencies ≥ 15 ; WT (wrong track) codes items that suggested a specific wrong rule-combination if only a subset of all 5 elements of the series were analyzed; Set denotes the item set (W: warmup, A: set A, B: set B); F: item family

4.3.3. Predictive power of the model

As illustrated in the introduction, both RM and LLTM lie on a continuum of IRT models with more or less item-explanatory variables with the maximum number of independent facet parameters in the LLTM defined as the number of item indicators in the RM. The more explanatory parameters a model has, the higher its explanatory power tends to be. By definition, adding additional explanatory variables will, in the worst case, leave the overall model prediction unchanged but will never result in a decline. In the random effects model, the item-side random effect gives the magnitude of error (i.e., unexplained variation in item difficulties) for each model. In the RM this error variance is zero as the model is fully parametrized (i.e., the model contains one parameter for every item). The virtual item model contains only half of the parameters and the LLTM further reduces the number of explanatory variables. One important question when evaluating the practical usefulness of any proposed LLTM model is whether the explanatory model can predict reasonable proportions of item difficulties based on a relatively sparser set of explanatory parameters. As was shown above, the LLTM with arithmetic rules only performs much worse than the model with the additional combination principle parameters.

In the current study, both the average differences and explained variation statistics are compared for the different models that were estimated. Results are summarized in Table 4.14.

Table 4.14.

Predictive power and sparseness of different explanatory IRT models for the NST (Study 2)

Model	Item-Explanatory Variables		Predictive Power		
	no. of item predictors	reduction compared to RM	R^2	$s_e^2(\text{Item})$	AAE
RM	22	−0.0%	1	0	0
Virtual item model	11	−50.0%	.996	0.024	0.178
LLTM (Model B)	7	−68.18%	.985	0.059	0.236
LLTM (Model A)	4	−81.82%	.716	1.644	1.046

Note. AAE = average absolute difference in logits between original sum-normed RM item parameters and predicted sum-normed item parameters based on explanatory model

While the RM has 22 explanatory variables, the virtual item model has only 11 item predictors, and the two LLTM variants further reduce the number of item predictors to 7 and 4, respectively. The number of parameters is reduced by up to −81.82% compared to the Rasch Model. Random effects variance components and average absolute differences in Logits between original sum-normed Rasch parameters and rescaled sum-normed LLTM parameters show only small differences between the virtual item model and the LLTM with 7 radicals (Model B), but a substantial increase for the LLTM with the sparser design matrix of only 4 item radicals (Model A). This indicates that the simple model without the combination principles does not sufficiently describe the data and predict item parameters. In contrast, Model B fits only slightly worse than the theoretically less strong (in terms of Drasgow et al.’s classification) virtual item model.

4.4. Discussion

Number series are ideally suited for rule-based item generation because their algorithmic nature makes it relatively easy to formalize their structure and create a large pool of items. However, previous attempts to generate psychometrically parallel number series items based on the same underlying structural parameters have not been successful. While the algorithmic structure of a number series item can be defined easily, the cognitive processes of test-takers completing such items are much harder to formalize in item-generation models. Based on a comprehensive literature review several challenges for test-developers regarding the generation of number series items have been identified.

One of the most important specifics of number series is that such items do not have the same composite character like typical figural item types (e.g., Matrices, Analogies), that allows for an additive item difficulty modeling based on the classical LLTM. As soon as two operations are applied to one number in order to calculate the next number, it is not possible to induce one rule first by inspecting the number sequence, hold it in mind and

then induce the next rule. Test-takers have to represent several possible rule-combinations and intermediate results in working memory while solving number series tasks. It has been shown that the sequence of the rules is an important factor for differences in item difficulties of number series. Further, different strategic approaches can reduce complexity for some items but not necessarily for all items of a given logical structure. This can cause large differences in solution probabilities of supposedly parallel items (e.g., Porsch, 2007). Ambiguity is not only present in terms of the processing strategies used to solve number series items; some number series items used in existing cognitive test batteries do not have unique solutions. This problem was introduced as the “non-uniqueness problem” of number series in the theoretical background of this study. Depending on the set of rules that is induced, different possible responses must be considered appropriate. It is without question that the prediction of item difficulties for automatically generated items will be less accurate when items, themselves, already contain ambiguity.

Two starting points for the current study were the ideas that, first, if complexity could be manipulated in a way that several cognitive operations have to be processed simultaneously and these operations stay the same across a whole number series, this would be a great advantage for rule-based generation of number series. Second, if complexity generated by the combination of single cognitive rules could be captured by additional item-predictors, a better alignment of true item difficulties and rescaled item difficulties might be possible. Based on this rationale, a new set of item-generative rules was developed.

Porsch’s study revealed some problems associated with the complexity parameters introduced by Holzman et al. (1983) and thereby helped to choose a set of meaningful item radicals for the current study. Changes to the item-generative framework in this study included a reduction of complexity in terms of the number size and extent of mathematical knowledge needed to solve items (cf. findings on the *Problem Size Effect* reviewed in the previous sections), an explicit focus on rather homogeneous rules all based on only two arithmetic operations, and the abandonment of the two previously malfunctioning design principles *Periodicity* and *Hierarchy*. Furthermore, the elsewhere successful approach to provide precise explanations of all rules to the test-takers in advance of the assessment (e.g., Beckmann, 2008; Carlstedt et al., 2000; Freund et al., 2008) was also followed in this study. The intent was not only to generate a new number series test, but to show that the new item generation approach with modified sets of cognitive rules was more beneficial from a construct validity point of view. In the following, results will be discussed along the two main research questions.

4.4.1. Conclusions regarding the research questions

The first major research question was whether the new item-generative framework provided a basis for the generation of truly parallel tests. Items sharing the same underlying

structure should not differ in terms of their statistical properties. It was hypothesized that structurally equal items of the new test were equally difficult. The feasibility of the item-generation approach to generate parallel items based on specific radical configurations was tested. Item difficulties were predicted based on the specific radical combination (i.e., theoretically identical items) of the respective item. The general psychometric characteristics of test items created based on the new item-generative framework were very satisfactory. Items covered a large range on the difficulty continuum and showed good Rasch-fit. The test also showed very satisfactory levels of internal consistency with Cronbach's $\alpha > .85$. While high internal consistencies had been shown for other number series tests as well, the creation of items covering the whole ability-difficulty continuum had turned out less successful before (e.g., Porsch, 2007). Despite for a comprehensive instruction of all solution principles, the test did not reach the ceiling for almost all participants, even under quasi-power conditions. Internal consistencies were higher for Set A and B compared to a warm-up set, and less items showed slight deviations from Rasch fit in the sets completed after the warm-up run. These findings are promising because they address a practical concern that familiarity with test principles is often considered a threat to the validity of the instrument. In line with the first hypothesis, a high alignment between true item difficulties and item difficulties predicted based on the item families was found for the comparison of two structurally parallel item sets A and B. Model comparisons of a RM with one parameter for each item and a "virtual item model" with one parameter for each structurally unique item only indicated that the sparser model captured item difficulties only a little less accurately than the full model. The high alignment between the item parameters for Set A and Set B is a strong indication that test-takers actually used the same cognitive processes as expected based on theoretical assumptions.

While, after completion of the warm-up run, item difficulties could be predicted very reliably for two parallel test forms A and B, item difficulties during warm-up turned out to differ considerably from these difficulties. This result in line with Anastasi's (1981) test sophistication hypothesis, that test-takers have to get used to an instrument while working on the first set of items, and that the psychometric quality of the test improves after initial practice. The same explanatory model that fitted very well to Set A and B data did not fit to the joint data from set A and a warm-up run. There was no general decrease or increase in item difficulties after the warm-up, but items comprising multiple rules tended to become easier after the warm-up set while very simple items comprising only 1 rule tended to increase in difficulty after the warm-up set. Test-takers might have had a better representation of the rules and their possible combinations after the warm-up set, therefore performing better on items that required a combination of rules. The slightly worse performance on very easy items could be a side-effect of the cognitive representation of all rules and the application of explicit strategies to induce the combination of these rules. That is, while subjects might have detected the simple addition and subtraction of a constant immediately in the warm-up set by calculating the differences between all consecutive elements of the series. Then, for set A and B this "difference-strategy" turned

out to be not efficient for the identification of rule-combinations in the more complex items, and therefore test-takers might have tried to solve the easy items in set A and B also with more advanced strategies. Huesmann and Cheng (1973) have shown for mathematical induction tasks that test-takers often retest a hypothesis that has been rejected if the hypothesis was partially supported during a previous test on a similar problem. Taking a look into the answering patterns for set A and B showed, for instance, that some test-takers were able to come up with the right solution for a majority of all complex multiple rule items while leaving the answer-field for one or several easy items blank.

The second research question was whether and how well item difficulties could be predicted based on the hypothesized underlying cognitive processes. Two drivers for relational complexity were distinguished, the complexity of each individual mathematical rule, and the principles of combining rules in one item. Both rule complexity itself, and the combination of rules were expected to increase the relational complexity, and thereby the difficulty, of a given number series.

From a construct-validity standpoint, all changes to the item-generative framework proved successful. Effects for all item-explanatory variables were inconsistent with theoretical models of numerical reasoning and the process models for number series. Holzmann's Framework of the information processing steps for number series items conceptualizes relations detection as one core process successful test-takers need to complete when answering number series items. Depending on the complexity of the relations between elements of the number series, the amount of cognitive resources needed to solve an item varies. In the new NST framework complexity levels could be manipulated considerably by combination of a set of relatively simple arithmetic rules requiring only addition and subtraction. The correlation of scores on the new measure with general cognitive ability was higher than the association with math grades. This finding is in line with the goal to generate a test that largely captures numerical reasoning and not primarily arithmetic skills. The arithmetic operations that had to be performed in order to solve the items were all simple, involving just the operations of addition and subtraction. Correlations were rather stable across the three item sets, which is additional support for the validity of the measure. Earlier studies on the rule-based generation of number series have shown problems with the robustness of such frameworks against uncontrolled influences such as surface patterns in a series caused by specific incidentals, effects of specific numbers used or specifics due to certain mathematical operations. The new item-generation framework presented here was carefully designed especially to address these common problems with number series items. The applied LLTM model could replicate true Rasch difficulties very well when both rules and combination principles were included as explanatory variables. The alignment of parameters was not only satisfactory in terms of the correlative relationship between Rasch and rescaled LLTM parameters; also the absolute differences between true and predicted parameters were very small. This is a very satisfactory result because it demonstrates that the new item type could be a feasible candidate for a fully comput-

Number Series		Response Format																
(a)	<table border="1"><tr><td>19</td><td>13</td><td>12</td><td>5</td><td>6</td><td>X</td></tr></table>	19	13	12	5	6	X	X= <input type="text"/>										
19	13	12	5	6	X													
(b)	<table border="1"><tr><td>19</td><td>13</td><td>12</td><td>5</td><td>6</td><td>X</td></tr></table>	19	13	12	5	6	X	X= <table><tr><td>7</td><td>4</td><td>9</td><td>2</td><td>11</td></tr><tr><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td></tr></table>	7	4	9	2	11	○	○	○	○	○
19	13	12	5	6	X													
7	4	9	2	11														
○	○	○	○	○														
(c)	<table border="1"><tr><td>19</td><td>13</td><td>12</td><td>5</td><td>6</td><td>X1</td><td>X2</td></tr></table>	19	13	12	5	6	X1	X2	X1= <input type="text"/> X2= <input type="text"/>									
19	13	12	5	6	X1	X2												
(d)	<table border="1"><tr><td>19</td><td>13</td><td>12</td><td>5</td><td>X</td><td>9</td></tr></table>	19	13	12	5	X	9	X= <input type="text"/>										
19	13	12	5	X	9													

Figure 4.10.

Possible alternative future versions of the NST (Study 2)

erized adaptive and generative test. However, the pre-specified item radicals were only good predictors for item difficulties after subjects had completed an extensive warm-up run. This finding points to potential problems of using AIG techniques without individual item calibrations in operational testing settings. Usually, limited testing time does not allow for the inclusion of comprehensive training sections under operational high-stakes conditions. The finding that item difficulties in the warm-up set differed substantially from difficulties in the consecutive sets shows that practice on reasoning measures has an influence not only on the level of scores (see Freund & Holling, 2011 for a recent study on re-test effects for measures of fluid intelligence) but also on the meaning of the measured construct. Therefore, providing test-takers with adequate training opportunities in advance of the actual test completion seem even more relevant.

The development of the new NST generative-framework represents an important contribution to research on number series type item. While previous test-development efforts have usually manipulated the difficulty-level of a series by increasing the period length of a series and using hierarchical overlap between rules, the current framework can generate item along the complete item difficulty continuum without making use of these strategies. This is an important improvement because only when the formal-logical operations are exactly the same between all elements in the series, can item difficulty be estimated on the actual item level. If a different rule accounts for the transition from, for instance, element 3 to 4 than from element 4 to 5, no unequivocal difficulty estimates can be derived for

the complete *series*. Strictly, difficulty can then only be estimated for the derivation of the specific missing *element* of the series. Many of the problems reported by previous studies (e.g., Porsch, 2007; Ebert & Tack, 1974; Hersh, 1974) can be solved by means of this change to the logical structure of the series. Furthermore, the fact that the same formal logical operations account for all relations between consecutive elements of the series opens up new possibilities for variations of the response format. In the current study only one response format (i.e., constructed response, CR) with only one response instruction (i. e., “Please continue the number series by one element”) were investigated. Future studies could investigate NSTs with different response formats, for instance a multiple-choice format, or — more importantly — alternative response instructions (e.g., continue by 2 elements, or fill in missing element) and test whether item difficulties for different variants of the NST can be predicted based on the same set of item radicals. Figure 4.10 illustrates possible different administration modes for the NST. Such research could contribute to the better understanding of the contribution to item difficulties of incidentals at the test level. If the cognitive model is valid, item difficulties should be determined by the relational complexity of a series and not by surface characteristics of the items (e.g., the actual numbers used) or the complete test (e.g., different types of response instruction).

4.4.2. Limitations and future prospects

The conclusions of this study are limited by several factors. Five categories of limitations are distinguished in the following. For each limitation, possible directions for future research are discussed.

First, the robustness of the findings of this study in general as well as in cross-cultural settings needs to be demonstrated by consecutive studies. Replications with larger samples are necessary to cross-validate the findings with regard to the item generation approach and the explanatory IRT modeling strategy. Only a selected sample of test-takers has been investigated and the test was administered in a low-stakes context. Investigations with other populations of test-takers and a comparison of low- and high-stakes testing settings are still pending. More emphasis should also be laid on the cross-cultural fairness of the instrument. While, in this study, it has been shown that the explanatory model can be applied to data from a culturally very heterogeneous sample, future studies should explicitly model cross-cultural effects on the functioning of the cloning approach and the design parameters. Instead of modeling item difficulties based on LLTM models that model difficulties based on the same principles for the whole sample, Differential Item Functioning models could be applied to model interactions between item characteristics and person variables such as cultural background, gender, or levels of previous experience with similar tests. The sample sizes in the current study were not large enough to focus explicitly on a cross-cultural research question. In order to make definite conclusions about the cross-cultural comparability of test scores for the new item type, further re-

search is needed that includes cultural background as an explicit covariate. Also, stronger experimental design principles should be applied to replicate the findings of this study regarding the equivalence of the two item sets A and B and the parameter-changes after the warm-up run. The question of warm-up effects was not central to this study. In order to draw clear conclusions regarding the role of warm-up effects, the sequence of test forms should be fully iterated (i.e., test should be administered in 6 different combinations: {warm-up, A, B}, {warm-up, B, A}, {A, warm-up, B}, {B, warm-up, A}, {A, B, warm-up}, {B, A, warm-up}).

A second class of limitations refers to the statistical modeling of difficulties for structurally identical items. Only 3 items for each radical combination were investigated in the current study. No advanced item cloning models were applied. Therefore, the promising results concerning the accordance of parameter estimates in the two sets A and B and the good recovery of Rasch item difficulties by rescaled LLTM difficulties have to be considered preliminary. The results for the “virtual item model” need to be interpreted with caution. While this model has been proposed as an adequate model to investigate parameter changes across time (Fischer and Ponocny (1995), for instance, wrote: “Clearly, change always resides in the persons; for formal convenience, however, we shall equivalently model change in terms of the item parameters. To that end, we again introduce the notion of virtual items, in contrast to real items”, p. 357), it is a very simplified model that especially does not account for the covariance between item and time variables. Local independence is assumed for all items across all time points. This is a very strong assumption. More complex longitudinal models should be applied instead of the simple virtual item model to account for these possible effects in the future. The suitability of the statistical item difficulty model to predict item difficulties of on-the-fly generated test items in computerized adaptive testing (CAT) needs to be demonstrated. A further test of the prediction of item parameters by Item Cloning Models (ICM) would be very useful. If a larger number of parallel item “clones” could be administered to a larger sample of test-takers, item family parameters and within-family variances could be estimated by Bayesian models. Also, sequence effects could be ruled out if more clones are administered to larger samples and the order of the clones is completely iterated across the samples. In the current study, all participants worked first on one set of clones and then proceeded with a second set of structurally equivalent items. Strictly speaking, there is a dependency of answers between the first and the consecutive sets that is not modeled in the IRT models applied here. Models with covariance parameters that account for shared variance across item sets would be a helpful extension of the models used in this study.

Third, the item generation framework itself must be considered only a first step towards the design of a fully automatic item generator. The demonstration of a fully automatic computerized item generation is still pending. The equivalence of items administered in paper-pencil and computer-based format needs to be demonstrated. Beyond that, not all possible item radical combinations were investigated in this study. The number of 11

items is by far not enough for CAT applications where large numbers of items covering the whole ability continuum are needed. It was shown that the combination of multiple rules in one item was sufficient to generate very easy to very difficult items. Future works should investigate a more fine graduation of item difficulties based on different radical combinations. It would also be very useful to investigate the significance of instruction for the difficulties of specific rule-combinations in future studies. Verguts and De Boeck (2002), for instance, showed that rules that are instructed prior to testing are facilitated against rules that are not instructed, and that test-takers tend to try to solve items predominantly with these rules. In the current study, all test-takers received the same instruction. Only the rules but not their possible combination principles were instructed. It is an open question whether it would change the validity of the NST when, instead of introducing the rules only one at a time, all possible rule-combinations would be instructed in advance of the assessment as well.

Forth, more research on the warm-up effects found in this study is needed. It was shown that there was considerable variation in item difficulties between a warm-up set of items and two subsequent parallel item sets. Though consistent with the assumptions of previous researchers, the evidence from this study is at most preliminary. This study builds an ideal starting point for a set of more complex studies that could focus explicitly on practice and training effects on the functioning of psychometric item-generation frameworks. This could include explicit analyses of item parameter drift across sets of parallel items. An extension of parameter drift from the item-level to the facet or family-level would be another aspect worth investigating. Recent studies have shown that practice and training effects are a serious threat to the validity of test results from cognitive assessments (e.g., Freund & Holling, 2011). If the findings from the current study that practice effects diminish after a warm-up and do not influence the difficulty of parallel item sets completed afterwards could be confirmed by other studies, this would be a tremendous benefit for cognitive diagnostic assessment in general.

Fifth, the criterion-related validity of the new measure has been investigated only at a minimum level in this study. Further criterion variables that should be included are, for instance, performance on other number series tests and performance on arithmetic tasks. Also, an inclusion of additional person variables, such as test motivation or strategy knowledge, could be beneficial. It is important to identify relevant person variables that determine whether a certain strategy is followed or not. Noncognitive person characteristics might play an important role here. A closer look at relationships of test performance with other tasks and additional person variables would help to answer the question of construct validity of the new measure in more detail.

5

A cross-cultural investigation of the Latin Square Task

This study investigates the cross-cultural validity of the Latin Square Task (LST), a figural reasoning measure that can be generated based on a set of item-generative rules. Performance differences in cross-cultural settings are a well documented finding, but still only little is known about the bias-generating processes on the item-level. Two research questions are addressed. First, it is asked whether relational complexity theory is a cross-culturally valid framework to generate figural reasoning items. Second, it is investigated whether item difficulties are comparable across countries or whether Bias on the item level (i.e. Differential Item Functioning) is present. Differential Facet Functioning analyses (i.e., person-by-facet interactions) are applied to test whether item level DIF can be explained by the underlying structure of item radicals. Further, qualitative analyses of DIF versus Non-DIF items were conducted to achieve a better understanding of the generating processes for DIF in the LST. Cultural background was investigated in a broad sense by comparing students from two countries representing traditionally individualistic (Germany, $N = 452$) versus collectivistic (Russia, $N = 201$) cultures. Countries of medium cultural distance and moderate differences in school systems and educational expenditures per child were chosen. Additionally, performance on the LST dependent on the (non-)existence of test-specific pre-knowledge was investigated. Knowledge of the number-placement game SUDOKU was assessed as a proxy of relevant pre-knowledge that might facilitate LST performance. Results confirm the cross-cultural validity of the LST in a broad sense but also point to problems with the functioning of individual items in a cross-cultural context. DFF analyses could help to some extent understanding item DIF effects but more research is needed to clarify on the relationships between DIF and DFF. Item surface characteristics could be identified that contribute to the emergence of DIF and should be controlled in future studies or applications of the LST.

Keywords. Cross-cultural Bias, Differential Item Functioning, Differential Facet Functioning, Relational Complexity Theory, Latin Squares, SUDOKU

5.1. Introduction

Studies 1 and 2 of this thesis described the development of new item generative frameworks for two of the most important item types in reasoning assessment, figural analogies and number series. As one limitation to the results presented it was mentioned that the validity of the item generation framework in one country or one population is only a first step in establishing fully-functioning AIG engines that are applicable in a wide range of scenarios, including assessment in the global educational and workforce assessments arena. Cross-cultural comparability problems for well-functioning cognitive tests have been reported in the literature with studies pointing to the biggest cross-cultural effects for exactly those item types that are –in theory– meant to function independently of cultural influences, that is figural language-free fluid intelligence measures (see e.g., Brouwers et al., 2009; Carroll, 1993; see also Jensen, 1998 or Hartmann et al., 2007; Lynn & Owen, 1994; Te Nijenhuis & van der Flier, 2001). At the same time, little is known about why certain items and item-types do while other do not show DIF. Nearly twenty years ago, Angoff (1993) wrote that “it has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values” (p. 19). Unfortunately only little has changed since then. In many situations no explanation can be given why some substantively sound items show large DIF values statistically whereas some other items expected to be biased from the substantive analysis do not display DIF at all. Even though DIF research cannot be considered a new analysis strategy anymore, up to now, the focus has generally been still on detecting (and excluding) DIF items, not on the identification of item-related sources of DIF. Detecting the causes of measurement bias presupposes a theory about why items would show bias for the various groups studied, or what item facets specifically contribute to the emergence of DIF. Unfortunately, many tests at use lack such a strong underlying theory that would enable an empirical test of differential effects on the level of underlying item facets. In many cases, there is no theoretical framework available concerning content-related sources of DIF. Consequentially, hardly any studies have investigated cross-cultural bias on the level of item facets so far, and if so, studies have produced inconsistent results (see e.g., Van de Vijver, 2002; Xie & Wilson, 2008). For instance, Van de Vijver (2002) used a LLTM to examine the relationship between item difficulties and the difficulties of their constituent item-generating rules across three countries. Analyses of equivalence provided strong evidence for structural equivalence, but only partial evidence for measurement unit equivalence. Full score equivalence was, however, not supported. Still, the instruments used in their study lacked the specific theoretical base that is necessary to link invariance findings to theoretical models of item bias (e.g., Spearman’s cognitive complexity theory). Explanatory IRT modeling with person-by-facet interactions (or “Differential Facet Functioning” (DFF), Meulders & Xie, 2004) offers, in theory, to go beyond a mere quantification of DIF and explain and predict DIF effects. The advantage of models that

try to explain cross-cultural differences by characteristics of the task (opposed to models that only describe relationships with global country-characteristics in a post-hoc fashion) is that they allow to test *cognitive* assumptions about the emergence of cross-cultural performance differences. Even though bias might be statistically associated with factors such as educational expenditure or very broad cultural differences (e.g. collectivistic vs. individualistic cultures), such variables are not at the root of measurable bias at the item level; a true understanding why individuals from different cultures perform differently on cognitive tests in general, or certain items in specific, can only be gained when cognitive variables are considered as well.

Investigations of the relationship between DIF and DFF findings have so far produced, at best, vague results (cf. e.g., Xie and Wilson (2008): “the significant DFF parameters reflect the patterns of the significant DIF parameters to some extent”, p. 412). Xie and Wilson (2008) recommend to “always check the individual DIF estimates first” (p. 414). This is inconsistent with idea that DFF is a means to predict and better understand DIF effects. Only if DFF and DIF methods flagged essentially the same items could DFF be used accordingly. Meulders and Xie (2004) derived the DFF model as a special case of explanatory IRT models including person-by-item interaction effects. No empirical study has actually demonstrated that DIF-effects can be in fact predicted by DFF analyses in actual testing contexts. It is not clear what role mis-specifications of the LLTM design matrix and the general explanatory power of the LLTM model, for instance, play for the prediction of DIF based on person-by-facet interaction. The practical use of DFF methods is therefore still in question. Clearly, more research focusing on the origins and the nature of DIF in fluid intelligence measures is needed.

The study presented in this chapter focuses on the issue of cross-cultural comparability of AIG frameworks across cultural borders and takes the different perspectives to this topic into account. A holistic approach is taken that includes classical DIF analyses, explanatory IRT modeling (i.e., DFF), and a qualitative analysis of test items. Cross-cultural item functioning is investigated on the Latin Square Task (LST; Birney et al., 2006), a language-free rule-based reasoning measure that has been described and validated in a number of previous studies. First, background on the LST will be given, followed by a summary of research questions of the current study. Then, the study design and results will be presented and discussed.

5.1.1. The Latin Square Task

The *Latin Square Task* (LST; Birney et al., 2006) is a figural reasoning measure that is based on RC theory. A Latin square of order n is a matrix of $n \times n$ cells, filled with n symbols such that the same symbol never appears twice in the same row or column. The origin of those grids dates back to ancient Greece. Later, Leonard Euler proposed the name “Latin Squares” and studied them. In a typical LST item, some cells are filled

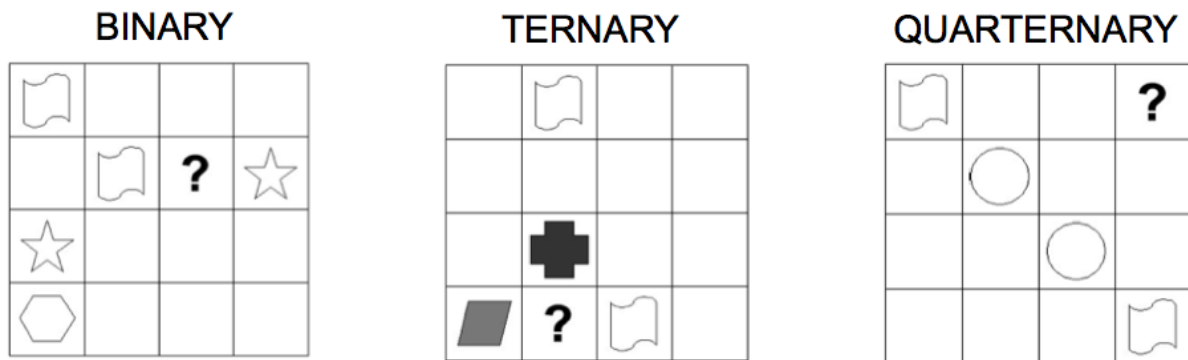


Figure 5.1.

Cognitive complexity determinants in the LST: Binary, ternary and quarternary processing (Study 3)

with geometric figures, whereas others are left blank. One of these empty cells contains a question mark. The test-taker has to select the correct geometric figure for this cell from a set of distractors. Figure 5.1 depicts three example LST items of different cognitive complexity. The cognitive complexity of a given LST can be described by three underlying processing variables:

- In *Binary processing* two sets of elements must be represented, i.e., the complete set of elements as displayed below the LST, as well as the given set of elements in a specific row (see Figure 5.1 for an illustration).
- In *Ternary processing* information from both a row and a column need to be integrated (see 5.1). Three distinct sets of elements must be cognitively represented, first the full set of elements, second the two elements in the lowest row, and third the element in the second column.
- In *Quarternary processing* elements across multiple rows and columns need to be integrated. A test-taker must take into consideration all possible symbols in one specific column or row while holding in mind all symbols in all rows or columns (see 5.1). This entails representing at least four pieces of information. It is assumed to be the most challenging operation because it has been shown that the upper limit of information processing in humans is met when four distinct elements must be simultaneously represented (cf. Cowan, 2010).

LST items can be designed in a very systematic and rule-based way, combining a strong prior theory based on cognitive psychology with growing empirical support. The LST minimizes the role of knowledge and storage capacity and thus refines the identification of a processing-capacity-related complexity effect in task performance. Therefore, its format offers optimal properties for the study of cross-cultural differences in reasoning performance.

Birney et al. (2006) found that cognitive complexity as defined by RC theory explained 64% of variance in item difficulties in the LST, showing that complexity as defined by RC theory is a powerful and theoretically sound predictor of item difficulty. Since the first presentation of the new task, the LST has been successfully applied in a number of studies as well as personnel selection scenarios (e.g., cut-e, 2011; Gold, 2008; Hoffmann, 2007; Kuhn, 2010) indicating that (a.) item difficulties can be, to a sufficient degree, explained by a manageable number of task parameters, and (b.) the relative weights of each of these parameters with regard to item difficulty are consistent with the assumptions of the underlying theory. The proposed explanatory models differ slightly but all include item-predictors for the three main complexity stages. For instance, (Zeuch, 2011) presented a model predicting item difficulties based on 7 item facets, namely “Binary: 1 Step”, “Binary: 2 Steps”, “Ternary: 1 Step”, “Ternary: 2 Steps”, “Ternary: 3 Steps”, “Ternary: 4 Steps”, and “Quarternary”. That is, she combined complexity stages with the number of processing steps when building item radicals. (Kuhn, 2010) used a slightly different modeling approach proposing 5 item predictors, “Binary”, “Ternary”, “Quarternary”, “Memory Load”, and “Size”. Memory load and Size were dichotomous indicators based on the number of processing steps necessary and the size of the grid (4x4 vs. 5x5). In terms of their predictive power, differences between these two modeling approaches appeared rather small ($R^2 = .86$ for Zeuch’s model versus $R^2 = .82$ for Kuhn’s model) with both “solutions” yielding results that could be interpreted clearly in line with the assumptions of RC Theory. Kuhn also analyzed what the three cognitive complexity parameters alone could explain, and how much additional variation in item difficulties could be accounted for by adding memory load and size to the model. His results showed a quite substantial improvement by including additional parameters. Cognitive complexity parameters alone accounted for only $R^2 = .42$ of variation in item difficulties, memory load alone reduced item random variance by $R^2 = .14$ (Kuhn, 2010, p. 101).

5.1.2. Similarities to the popular number placement game SUDOKU

The number-placement game “SUDOKU” (see Figure 5.2 for an example) has become one of the most popular cognitive puzzles during the last decade. For instance, a search on Amazon.com for the word “SUDOKU” reveals more than 8,000 results. SUDOKU puzzles share many characteristics with the LST as they derive directly from Latin squares. A standard SUDOKU puzzle is a 9×9 Latin Square filled with the digits 1 to 9, which is divided into nine 3×3 sub-matrices. Compared to the LST, one additional rule is added: each sub-matrix also must contain all digits 1 to 9. SUDOKU requires no arithmetic, no language skills or specific educational competencies. That is why, SUDOKU puzzles have been classified as tasks “of pure deduction [whose] solution depends ultimately on the ability to make valid deductive inferences” (Lee et al., 2008, p. 343). An important

		1		9		2	7
		9			2		5
2					3		
3				1	4		2
	8						4
1			2	8			5
			9				7
	1		3			9	
	4	6		7		5	

Figure 5.2.

Example SUDOKU puzzle (Study 3)

difference between the LST and SUDOKU is that the SUDOKU player is to fill *all* vacant cells, not only one cell marked by a question mark. Because of this, test-takers have more freedom to decide themselves how to proceed and what strategies to follow, whereas in the LST it can be clearly described what processes are needed to fill a certain empty cell, given the current configuration of filled and empty cells. In comparison to LST items, large SUDOKU puzzles rapidly become intractable.

As described in (Lee et al., 2008) Binary, ternary and quarternary processing are also key factors for the solution of SUDOKU puzzles. In order to solve a SUDOKU puzzle, individuals have to follow systematic sequences of elementary mental steps or so-called tactics. *Exclusion tactics* directly exclude possible digits from a particular target cell so that it can only contain one specific digit. *Inclusion tactics* use the occurrence of a digit in other cells in a set to infer that it must be included in the target cell.

Simple tactics rely on binary processing, that is if any set contains eight digits, then the empty cell in the set has the missing ninth digit. When simple tactics are used, individuals need to consider only one set to deduce the value of the target cell (Lee et al., 2008). Lee et al. showed in a number of experiments that with no instruction whatsoever individuals were able to discover most of the simple tactics for themselves. Such tactics are, however, not appropriate to solve any real SUDOKU puzzle, because the initial state of the puzzle never has eight digits in the same set. According to Lee et al., a crucial shift in strategy is therefore necessary to solve SUDOKU puzzles: individuals have to keep a record of the possible digits in cells, and to use advanced tactics that enable them to eliminate possible digits. Advanced tactics can be described as a two-step process. The first step is to infer a set of digits as the only possibilities for certain cells, and the second step is to use these possibilities to eliminate possibilities from other cells. The same is true for LST items involving ternary and quarternary processing: in order to find the missing

element, subjects have to represent the elements of some or all other cells in mind, i.e. keep a record of the sets of possible elements in the other empty cells.

Lee et al. (2008) showed that the greater the relational complexity, the more difficult a SUDOKU puzzle is, the longer it takes to solve a puzzle, and individuals are more likely to make mistakes. It has been demonstrated in a number of studies that SUDOKU-knowledge relates positively to LST performance (e.g., Kuhn, 2010; Zeuch, 2011).

However, it remains unclear whether SUDOKU knowledge changes the difficulties of the cognitive processes necessary to solve an item differently, thereby causing changes in construct validity. One of the major disagreements of the interpretation of the evidence presented in support of RC theory centers on the role of knowledge in processing and the determination of task complexity. For instance, taught or acquired processing strategies might change the effective RC of a task, thereby rendering items to be less difficult for SUDOKU players. This effect might also depend on the complexity of the processing step, with more complex rules (such as Quarternary) possibly influenced to a large extent. Specifically, because SUDOKU shares some important solution principles with the LST (cf. Lee et al., 2008) it is reasonable to assume that playing SUDOKU a lot could be an advantage when being confronted with LST for the first time or there might be a lasting advantage for SUDOKU players on the LST. That said, SUDOKU knowledge must not be ignored when performance on LST items is analyzed.

5.1.3. Cross-cultural validity of the LST

The cross-cultural validity of the LST has not been investigated so far. This is surprising because its strong foundation in RC theory and its rule-based structure makes the LST a prototypical instrument to investigate sources for Bias. In a recent study with German children, Kuhn (2010) investigated the structure of cognitive processes in the LST. Kuhn (2010) reports that quarternary processing was especially difficult, requiring the integration of four different information pieces into a temporary representation. Also, large inter-individual differences with regard to quarternary processing were found. Individuals use heterogeneous strategies with respect to quarternary processing (Kuhn, 2010). According to the cognitive complexity model, quarternary processing should be especially affected by cross-cultural factors.

From the cultural complexity perspective, surface characteristics of LST defined by incidentals seem of at least equal importance. While the contributions of the three processing variables (i.e., Binary, Ternary and Quarternary) have been analyzed in several studies, little attention was devoted to the role of item incidentals so far. The specific figural elements used in a LST, their color and size, or the visual distinctiveness of different possible patterns of partly filled Latin Squares have not been systematically investigated. Usually these characteristics have been conceptualized as incidental item features that are not expected to influence item difficulties. However, item incidentals might influence

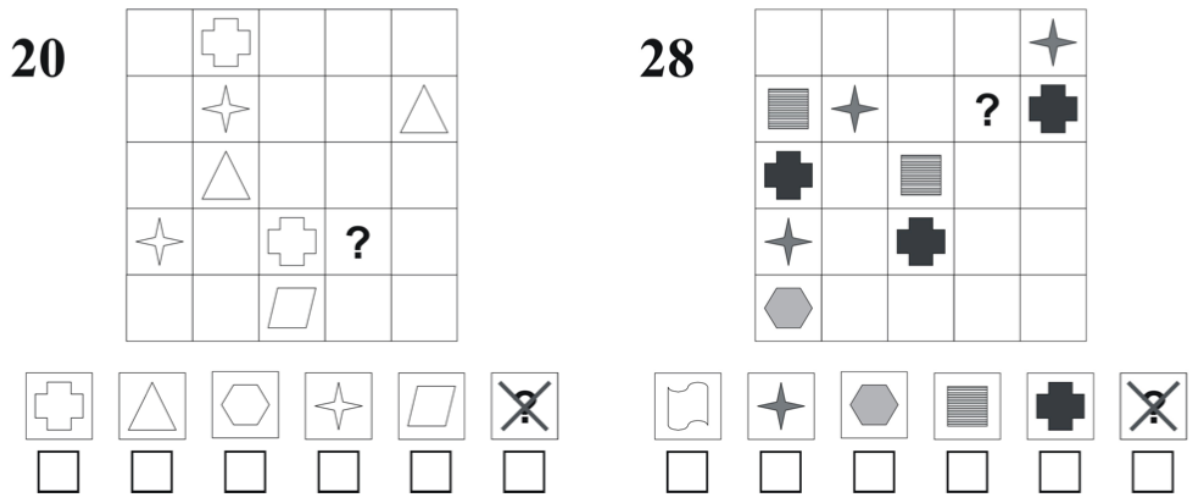


Figure 5.3.

Incidental item surface characteristics in the LST: example of two items based on the same item radicals (Study 3)

the cultural complexity of the task, independent of its cognitive complexity, and might be potential factors for the emergence of DIF. Figure 5.3 shows two LST items used in Holling et al. (2010) and Zeuch (2011) that share the same radicals but differ in terms of their incidentals. While the solution requires two Binary- and one Ternary-step for both items, the two LSTs differ in terms of multiple other characteristics. Elements are all without filling in item 20 whereas elements in item 28 are filled with different patterns. The shapes of the five elements used partly differ as well. 7 cells in LST20 are already filled, in LST28 this number is 9. The question mark is in a different position and the number of empty fields in the same row as the question mark differs between the two items (3 vs. 2). In LST20 no element appears more than twice in the LST, in LST28 two elements appear three times each. This list (which could still be extended) shows that surface characteristics might play a larger role for item difficulties in the LST than one could expect. Indications that this is the case can be found in previous studies as well. For example, Zeuch (2011) reported that “Obviously there are additional factors with impact on item difficulty apart from the investigated basic parameters.” (p. 57) and “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.” (p. 64). In addition, Zeuch, 2011 reported Rasch model misfit for some LST items while other items based on exactly the same structural components showed no indication of model-misfit. Two of these items are the two LST items shown in Figure 5.3. Zeuch (2011) excluded LST28 from further analyses due to poor Rasch fit. LST20 did not show any considerable misfit.

When LST items are generated based on AIG and Item Cloning models, reducing the influence of item surface characteristics on item properties is a goal of key importance. For high-stakes applications of the LST in selection contexts it is important to be able to draw test items from a large pool of items with known difficulties that are as well valid and fair test items for all possible test-takers. This includes the lack of DIF as well as the fit of newly generated items to the underlying test model. Especially when test results are used to justify selection decisions, the equivalence of item characteristics and the related statistics across different cultural groups is a key variable to test fairness. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) indicate that DIF in a test also diminishes the practical value of the assessment. As pointed out in a recent cross-cultural study on PISA data, ‘it would be economically and technically worthwhile if it were possible to detect items with potential of DIF *before* any test administration process. This might bring more effective test item writing practices and consequently more fair and valid cross cultural tests could be constructed.’ (Yildirim & Berberoglu, 2009, p. 110). One might think that the similarities between the LST and SUDOKU might be counterproductive for the use of the LST in selection practice because some individuals might have gathered strategy knowledge useful in LST as well, while others do not possess this advantage. Otherwise, both LST and SUDOKU puzzles build a strong theoretical basis for the investigation of test fairness and bias both across groups. For the most well-known reasoning measures, such as the Advanced Progressive Matrices (APM, Raven, 1962), there is no such an “equivalent” in popular everyday culture that would enable to test pre-knowledge effects in an ecological setting.

5.1.4. Research questions

The current study attempts to further investigate the cross-cultural comparability of test scores on reasoning measures using the Latin Square Task. Special attention is devoted to the identification of potential factors for DIF related to the radicals and incidentals in the AIG framework of the LST. Two broad research questions are addressed:

1. First, the cross-cultural validity of the LST is investigated in terms of its measurement properties, the internal structure of item difficulties based on the explanatory model applied, and the pattern of relationships with other cognitive and noncognitive variables. If reasoning processes are universal across countries, reasoning items based on an AIG framework grounded in RC-Theory generated in one country should not function considerably different in another country. That is, item difficulties should reside from the same underlying cognitive task principles in samples from one or another cultures. Furthermore, in order to be cross-culturally construct valid, there should be no substantial differences in the nomothetic span of the measures.

2. Second, it is investigated whether LST items function differentially across culturally diverse groups. Previous studies have shown that it is unreasonable to assume full measurement invariance for instruments administered in cross-cultural settings. It is investigated whether items requiring operations of higher cognitive complexity are more susceptible to cross-cultural bias than low-complexity items. Also, item surface characteristics will be investigated as one potential source of cross-cultural bias. Two different conceptualizations of “culture” are compared, a broad definition based on the country of living, and a very narrow conceptualization as the amount of shared prior knowledge that might facilitate reasoning test performance. Previous studies have shown that prior knowledge and experience with similar tests leads to improved performance on cognitive tests. An important research goal of this study is to identify factors in LST items that determine their cross-cultural comparability. The demonstration of the cross-cultural comparability of LST items is an important requirement for future applications of the LST in real-world assessment settings. Especially when AIG or Item Cloning approaches should be used to generate large numbers of items, knowledge about such factors could enhance the quality of the generated items.

In contrast to the first two studies, this study does not involve the derivation of a new item-generation framework but is based on a framework for figural reasoning items that has been described and validated previously. Consequentially, the focus of this study will mostly lie on the second general research goal outlined in the general introduction, that is, the value of item-generation models for a better understanding and improvement of construct validity of reasoning measures. Also, the goal of this study is *not* to study differences in reasoning ability between countries or cultures but to study item properties, especially their susceptibility for cross-cultural bias and the benefit of rule-based item generation models for enhancing the cross-cultural validity of reasoning measures.

5.2. Method

5.2.1. Sample

Participants were recruited at two universities in Russia and Germany and received feedback of their results as an incentive. The total sample consists of $N = 653$ persons (452 German students and 201 Russian students; 66.5% female). The mean age in the Russian subsample was 19.42 years ($SD = 1.42$), and in the German subsample 18.83 years ($SD = 0.98$). 21.4% of the Russian participants reported prior experience with IQ tests, and 46.8% reported that they had played Sudoku puzzles before. Among the German participants, these percentages were 48.5% and 66.3%, respectively. The percentage of prior test experience is representative (cf. meta-analytic findings by Hausknecht et al., 2007). All

participants gave consent that their data be used for scientific purposes. The countries were chosen because of their different cultural backgrounds, Germany representing an individualistic culture, Russia representing a more collectivistic culture. Also, differences in GDP and school life expectancy were considered when choosing the two countries. Meta-analytic findings show that educational expenditure is a significant predictor of country differences in mental test performance (Van de Vijver, 1997). Schooling does not have a formative influence on higher-order forms of thinking but tends to broaden the domains in which these skills can be successfully applied. Schooling facilitates the usage of skills by their training and by exposure to psychological and educational tests. Germany and Russia show considerable differences in educational systems and GDP (per capita). The GDP figures per capita for 2009 were US\$ 34200 and US\$ 15100 (Central Intelligence Agency, 2009), respectively. School life expectancy (i.e. the expected number of years of schooling that will be completed, including years spent repeating one or more grades) in the two countries is 16 and 14 years. That is, countries with considerable differences in school systems and educational expenditures per child were chosen. Note that the two countries were chosen as examples of traditionally individualistic vs. traditionally collectivistic societies. This study aimed *not* at the study of differences in reasoning ability between countries or cultures but at investigating item properties and their susceptibility for cross-cultural bias. Countries of medium cultural distance were chosen in order to guarantee an also practically meaningful comparison.

5.2.2. Instruments and procedure

Figural reasoning measures

Two figural reasoning measures were administered, one of them a rule-based generated test, the *Latin Square Task* (LST), the other one an established and cross-culturally applicable test, Cattell's *Culture Fair Test* (CFT). The LST was included as the primary measure that is focus of the research questions investigated. It allows to relate item difficulties to an item's relational complexity. The CFT was used as a control measure of fluid intelligence that is not core of the research questions. It was used to assure the comparability of the investigated groups and a necessary measure to investigate the nomothetic span of the LST across cultures. Testing conditions were hold exactly constant across the two samples. All tests were administered by the authors of this study under the assistance of local research assistants. The number of testers in one classroom was at least two.

Culture Fair Test (CFT 20) As a measure of general cognitive ability g , the four subtests from the Culture-Fair Test (CFT 20; Weiß, 2007), an adaptation of the Culture Fair Intelligence Test, Scale 2 (Cattell, 1973), were administered. The CFT 20 is a paper-and-

pencil test which provides high loadings on fluid intelligence and has good psychometric properties. It consists of four different subtests: Series completion, Classifications, Matrices and Topologies. Items were administered under timed conditions in line with the guidelines in the test manual. 56 items had to be completed in a total testing time of 14 min plus instruction time.

The Latin Square Task (LST) The same 30 item version of the LST as described in Holling et al. (2010) and Zeuch (2011) was administered to test-takers from both samples. Items were administered under power conditions, i.e. no time limit was given for the completion of the test. Time used by each test taker was registered by the test administrators. The rule-based item generation of the LST is described in detail elsewhere (e.g., Birney et al., 2006; Zeuch, 2011). The rule-based generation of the two item types used in study 1 and study 2 of this thesis were described in detail because they were newly developed as part of this thesis. The LST are not developed as a result of this thesis and, therefore, it is referred to the relevant literature regarding the specifics and details of the underlying item-generation.

Preceding the actual assessment, a detailed description of the item type at hand was given. All logical principles that determined the structure of the items (i.e. Binary, Ternary and Quarternary Processing) were explained in detail including examples for each rule. Participants were informed that several rules can be combined in one item, and this was also shown in a sample item. The procedure of the testing was explained and two warm-up items were given before the start of the actual test.

Experiences as relevant test-specific pre-knowledge

In addition to these two reasoning measures, relevant prior knowledge was assessed by asking subjects about their previous experiences with intelligence tests in general and cognitive SUDOKU puzzles in specific. SUDOKU puzzles share core characteristics with LSTs and must therefore be considered a relevant prior experience. As described above, binary, ternary and quarternary processing are also key factors for the solution of SUDOKU puzzles. Test-takers were asked to indicate whether they (a) had participated in an intelligence test before and (b) had played SUDOKU before.

Noncognitive variables

Personality variables were assessed to assure the comparability of the samples and provide variables to investigate the discriminant validity of the LST across the two cultural samples. The NEO-FFI personality inventory (Bodunov, Bezdenezhnykh, & Akexandrov, 1996; Borkenau & Ostendorf, 1993) was administered as an instrument measuring the big

Table 5.1.
Means and standard deviations for the German and Russian sample (Study 3)

	Germany		Russia		T	p^*
	Mean	SD	Mean	SD		
LST (raw score)	20.16	(5.414)	17.15	(4.889)	7.035	< .001
LST (time used in minutes)	43.16	(10.783)	38.96	(8.092)	5.179	< .001
CFT20 (raw score)	37.95	(3.980)	37.28	(4.463)	1.563	.120
Openness	2.47	(0.572)	2.45	(0.464)	0.472	.637
Conscientiousness	2.70	(0.568)	2.52	(0.593)	3.605	< .001
Extraversion	2.48	(0.443)	2.68	(0.490)	-5.029	< .001
Agreeableness	2.65	(0.508)	2.42	(0.541)	5.214	< .001
Neuroticism	1.68	(0.630)	1.70	(0.594)	-0.492	.623

Note. * p -value for significance of mean differences between the two samples; Cronbach's α for the 30-item LST version is $\alpha = .813$.

five personality dimensions. These are neuroticism, extraversion, openness, agreeableness, and conscientiousness. Subjects had to rate 60 statements on a five-point rating scale.

5.3. Results

There are five parts to the results section that address the research questions outlined above. Parts 1 and 2 present analyses conducted separately for Russian and German students, parts 3, 4, and 5 present analyses conducted on the joint sample with country as a grouping variable included in the statistical models. First, analyses regarding the general comparability of the two samples, the existence of potentially disturbing third variables, and the psychometric properties of the LST in both samples are presented. This part includes also results regarding the validity of the LST in terms of relationships with other measures in the two samples. Second, results from explanatory IRT modeling applied to each sample separately are presented, addressing the research question whether item difficulties can be explained based on the same set of item explanatory rules in the two countries. Third, results from classical and IRT-based DIF analyses are presented. Forth, results from DFF analyses, i.e. investigating country-by-facet interactions are presented. Fifth, findings from a qualitative analysis of LST items, specifically of potential explanations for DIF are presented.

5.3.1. Psychometric properties of the LST for the two samples

Table 5.1 shows mean scores and standard deviations for the Russian and the German sample. T statistics indicate whether mean scores differ between the two samples. On average, test-takers in the German sample were able to solve about three thirds of the 30 items correctly, with considerable variation among test scores. Test performance of the Russian test-takers is significantly lower with an average score of 17.15 items ($d_{\text{Score}} = -0.58, p < .001$). Russian students also spent significantly less time to work on the items ($d_{\text{Time}} = -0.52, p < .001$). With regard to general cognitive ability g as measured by the CFT20, no differences between the two samples were found ($d_{\text{CFT}} = -0.16, p = .120$). Scores on the Big Five personality factors partly differed between Germans and Russians, but these effect sizes are very small and practically negligible. Cronbach's α for all tests is satisfactory, exceeding the value of .80 (cf. Anastasi, 1981).

Category frequencies for all items for the Russian and German students were investigated to test the general comparability of the two samples (frequency-plots for all items are provided in the Appendix). The response patterns were very similar for most items. However, Russian test takers chose the option “not solvable” significantly more often than test-takers from the German sample ($t = -5.80, p < .001$). On average, German students ticked the “not solvable” option around 4 times ($M = 3.92$), which exactly corresponds to the actual number of not solvable items in the test, whereas Russian students chose the option around 1.5 times more frequently ($M = 5.43$). Figure C.8 in the Appendix illustrates the frequency distributions for choice of the “not solvable”-option for both samples. Correlations of “not solvable” choice with other measures, specifically CFT scores, values on the Big Five, and prior Sudoku experiences, were investigated to achieve a better understanding why students in the Russian sample might have picked this option more frequently. These analyses extend analyses presented in previous studies (e.g., Kuhn, 2010; Zeuch, 2011) that have used a “not solvable” answer category as well but mostly did not analyze answer behavior specifically on these items. While the number of “not solvable” choices was negatively related to performance on the LST ($r = -.239, p < .001$), it did not correlate significantly with any of the other measures (correlations ranged from $-.075$ to $.034$; all p -values $> .10$). That is, choice of this response category seemed to be unrelated to general cognitive ability, personality attributes, and pre-knowledge. Responses might be related to lower motivation of some Russian test takers as suggested by the negative correlation with response times ($r = -.173, p = .015$) in the Russian sample.

In order to rule out that any of the results regarding the research questions of this study could be biased because of the differences in relative frequencies for the endorsement of the “not solvable” option, it was decided to run central analyses not only for the full Russian sample but as well for a reduced sample. Two different “reduced” samples were created using different cut-off values for the choice of the not-solvable option, one representing a lenient, the other a strict cut-off (see also Figure C.8). First, students who ticked the “not

Table 5.2. Rasch parameters and item fit statistics for the LST in both samples (Study 3)

Item	Russia						Germany						
	σ	SE	Q	Z _Q	P _Q	Infit	σ	SE	Q	Z _Q	P _Q	Outfit	Infit
LST01	3.149	0.376	.410	0.592	.277	1.490	0.984	0.208	.256	0.191	.424	1.013	0.957
LST02	1.741	0.214	.476	2.095	.018	1.743	1.111	0.216	.418	1.653	.049	2.237	1.047
LST03	2.188	0.251	.256	-0.233	.592	1.286	0.901	0.120	.099	-2.784	.997	0.601	0.799
LST04	0.870	0.168	.183	-1.331	.908	0.760	0.889	0.219	.143	-1.959	.975	0.748	0.863
LST05	0.087	0.149	.288	0.642	.260	1.061	1.042	0.297	.284	1.468	.071	1.068	1.110
LST06	-0.071	0.148	.229	-0.466	.680	0.958	0.960	0.258	.263	0.973	.165	1.090	1.060
LST07	0.064	0.149	.229	-0.446	.672	0.955	0.960	0.405	.266	1.018	.154	1.071	1.065
LST08	0.393	0.155	.289	0.509	.305	0.998	1.043	0.297	.248	0.610	.271	0.963	1.053
LST09	-1.081	0.153	.297	0.591	.277	1.028	1.045	-0.777	.421	5.420	.001	1.544	1.376
LST10	-0.578	0.147	.281	0.455	.325	1.052	1.026	-0.840	.227	0.225	.411	1.040	1.011
LST11	-0.004	0.148	.145	-2.023	.978	0.793	0.849	0.057	.086	-3.445	.001	0.648	0.751
LST12	0.019	0.149	.177	-1.417	.922	0.853	0.892	0.518	.220	-0.112	.544	0.918	0.987
LST13	-0.204	0.147	.285	0.596	.276	1.053	1.040	-0.424	.312	2.457	.007	1.205	1.176
LST14	-0.248	0.147	.117	-2.575	.995	0.760	0.806	0.576	.179	-0.997	.841	0.802	0.929
LST15	0.041	0.149	.176	-1.430	.924	0.868	0.886	-0.246	.089	-3.475	.999	0.679	0.750
LST16	-0.204	0.147	.243	-0.211	.584	0.985	0.979	-0.358	.176	-1.192	.883	0.898	0.911
LST17	0.729	0.163	.249	-0.271	.607	0.958	0.960	0.878	.286	1.250	.106	1.106	1.079
LST18	0.297	0.153	.253	-0.104	.542	0.971	0.989	-0.132	.176	-1.153	.875	0.893	0.913
LST19	-1.941	0.181	.323	0.615	.269	1.072	1.024	-1.792	.150	-1.559	.941	0.782	0.870
LST20	0.110	0.150	.260	0.111	.456	0.985	1.009	-0.280	.138	-2.187	.986	0.773	0.847
LST21	-0.314	0.147	.259	0.107	.457	0.988	1.010	-0.445	.140	-2.135	.984	0.788	0.851
LST22	-0.424	0.147	.271	0.310	.378	1.007	1.026	-0.489	.174	-1.226	.890	0.918	0.908
LST23	-0.248	0.147	.351	1.829	.034	1.139	1.140	0.219	.262	0.963	.168	1.061	1.066
LST24	-0.071	0.148	.208	-0.856	.804	0.905	0.935	0.561	.287	1.419	.078	1.176	1.081
LST25	-0.314	0.147	.284	0.581	.280	1.050	1.040	-0.532	.273	1.413	.079	1.088	1.108
LST26	-1.659	0.169	.436	2.592	.005	1.318	1.188	-1.264	.321	2.681	.004	1.230	1.204
LST27	0.064	0.149	.204	-0.909	.818	0.894	0.932	-0.086	.136	-2.188	.986	0.764	0.846
LST28	-0.556	0.147	.321	1.207	.114	1.103	1.086	-0.456	.321	2.707	.003	1.252	1.189
LST29	-0.986	0.152	.259	-0.092	.537	0.965	0.989	-0.735	.194	-0.652	.743	0.935	0.953
LST30	-0.847	0.150	.250	-0.241	.595	0.961	0.979	-0.682	.236	0.467	.320	1.032	1.033

Note. σ : Difficulty parameter according to RM, Z_Q: z-standardized Q index; P_Q: probability P(Z > Z_Q).

Table 5.3.
Correlations between the LST and other variables (Study 3)

	Gender	Time	CFT20	O	C	E	A	N	Sudoku	other
RUS	.003	.360**	.360**	.125	-.055	.064	.067	-.113	.185**	.014
GER	-.044	.305**	.358*	.051	-.046	.001	.146*	-.074	.337**	.006

Note. * $p < .05$, ** $p < .01$; other=prior experience with other cognitive tests

solvable”-option more than 10 times were excluded from the sample. This corresponds to deleting cases where the “not solvable” option was chosen for more than one third of all items. This applied to 23 students in the Russian, and 8 student in the German sample. Second, as a more strict case, all students who ticked the “not solvable”-option more than 7 times were excluded from the sample. Four items were not solvable in the test, so under this procedure all students that chose the category at least twice as often as expected were removed from the sample. This applied to 50 students in the Russian, and 25 student in the German sample. The mean difference between Russian and German test-takers based on the lenient reduction still reached statistical significance (Difference = 0.49; $p = .022$), whereas no significant mean differences for the strict reduction could be found (Difference= 0.08, $p = .64$).

The applicability of the Rasch model to the LST in general, and the specific test version used here has been demonstrated in full detail in previous studies (Hoffmann, 2007; Holling et al., 2010; Kuhn, 2010). The same modeling strategy was applied here. However, separate RMs for each sample were run to make sure that the data did not deviate considerably from the previously reported findings. Rasch models for both samples were estimated using a conditional maximum likelihood estimator. For the assessment of item fit, z -transformed Q -indices (Rost & Davier, 1994) as well as Infit and Outfit statistics (Linacre, 2010) were applied. Item difficulties and fit statistics for the two samples are summarized in Table 5.2. Three items in the German sample and 8 items in the Russian sample showed some misfit on at least one of the three fit indices. Especially overfit in terms of the Q statistic was detected, i.e. items fit the Rasch model “too well” in sense of an almost Guttman-like deterministic answer pattern for theses items. No deviations from Rasch-scalability were detected based on the Infit statistic. All values lie in the range [0.5, 1.5] and can, therefore, be considered productive for measurement (Linacre, 2010). Interestingly, the items that show some degree of Rasch misfit are largely not the same items that were flagged in the study by Zeuch (2011).

Item difficulty parameters for the total sample of Russian test-takers were compared with parameters for the previously described subsamples of test-takers to account for possible differences due to frequent choices of the “not solvable” option. Sum-normed item difficulty parameters for the two subsamples were highly correlated with the parameters from the full sample ($r = .992$ for the lenient cut-off, $r = .982$ for the strict cut-off) with

no indication of systematic bias in parameters for the full compared to the two reduced samples (see Appendix for the exact values and a scatterplot). The average absolute difference in in sum-normed item difficulty parameters was 0.081 logits for the lenient cut-off and 0.126 logits for the strict cut-off. Based on these results it seems very unlikely that there are any systematic differences regarding the item difficulty structure based on the more frequent choice of the “not solvable” option in the Russian sample.

The correlations displayed in Table 5.3 show that no gender differences were present for the LST, whereas response time was substantially related to test performance in both samples, indicating that high achieving subjects were spending more time to answer the items than subjects pertaining lower reasoning ability. Performance on the LST was associated with general cognitive ability and not correlated substantially with any of the Big Five personality factors in both samples. Subjects with higher CFT scores also solved more LST items correctly. In general, the patterns of correlations were very similar in the German and the Russian sample. The impact of SUDOKU knowledge on LST performance differed between the two samples. The correlation of prior SUDOKU experience with test performance was high for the German test-takers, but lower in the Russian sample. All statistics were calculated again for the subsample of the Russian test-takers. Results showed that there were no considerable differences in the pattern of correlations between this subsample and the full Russian sample. Based on this finding, subsequent analyses were run on the full Russian sample. Note that, the analysis summarized in the previous section was to estimate the degree of cross-cultural validity of the LST, not to identify or exclude misfitting items in any of the two samples. If automatically and computerized and “on-the-fly” generated new LST items were to be administered as part of an operational assessment, a necessary assumption would be that all items that passed the generation process had sufficient statistical properties.

5.3.2. Item difficulty modeling for the two samples

In order to test whether the same cognitive model was suitable to model item difficulties in both samples (i.e., whether construct representation was given in both samples), an LLTM modeling strategy based on item facets defined by RC Theory was applied in line with previous LST applications (e.g., Zeuch, 2011). Two different LLTM model variants were estimated. The first model (LLTM 1) was an LLTM that included only the three relational complexity parameters (i.e., binary, ternary, and quarternary processing). This model could be directly derived from RC theory and included only the three predictors that are the core difficulty drivers if RC theory is applied. The sparseness of the model, however, restricts its explanatory power. Previous applications of the LST (e.g., Zeuch, 2011) have used more complex models that distinguish not only between the three RC parameters but also take the number of cognitive steps into account. Model 2 is an LLTM with an extended design matrix that includes 6 item covariates, namely two binary

Table 5.4.
LLTM with basic design matrix for each sample (Study 3)

<i>Fixed Effects</i>	Germany		Russia	
	Est.	SE	Est.	SE
Intercept	3.518**	0.472	3.055**	0.573
bin	-0.928**	0.293	-0.746**	0.351
ter	-1.763**	0.343	-1.913**	0.420
quar	-2.074**	0.394	-2.336**	0.478
<i>Random Effects</i>	VAR	SE	VAR	SE
$s_e^2(I)$	0.321	0.088	0.461	0.131
$\Delta s_e^2(I)$	-53.9%		-49.7%	
$s_e^2(P)$	0.873	0.077	0.519	0.073
$\Delta s_e^2(P)$	$\pm 0.0\%$		$\pm 0.0\%$	
<i>Model fit</i>				
n	454		203	
ll	-7527.189		-3669.660	
df	6		6	
AIC	15066.377		7351.321	
BIC	15091.086		7371.200	

Note. * $p < .05$, ** $p < .01$; Variance components in empty model containing only a constant and neither person nor item predictors: German sample, item variance: $s_e^2(I) = 0.698$ person variance: $s_e^2(P) = 0.873$; Russian sample, item variance: $s_e^2(I) = 0.917$ person variance: $s_e^2(P) = 0.519$

parameters (one step vs. two steps), three ternary parameters (one vs. two vs. three or more steps), and one parameter for quarternary processing (none of the items in the test combined more than one quarternary operation in one item, therefore only one parameters was included here). Results for both models are shown in Table 5.4 and Table 5.5.

Table 5.4 displays parameter estimates for LLTM 1 separately for each country. All three item facets contributed significantly to item difficulties and the order of the item facet difficulties was the same across the two data sets. Binary processing was the easiest item facet. When the relational complexity of a given LST was binary, the logit for a correct response to the respective item decreased by -0.928 in the German and -0.746 in the Russian sample. Ternary processing was more difficult, i.e. the logit changed by -1.763 and -1.913 , respectively. Relations of the highest complexity level, i.e. Quarternary, constituted the largest item facet parameters with -2.074 and -2.336 . Altogether, the three item facets explained $R^2 = 53.9\%$ (Germany) and $R^2 = 49.7\%$ (Russia) of variation

Table 5.5.
LLTM with extended design matrix for each sample (Study 3)

<i>Fixed Effects</i>	Germany		Russia	
	Est.	SE	Est.	SE
Intercept	3.384**	0.462	2.956**	0.569
bin1	-0.421**	0.280	-0.260**	0.340
bin2	-0.571**	0.313	-0.233**	0.380
ter1	-1.752**	0.366	-1.857**	0.453
ter2	-1.960**	0.424	-2.336**	0.526
ter3	-2.631**	0.444	-2.933**	0.551
quar	-2.114**	0.410	-2.427**	0.569
<i>Random Effects</i>	VAR	SE	VAR	SE
$s_e^2(I)$	0.287	0.079	0.419	0.118
$\Delta s_e^2(I)$	-58.89%		-54.31%	
$s_e^2(P)$	0.873	0.077	0.519	0.073
$\Delta s_e^2(P)$	$\pm 0\%$		$\pm 0\%$	
<i>Model fit</i>				
n	454		203	
ll	-7525.622		-3668.086	
df	9		9	
AIC	15069.244		7354.172	
BIC	15106.307		7383.991	

Note. * $p < .05$, ** $p < .01$; Variance components in empty model containing only a constant and neither person nor item predictors: German sample, item variance: $s_e^2(I) = 0.698$ person variance: $s_e^2(P) = 0.873$; Russian sample, item variance: $s_e^2(I) = 0.917$ person variance: $s_e^2(P) = 0.519$

in item difficulties. Item random effect variance could be reduced from $s_G^2(I) = 0.698$ ($s_R^2(I) = 0.917$) in a model without any item predictor to $s_G^2(I) = 0.321$ ($s_R^2(I) = 0.461$) in a model with the 3 relational complexity rules. This amount of variation explained by the cognitive model is comparable to that reported for some other automatically generated task types (e.g., Figural Matrices; Freund et al., 2008) but smaller than R^2 values reported previously for the LST when more complex models were used (e.g., Kuhn, 2010).

Table 5.5 shows results for the application of the LLTM in both countries. Not all 6 item facets contributed significantly to item difficulties in the two countries. Consistent across the two samples, only the parameters for ternary and quarternary processing reached statistical significance. The direction of all parameters (also the two non-significant pa-

rameters for binary) are in line with the hypothesis. Because of the random-effects LLTM model applied here, standard errors for the facet parameters are very large. The magnitudes of the parameters for Ternary and Quarternary processing are in line with the assumptions from RC theory. Difficulties increase with the relational complexity (i.e., one-step quarternary processing is more difficult than one-step ternary processing), but also with the number of processing steps (i.e., facet difficulties increase when operations have to be repeated two or three times). Altogether, the three item facets explained $R^2 = 59\%$ (Germany) and $R^2 = 54\%$ (Russia) of variation in item difficulties. .

In addition to these classical model comparisons of the LLTM and Rasch Model, standardized absolute errors were analyzed as proposed by Zeuch, 2011: Absolute differences between Rasch and rescaled LLTM item difficulty parameters were computed and standardized by dividing them by the standard errors of the Rasch parameters. Detailed results for the two alternative LLTM models for both samples are provided in the Appendix. On average, standardized absolute errors denoted to $M_{basic} = 2.902$ and $M_{ext} = 2.731$ for the Russian sample, and $M_{basic} = 4.096$ and $M_{ext} = 3.649$ in the German sample. Larger values in the German sample are due to the smaller standard errors in this, compared to the Russian sample, larger sample. These values are in the same range as values reported by Zeuch (2011) for the same LST items. Unstandardized absolute differences denote to 0.512 (LLTM 1) and 0.468 (LLTM 2) logits for the Russian sample and to 0.480 and 0.426 logits in the German sample, respectively.

5.3.3. Differential item functioning analyses

Uniform as well as non-uniform DIF was analyzed using both IRT and non-IRT methods. In total, five different methods were applied: Mantel-Haenszel (Mantel & Haenszel, 1959), Breslow-Day (Breslow & Day, 1980), uni- and non-uniform DIF in the Logistic Regression framework, and uniform DIF applying Lord’s χ^2 statistic (Lord, 1980) based on a 1PL model. These methods are summarized in detail in the Theoretical Background of this thesis. These statistics represent the whole range of possible DIF statistics. It was decided to include multiple different approaches to guarantee that DIF results are not driven by the assumptions of one specific model used. For instance, Lord’s method based on a 1PL is based on much stronger assumptions than the contingency-based methods.

All DIF analyses were run for two different grouping criteria. First, item responses for test-takers from Germany and Russia were compared. This represents the typical DIF approach where culturally diverse samples are compared (in the following labelled “cDIF”). Second, item responses for test-takers with and without self-reported SUDOKU knowledge were compared (in the following labelled “sDIF”). SUDOKU experiences were chosen as a very narrow operationalization of culture as the amount of shared prior knowledge that might facilitate reasoning test performance. Results for these analyses are summarized in Table 5.6 (cDIF) and Table 5.6 (sDIF). In addition to the statistics for five methods,

Table 5.6.
Results for country-dependent DIF in the LST (Study 3)

Item	Contingency-Table based			Logistic Regression						IRT Model based			
	Mantel-Haenszel		Breslow-Day	Uniform DIF			Non-Uniform DIF			Lord (IPL)			
	MH	ETS	BD	Log	R ²	ZT	JG	Log	R ²	ZT	JG	Lord	ETS
LST01	3.64	C	25.87*	3.29	.047	A	B	2.26	.032	A	A	3.95	C
LST02	7.14*	C	8.89	10.34**	.205	B	C	0.00	.000	A	A	3.69	B
LST03	49.71**	C	102.03**	62.12**	.212	B	C	4.77	.016	A	A	31.03**	C
LST04	19.06**	C	14.99	19.70**	.074	A	C	0.03	.000	A	A	9.47**	C
LST05	1.66	A	17.39	2.70	.021	A	A	0.37	.003	A	A	1.15	A
LST06	2.18	A	23.48	3.36	.022	A	A	1.79	.012	A	A	2.44	A
LST07	2.05	A	21.48	3.93	.027	A	A	2.47	.017	A	A	2.86	A
LST08	0.11	A	11.28	0.04	.000	A	A	0.01	.000	A	A	0.19	A
LST09	10.60**	B	40.17*	10.88**	.101	A	C	10.51**	.096	A	C	2.85	A
LST10	1.94	A	17.57	1.87	.011	A	A	0.52	.003	A	A	1.36	A
LST11	1.45	A	19.50	1.88	.005	A	A	2.51	.006	A	A	0.06	A
LST12	4.99*	B	27.25	5.71*	.027	A	A	3.49	.016	A	A	5.94*	B
LST13	0.06	A	35.22*	0.16	.001	A	A	1.11	.009	A	A	1.24	A
LST14	12.19**	C	19.94	15.28**	.052	A	B	8.08**	.027	A	A	16.15**	C
LST15	10.20**	C	37.96*	12.26**	.030	A	A	5.95*	.014	A	A	2.15	A
LST16	1.46	A	35.85*	2.16	.008	A	A	1.60	.006	A	A	0.61	A
LST17	1.45	A	30.22	1.05	.009	A	A	2.88	.024	A	A	0.42	A
LST18	5.94*	B	18.27	8.32**	.040	A	B	1.92	.009	A	A	4.64*	B
LST19	0.01	A	24.45	0.00	.000	A	A	9.33**	.038	A	B	0.65	A
LST20	7.09*	B	22.30	9.27**	.035	A	A	7.92**	.029	A	A	3.89	A
LST21	2.80	A	22.54	2.52	.009	A	A	8.15**	.029	A	A	0.43	A
LST22	0.54	A	31.78	0.80	.004	A	A	4.87*	.021	A	A	0.10	A
LST23	7.97**	B	24.90	9.27**	.065	A	B	1.71	.012	A	A	5.65*	B
LST24	12.06**	C	21.74	11.72**	.081	A	C	5.31*	.036	A	B	8.83**	B
LST25	0.58	A	20.77	0.46	.003	A	A	0.07	.001	A	A	1.19	A
LST26	8.79**	B	29.81	10.01**	.081	A	C	1.61	.013	A	A	3.81	B
LST27	4.69*	B	25.46	3.53	.012	A	A	2.02	.007	A	A	0.62	A
LST28	2.04	A	37.03*	2.14	.018	A	A	0.12	.001	A	A	0.27	A
LST29	0.71	A	17.01	1.00	.004	A	A	0.94	.004	A	A	1.73	A
LST30	0.36	A	14.90	0.81	.004	A	A	0.10	.001	A	A	0.78	A

Note. * $p < .05$, ** $p < .01$

Table 5.7.
Results for pre-knowledge dependent DIF in the LST (Study 3)

Item	Contingency-Table based				Logistic Regression						IRT Model based			
	Mantel-Haenszel		Breslow-Day		Uniform DIF			Non-Uniform DIF			Lord (1PL)			
	MH	ETS	BD		Log	R ²	ZT	JG	Log	R ²	ZT	JG	Lord	ETS
LST01	0.01	A	17.50		0.01	.000	A	A	0.08	.001	A	A	0.17	A
LST02	0.44	A	18.06		0.68	.013	A	A	0.07	.001	A	A	0.47	A
LST03	6.58*	C	20.28		7.41**	.033	A	A	0.96	.004	A	A	3.01	A
LST04	0.01	A	31.83*		0.17	.001	A	A	3.71	.014	A	A	0.08	A
LST05	2.76	A	20.35		4.76*	.036	A	B	3.18	.024	A	A	1.68	A
LST06	4.53*	A	26.45		6.04*	.039	A	B	0.11	.001	A	A	3.74	A
LST07	0.29	A	18.62		1.04	.007	A	A	0.09	.001	A	A	0.43	A
LST08	2.41	A	17.38		6.35*	.041	A	B	2.41	.015	A	A	3.35	A
LST09	0.91	A	24.96		1.71	.022	A	A	2.12	.028	A	A	0.61	A
LST10	0.24	A	17.96		0.40	.002	A	A	0.10	.001	A	A	0.49	A
LST11	2.83	A	12.89		3.36	.008	A	A	0.07	.000	A	A	0.01	A
LST12	0.25	A	12.26		0.06	.000	A	A	0.11	.001	A	A	0.01	A
LST13	0.10	A	14.34		0.41	.004	A	A	1.02	.009	A	A	2.83	A
LST14	0.46	A	22.96		1.06	.004	A	A	0.11	.000	A	A	2.86	A
LST15	7.28**	B	20.21		8.11**	.021	A	A	3.33	.008	A	A	0.62	A
LST16	2.69	A	19.97		3.36	.014	A	A	0.15	.001	A	A	1.25	A
LST17	0.72	A	29.19		0.85	.007	A	A	0.09	.001	A	A	0.14	A
LST18	0.01	A	25.60		0.04	.000	A	A	0.35	.002	A	A	0.00	A
LST19	8.45**	C	25.02		8.01**	.031	A	A	3.32	.013	A	A	2.17	A
LST20	0.00	A	29.54		0.02	.000	A	A	4.33*	.016	A	A	0.44	A
LST21	0.00	A	27.95		0.19	.001	A	A	9.43**	.034	A	A	1.03	A
LST22	0.10	A	19.19		0.16	.001	A	A	0.09	.000	A	A	0.52	A
LST23	0.85	A	25.24		1.27	.009	A	A	0.53	.004	A	A	0.13	A
LST24	0.27	A	22.25		0.69	.005	A	A	0.03	.000	A	A	0.20	A
LST25	5.58*	B	20.00		6.93**	.049	A	B	0.01	.000	A	A	2.50	A
LST26	0.27	A	10.89		0.13	.001	A	A	0.19	.002	A	A	3.29	A
LST27	0.53	A	16.39		0.71	.003	A	A	0.67	.002	A	A	0.01	A
LST28	0.43	A	20.86		0.27	.003	A	A	1.58	.015	A	A	2.85	A
LST29	0.08	A	19.05		0.39	.002	A	A	3.71	.016	A	A	0.81	A
LST30	0.11	A	31.16		0.18	.001	A	A	0.92	.005	A	A	0.20	A

Note. * $p < .05$, ** $p < .01$

cDIF overview

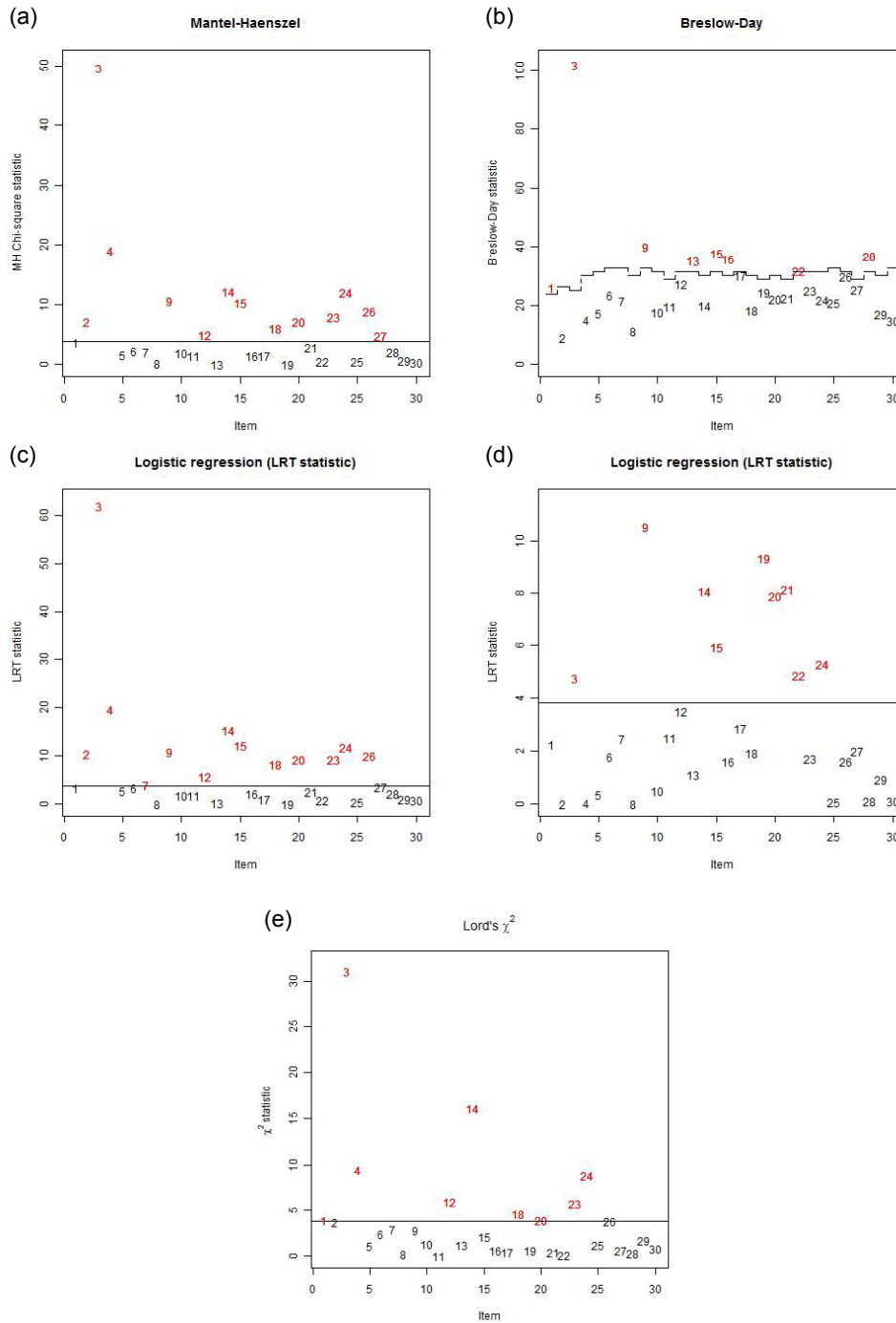


Figure 5.4. Overview of country-dependent DIF in the LST; (a) MH, (b) BD, (c) Logistic Regression, Uniform DIF, (d) Logistic Regression, Non-Uniform DIF, (e) Lord

sDIF overview

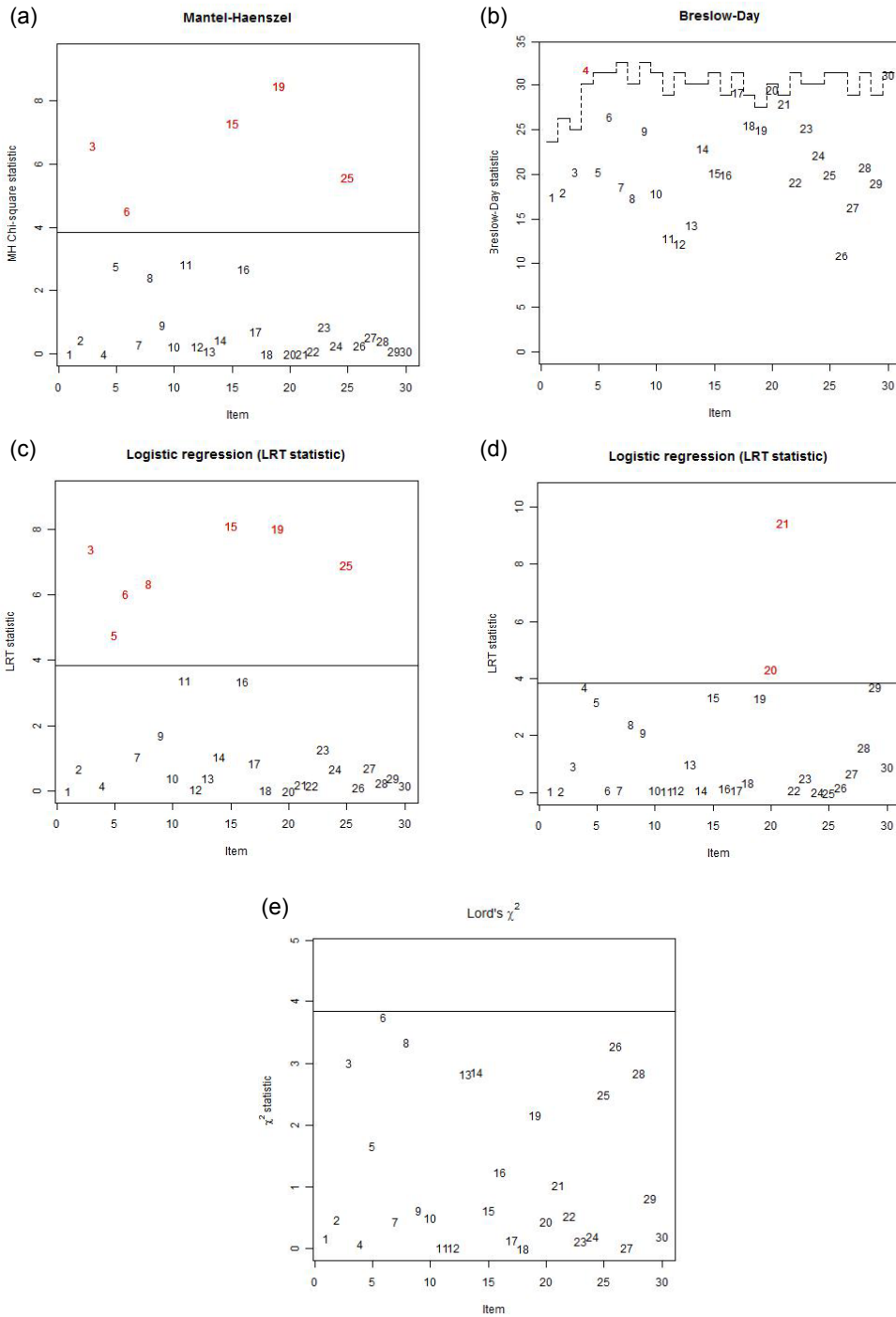


Figure 5.5. Overview of pre-knowledge dependent DIF in the LST; (a) MH, (b) BD, (c) Logistic regression, Uniform DIF, (d) Logistic regression, Non-uniform DIF, (e) Lord

several effect size measure are given in these Tables. The Delta scale is used with the ETS classification for MH and Lord (Holland & Thayer, 1985) and the two alternative classifications by Zumbo and Thomas (1997) (ZT) and Jodoin and Gierl (2001) (JG) for the Logistic Regression method. Here, JG represents a less conservative classification than ZT. Three categories of DIF effects are proposed, i.e. negligible effects ($\Delta = A$), moderate effects ($\Delta = B$) and large effects ($\Delta = C$). In addition, Nagelkerke's R^2 (Nagelkerke, 1991) is reported for the Logistic Regression methods. Figures 5.4 and 5.4 provide graphical summaries of the DIF effects found. For each method, a scatterplot is shown that displays Items on the horizontal axis and the respective DIF statistic on the vertical axis. The horizontal line indicates the cut-off value for this method to flag an item as DIF.

As a general result, at least one third of the 30 LST items is flagged for DIF on at least one method for both group comparisons. That is, the amount of DIF in the LST is substantial. 7 out of 30 items showed large country-dependent DIF effects according to the ETS Delta scale based on the MH statistic, one of the DIF statistics that is mostly used in testing practice. 2 out of 30 items showed large pre-knowledge dependent DIF according to the same criterion. The result that more items are flagged as cDIF than sDIF was consistently found for the other methods as well. Figure 5.6 plots the number of methods that flagged an item as DIF for all items for cDIF and sDIF together in one chart. Items are plotted on the vertical axis, the number of statistically significant DIF statistics is plotted on the horizontal axis. 10 Items clearly show DIF. These items are flagged for DIF by three or more of the five methods. None of the items shows sDIF. If a more strict cut-off is used, almost half of the items, 14, are flagged for cDIF and 5 items indicate sDIF.

5.3.4. Differential facet functioning analyses

Additionally to the two LLTM models described in the previous section, analyses were extended to include models with facet*person interaction parameters, i.e. DFF effects were estimated. In total, 7 different model variants were compared. All analyses were based on the LLTM with extended design matrix including both cognitive complexity and processing steps given the better prediction of item difficulties based on this model ($R^2 = .539$ vs. $R^2 = .589$ for the German sample; $R^2 = .497$ vs. $R^2 = .543$ for the Russian sample). Estimation results and fit indices for all models are summarized in Table 5.9. As the interest was in the explicit modeling of specific facet*person interaction effects, a fixed effects modeling approach was chosen.

Model 1 served as a baseline model. Models 2 and 3 were used to test the cross-cultural equivalence of item facet difficulties in a broad sense, i.e. whether facet difficulties varied between the two countries investigated. Models 4 and 5 were used to test the equivalence of item facet difficulties in a narrow sense, i.e. whether facet difficulties vary between

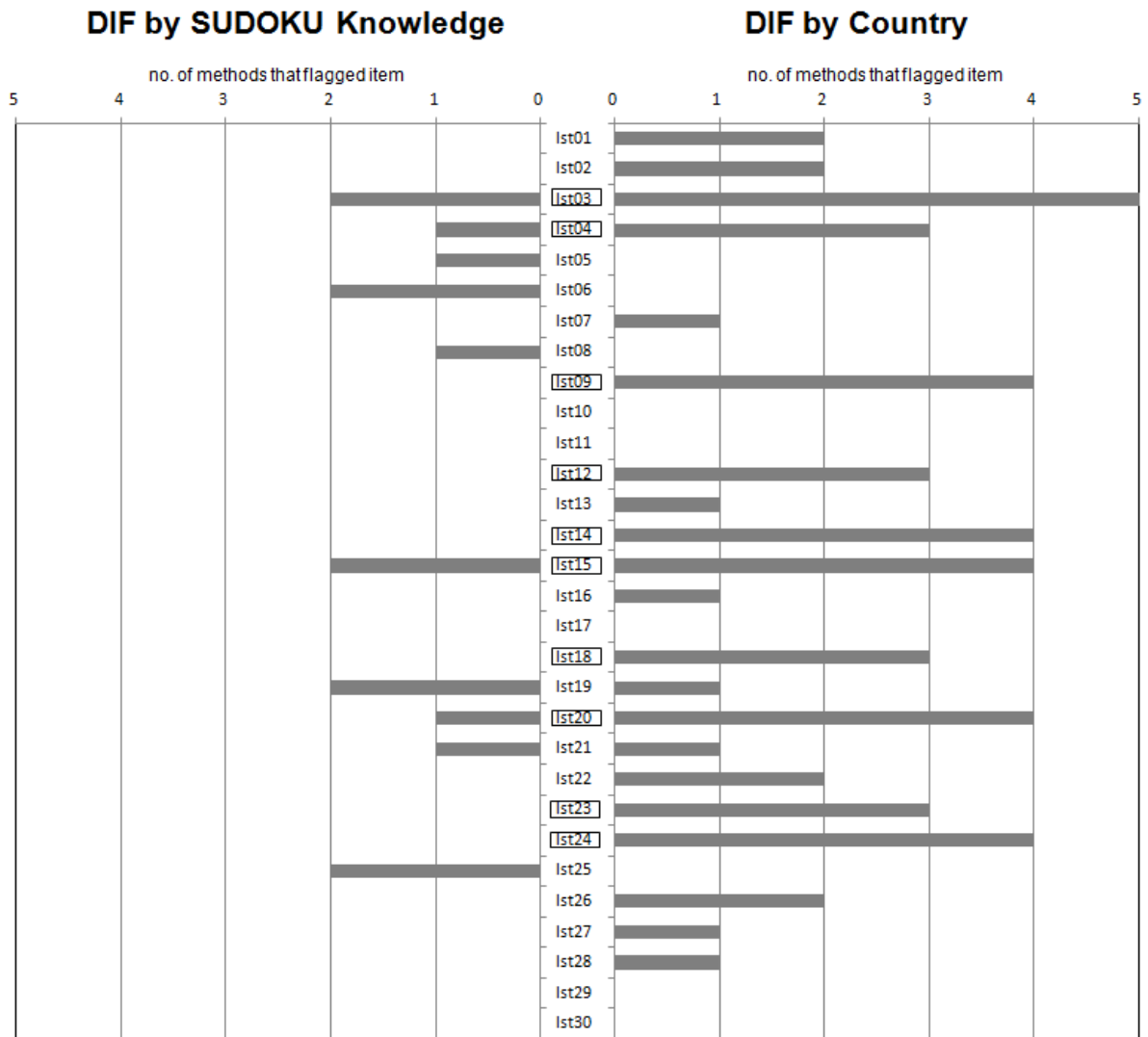


Figure 5.6.
 Number of items flagged as cDIF and sDIF by the five methods used (Study 3)

Table 5.8. *Surface characteristics and applicability of two simple solution heuristics for LST items flagged as DIF*

Item	Items flagged as DIF		Predicted DIF	Complexity Parameters				Incidentals			Heuristic	
	total	lenient strict		BIN	TER	QUAR	Color	NS	Size	H1	H2	
LST01	2	DIF	0.057	2	0	0	0	0	0	0	1	1
LST02	2	DIF	0.104	0	1	0	0	1	0	0	1	1
LST03	5	DIF	0.047	1	1	0	0	0	0	0	1	0
LST04	3	DIF	0.047	2	1	0	0	1	0	0	1	0
LST05	0		0.047	1	2	0	0	1	0	0	0	0
LST06	0		0.047	1	2	0	0	0	0	0	0	0
LST07	1		0.047	2	2	0	0	0	0	0	0	0
LST08	0		0.305	0	0	1	1	1	1	0	0	0
LST09	4	DIF	0.047	1	1	0	0	1	0	0	1	0
LST10	0		0.047	1	3	0	0	1	0	0	0	0
LST11	0		0.305	0	0	1	0	0	0	0	0	1
LST12	3	DIF	0.305	0	0	1	0	0	0	0	0	1
LST13	1		0.047	1	3	0	0	1	0	0	0	0
LST14	4		0.305	0	0	1	0	0	0	0	0	1
LST15	4	DIF	0.305	0	0	1	1	1	0	0	0	1
LST16	1		0.047	2	2	0	0	1	0	0	0	0
LST17	0		0.104	0	2	0	0	0	0	1	0	0
LST18	3	DIF	0.047	2	1	0	0	1	0	1	1	0
LST19	1		0.409	0	1	1	0	0	0	1	0	0
LST20	4		0.047	2	2	0	0	0	0	1	0	1
LST21	1		0.248	1	0	1	1	1	0	1	1	0
LST22	2		0.047	2	3	0	0	0	0	1	0	0
LST23	3	DIF	0.248	1	0	1	0	0	0	1	1	0
LST24	4	DIF	0.047	1	1	0	0	1	1	1	1	1
LST25	0		0.047	2	3	0	0	1	0	1	0	0
LST26	2		0.047	2	4	0	0	0	0	1	0	0
LST27	1		0.104	0	2	0	0	1	0	1	0	0
LST28	1		0.047	2	2	0	0	1	0	1	1	0
LST29	0		0.409	1	1	0	0	0	1	1	1	1
LST30	0		0.409	0	1	1	1	1	0	1	0	0

Note. strict: item is flagged if > 1 methods flag item; lenient: item is flagged if > 2 methods flag item; Predicted DIF: linear combination of item facets and respective DFF parameters; NS: not solvable; H1: Reduction of Alternatives; H2: Easy Falsification)

test-takers with or without previous experiences with similar task types. Models 6 and 7 were used to test the equivalence of item facet difficulties both in a broad cross-cultural and a more narrow pre-knowledge dependent sense. They allow for a direct comparison of the two sources for possible bias in one model.

1. *Model 1* neither included person main effects nor DFF parameters. It was identical with the respective LLTM model applied to the two samples separately. This model represented the (null-hypothesis) assumption that DFF was neither present due to cultural background nor to test-specific prior experiences. Model 1 was considered the baseline model for all other models that included further predictor variables. Results were largely in line with the results found in the separate samples. As summarized in Table 5.9, binary processing overall was less difficult than Ternary processing), and Quarternary processing had the largest contribution to item difficulties. All item covariates reached statistical significance.
2. *Model 2* included an additional dichotomous country indicator as a proxy of cultural background. This country main effect captured overall mean differences on the latent ability between the two samples. In alignment with the findings based on sum scores reported in the previous section, results indicated that solution probabilities differed significantly between the two samples: the difference in the logit for a correct response is -0.529 ($p < .001$). Model fit improved by adding the latent regression parameter to the LLTM model: AIC and BIC statistics are consistently smaller compared to model 1. The direction of the country effect is consistent with well-documented findings in the literature that DIF effects usually favor test-takers from the country where the test was designed and initially tested and calibrated. There have been numerous studies using the LST in Germany and the items used in this study were initially developed for a German sample, but no studies have been conducted in Russia before.
3. *Model 3* tested DFF based on the comparison of individuals from the two culturally diverse samples. In addition to the country main effect in Model 2, this model contained additional facet-by-country interaction effects for each item facet, i.e. 6 additional parameters. When interaction effects are included in the model, the country main effect is reduced but still statistically significant. The effects are rather small and only partly significant. The direction of these DFF effects is less clear than expected based on the cognitive complexity model. While two-step binary processing is facilitated for Russian test-takers, multiple-step Ternary processing and Quarternary processing are facilitated for German test-takers. Information criteria and LR tests for the comparison of models 2 and 3 showed that the inclusion of interaction parameters could only partly improve model fit. A model without DFF does not necessarily fit worse than a model comprising DFF effects. According to these results, the functioning of most item facets is not or only minor affected by

the broad cultural factors captured in the country indicator. While AIC was smaller for model 3 compared to models 1 and 2, BIC favored the sparser model 2.

4. *Model 4* had the same number of parameters as Model 2 but included Sudoku experience as a proxy of prior experience with similar tasks instead of the country indicator. This effect is significant ($p < .001$) and slightly larger than the country-related main effect. Results show that solution probabilities depend considerably on the presence or absence of Sudoku experience: the value of the logit increased by 0.637 when a test-taker has played Sudoku puzzles before. A comparison of this model with Model 2 indicated that Model 4 showed better model fit both in terms of AIC as well as BIC, i.e. prior test experience appeared as a better predictor for solution probabilities in the LST than broad cultural background.
5. *Model 5* contained additional facet-by-Sudoku interaction parameters for each item facet, i.e., six additional parameters. This model was specified to test DFF based on the comparison of individuals with and without Sudoku experience, i.e. individuals with and without prior experiences with similar tests. Whereas the main person effect for Sudoku experience remained significant, none of the facet-by-Sudoku interaction parameters reached statistical significance or any meaningful effect sizes. Accordingly, model fit did not clearly improve when DFF (i.e. interaction) parameters were added to the model comprising only a Sudoku main effect.
6. *Model 6* estimated solution probabilities as dependent on both pre-knowledge and cultural background, i.e. two latent-regression parameters were added to Model 1. Results indicated that item performance depended on both person variables; the unique contributions of both predictors were smaller than in Models 2 and 4, i.e. Sudoku experience and cultural background were not uncorrelated. In direct comparison, the linear predictor for Sudoku experience was higher than for culture. This is the case for the two LLTM models with different complex design matrices. In terms of information criteria, model fit statistics for model 6 were superior to both models 2 and 4.
7. *Model 7* tested DFF based on the combination of the two person predictor variables, Sudoku experience and cultural background. When both factors were included simultaneously, only 3 out of the 14 additional parameters reached statistical significance: Sudoku experience had a major impact on item difficulties with a facet parameter > 0.5 on the logit scale. Cultural background per se had no significant impact on solution probabilities, two-step Binary processing favored Russian test-takers and Quarternary processing was facilitated for German test-takers.

Out of all 7 models, Model 6 showed the best fit in terms of both AIC and BIC. A LR test for the comparison of model 6 and model 7 misses statistical significance at the 1% level ($\Delta\chi^2(12)=24.460, p = .018$) That is, in relation to the large number of parameters in this model, the increase in model fit is rather small. Model 6 represents the assumption

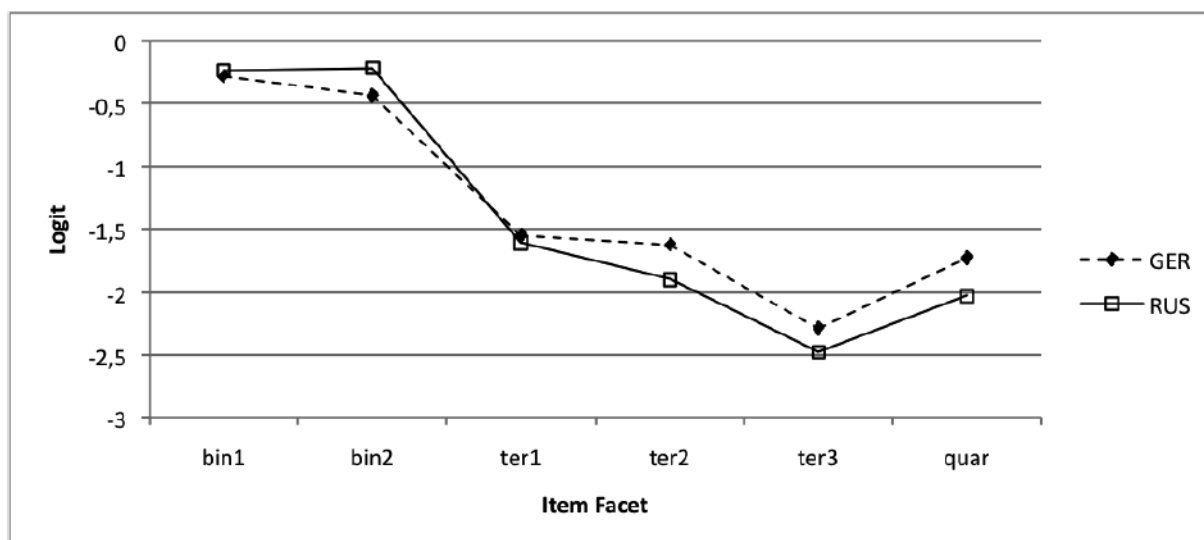


Figure 5.7.

Cross-cultural DFF in the LST (Study 3)

Note. Logits are facet difficulties for individuals from each of the two cultural groups based on joint DFF-LLTM; bin1 = Binary processing, one step; bin2 = Binary processing, two steps; ter1 = Ternary processing, one step; ter2 = Ternary processing, two steps, ter3 = Ternary processing, three or more steps, quar = Quarternary processing

of existing country and pre-knowledge main effects but no DFF; Model 7 estimates main effects *and* DFF effects. Table 5.10 summarizes all fit indices for the 7 different models that were estimated. Figure 5.7 displays the DFF effects in the current study based on the most complex DFF-model (Model 7). In order to focus on the DFF effects, main effects are not plotted in this Figure. It can be seen that DFF effects are relatively small. The pattern of item facet difficulties is very similar in the two countries. It can be seen that there is no clear direction of DFF effects, i.e. not all facets are facilitated for the same country.

In order to test whether facet level analyses could help to identify DIF, DFF results were compared to the “classical” DIF-analyses presented in the previous section. While the DFF effects reported above indicated mostly the absence of considerable bias due to specific item facets, DIF parameters indicate a considerable amount of DIF in the LST scores of German and Russian test-takers. 8 to 12 (depending on whether lenient or strict cut-offs were used) out of 30 items (at least 25%) showed DIF. Based on the DFF effects and the design matrix underlying the LST version applied here (see Appendix), rescaled DIF effects were calculated in a similar way as rescaled item difficulties can be calculated based on LLTM parameters. The vector of DFF parameters is multiplied with the design

matrix so that, for each item, the resulting DIF parameter reflects the sum of all DFF effects present in the respective item:

$$\text{Predicted DIF}_i = \sum_{k=1}^K \gamma_k X_{ik} \quad (5.1)$$

Here, K is the number of item radicals or facets specified in the explanatory model. γ_k denote weights for the interaction of item facet difficulties with group membership. The person group-membership variable Z_p is not included in the calculation of predicted DIF values as DIF is predicted for items and not for persons. These rescaled values are values on the same logit scale as the “original” item difficulty parameters. That is, a value of Predicted DIF_{*i*} = 0.5 means that the difficulty parameter for this item is shifted by half a logit for the focal relative to the reference group. If DFF “helps to explain the DIF effects more substantively” (Xie & Wilson, 2008, p. 414), rescaled DIF parameters should be higher for items that are flagged as DIF items when classical approaches are used. Table 5.8 includes the predicted DIF effects based on the LLTM with extended design matrix (LLTM 2). While, for some items (e.g. Items 14 and 15) higher values go along with DIF results on the other methods, this is not the case for other items. For instance, DIF should be expected for item 8 based on DFF parameters, but no DIF is found. On the contrary, item 3 pertains DIF according to all traditional methods, and no DIF would be expected based on the linear combination of DFF parameters. For practical purposes, it must be concluded that actual DIF statistics could not be predicted by DFF parameters or a combination of these parameters. This finding indicates that important item characteristics that are causal for DIF have not (or not sufficiently) been captured in the design matrix used here. This might point to more general problems with the item difficulty model for the LST.

5.3.5. Qualitative analyses of DIF in LSTs

In order to identify sources that might have caused DIF in the current application of the LST, all 30 items of the LST were investigated qualitatively in detail. Items that were flagged as DIF items were inspected with special attention and compared to items that did not manifest DIF. Here, only country-dependent DIF effects were analyzed, given the almost complete absence of DIF for the groups differing in their SUDOKU experiences. In several cases, DIF and non-DIF items shared exactly the same item radicals as specified in the item-difficulty model. Table 5.8 summarizes information on the three item complexity parameters and the incidentals that were manipulated explicitly during item generation (cf. previous publications on the generation of LST items, e.g., Birney et al., 2006) along with an indicator of DIF for each item. These item characteristics are color, size, and solvability. The color of the LST could vary in the sense that some items contained

figural shapes that were filled in different shades of grey while other items contained only shapes with no/white filling (see e.g., Figure 5.8). The parameter size was varied by including LSTs of dimensionality four as well as LSTs of dimensionality five. LSTs of higher dimensionality have been shown to be more difficult than lower dimensional LST items (e.g., Gold, 2008). LSTs of higher dimensionality allow for the combination of more rules. Solvability refers to the distinction of items with a correct solution and items that do not have a correct solution. Two different criteria for DIF are given in Table 5.8: based on the number of DIF statistics that reached statistical significance, a lenient flag was assigned when three or more methods indicated DIF in an item, a more strict flag was assigned when two or more methods indicated DIF in an item.

Results from this table indicate that neither the three item radicals nor the three incidentals are linked to the presence or absence of DIF in the LST. From Table 5.8 it can be seen that the distribution of Binary, Ternary, and Quarternary items is very similar across items flagged as DIF and non-DIF items. On average, non-DIF items contain 0.88 Bin, 1.63 Ter, and 0.31 Quar steps, For DIF items these numbers are 1.14 Bin steps, 1.07 Ter steps, and 0.29 Quar steps (based on the strict criteria for DIF). That is, DIF items contain more Binary, but less Ternary and Quarternary steps. This finding does not go in line with the assumption that cognitively more complex operations are more prone to DIF as expected based on the cognitive complexity model. Also, no clear pattern of differences in the incidental parameters *Solvability* and *Size* given in Table 5.8 seemed apparent from the data. 13% of the non-DIF items and 13% of the DIF items were not solvable, 50% of the non-DIF items and 43% of the DIF items were 5×5 matrices instead of 4×4 . The proportion of items with white elements (vs. filled elements) among the items flagged as DIF was higher, though (64% vs. 37%). This might indicate that items with clearly distinct figural elements might enhance the psychometric quality of the items. However, this trend is, as all other effects above, not statistically significant ($p > .05$).

However, the qualitative analysis of the LST items revealed other potentially causal factors for DIF. Two surface characteristics were identified that apply to most of the DIF items but not to most of the non-DIF items. First, some items allowed for the quick exclusion of all but two response alternatives, thereby increasing the chance of picking the right answer by chance considerably. Second, some items allowed for the identification of the correct response alternative without considering any other element, simply by using a falsification strategy. These two characteristics, labelled “reduction of response alternatives” (H1) and “easy falsification” (H2), are included at the righthand side of Table 5.8. The proportion of DIF vs. Non-DIF items allowing for the application of either of the strategies is three times higher for H1 and four times higher for H2. Both effects are statistically significant (H1: $\chi^2(1) = 4.739$, $p = .029$; H2: $\chi^2(1) = 5.000$, $p = .025$). The underlying qualitative analyses are summarized in more detail in the following section.

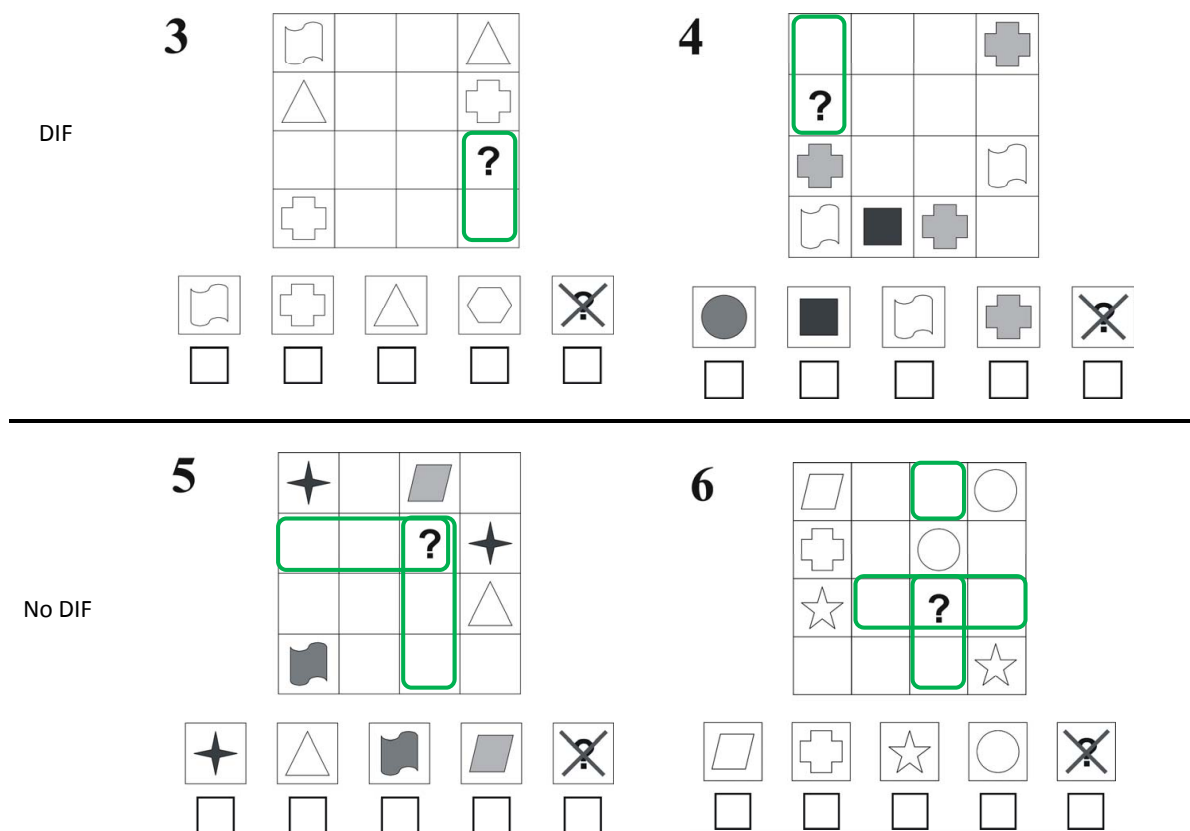


Figure 5.8.

LST items that allow or not allow for the application of a quick exclusion of response alternatives strategy (Study 3)

All items flagged as DIF were analysed qualitatively and compared with items not flagged as DIF. Here, the ETS delta criterion based on the MH statistic was used as a criterion for DIF. Items were flagged as DIF when $\Delta = C$.

First, some items allow for the quick exclusion of all but two response alternatives, thereby increasing the chance of picking the right answer by chance considerably. If some test-takers show a higher willingness to guess than others, for instance, independent of their ability, it is likely that this might cause specific response patterns especially in such items but not in items with a larger number of reasonable response alternatives. As shown in Figure 5.8, items 3 and 4 both allow for such a quick reduction of the number of response alternatives, whereas items 5 and 6 do not. The green highlighted cells indicate the minimum number of elements that are viable possible responses when only the already filled cells in the respective row or column or inspected. That is, if a test taker completely ignores the complete matrix and just focuses on this one row or column, the chance of

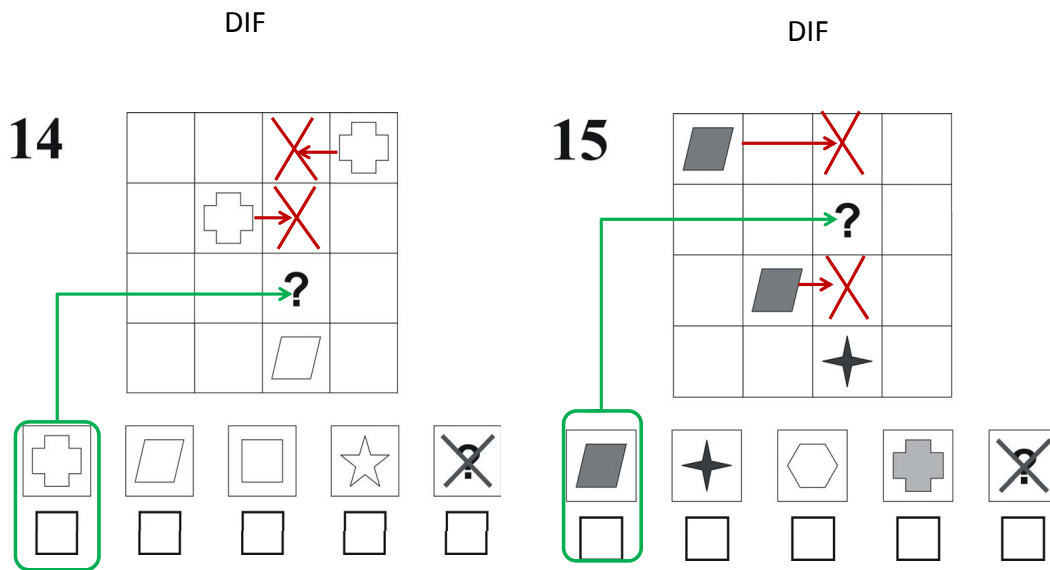


Figure 5.9.

LST items that allow for application of a quick falsification heuristic (Study 3)

solving an item correctly just by guessing is not the same for items 3 and 4 versus items 4 and 6. The ETS delta statistic indicates DIF for Items 3 and 4 but not for items 5 and 6. In terms of their cognitive complexity as determined based on RC theory in the item difficulty model applied here, the cognitive complexity of all four items should be very similar, that is, if cognitive complexity is a causal factor for DIF, all four items should demonstrate comparable amounts of DIF.

Second, some items allow for the identification of the correct response alternative without considering any other elements, simply by using a falsification strategy. For instance, items 14 and 15, both flagged for DIF, can be solved by identifying the correct alternative as the only element that can possibly fill the empty cell with the question mark (see Figure 5.9), notably while at the same time ignoring the possibility that the item might be an item that has no correct solution. It is plausible that these items draw the test-taker's attention to a falsification strategy focusing on one shape only because only few cells in the matrix are filled out and most shapes are arranged in one part of the matrix (e.g. all in the right lower corner in item 24). If a test-taker, on the other hand, tries to solve these items by mentally filling in all elements in the row or column with the question mark, solving this item will be much harder; moreover, if a verification strategy is followed, it is sheer not possible to fill in all cells unequivocally. This distinguishes items 14 and 15 from most of the other items in the test and might be one cause for DIF in these items. Because of the availability of this rather simple strategy, the cultural complexity of these items is

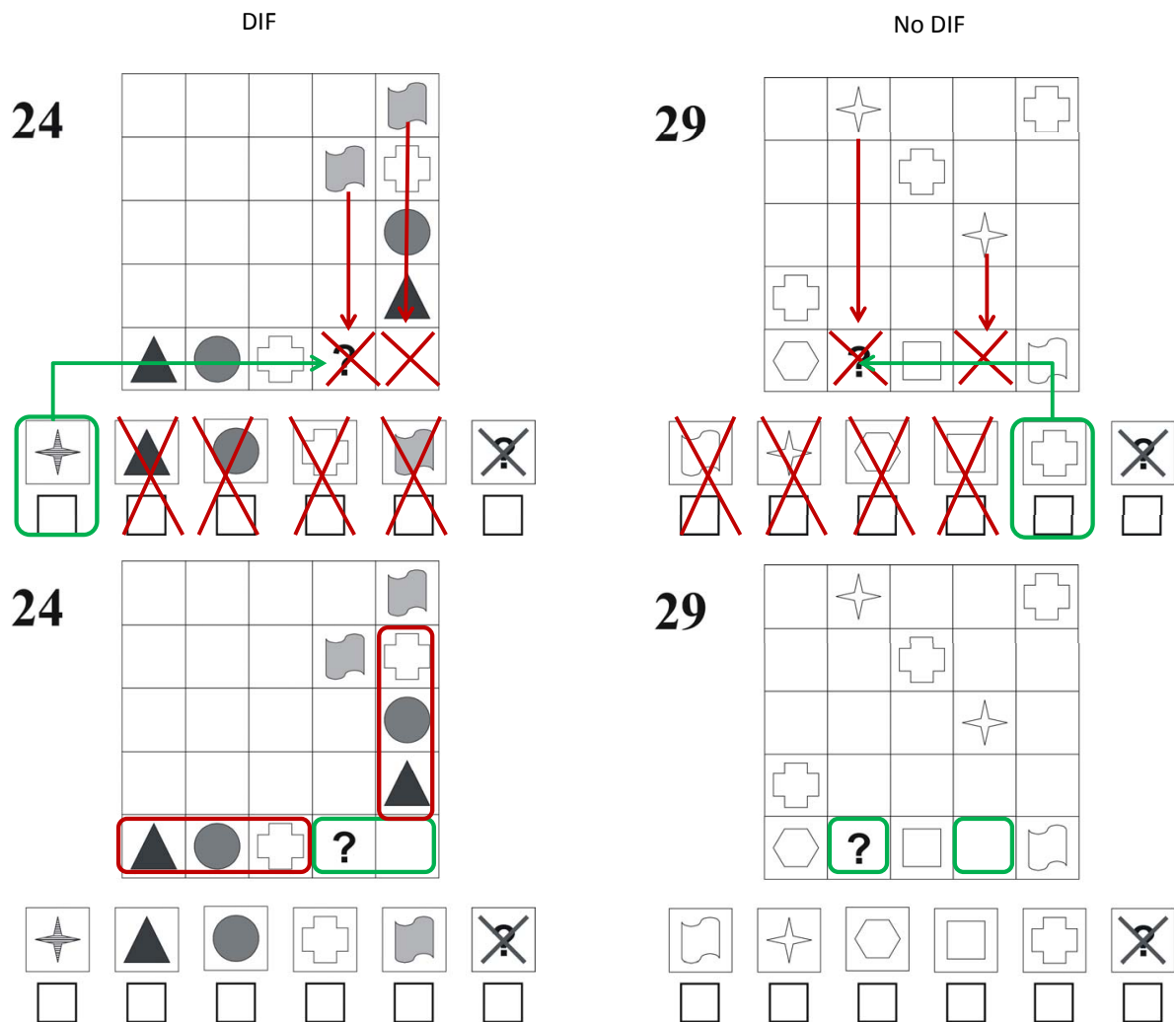


Figure 5.10. *LST items that allow both for a quick reduction of response alternatives and for application of a falsification heuristic (Study 3)*

stronger relative to their cognitive complexity compared to items where such strategies are unavailable.

Items 24 and 29 in Figure 5.10 fulfill both of the aforementioned surface characteristics. Only two response alternatives need to be considered for the solution even though a LST of dimensionality five is used here. In combination with a falsification strategy, the correct element can be identified without mentally “filling in” other elements in other columns or rows. According to the design matrix based on RC theory for the LST, item 24 requires one ternary and one binary step. As illustrated in Figure 5.10, however, this item can

be solved by simply crossing out response alternatives. The availability of completely different, equally successful, solution strategies might be one generating factor for DIF that is not reflected in the theoretically assumed structure of item difficulties based on the RC approach only. The same is true for item 29, yet this item is not flagged for DIF. A deeper comparison of item 24 and item 29 reveals that item 24 might draw the test-taker much stronger towards quickly excluding a majority of response alternatives. Here, not only the row with the question mark contains only 2 empty cells; also a neighboring column is filled out completely but for one cell. This is not the case in item 29. Here, the four shapes that are present in the LST are also distributed more widely across the whole matrix, making it less obvious for the test-taker what options to exclude in a first step.

Items 22 and 25 share the same radicals as items 24 and 29 but show neither of the aforementioned surface characteristics. Both items are free from DIF. Here, the maximum number of response alternatives that can be excluded by considering the number of already filled out cells is 2, leaving at minimum 3 response alternatives that need to be further considered during the solution process. Also, the shapes that are present in the LST are distributed across the whole matrix. Using one of the simple heuristic strategies cannot lead to the correct solution easily here.

As these additional qualitative analyses of the DIF and non-DIF items shows, many characteristics of the items that determine their visual appearance and thereby also their overall complexity are not covered by the item-radicals specified in the RC-theory based item difficulty model. This misspecification of the design matrix might be an explanation for the poor explanatory power of the LLTM models applied, both for the prediction of item difficulties and for the prediction of differential item functioning. While the item-by-facet interaction parameters can account for less than 10% of variation in DIF statistics, a classification of LST items regarding the applicability of the two rather simple heuristics described above each can explain considerable amounts of variation in DIF operationalized by the ETS delta statistic ($\chi^2_{\text{reduction}}(2) = 5.970, p = .05, \eta = .446$; $\chi^2_{\text{falsification}}(2) = 8.061, p = .02, \eta = .504$). The cross-tables for these analyses can be found in the Appendix.

5.4. Discussion

The present study was designed to fill the gap between studies on cross-cultural bias and rule-based AIG by testing the cross-cultural applicability of an item-generative framework based on relational complexity theory. One reason that only very few studies have examined content-related causes of cross-cultural differences in test performance is that most instruments used in educational and vocational assessment lack a strong underlying theoretical framework available concerning content-related sources of DIF. The generation of reasoning items “on-the-fly” based on a set of item radicals or item templates that are cloned by varying surface characteristics pre-assumes (a.) that item difficulties

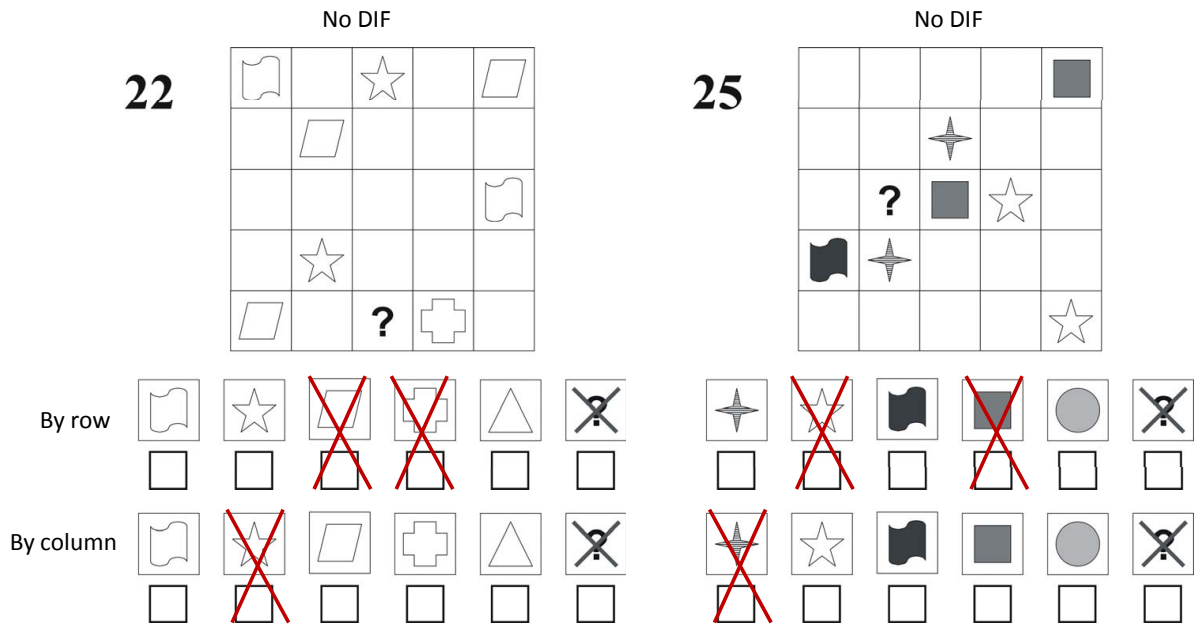


Figure 5.11.

LST items that allow neither for a quick reduction of response alternatives nor for application of a falsification heuristic; ‘by row’ indicates what response alternatives can be excluded when only the currently present elements in the row with the question mark are inspected, ‘by column’ indicates what response alternatives can be excluded when only the currently present elements in the column with the question mark are inspected

for structurally identical items are identical, and (b.) that all clones based on a certain item type are equally valid for different groups of potential test-taker. Especially when test-takers from different countries or different cultural background are tested, it is important that items do not show DIF. The capability of an item-generation framework to generate DIF-free items is especially important for measures of fluid intelligence given that the largest cross-cultural differences on cognitive tests have been reported in fluid reasoning measures (e.g., Brouwers et al., 2009; Carroll, 1993; see also Jensen, 1998 or Hartmann et al., 2007; Lynn & Owen, 1994; Te Nijenhuis & van der Flier, 2001). At the same time reasoning measures are among the most relevant psychological tests used in the workforce (e.g., Ones et al., 2005; Schmidt & Hunter, 1998). If factors could be identified that increase the likelihood of items of a certain type to pertain or not to pertain DIF, this could enhance the feasibility of fully computerized automatic item generation and assessment systems tremendously. Such factors for DIF can, theoretically, refer to either typical structural complexity parameters (i.e., radicals) of test items (representing the assumption that *cognitive* complexity is one important factor for DIF) or to surface differences in the layout of the items (i.e., incidentals) that are not directly linked to the

structural complexity of the items (representing the assumption that *cultural* complexity is an important factor for DIF). The current study investigated how both structural and surface characteristics of a widely used item type, the Latin Square Task (LST), could be linked to the emergence of DIF in these items.

5.4.1. Conclusions regarding the research questions

Two main research questions were addressed, first the cross-cultural validity of the LST was investigated in terms of its measurement properties and the pattern of relationships with other cognitive and noncognitive variables. Second, Differential Item Functioning as an indicator of Item Bias was tested for the LST, along with a detailed qualitative analysis of DIF items in terms of their structural and incidental item features. Data was collected from two culturally diverse populations of Russian and German university students. In order to answer whether factors for performance differences between cultures could be run down to specific content-related factors or were more related to broad cultural variables, it was made use of the similarity of the LST and the worldwide popular SUDOKU puzzles. The resemblance of the two tasks allowed to compare pre-knowledge dependent DIF with culture-dependent DIF in an ecological quasi-experimental setting.

First, it was investigated whether the same construct was measured with the LST in both countries, i.e. whether there was a general construct-bias that would render any subsequent DIF analyses very complicated. The results for Rasch model and correlational analyses for both countries supported the absence of considerable construct bias. Overall performance on the LST was substantially related to general cognitive ability as measured by the CFT. Correlations were nearly identical in the two samples. Correlations with grades and response times were all substantial and nearly identical in both samples as well. As expected, SUDOKU knowledge was positively associated with LST performance. However, this correlation was much stronger in the German sample ($r = .34$) than in the Russian Sample ($r = .19$), indicating that SUDOKU knowledge *per se* might be qualitatively not identical in the two samples. At the time of test administration, in Germany, SUDOKU was much more popular than in Russia with for instance considerably more newspapers and magazines providing free Sudoku puzzles. Also the proportion of test-takers that had played Sudoku before was significantly higher in the German sample (i.e., there was a wider range of pre-knowledge in this sample). LST performance was not related to unspecific experiences with cognitive ability tests. The same applies to associations between the LST and the “Big Five” dimensions of personality; this finding is consistent with the assumption that there are no major differences in the structure of personality factors across cultures (e.g., McCrae & Costa, 1997). In sum, these findings show that, in terms of criterion-related validities, the same construct was measured with the LST in the two cultural samples as well. That is, there was no indication for predictive bias (cf. Van de Vijver, 2002) of the LST.

Second, DIF analyses representing both IRT and non-IRT approaches were applied to test the cross-cultural comparability of the LST. Different models were estimated to test the proneness of the item-generative framework to both broad cultural biases as well as bias caused by different levels of task-relevant pre-knowledge. Pre-knowledge-dependent DIF effects were marginal and mostly not statistically significant. This is an important finding because it indicates that practice with similar item types as item types used in actual assessments, in this case the popular number-placement game SUDOKU, does not diminish the validity of the cognitive measures. Higher scores on the LST for students pertaining previous SUDOKU experiences do, in fact, point to higher ability levels of these students, and not to item bias that is unrelated to the underlying ability. While items showed favorable characteristics when test-takers with and without SUDOKU experiences were compared, results for country-specific DIF revealed serious problems with the cross-cultural applicability of the LST. Around one third of all test-items showed substantial DIF effects, meaning that comparisons of scores on these items from test-takers with different cultural backgrounds cannot be interpreted as valid indicators of differences in the underlying ability. Based on the cognitive complexity model, it was expected to find that DIF effects would be related to the cognitive complexity of the items as defined by the pre-specified item radicals, Binary, Ternary, and Quarternary processing. However, results did not support this expectation.

Differential facet functioning (DFF) analyses have been proposed as a means to understand the generating processes for DIF better (Xie & Wilson, 2008). Yet, no published studies have actually demonstrated this benefit in a practical sense, i.e. for the prediction of DIF items in real-world testing settings. Results from the few existing studies have been very vague and not satisfactory. The latter is also true for the current study. DIF effects could not be predicted by any of the DFF models applied to data from two culturally diverse samples. Whereas a considerable number of items showed DIF according to classical DIF indicators, DFF effects were (a) very small in magnitude and (b) inconsistent with the results of the former. While DFF effects seemed to show a clear picture when inspected isolated, their correspondence with DIF findings was not given. When DFF parameters were used to calculate expected DIF based on the design matrix, items showing higher values on this expected score were not the items actually showing DIF in terms of both IRT and non-IRT methods. Based on the current data, it cannot be concluded whether this results indicate a general lack of predictive power of the DFF model, or whether the lack of predictive power is due to the relatively poor explanation of variation in item difficulties of the LST by the LLTM, i.e., a design matrix mis-specification. In any of these two cases, DFF results should be interpreted only with great caution. Given that the predictive power of the LLTM models in this study was low, but not lower than those reported for other reasoning item types (cf. e.g., Freund et al., 2008), it must be concluded that the practical value of DFF analyses for applied testing and item generation settings seems not to hold up to the expectations.

The additional qualitative analyses of DIF items in this study further indicated that it might be item incidentals (here, certain patterns in the surface structure of the items) rather than radicals in terms of theoretically underlying cognitive processes that drive DIF effects. First, some items allowed for the quick exclusion of all but two response alternatives, thereby increasing the chance of picking the right answer by chance considerably. Second, some items allowed for the identification of the correct response alternative without considering any other element, simply by using a falsification strategy. Both of the two in a post-hoc way identified solution heuristics could explain DIF better than any of the cognitive facet-by-person interaction parameters included in the DFF models. This means that special attention should be paid on the design of item surface characteristics when structurally equivalent items are to be constructed based on item-generation models. Similar to findings reported about other task types (e.g., Irle, 1969; Mittring & Rost, 2008) the use of test-taking and answer strategies that deviate from the cognitive processes assumed by information processing models was found here as well. For the LST, the underlying cognitive model assumes that test-takers are actually following the logical reasoning processes defined by Relational Complexity Theory, and not that test-takers might simply analyze the sets of distractor stimuli, make decisions based on distinct patterns of elements in the partly filled matrices, or simply guess what the correct solution might be. The current study showed that the availability of such simple solution heuristics might be one reason for the emergence of DIF effects. This is an important finding for future applications of the LST. AIG or Item Cloning engines should include control mechanisms for the identification of items allowing for the application of such heuristics to make sure that test-takers cannot simplify the solution process by deviating from the statistically modeled cognitive steps. Also, note that the importance of cultural complexity in explaining cross-cultural differences does not imply that cognitive complexity does not matter. When one of the studied groups has very little experience with certain cognitive tasks and less training than the other groups in the cognitive abilities reflected in those tasks, performance of its members is negatively influenced. As a consequence, cross-cultural score differences are not merely a reflection of differences in familiarity with test content but also of differences in skills as a result of differences in cognitive ability training. This implies that research aimed at addressing cross-cultural score differences should take both explanations (cultural and cognitive complexity) into account and should be careful in drawing conclusions on the importance of one as compared to the other.

5.4.2. Limitations and future prospects

The findings of this study are only a first step in the full understanding of cross-cultural differences in test performance of the LST. In order to gain a full understanding of the cross-cultural fairness of the LST, samples from further cultural populations with greater differences with regard to educational and economic characteristics could be investigated.

The mean raw score differences between the German and the Russian samples must be viewed with caution. The two samples investigated are not representative samples representing the whole cultural groups. Russian test takers used significantly less time to work on the LST administered to them. This might point to differences in motivational states. In addition to this, a robust finding from the data of the current study was that test-takers from the Russian sample chose, on average, the response category “not solvable” more often than test-takers from the German sample. A recent study showed evidence for bias caused by similar response tendencies: Test-takers with a tendency to skip questions performed significantly worse even if they had the same underlying ability (Baldiga, 2011). Even though, in the current study, no indication for bias caused by this pattern among wrong responses could be found, this issue should be investigated more closely in future studies. The finding demonstrates that reaction to distractor stimuli can also provide meaningful insights into person characteristics (cf. also Study 1 in this thesis).

In addition to the methods applied so far, it would be interesting to use specific methods (e.g. think-aloud protocols) to take a deeper look into the strategies, that test-takers use explicitly during test completion in order to gain a better understanding of why SUDOKU knowledge has such strong impact on LST test performance. As shown by Lee et al. (2008), performance on complex SUDOKU items strongly depends on the familiarity with the relevant strategies how to solve that type of item. Correlations between actual Sudoku performance and LST performance could provide a more comprehensive picture of similarity and differences between the two task types as well.

Replications with larger samples and more than two cultural groups could be beneficial to cross-validate our findings with regard to the ordering and absolute height of facet*person interaction effects. This might also comprise the investigation of LST items of higher difficulty level. The 30-item LST used in this study is a relatively easy test version. Difficulty can be increased by using Latin squares with higher order (i.e. more rows and columns), thereby increasing the working memory load and allowing for more combinations of item facets. Successful construction of bigger LSTs was described by Gold (2008) who used order-6 Latin Squares and could show that item difficulty increased while item quality was still good.

The two identified solution heuristics related to item surface characteristics should be further investigated. In the current study, these heuristics have been identified in a post-hoc way by a qualitative analysis of items that showed DIF and a comparison with non-DIF items. Future studies could study these item characteristics more thoroughly. This could involve a systematic manipulation and inclusion in the item-difficulty model (i.e., the design matrix) or a consequent elimination and empirical comparison of test forms with and without such items. Other person predictors could be included in the item-difficulty model as well, such as specific strategy knowledge and the tendency and ability to actually use relevant strategies in the testing situation.

In sum, the findings of this study contribute to future studies using LST type items. This study marks an important step on the way to establishing fully computerized adaptive test systems. If DIF-generating item surface characteristics can be controlled during the item-generation process, LST items seem to be a good candidate item type for such a test system. Future studies need to cross-validate the findings of this study and should investigate the impact of the identified heuristics on statistical item properties more closely.

Table 5.9.
LLTM & DFF modeling for the total sample (Study 3)

Fixed Effects	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6		Model 7	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Intercept	2.763**	0.089	2.927**	0.092	2.893**	0.107	2.380**	0.097	2.336**	0.133	2.554**	0.101	2.456**	0.153
<i>(item facets)</i>														
bin1	-0.266**	0.046	-0.266**	0.046	-0.283**	0.057	-0.266**	0.046	-0.273**	0.069	-0.266**	0.046	-0.295**	0.081
bin2	-0.360**	0.051	-0.360**	0.051	-0.434**	0.063	-0.360**	0.051	-0.353**	0.077	-0.360**	0.051	-0.447**	0.090
ter1	-1.554**	0.064	-1.554**	0.064	-1.545**	0.077	-1.554**	0.064	-1.456**	0.099	-1.554**	0.064	-1.427**	0.114
ter2	-1.703**	0.076	-1.703**	0.076	-1.618**	0.092	-1.703**	0.076	-1.687**	0.114	-1.703**	0.076	-1.572**	0.132
ter3	-2.341**	0.078	-2.341**	0.078	-2.286**	0.094	-2.342**	0.078	-2.260**	0.119	-2.342**	0.078	-2.177**	0.137
quar	-1.816**	0.071	-1.816**	0.071	-1.724**	0.086	-1.817**	0.071	-1.826**	0.109	-1.817**	0.071	-1.705**	0.126
<i>(pre-knowledge/country main effects)</i>														
country	--	--	-0.529**	0.078	-0.397*	0.188	--	--	--	--	-0.418**	0.076	-0.271	0.189
sudoku	--	--	--	--	--	--	0.637**	0.072	0.714**	0.176	0.564**	0.072	0.666**	0.178
<i>(broad/unspecific DFF effects (country*facet interactions))</i>														
country*bin1	--	--	--	--	0.046	0.096	--	--	--	--	--	--	0.050	0.098
country*bin2	--	--	--	--	0.215*	0.108	--	--	--	--	--	--	0.220*	0.110
country*ter1	--	--	--	--	-0.059	0.140	--	--	--	--	--	--	-0.091	0.142
country*ter2	--	--	--	--	-0.277	0.163	--	--	--	--	--	--	-0.288	0.165
country*ter3	--	--	--	--	-0.185	0.168	--	--	--	--	--	--	-0.213	0.170
country*quar	--	--	--	--	-0.301*	0.154	--	--	--	--	--	--	-0.315*	0.156
<i>(narrow/specific DFF effects (pre-knowledge*facet interactions))</i>														
sudoku*bin1	--	--	--	--	--	--	--	--	0.011	0.092	--	--	0.018	0.094
sudoku*bin2	--	--	--	--	--	--	--	--	-0.012	0.103	--	--	-0.022	0.105
sudoku*ter1	--	--	--	--	--	--	--	--	-0.172	0.130	--	--	-0.188	0.132
sudoku*ter2	--	--	--	--	--	--	--	--	-0.025	0.152	--	--	-0.072	0.155
sudoku*ter3	--	--	--	--	--	--	--	--	-0.138	0.157	--	--	-0.173	0.160
sudoku*quar	--	--	--	--	--	--	--	--	-0.023	0.144	--	--	-0.026	0.146

Note. * $p < .05$, ** $p < .01$

Table 5.10.
Model fit indices for the different DFF-models (Study 3)

	Model 1	Model 2	Model 3	Model 4	Model5	Model 6	Model 7
<i>N</i>	657	657	657	657	657	657	657
<i>ll</i>	-11648.118	-11625.685	-11615.635	-11610.796	-11607.844	11595.804	-11583.574
<i>df</i>	6	7	13	7	13	8	20
AIC	23308.236	23265.370	23257.270	23235.592	23241.688	-23175.608	23207.148
BIC	23335.162	23296.784	23315.610	23267.006	23300.028	-23139.707	23296.902

Note. Fit indices printed in boldface indicate best model fit across the competitive models.

6

Epilogue

With this thesis, “Measuring Reasoning Ability: Applications of Rule-Based Item Generation”, two general research goals were addressed through three empirical studies. First, this thesis aimed at further investigating the usefulness of item-generation models, specifically the class of explanatory IRT models, in predicting item difficulties under various conditions and for different types of reasoning measures. The second goal was to show the benefits and limitations of item-generation models for a deeper understanding and improvement of construct validity. The first two studies concerned the construction and validation of two new instruments, the Figural Analogy Test (FAT), and the Number Series Test (NST). The third study represents an application of a previously validated reasoning measure, the Latin Square Task (LST), in a cross-cultural setting. Each of the three studies contributes to different extents to the two main research goals. An overall goal of this thesis was to develop new item-generative frameworks that provide a basis for automatic generation of test items to be included in future computerized adaptive testing software. Future studies might use a currently being developed software that allows for the automatic on-the-fly generation of items for all three instruments (i.e., FAT, NST, and LST) described in this thesis.

6.1. Prediction of item difficulties by means of explanatory IRT models

From an applied point of view, rule based AIG comes along with the promise to provide a framework allowing for the prediction of item difficulties in practical assessment situations. When explanatory IRT models could be used to predict item difficulties based on a pre-specified matrix of task parameters calibration of individual items is, in theory, not necessary anymore and a (more or less) endless universe of possible test items with known psychometric properties could be generated on-the-fly (i.e., during test administration). Here, two aspects need to be highlighted: first, predictions of item difficulties for actual rule-based generated tests have turned out by far less perfect than expected “in theory”. Second, the large number of constraints that apply to the free combination of item radicals to generate new items makes the universe of possible test items by far less “endless” than one might hope for. The number of constraints of the generation model often limits the sheer number of possible radical combinations and thereby the number of theoretically differently difficult item families.

The two new item generation-frameworks devoted special attention to reduce the number of constraints regarding the possible combinations of item radicals in order to allow for generation of items covering the full item difficulty continuum. With sets of, when considered alone, relatively simple logical rules both new item-generation frameworks succeeded in allowing for the generation of items of low, medium as well as high difficulty. This is an important requirement for the use of these frameworks in high-stakes and/or adaptive testing situations. Paper-pencil administrations were conducted for the studies presented here. A computer-based item-generation and test-administration will be possible with a software package that is currently under development. Item-generative rules for the FAT and the NST were outlined in the necessary detail to allow future fully automatic item generation based on the specified radicals and incidentals. The two empirical studies presented can be considered important pilot-studies that demonstrate the general feasibility of the generation approach.

The two new frameworks provide an excellent basis for generating items and show reasonable to good — depending on how complex the design matrices specified are — performance in terms of prediction of item difficulty parameters. The extent to which facet difficulties are good predictors of item difficulties, ultimately, determines how useful item-generation approaches are for actual testing settings and how much they can increase the efficiency of the assessment process. A major factor for the explanatory value and the power in predicting and explaining item difficulties of the model is the structure of the design matrix chosen. Classical LLTM applications build on the additivity assumption, i.e., item difficulty is given by a linear combination of item facet difficulties. Item cloning approaches focus on the difficulties of certain families of structurally identical items. Results from study 1 and study 2 showed that the item-generating rules underlying a test

item alone were not sufficient to achieve an accurate prediction of “true” item difficulty parameters. Although predicted difficulties and true difficulties correlated substantially (some authors have introduced the correlation of predicted and true parameters as a goodness-of-fit indicator for the item-generative model; see e.g., Arendasy et al., 2007; Freund et al., 2008; Preckel, 2003), the error in terms of the absolute difference in parameters were tremendous. For some items, errors exceeded values of one logit, meaning that an item with a true difficulty of 1 could be wrongly classified as an item with a negative parameter. It has been shown that these values are in the range that have been reported by other item-generation studies as well (see e.g., Zeuch, 2011), and not due to extremely bad model fit for the current applications. When automatically generated and uncalibrated items should be used in a CAT context, this could lead to serious errors in the estimation of person abilities. However, these results could be substantially improved for both new tests when not only rules alone but also other complexity parameters, such as the combination principles that were applied to combine rules in one item, were added to the explanatory models. Here, both correlative relationships between true and predicted item parameters as well as errors in terms of absolute differences between parameters indicated that parameter recovery was very satisfactory. Results from study 2 also showed that a “virtual item model” with a less sparse design matrix that included one item predictor for each radical configuration instead of only one parameter per radical (i.e., the design matrix was set up in line with the item-cloning idea) could lead to very accurate predictions of item difficulties. The results of this study demonstrate that parallel number series test forms could be constructed based on a generative framework if sources for heterogeneity in item difficulties were carefully controlled.

These findings point to the need of a more precise definition of the design matrix, including surface characteristics and specific item features present in individual item types in addition to the generating and logical rules. By definition, adding item explanatory variables to the model must increase the predictive value of the model, in the most extreme case leading to a model with one parameter for each item yielding perfect prediction (or “description”, cf. De Boeck & Wilson, 2004a). Future studies should investigate effects of this trade-off between a sparse explanatory model for item difficulties and an accurate prediction of true item difficulties on the estimation of person abilities more systematically. In order to achieve item parameter predictions based on explanatory IRT modeling that are accurate enough to be used in practical testing settings, such as computerized adaptive testing, there seems no way around specifying complex design matrices that incorporate more parameters than “just” the very basic set of logical rules that determine the structure underlying an item. Study 1 showed that spatial displacement parameters alone could explain around 70 percent of variation in item difficulty parameters for figural analogy items. Results from study 1 showed that the inclusion of additional complexity parameters (parameters that could have been also conceptualized as incidentals as they were not related to the underlying logical structure of the analogy) could add significantly to the prediction of item difficulties, yielding an explanatory power of 86

percent. By investigating both correlational relationships between rescaled LLTM and true RM parameters and absolute parameter differences, and by demonstrating considerable discrepancies in results depending on the method applied, the studies in this thesis demonstrated the need of development of better criteria of what “good predictions” of item difficulties are and when item-generation models provide a sufficient basis for “on-the-fly” generation of items. The extreme absolute errors in predicting item difficulties for models that, at the same time, fulfilled criteria for sufficient construct representation mentioned in the literature (e.g., Arendasy, 2005), point that a new answer to the question what “good enough” prediction means and how it can be operationalized, needs to be found.

Study 3 looked from a different angle on an established rule-based generated figural reasoning measure by investigating the degree of Differential Item Functioning (DIF) in items with defined underlying structure. Specifically the finding was further investigated that items with identical design vectors did not necessary function the same in a cross-cultural context. some items showed DIF while others did not. These differences could not be explained by the underlying cognitive structure of an item. Qualitative analyses showed that other item features than the ones specified in the design matrix seemed to be causal for the DIF effects.

6.2. Understanding and enhancing construct validity

The three studies presented in this thesis indicated that explanatory IRT models with comprehensive design matrices can achieve reliable predictions of item difficulties. However, If accurate item difficulty predictions are needed, individual item calibrations are still clearly superior than predictions based on calibrated item facets. However, even imperfect predictions of item difficulties can be of great benefit to understand item-response processes better and enhance construct validity. Whereas classical approaches have defined reasoning rather idiosyncratically and tested the construct-validity of measures supposed to measure Reasoning ability only in a post-hoc way by investigating correlations of test scores with other tests and variables, rule-based item generation approaches allow (and also force) the test developer to break down test items into their constituting parts, i.e. item facets. Facet level analyses can be used to test the construct validity of a test by modeling the internal structure of the tasks. Thereby, item-difficulty models can as well help to test cognitive theories about human performance (cf. Embretson, 1983).

Theories and findings about analogical and spatial reasoning were used to derive item-generative rules for a new figural-spatial analogy test, the FAT. Information processing theories for series completion tasks, specifically previous works regarding number series tasks, were the basis for the derivation of the item-generative framework for a new type of number series items, the NST. From the perspective of rule-based item generation as a means to test and enhance construct validity, both test development attempts clearly

were successful. For both tests, hypotheses about difficulty-generating item features and rules could be confirmed.

Spatial displacement rules in the FAT were main drivers of item difficulty in figural analogy items, and gender-effects were driven by specific item-facets that are related to more narrow gender-specific abilities: in line with the specific hypotheses item facets related to mental rotation and application of holistic processing strategies were facilitated for men. Item facets related to the application of analytic processing strategies were facilitated for female test-takers. The distinction between radicals representing structural-logical rules and radicals representing general complexity parameters showed that a major proportion of item difficulties is produced by the difficulty of the underlying logical rules. However, general complexity parameters turned out to be of considerable importance as well. This adds an important facet to the knowledge about difficulty generating processes for figural-spatial reasoning items. Given the inconsistencies between the height of facet difficulties and underlying information processing theories reported by previous studies (e.g., Beckmann, 2008; Porsch, 2007) the results of the studies in this thesis are very promising. The very detail-focused derivation of item-generative rules and the careful definition of radicals and incidentals turned out to be necessary in order to describe the underlying construct accurately. The fact that gender differences could be replicated in line with theories about gender differences in figural-spatial tasks adds to the validity of the new framework. Correlations with other measures showed that both general figural reasoning and specific aspects of spatial reasoning could be captured by the new measure. Even though the FAT is based on strictly two-dimensional stimuli correlations with the 3DW, a three-dimensional mental rotation test, showed substantial overlap between the two measures. Regarding the prediction of grades, the new measure showed incremental validity beyond the explanation based on the CFT and the 3DW.

Item difficulties of the NST were predominantly determined by the relational complexity of two consecutive numbers. Complexity levels could be manipulated considerably by combination of a set of relatively simple arithmetic rules requiring only addition and subtraction. This helped to solve several of the problems described in previous studies. By relying on simple arithmetic operations only, a true reasoning measure (instead of a test of arithmetics) could be developed. Also, problems such as the problem size effect (cf. Ashcraft, 1992) could be avoided by constraining the range of possible numbers and mathematic operations. Most notably, the new item-generative framework presented a solution to how number series can be generated that are built based on exactly the same logical operations for each pair of neighboring numbers. Existing number series tests have used a sequential combination of rules to produce complexity (e.g., Amthauer et al., 2001), thereby — as a side-product — creating number series with increased period lengths and breaking points that allow the experienced test-taker to apply a number of facilitating test-taking strategies. However, the difficulty of a number series can only be defined unequivocally if the operations that need to be performed are exactly the same,

no matter whether the sixth or seventh, or eighth element has to be derived. Only in this case do test-takers need to represent all logical rules and cannot solve a series by relying on heuristics, such as difference- or ratio-strategies (cf. Irle, 1969). In the new item-generative framework, the logical rules between every pair of neighboring elements are exactly the same. This also allows for alternative, innovative item presentation modes. For instance, a series could be presented with any (and not necessarily the rightmost) element missing — an idea that could be investigated empirically by future studies. The item-generative framework was shown to be relatively robust against irrelevant surface patterns in the numbers caused by random incidentals. After a warm-up run, item difficulties could be predicted very reliably for two parallel test forms. The demonstration of psychometric equivalence of items with the same underlying theoretical structure is an important accomplishment for the construct validity of the NST.

Results from the third study confirmed the cross-cultural validity of the LST in a broad sense but also pointed to problems with the functioning of individual items in a cross-cultural context. Analyses indicated that the same construct was measured with the LST in both populations. However, a considerable number of items showed DIF, indicating that test-takers with the same value on the latent ability continuum did not have equal chances of answering such items correctly. Analyses indicated that bias might be caused by item incidental parameters rather than by item radicals. The expectations of the cognitive complexity approach, that more complex operations were more prone to DIF, could not be confirmed. With regard to the bias-generating processes, the DIF results suggested that bias was caused by broad cultural variables that cause differences between the test-takers from the two countries, and not predominantly by test-specific context variables, here operationalized as experiences with the number-placement game SUDOKU. Results from Differential Facet Functioning models were used to evaluate cognitive theories on the emergence of item bias. The predictive power of this approach for the identification of DIF items before the actual test administration was investigated in an empirical study with data from two culturally diverse samples. Results indicate that more research is needed to clarify on the benefit of facet-level DIF analyses.

6.3. Limitations

In addition to the limitations mentioned in each of the three studies, the conclusions of this thesis are limited in several ways. Not all tests involved long warm-up runs. Results of study 2 show that the role of warm-up runs seems to be by far more important than one might expect with large effects after the initial warm-up run. The fact that only few warm-up items were used in studies 1 and 3 constitutes a major limitation for these studies. Future studies need to investigate this and test whether findings reported here might be biased due to the lack of long warm-up runs. This should involve also a more extensive study of the effects of prior relevant knowledge and an investigation of the specific transfer

effects that occur when knowledge gained through certain everyday experiences is applied to tasks in a testing situation. With regards to the cross-cultural research questions investigated, this thesis only made a first step towards validating AIG models in multinational applications. Studies with more than two cultural groups are needed and would constitute important extensions of the work presented here. The DFF results presented need to be extended to fully clarify on the relationships between DIF and DFF. Simulation studies might be a way to proceed here. Further, future studies should investigate the technical feasibility of the new item generation approaches for large-scale or high-stakes assessments of reasoning ability. The suitability of the statistical item difficulty model to predict item difficulties of on-the-fly generated test items in computerized adaptive testing (CAT) needs to be demonstrated. A further test of the prediction of item parameters by the recently proposed variants of Item Cloning Models (ICM; e.g. Geerlings et al., 2011) would be very useful. If a larger number of parallel item “clones” could be administered to a larger sample of test-takers, item family parameters and within-family variances could be estimated by Bayesian models. A test of the robustness of the generation approach to practice and training effects would be valuable to determine the suitability of the new Reasoning measures for high-stakes testing applications. After all, a direct comparison of items generated manually based on a rule-based framework (as the case in the current thesis) and fully computerized item generation is still pending. An important step will be to implement relevant quality control mechanisms when fully computerized generation is used.

While this thesis illustrated some of the most important advantages and issues regarding the use of AIG, the major result is probably that it must be said that lots of additional work is needed to solve all remaining questions and establish a sufficient research base for operational use of AIG in high-stakes cross-cultural testing scenarios. But it is clear from the findings presented that this additional research seems worthwhile and is not unlikely, if successful, to fundamentally change the ways cognitive tests are developed and implemented in the future.

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.
- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47-76.
- Aguerri, M. E., Galibert, M. S., Attorresi, H. F., & Maranon, P. P. (2009). Erroneous detection of nonuniform DIF using the Breslow-Day test in a short test. *Quality & Quantity*, *43*, 35-44.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing* (2nd ed.). Washington, DC: American Psychological Association.
- Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R*. Göttingen: Hogrefe.
- Anastasi, A. (1981). Coaching, test sophistication and developed abilities. *American Psychologist*, *36*, 1086-1093.
- Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*, 153 - 219.
- Angoff, W. (1993). *Perspectives on differential item functioning methodology*. Mahwah, NJ: Erlbaum.
- 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., Hergovich, A., Sommer, M., & Bogner, B. (2005). Dimensionality and construct validity of a video-based, objective personality test for the assessment of willingness to take risks in road traffic. *Psychological Reports*, *97*, 309-320.
- Arendasy, M., Sommer, M., & Hergovich, A. (2007). Psychometrische Technologie: Automatische Zwei-Komponenten-Itemgenerierung am Beispiel eines neuen Aufgabentyps zur Messung der Numerischen Flexibilität. *Diagnostica*, *53*, 119-130.
- Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*, 75-106.
- Baddely, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

- Baenninger, M., & Newcombe, N. (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles, 20*, 327-344.
- Baenninger, M., & Newcombe, N. (1995). Environmental input to the development of sex-related differences in spatial and mathematical ability. *Learning and Individual Differences, 7*, 363-379.
- Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement, 17*, 201-210.
- Baldiga, K. (2011). *Gender differences in willingness to guess and the implications for test scores (Harvard University Job Market Candidate Paper Series)*. <http://www.people.fas.harvard.edu/~kbaldiga/Gender.pdf>. (Retrieved 12/03/2011)
- Bechger, T., Verstralen, H., & Verhelst, N. (2002). Equivalent linear logistic test models. *Psychometrika, 67*, 123-136.
- Beckmann, B. (2008). *Reasoning ability: Rule-based test construction of a figural analogy test*. Münster, Germany: University of Münster.
- Bejar, I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*, 237-245.
- Bejar, I. (1993). A generative approach to psychological and educational measurement. In R. J. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (p. 323-357). Hillsdale: Erlbaum.
- Bejar, I. (2012). Item generation: Implications for a validity argument. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (p. 40-56). New York: Routledge.
- Bejar, I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of on-the-fly item generation in adaptive testing [ETS Research Report 02-23]*. Princeton, NJ: Educational Testing Service.
- Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002a). *Cross-cultural psychology: Research and applications* (2nd ed.). New York: Cambridge University Press.
- Berry, J. W., Poortinga, Y. H., Segall, M. H., & Dasen, P. R. (2002b). *Cross-cultural psychology: Research and applications* (2nd ed.). New York: Cambridge University Press.
- Bertling, J. P. (2007). *Prediction of person ability based on imprecisely estimated item parameters: A simulation approach to rule-based item generation*. Münster, Germany: Westfälische Wilhelms-Universität.
- Bertling, J. P., & Holling, H. (2009). *Figural Analogy Test (FAT): Item Construction Manual [Internal working paper]*. Muenster, Germany: University of Muenster.
- Bethell-Fox, C. E., Lohman, D. F., & Snow, R. E. (1984). Adaptive reasoning: Componential and eye movement analysis of geometric analogy performance. *Intelligence, 8*, 205 - 238.
- Binet, A. (1903). *L'étude expérimentale de l'intelligence [Experimental studies of intelligence]*. Paris: Schleicher.

- Birenbaum, M., Kelly, A. E., & Levi-Keren, M. (1994). Stimulus features and sex differences in mental rotation test performance. *Intelligence*, *19*, 51-64.
- 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.
- Blaira, C., Gamsonb, D., Thornec, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence*, *33*, 93-106.
- Bodunov, V., M., Bezdenezhnykh, B. N., & Akexandrov, Y. I. (1996). Peculiarities of psychodiagnostic test item responses and the structure of individual experience. *Psikhologicheskii Zhurnal*, *17*, 87-96.
- Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren-Inventar (NEO-FFI)*. Göttingen, Germany: Hogrefe.
- Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. Thousand Oaks, CA: Wiley.
- Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Vol. 1. The analysis of case-control studies (Scientific Publication No. 32)*. Lyon, France: International Agency for Research on Cancer.
- Brouwers, S. A., Van de Vijver, F. J. R., & van Hemert, D. (2009). Variation in Raven's progressive matrices scores across time and place. *Learning and Individual Differences*, *19*, 330-338.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Carlstedt, B., Gustafsson, J.-E., & Ullstadius, E. (2000). Item sequencing effects on the measurement of fluid intelligence. *Intelligence*, *28*, 145 - 160.
- 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. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology*, *31*, 161-179.
- Cattell, R. B. (1949). *Test of "g": Culture fair*. Savoy, IL: Institute for Personality and Ability Testing.
- Cattell, R. B. (1968). Are IQ-tests intelligent? *Psychology Today*, *2*, 56-62.
- Cattell, R. B. (1971). *Abilities: their structure, growth and action*. Boston: Houghton Milton.

- Cattell, R. B. (1973). *Measuring intelligence with the Culture Fair Test*. Champaign, IL: Institute for Personality and Ability Testing.
- Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? a reassessment of the evidence. *Developmental Psychology*, *27*, 703-722.
- Ceci, S. J. (1999). How much does schooling effect general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology*, *27*, 703-722.
- Central Intelligence Agency. (2009). *The World Factbook 2009*. Washington, DC: Central Intelligence Agency.
- Colom, R., Juan-Espinosa, M., Abad, F., & Garcia, L. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*, 57-68.
- Cooke-Simpson, A., & Voyer, D. (2007). Confidence and gender differences on the mental rotations test. *Learning and Individual Differences*, *17*, 181-186.
- Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*, 51-57.
- Cronbach, L. J. (1957). The two disciplines of scientific Psychology. *American Psychologist*, *12*, 671-684.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281 - 302.
- Crone, E. A., Wendelken, C., van Leijenhorst, L., Honomichl, R. D., Christoff, K., & Bunge, S. A. (2009). Neurocognitive development of relational reasoning. *Developmental Science*, *12*, 55-66.
- cut-e. (2011). *Logical consequences with legal quadrants: Scales 1st*. <http://www.cut-e.com/solutions/our-products/online-tests/experts/scales-1st-logical-thinking/>. (Retrieved 07/19/2011)
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533-559.
- De Boeck, P., & Wilson, M. (2004a). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
- De Boeck, P., & Wilson, M. (2004b). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 3-39). New York: Springer.
- Deary, I. J. (2001). Human intelligence differences: Towards a combined experimental-differential approach. *Trends in cognitive science*, *5*, 164-170.
- Demetriou, A., Kui, Z. X., Spanoudis, G., Christou, C., Kyriakides, L., & Platsidou, M. (2005). The architecture, dynamics, and development of mental processing: Greek, Chinese or universal? *Intelligence*, *33*, 109-141.
- Domino, G., & Domino, M. L. (2006). *Psychological testing*. Cambridge: Cambridge University Press.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic

- Aptitude Test. *Journal of Educational Measurement*, 23, 355-368.
- Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., p. 471-516). Washington, DC: American Council on Education.
- Dykiert, D., Gale, C. R., & Deary, I. J. (2009). Are apparent sex differences in mean IQ scores created in part by sample restriction and increased male variance? *Intelligence*, 37, 42-47.
- Ebert, H., & Tack, W. (1974). Einige Lerneffekte bei Aufgaben zur Zahlenfolgen-Induktion. *Zeitschrift für experimentelle und angewandte Psychologie*, 11, 511-529.
- Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.
- 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.
- Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, 38, 343-368.
- Embretson, S. E., & Schmidt-McCollam, K. M. (2000). Psychometric approaches to understanding and measuring intelligence. In R. J. Sternberg (Ed.), *Handbook of Intelligence* (p. 423-444). New York: Cambridge University Press.
- Engelhard, G. J. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5, 171-191.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128(3), 309 - 331.
- English, L. D., & Halford, G. S. (1995). *Mathematics Education Models and Processes*. Mahwah, NJ: Erlbaum.
- 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.
- Feng, J., Spence, I., & Pratt, J. (2007). Playing an action video game reduces gender differences in spatial cognition. *Psychological Science*, 18, 850-855.
- 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.
- Fischer, G. H. (1995). Linear logistic models for change. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (p. 157-180). New York: Springer.
- Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement*, 6,

- 397-416.
- Fischer, G. H., & Ponocny, I. (1995). Extended rating scale and partial credit models for assessing change. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: foundations, recent developments, and applications* (p. 353-370). New York: Springer.
- Fiske, A. P. (2002). Using individualism and collectivism to compare cultures—a critique of the validity and measurement of the constructs: Comment on oyserman et al. (2002). *Psychological Bulletin*, *128*, 78D88.
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171-191.
- 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.
- Freund, P. A., & Holling, H. (2011). How to get really smart: Modeling retest and training effects in ability testing using computer-generated figural matrix items. *Intelligence*, *39*, 233-243.
- Frijda, N., & Jahoda, G. (1966). On the scope and methods of cross-cultural research. *International Journal of Psychology*, *1*, 109-127.
- Geerlings, H., Glas, C. A. W., & Van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*, 337-359.
- Geiser, C., Lehmann, W., & Eid, M. (2006). Separating “Rotators” from “Nonrotators” in the Mental Rotations Test: A multigroup latent class analysis. *Multivariate Behavioral Research*, *41*(3), 261 - 293.
- Geiser, C., Lehmann, W., & Eid, M. (2008). A note on sex differences in mental rotation in different age groups. *Intelligence*, *36*, 556-563.
- Georgas, J., Van de Vijver, F. J. R., & Berry, J. W. (2004). The ecocultural framework, ecosocial indicators and psychological variables in cross-cultural research. *Journal of Cross-Cultural Psychology*, *35*, 74-96.
- Gierl, M. J., & Haladyna, T. M. (2012). *Automatic Item Generation: Theory and practice*. New York: Routledge.
- Gierl, M. J., & Lai, H. (2012). Using weak and strong theory to create item models for automatic item generation: Some practical guidelines with examples. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (p. 26-39). New York: Routledge.
- Gilinsky, A. S., & Judd, B. B. (1993). Working memory and bias in reasoning across the life span. *Psychology and Aging*, *9*, 356-371.
- Gittler, G. (1990). *3 DW Dreidimensionaler Würfeltest [Three-dimensional Cube Test]*. Göttingen, Germany: Hogrefe.

- Gittler, G. (1999). Sind Raumvorstellung und Reasoning separierbare Fähigkeitsdimensionen? Dimensionalitätsanalysen zweier Rasch-skaliertes Tests: 3DW und WMT. *Diagnostica*, *45*, 69-81.
- Glas, C. A. W., & Van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247-261.
- Glaser, R. (1982). Analyzing aptitudes for learning: Inductive Reasoning. In J. W. Pellegrino & R. Glaser (Eds.), *Advances in instructional psychology* (p. 269-345). Mahwah, NJ: Erlbaum.
- 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]*. Münster, Germany: University of Münster.
- Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*, 477-485.
- Gorin, J. S., & Embretson, S. E. (2012). Using cognitive psychology to generate items and predict item characteristics. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (p. 136-156). New York: Routledge.
- Gottfredson, L. J. (1997). Why g matters: The complexity of everyday life. *Intelligence*, *24*, 79-132.
- Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*, 316-322.
- Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement*, *26*, 147-160.
- Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.
- Guilford, J. P. (1985). The structure-of-intellect model. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (p. 225-266). New York: Wiley.
- Gustafsson, J.-E. (2001). Schooling and intelligence: Effects of track of study on level and profile of cognitive abilities. *International Education Journal*, *2*, 166-186.
- Gustafsson, J.-E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (p. 186-242). New York: Simon & Schuster Macmillan.
- Hackett, G., Betz, N. E., O'Halloran, M., & Romac, D. S. (1990). Effects of verbal mathematics task performance on task and career self-efficacy and interest. *Journal of Counseling Psychology*, *37*, 169-177.
- Hakstian, A. R., & Vandenberg, S. G. (1979). The cross-cultural generalizability of a higher-order cognitive structure model. *Intelligence*, *3*, 73-103.
- 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.
- Halpern, D. F., & La May, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review*, 12, 229-246.
- Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology*, 44, 339-384.
- Hartmann, P., Kruuse, N. H., & Nyborg, H. (2007). Testing the cross-racial generality of Spearman's hypothesis in two samples. *Intelligence*, 35, 47-57.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. O. M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92, 373-385.
- Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, 32, 175 - 191.
- Heil, M., & Jansen-Osmann, P. (2008). Sex differences in mental rotation with polygons of different complexity: Do men utilize holistic processes whereas women prefer piecemeal ones? *The Quarterly Journal of Experimental Psychology*, 61, 683-689.
- Heller, K. A., Schön-Gaedike, A.-K., & Weinläder, H. (1976). *Kognitiver Fähigkeits-Test für 4. bis 13. Klassen (KFT 4-13+)*. Weinheim: Beltz.
- Helms-Lorenz, M., Van de Vijver, F. J. R., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: g or c? *Intelligence*, 31, 9-29.
- Hennessy, J. J., & Merrifield, P. R. (1976). A comparison of the factor structures of mental abilities in four ethnic groups. *Journal of Educational Psychology*, 68, 754-759.
- Herkovits, M. J. (1948). *Man and his works: The science of cultural anthropology*. New York: Alfred Knopf.
- Herrnstein, R. J., Nickerson, R. S., de Sanchez, M., & Swets, J. A. (1986). Teaching thinking skills. *American Psychologist*, 41, 1279-1289.
- Hersh, H. M. (1974). The effects of irrelevant relations on the processing of sequential patterns. *Memory & Cognition*, 2, 771-774.
- Hochberg, J., & Gellman, L. (1977). The effect of landmark features on mental rotation times. *Memory & Cognition*, 5, 23-26.
- Hoffmann, N. (2007). *Lateinische Quadrate: Psychometrische Eigenschaften und Lerneffekte [unpublished diploma thesis]*. Münster, Germany: Westfälische Wilhelms-Universität.
- Hohensinn, C., Kubinger, K. D., Reif, M., S., H.-E., 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.
- Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ets delta scale of item difficulty [ETS Research Report 85-43]*. Princeton, NJ: Educational Testing Service.

- 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., Bertling, J. P., Zeuch, N., & Kuhn, J.-T. (2010). Automatische Itemgenerierung: Grundlagen und Darstellung automatisch generierter Items anhand Lateinischer Quadrate. In F. Preckel, W. Schneider, & H. Holling (Eds.), *Tests und Trends: Hochbegabung* (p. 211-231). Berlin: Hogrefe.
- Holzman, T. G., Pellegrino, J. W., & Glaser, R. (1983). Cognitive variables in series completion. *Journal of Educational Psychology*, *75*, 603-618.
- Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*, 107-129.
- Huesmann, L. R., & Cheng, C.-M. (1973). A theory for the induction of mathematical functions. *Psychological Review*, *80*, 126-138.
- Hugdahl, K., Thomsen, T., & Ersland, L. (2006). Sex differences in visuo-spatial processing: An fMRI study of mental rotation. *Neuropsychologia*, *44*, 1575-1583.
- Hunt, E. (1976). Varieties of cognitive power. In L. Resnick (Ed.), *The nature of intelligence* (p. 237-260). Hillsdale, NJ: Erlbaum.
- Imbo, I., & Vandierendonck, A. (2008). Effects of problem size, operation, and working-memory span on simple-arithmetic strategies: Differences between children and adults? *Psychological Research*, *72*, 331-346.
- Irle, M. (1969). Beiträge zur Computersimulation kognitiver Prozesse auf dem Niveau interindividueller Differenzen. In R. Groner (Ed.), *Bericht über den 26. Kongress der DGfP* (p. 279-286). Göttingen: Hogrefe.
- Irvine, S. H. (1969). Factor analysis of african abilities and attainments: Constructs across cultures. *Psychological Bulletin*, *71*, 20-32.
- Irvine, S. H., & Berry, J. W. (1988). The abilities of mankind: A reevaluation. In S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context*. Cambridge: Cambridge University Press.
- Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Mahwah, NJ: Erlbaum.
- Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, *96*, 505-524.
- Irwing, P., & Lynn, R. (2006). Intelligence: Is there a sex difference in IQ scores? *Nature*, *438*, 31-32.
- Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, *28*, 195-226.
- Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., et al. (2006). *Berliner Intelligenzstruktur-Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB)*. Göttingen, Germany: Hogrefe.

- Janssen, A. B., & Geiser, C. (2010). On the relationship between solution strategies in two mental rotation tasks. *Learning and Individual Differences, 20*, 473-478.
- Janssen, A. B., & Geiser, C. (in press). Cross-cultural differences in spatial abilities and solution strategies: An investigation in Cambodia and Germany. *Journal of Cross-Cultural Psychology*.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 189-210). New York: Springer.
- Jensen, A. R. (1977). Cumulative deficit in IQ of Blacks in the rural South. *Developmental Psychology, 13*, 184 - 191.
- Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Free Press.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349.
- Johnson, D. M. (1962). Serial analysis of verbal analogy problems. *Journal of Educational Psychology, 53*, 86-88.
- Johnson, M. S., & Sinharay, S. (2003). *Calibration of polytomous item families using bayesian hierarchical modeling (ETS RR-03-23)*. Princeton, NJ: Educational Testing Service.
- Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence, 33*, 393 - 416.
- Johnson, W., & Bouchard, T. J. (2007). Sex differences in mental abilities: g masks the dimensions on which they lie. *Intelligence, 35*, 23-39.
- Jolicoeur, P. (1985). The time to name disoriented natural objects. *Memory & Cognition, 13*, 289-303.
- Jordan, K., Wüstenberg, T., Heinze, H. J., Peters, M., & Jäncke, L. (2002). Women and men exhibit different cortical activation patterns during mental rotation tasks. *Neuropsychologia, 40*, 2397-2408.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.
- Kanamori, N., & Yagi, A. (2002). The difference between flipping strategy and spinning strategy in mental rotation. *Perception, 31*, 1456-1466.
- Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review, 9*, 637-671.
- Kenrick, D. T., Maner, J. K., Butner, J., Li, N. P., Becker, D. V., & Schaller, M. (2002). Dynamic evolutionary psychology: Mapping the domains of the new interactionist paradigm. *Personality and Social Psychology Review, 6*, 347-356.

- Klahr, D., & Wallace, J. (1970). An information processing analysis of some plagetian experimental tasks. *Cognitive Psychology*, *1*, 358-387.
- Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, *107*, 191-206.
- Korossy, K. (1998). Solvability and uniqueness of linear-recursive number sequence tasks. *Methods of Psychological Research Online*, *3*, 43-68.
- Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, *4*, 399-424.
- Kozhevnikov, M., Kosslyn, S., & Shephard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory & Cognition*, *33*, 710-726.
- Kroeber, A. L., & Kluckhohn, C. K. (1952). *Culture: A critical review of concepts and definitions*. Cambridge, MA: Harvard University Press.
- Kuhn, J.-T. (2010). *Reasoning ability and working memory capacity in children*. Münster, Germany: University of Münster.
- Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*, 162-181.
- Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59-81.
- Kyllonen, P. C., & Christal, R. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389-433.
- Lai, H., Alves, C., & Gierl, M. J. (2009). *Using automatic item generation to address item demands for CAT*. www.psych.umn.edu/psylabs/CATCentral/. (Retrieved 09/10/2011)
- Lawson, R., Humphreys, G. W., & Jolicoeur, P. (2000). The combined effects of plane disorientation and foreshortening on picture naming: One manipulation or two? *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 568-581.
- Lee, N. Y., Goodwin, G. P., & Johnson-Laird, P. N. (2008). The psychological puzzle of Sudoku. *Thinking and Reasoning*, *14*, 342-364.
- LeFevre, J.-A., & Bisanz, J. (1986). A cognitive analysis of number-series problems: Sources of individual differences in performance. *Memory & Cognition*, *14*, 287-298.
- LeFevre, J.-A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 216-230.
- Lemke, M., & Gonzales, P. (2006). *U.S. Student and Adult Performance on International Assessments of Educational Achievement: Findings from The Condition of Education 2006 (NCES 2006-073)*. U.S. Department of Education. Washington, DC: National Center for Education Statistics (NCES).

- Leung, K., & Van de Vijver, F. J. R. (2008). Strategies for strengthening causal inferences in cross cultural research: The consilience approach. *International Journal of Cross Cultural Management*, 8, 145-169.
- Lievens, F., Reeve, C. L., & Heggstad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, 92, 1672-1682.
- Linacre, J. M. (2010). *A user's guide to Winsteps: Rasch-model computer programs (Version 3.70.1)*. Chicago: John M. Linacre.
- Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development*, 56, 1479-1498.
- Lonner, W. J. (1980). The search for psychological universals. In H. C. Triandis & W. W. Lambert (Eds.), *Handbook of cross-cultural psychology: Perspectives (Vol. 1)* (p. 143-204). Boston, MA: Allyn & Bacon.
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Luria, A. (1976). *Cognitive development: Its cultural and social foundations*. Cambridge, MA: Harvard University Press.
- Lynn, R., & Owen, K. (1994). Spearman's hypothesis and test score differences between Whites, Indians, and Blacks in South Africa. *Journal of General Psychology*, 121, 27-36.
- Lynn, R., & Vanhanen, T. (2002). *IQ and the wealth of nations*. Praeger: Westport, CT.
- Macionis, J., & Plummer, K. (1998). *Sociology*. New York: Prentice Hall.
- Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42, 847-862.
- Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20, 1-20.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, 7, 107 - 127.
- Masters, M. S., & Sanders, B. (1993). Is the gender difference in mental rotation disappearing? *Behavior Genetics*, 23, 337-345.
- Matsumoto, D. (2001). *The handbook of culture and psychology*. New York: Oxford University Press.
- Matsumoto, D., & Yoo, S. H. (2006). Toward a new generation of cross-cultural research. *Perspectives on Psychological Science*, 1, 235-250.
- McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, 52, 509-516.

- Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models* (p. 214-240). New York: Springer.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.
- Misra, G., Sahoo, F. M., & Puhan, B. N. (1997). Cultural bias in testing: India. *European Review of Applied Psychology*, *47*, 309-317.
- Mittring, G., & Rost, D. H. (2008). Die verflixten Distraktoren: Über den Nutzen einer theoretischen Distraktorenanalyse bei Matrizen tests (für besser Begabte und Hochbegabte). *Diagnostica*, *54*, 193-201.
- Moe, A., Meneghetti, C., & Cadinu, M. (2009). Women and mental rotation: Incremental theory and spatial strategy use enhance performance. *Personality and Individual Differences*, *46*, 187-191.
- Mulholland, T. M., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. *Cognitive Psychology*, *12*, 252-284.
- Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Mathematics Framework*. Amsterdam, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA). Available from http://timss.bc.edu/timss2011/downloads/TIMSS2011_Frameworks-Chapter1.pdf
- Murray, J. E. (1997). Flipping and spinning: Spatial transformation procedures in the identification of rotated natural objects. *Memory & Cognition*, *25*, 96-105.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691-692.
- Naglieri, J. A., & Jensen, A. R. (1987). Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence*, *11*, 21-43.
- Novick, L. R., & Tversky, B. (1987). Cognitive constraints on ordering operations: The case of geometric analogies. *Journal of Experimental Psychology: General*, *116*, 50-67.
- Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*, 641-652.
- Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in selection decisions. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (p. 431-468). Thousand Oaks, CA: Sage.
- Organisation for Economic Co-Operation and Development. (2004). *Problem solving for tomorrow's world: First measures of cross-curricular competencies from PISA 2003*. Paris: OECD.
- Organisation for Economic Co-Operation and Development. (2010). *Pisa 2012 mathematics framework*. Paris: OECD. Available from <http://www.oecd.org/dataoecd/8/38/46961598.pdf>
- Porsch, T. (2007). *Theoriegeleitete Schwierigkeitsvariation von Zahlenreihenaufgaben [Unveröffentlichte Diplomarbeit]*. Münster, Germany: Westfälische Wilhelms-

- Universität.
- Preckel, F. (2003). *Diagnostik intellektueller Hochbegabung. Testentwicklung zur Erfassung der fluiden Intelligenz*. Göttingen, Germany: Hogrefe.
- Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*, 41-70.
- Quereshi, M. Y., & Seitz, R. (1993). Identical rules do not make letter and number series equivalent. *Intelligence*, *17*, 399-405.
- Quereshi, M. Y., & Smith, H. (1998). Reasoning ability in older adults measured through letter and number series. *Current Psychologia*, *17*, 20-27.
- Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495-502.
- Rasch, G. (1860). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.
- Raven, J. C. (1962). *Advanced Progressive Matrices, set II*. London: H. K. Lewis.
- Ree, M. J., & Carretta, T. R. (1995). Group differences in aptitude factor structure on the ASVAB. *Educational and Psychological Measurement*, *55*, 268-277.
- Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 271-285.
- Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: the homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, *21*, 667-706.
- Rohner, R. P. (1984). Toward a conception of culture for cross-cultural psychology. *Journal of Cross-Cultural Psychology*, *15*, 111-138.
- Roid, G., & Haladyna, T. (1981). *A technology of test-item writing*. New York: Academic Press.
- Rost, J., & Davier, M. v. (1994). A conditional item-fit index for Rasch Models. *Applied Psychological Measurement*, *18*, 171-182.
- Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, *68*, 78-96.
- 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.
- Schmitt, D. P. (2006). Cultural influences on human mating strategies: Evolutionary theories, mechanisms, and explanations of change. *Psychological Inquiry*, *17*, 116-117.
- Segall, M. H., Lonner, W. J., & Berry, J. W. (1998). Cross-cultural psychology as a scholarly discipline: On the flowering of culture in behavioral research. *American Psychologist*, *53*, 1101-1110.
- Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*, 701-703.

- Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development*. Hillsdale, NJ: Erlbaum.
- 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.
- Smith, P. B., & Bond, M. H. (1998). *Social psychology across cultures*. London: Prentice Hall.
- Snow, R. E., Federico, P., & Montague, W. E. (1980). Components of inductive reasoning. In J. W. Pellegrino & R. Glaser (Eds.), *Aptitude, learning, and instruction* (p. 177-217). Mahwah, NJ: Erlbaum.
- Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, 50, 345-362.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, 15, 201-203.
- Spearman, C. (1923). *The nature of intelligence and the principles of cognition*. London: Macmillan.
- Spearman, C. (1927). *The abilities of man*. London: Macmillan.
- Spilisbury, G., Stankov, L., & Roberts, R. (1990). The effect of a test's difficulty on its correlation with intelligence. *Personality and Individual Differences*, 11, 1069 - 1077.
- Stankov, L. (2005). g factor: Issues of design and interpretation. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (p. 279-294). Thousand Oaks, CA: Sage.
- Sternberg, R. J. (1977a). Component processes in analogical reasoning. *Psychological Review*, 84, 353-378.
- Sternberg, R. J. (1977b). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.
- Sternberg, R. J. (1984). Analogical thinking and human intelligence. In K. J. Holyoak (Ed.), *Advances in the Psychology of human intelligence* (p. 199-230). Mahwah, NJ: Erlbaum.
- Sung, Y. H., & Dawis, R. V. (1981). Level and factor structure differences in selected abilities across race and sex groups. *Journal of Applied Psychology*, 66, 613-624.
- Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability, and a little bit more. *Intelligence*, 30, 261 - 288.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.
- Te Nijenhuis, J., & van der Flier, H. (2001). Group differences in mean intelligence for the Dutch and Third World immigrants. *Journal of Biosocial Science*, 33, 469-475.

- Terlecki, M., Newcombe, N. S., & Little, M. (2008). Durable and generalized effects of spatial experience on mental rotation: Gender differences in growth patterns. *Applied Cognitive Psychology, 22*, 996-1013.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In D. Wainer & H. Braun (Eds.), *Test validity* (p. 147-170). Hillsdale, NJ: Erlbaum.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence*. Chicago: University of Chicago Press.
- Tooby, J., & Cosmides, L. (1995). The psychological foundations of culture. In J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *Adapted mind: Evolutionary psychology and the generation of culture* (p. 19-136). New York: Oxford University Press.
- Triandis, H. C. (1972). *The analysis of subjective culture*. New York: Wiley.
- Triandis, H. C., Bontempo, R., Villareal, M. J., Asai, M., & Lucca, N. (1988). Individualism and collectivism: Cross-cultural perspectives on self-ingroup relationships. *Journal of Personality and Social Psychology, 54*, 323-338.
- Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science, 14*, 623-628.
- Tylor, E. B. (1871). *Primitive culture*. London: John Murray.
- Undheim, J. O., & Gustafsson, J.-E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations. *Multivariate Behavioral Research, 22*, 149-171.
- Van de Vijver, F. J. R. (1997). Meta-analysis of cross-cultural comparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28*, 678-709.
- Van de Vijver, F. J. R. (2002). Inductive reasoning in Zambia, Turkey, and the Netherlands: Establishing cross-cultural equivalence. *Intelligence, 30*, 313-351.
- Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.
- Van de Vijver, F. J. R., Helms-Lorenz, M., & Feltzer, M. J. A. (1999). Acculturation and cognitive performance of migrant children in the Netherlands. *International Journal of Psychology, 34*, 149-162.
- Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research. In J. W. Berry, Y. Poortinga, & J. Pandey (Eds.), *Handbook of Cross-cultural Psychology* (2nd ed., p. 257-300). Boston: Allyn & Bacon.
- Van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33-51.
- Van de Vijver, F. J. R., & Tanzer, N. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*, 119-135.

- 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.
- Van Hemert, D. A., Van de Vijver, F. J. R., & Poortinga, Y. H. (2004). Models for explaining cultural differences: A meta-analysis. *International journal of psychology*, *39*, 306-306.
- Verguts, T., & De Boeck, P. (2002). On the correlation between working memory capacity and performance on intelligence tests. *Learning and Individual Differences*, *13*, 37-55.
- Vernon, P. A. (1979). *Intelligence: Heredity and environment*. San Francisco: W. H. Freeman & Company.
- Vernon, P. A., & Jensen, A. R. (1984). Individual and group differences in intelligence and speed of information processing. *Personality and Individual Differences*, *5*, 411 - 423.
- Vock, M., & Holling, H. (2008). The measurement of visuo-spatial and verbal-numerical working memory: Development of IRT based scales. *Intelligence*, *36*, 161-182.
- Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250-270.
- Weiß, R. H. (2007). *Grundintelligenztest Skala 2 mit Wortschatztest und Zahlenfolgentest*. Göttingen: Hogrefe.
- Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, *5*, 383-397.
- Wilhelm, O. (2005). Measuring reasoning ability. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence*. Thousand Oaks, CA: Sage.
- Wilhelm, O., & Engle, R. W. (2005). *Handbook of understanding and measuring intelligence*. Thousand Oaks: Sage.
- Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory Item Response Models: A generalized linear and nonlinear approach* (p. 44-74). New York: Springer.
- Winship, C., & Korenman, S. (1997). Does staying in school make you smarter? The effect of education on IQ in the bell curve. In B. Devlin, S. E. Fienberg, D. P. Resnick, & K. Roeder (Eds.), *Intelligence, genes, and success* (p. 215-235). New York: Springer.
- Xie, X., & 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.
- Yang, Y., & Johnson-Laird, P. (2001). Mental models and logical reasoning problems in the GRE. *Journal of Experimental Psychology: Applied*, *7*, 308-316.
- Yildirim, H. H., & Berberoglu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing*, *9*,

108-121.

- Zeuch, N. (2011). *Rule based item construction: Rule-based item construction, analysis with and comparison of linear logistic test models and cognitive diagnostic models with two item types*. Münster, Germany: University of Münster.
- Zeuch, N., Geerlings, H., Holling, H., Van der Linden, W. J., & Bertling, J. P. (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]. *Zeitschrift für Pädagogik [German journal of education]*, 52-63.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223-233.
- Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF [Working Paper of the Edgeworth Laboratory for Quantitative Behavioral Science]*. Prince George, BC: University of Northern British Columbia.

Appendix

This Appendix contains all actual test items for the newly developed tests as well as additional figures and tables that summarize analyses that were done in addition to the main analyses necessary for answering the central research questions. Mostly these are analyses that were added to investigate the general psychometric quality of the measures used and the comparability of test scores from the different samples. The Appendix is structured along the order of the three studies presented.



Study 1 – Additional Materials

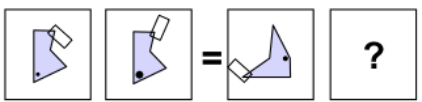










































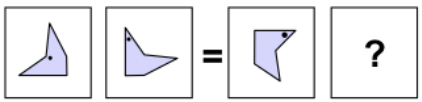










































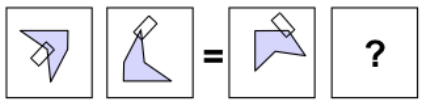










































<p>FAT01</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td style="padding: 5px;">Keine der Antworten ist richtig</td> </tr> </table>	a	b	c	d	e						f	g	h	i	j						k	l	m	n	o					Keine der Antworten ist richtig
a	b	c	d	e																											
																															
f	g	h	i	j																											
																															
k	l	m	n	o																											
				Keine der Antworten ist richtig																											
<p>FAT02</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td style="padding: 5px;">Keine der Antworten ist richtig</td> </tr> </table>	a	b	c	d	e						f	g	h	i	j						k	l	m	n	o					Keine der Antworten ist richtig
a	b	c	d	e																											
																															
f	g	h	i	j																											
																															
k	l	m	n	o																											
				Keine der Antworten ist richtig																											
<p>FAT03</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> <tr> <td></td> <td></td> <td></td> <td></td> <td style="padding: 5px;">Keine der Antworten ist richtig</td> </tr> </table>	a	b	c	d	e						f	g	h	i	j						k	l	m	n	o					Keine der Antworten ist richtig
a	b	c	d	e																											
																															
f	g	h	i	j																											
																															
k	l	m	n	o																											
				Keine der Antworten ist richtig																											

Figure A.1.
Figural Analogy Test: Items 1-3

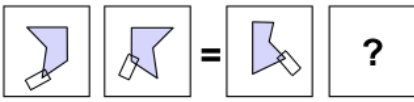










































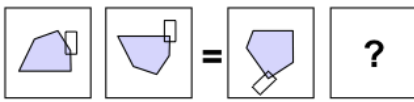










































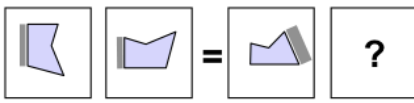










































<p>FAT04</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="width: 20px;">a</td><td></td> <td style="width: 20px;">b</td><td></td> <td style="width: 20px;">c</td><td></td> <td style="width: 20px;">d</td><td></td> <td style="width: 20px;">e</td><td></td> </tr> <tr> <td>f</td><td></td> <td>g</td><td></td> <td>h</td><td></td> <td>i</td><td></td> <td>j</td><td></td> </tr> <tr> <td>k</td><td></td> <td>l</td><td></td> <td>m</td><td></td> <td>n</td><td></td> <td>o</td><td style="text-align: left; padding-left: 5px;">Keine der Antworten ist richtig</td> </tr> </table>	a		b		c		d		e		f		g		h		i		j		k		l		m		n		o	Keine der Antworten ist richtig
a		b		c		d		e																							
f		g		h		i		j																							
k		l		m		n		o	Keine der Antworten ist richtig																						
<p>FAT05</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="width: 20px;">a</td><td></td> <td style="width: 20px;">b</td><td></td> <td style="width: 20px;">c</td><td></td> <td style="width: 20px;">d</td><td></td> <td style="width: 20px;">e</td><td></td> </tr> <tr> <td>f</td><td></td> <td>g</td><td></td> <td>h</td><td></td> <td>i</td><td></td> <td>j</td><td></td> </tr> <tr> <td>k</td><td></td> <td>l</td><td></td> <td>m</td><td></td> <td>n</td><td></td> <td>o</td><td style="text-align: left; padding-left: 5px;">Keine der Antworten ist richtig</td> </tr> </table>	a		b		c		d		e		f		g		h		i		j		k		l		m		n		o	Keine der Antworten ist richtig
a		b		c		d		e																							
f		g		h		i		j																							
k		l		m		n		o	Keine der Antworten ist richtig																						
<p>FAT06</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="width: 20px;">a</td><td></td> <td style="width: 20px;">b</td><td></td> <td style="width: 20px;">c</td><td></td> <td style="width: 20px;">d</td><td></td> <td style="width: 20px;">e</td><td></td> </tr> <tr> <td>f</td><td></td> <td>g</td><td></td> <td>h</td><td></td> <td>i</td><td></td> <td>j</td><td></td> </tr> <tr> <td>k</td><td></td> <td>l</td><td></td> <td>m</td><td></td> <td>n</td><td></td> <td>o</td><td style="text-align: left; padding-left: 5px;">Keine der Antworten ist richtig</td> </tr> </table>	a		b		c		d		e		f		g		h		i		j		k		l		m		n		o	Keine der Antworten ist richtig
a		b		c		d		e																							
f		g		h		i		j																							
k		l		m		n		o	Keine der Antworten ist richtig																						

Figure A.2.
Figural Analogy Test: Items 4-6

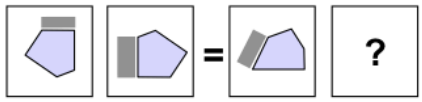
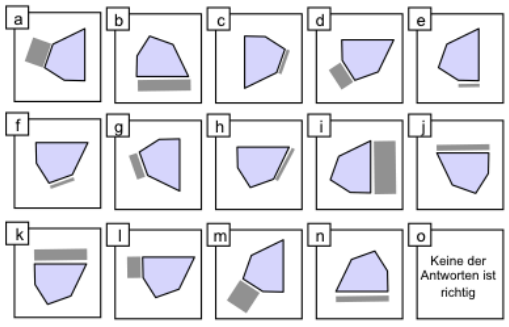
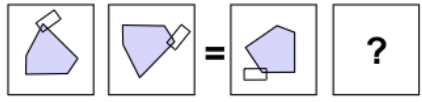
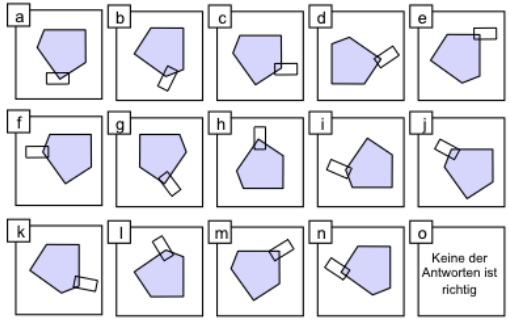
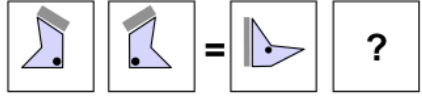
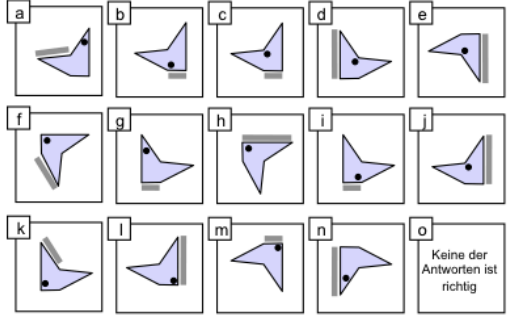
<p>FAT07</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> 
<p>FAT08</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> 
<p>FAT09</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> 

Figure A.3.
Figural Analogy Test: Items 7-9

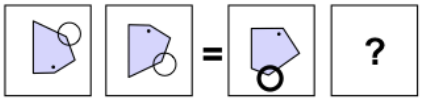
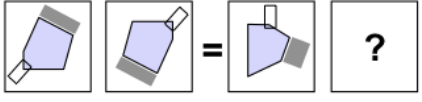
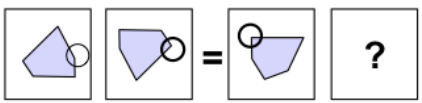
<p>FAT10</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT11</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT12</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												

Figure A.4.
Figural Analogy Test: Items 10-12

FAT13

=

?

Welche Antwortalternative ist die richtige Lösung?

a		b		c		d		e	
f		g		h		i		j	
k		l		m		n		o	Keine der Antworten ist richtig

FAT14

=

?

Welche Antwortalternative ist die richtige Lösung?

a		b		c		d		e	
f		g		h		i		j	
k		l		m		n		o	Keine der Antworten ist richtig

FAT15

=

?

Welche Antwortalternative ist die richtige Lösung?

a		b		c		d		e	
f		g		h		i		j	
k		l		m		n		o	Keine der Antworten ist richtig

Figure A.5.
Figural Analogy Test: Items 13-15

FAT16

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
				Keine der Antworten ist richtig

FAT17

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
				Keine der Antworten ist richtig

FAT18

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
				Keine der Antworten ist richtig

Figure A.6.
Figural Analogy Test: Items 16-18

<p>FAT19</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> 
<p>FAT20</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> 
<p>FAT21</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> 

Figure A.7.
Figural Analogy Test: Items 19-21

FAT22

=

?

Welche Antwortalternative ist die richtige Lösung?

a		b		c		d		e	
f		g		h		i		j	
k		l		m		n		o	Keine der Antworten ist richtig

FAT23

=

?

Welche Antwortalternative ist die richtige Lösung?

a		b		c		d		e	
f		g		h		i		j	
k		l		m		n		o	Keine der Antworten ist richtig

FAT24

=

?

Welche Antwortalternative ist die richtige Lösung?

a		b		c		d		e	
f		g		h		i		j	
k		l		m		n		o	Keine der Antworten ist richtig

Figure A.8.
Figural Analogy Test: Items 22-24

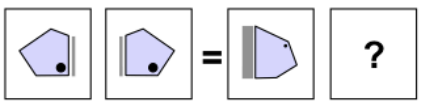
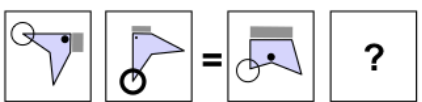
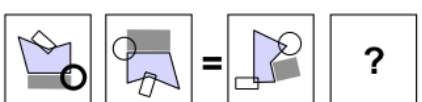
<p>FAT25</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT26</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT27</p> 	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												

Figure A.9.
Figural Analogy Test: Items 25-27

<p>FAT28</p>	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td> </tr> <tr> <td>f</td><td>g</td><td>h</td><td>i</td><td>j</td> </tr> <tr> <td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> </table> <p style="text-align: right; font-size: small;">Keine der Antworten ist richtig</p>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT29</p>	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td> </tr> <tr> <td>f</td><td>g</td><td>h</td><td>i</td><td>j</td> </tr> <tr> <td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> </table> <p style="text-align: right; font-size: small;">Keine der Antworten ist richtig</p>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT30</p>	<p>Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td>a</td><td>b</td><td>c</td><td>d</td><td>e</td> </tr> <tr> <td>f</td><td>g</td><td>h</td><td>i</td><td>j</td> </tr> <tr> <td>k</td><td>l</td><td>m</td><td>n</td><td>o</td> </tr> </table> <p style="text-align: right; font-size: small;">Keine der Antworten ist richtig</p>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												

Figure A.10.
Figural Analogy Test: Items 28-30

<p>FAT31</p> <div style="text-align: center; margin-top: 20px;"> </div>	<p style="text-align: center;">Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT32</p> <div style="text-align: center; margin-top: 20px;"> </div>	<p style="text-align: center;">Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												
<p>FAT33</p> <div style="text-align: center; margin-top: 20px;"> </div>	<p style="text-align: center;">Welche Antwortalternative ist die richtige Lösung?</p> <table border="1" style="width: 100%; text-align: center; border-collapse: collapse;"> <tr> <td style="padding: 5px;">a</td><td style="padding: 5px;">b</td><td style="padding: 5px;">c</td><td style="padding: 5px;">d</td><td style="padding: 5px;">e</td> </tr> <tr> <td style="padding: 5px;">f</td><td style="padding: 5px;">g</td><td style="padding: 5px;">h</td><td style="padding: 5px;">i</td><td style="padding: 5px;">j</td> </tr> <tr> <td style="padding: 5px;">k</td><td style="padding: 5px;">l</td><td style="padding: 5px;">m</td><td style="padding: 5px;">n</td><td style="padding: 5px;">o</td> </tr> </table>	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
a	b	c	d	e												
f	g	h	i	j												
k	l	m	n	o												

Figure A.11.
Figural Analogy Test: Items 31-33

FAT34

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o

FAT35

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o

FAT36

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o

Figure A.12.
Figural Analogy Test: Items 34-36

FAT37

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
				Keine der Antworten ist richtig

FAT38

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
				Keine der Antworten ist richtig

FAT39

Welche Antwortalternative ist die richtige Lösung?

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
				Keine der Antworten ist richtig

Figure A.13.
Figural Analogy Test: Items 37-39

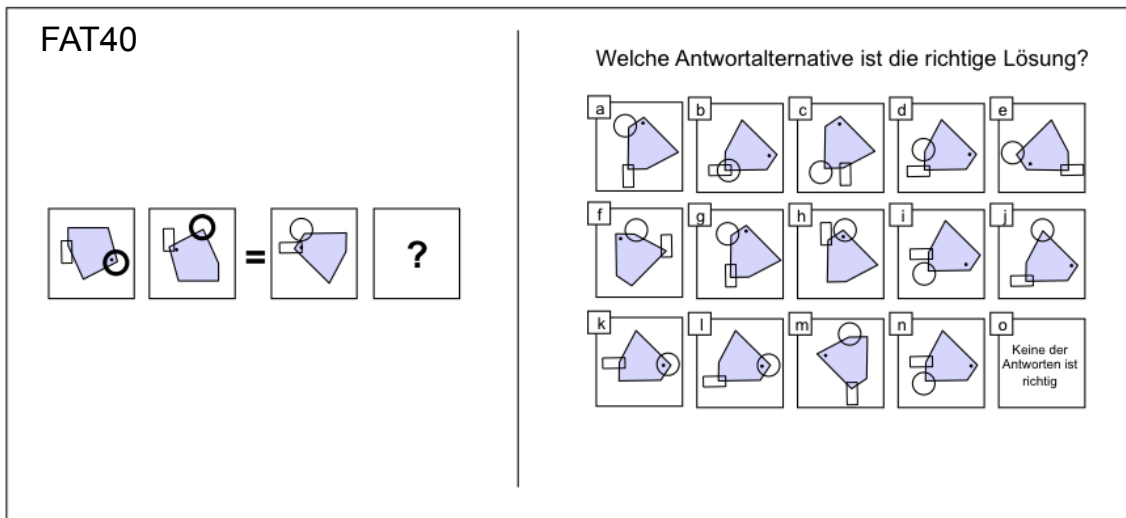


Figure A.14.
Figural Analogy Test: Item 40

Figure A.15.
Optimal design SAS input file (Syntax)

```

proc factex;
factors formtyp xachse yachse pm90 pm180 zweifeat dreifeat pm1 pm2 randomvar /nlev=2;
output out=can
formtyp nvals=(0 1)
xachse nvals=(0 1)
yachse nvals=(0 1)
pm90 nvals=(0 1)
pm180 nvals=(0 1)
zweifeat nvals=(0 1)
dreifeat nvals=(0 1)
pm1 nvals=(0 1)
pm2 nvals=(0 1)
randomvar nvals=(0 1);
data can;
set can;
if xachse + yachse + pm90 + pm180 <2;
if zweifeat + dreifeat < 2;
if zweifeat + dreifeat=1 then if pm1+pm2 < 3;
if zweifeat + dreifeat=0 then if pm1+pm2 < 2;
if xachse + yachse + pm90 + pm180 =0 then if pm1+pm2+randomvar>0;
if formtyp + xachse + yachse + pm90 + pm180 + zweifeat + dreifeat + pm1 + pm2 + randomvar >0;
proc optex data=can seed=57922;
model formtyp xachse yachse pm90 pm180 zweifeat dreifeat pm1 pm2 randomvar/noint;
generate n=40 method=fedorov;
output out=des;
proc print data=des;
run;

```

Table A.1.
Design matrix for the 40-item FAT investigated in this study

Item	Spatial Displacement Rules						Compl. Param.		
	mx	my	r90	r180	cp1	cp2	tof	fp1	rcf
01	0	1	0	0	0	0	0	1	1
02	0	0	1	0	1	0	0	0	0
03	0	0	0	1	1	0	0	0	0
04	0	1	0	0	1	0	0	0	1
05	1	0	0	0	1	0	1	0	0
06	0	0	1	0	1	0	0	0	0
07	0	0	0	1	1	0	1	0	1
08	1	0	0	0	1	0	1	0	1
09	0	1	0	0	0	0	0	1	0
10	1	0	0	0	0	0	1	1	0
11	0	0	0	1	0	0	1	1	0
12	0	0	0	1	0	1	1	0	1
13	1	0	0	0	0	1	0	0	0
14	0	0	0	1	0	1	0	0	1
15	0	1	0	0	0	1	1	0	1
16	0	0	1	0	0	1	0	0	0
17	0	0	0	0	1	1	0	1	0
18	1	0	0	0	0	0	0	0	0
19	0	1	0	0	1	0	1	0	1
20	0	1	0	0	0	0	1	0	0
21	1	0	0	0	0	0	1	1	1
22	0	0	1	0	0	0	1	1	1
23	0	1	0	0	0	1	1	0	0
24	0	0	1	0	0	1	1	0	1
25	0	1	0	0	1	0	1	1	0
26	0	0	1	0	0	0	0	0	1
27	0	0	0	1	0	0	0	0	1
28	0	0	0	1	0	0	1	0	0
29	1	0	0	0	0	1	0	1	1
30	0	0	1	0	1	0	0	0	1
31	0	0	0	1	1	1	0	1	0
32	0	0	1	0	1	1	1	1	1
33	0	1	0	0	1	1	0	1	1
34	0	0	1	0	1	1	1	1	0
35	0	1	0	0	0	1	0	0	0
36	0	0	1	0	0	1	1	0	0
37	0	0	0	0	1	1	1	0	1
38	1	0	0	0	1	1	1	0	0
39	1	0	0	0	1	1	0	0	1
40	0	0	0	1	1	1	1	0	0

Note. mx= mirroring at horizontal axis; my= mirroring at vertical axis; r90 = rotation by 90 degrees clockwise or counter-clockwise; r180 = rotation by 180 degrees; cp1 = change in feature-position by one corner/edge; cp2= change in feature-position by two corners/edges; tof= type of form (0=concave, 1=convex); fp1=additional feature; rcf = random change in feature surface characteristics

Figure A.16.
Optimal design SAS output file

```

Das SAS System      16:37 Saturday, May 17, 2008   1

The OPTEX Procedure

Factor Ranges

Faktor      Unterer Wert   Oberer Wert

formtyp      0             1.000000
xachse       0             1.000000
yachse       0             1.000000
pm90         0             1.000000
pm180        0             1.000000
zweifeat     0             1.000000
dreifeat     0             1.000000
pml          0             1.000000
pm2          0             1.000000
randomvar    0             1.000000

Das SAS System      16:37 Saturday, May 17, 2008   2

The OPTEX Procedure

                Durchschnitt
Plannummer    D-Effizienz   A-Effizienz   G-Effizienz   Vorhersagestandardfehler
-----
1             95.2284      89.1046       95.8097      0.4989
2             95.1363      89.5216       95.4382      0.4986
3             95.0688      88.8462       94.2670      0.4989
4             95.0002      88.7365       95.6217      0.4985
5             94.9944      88.7199       94.8255      0.4989
6             94.9742      88.6171       94.8300      0.4985
7             94.9732      88.4983       93.6232      0.4987
8             94.9502      88.7332       95.2874      0.4971
9             94.8771      88.3452       93.2435      0.4986
10            94.8706      88.4267       93.7134      0.4989

Das SAS System      16:37 Saturday, May 17, 2008   3

Beob.   formtyp   xachse   yachse   pm90   pm180   zweifeat   dreifeat   pml   pm2   randomvar

1       0         0         0         0         0         1         0         0         0         1
2       0         0         0         0         0         1         0         1         1         0
3       0         0         0         0         1         0         0         0         1         1
4       0         0         0         0         1         0         0         1         0         0
5       0         0         0         0         1         0         1         0         0         1
6       0         0         0         0         1         1         0         1         1         0
7       0         0         0         1         0         0         0         0         1         0
8       0         0         0         1         0         0         0         1         0         0
9       0         0         0         1         0         0         0         1         0         0
10      0         0         0         1         0         0         1         0         0         1
11      0         0         0         1         0         0         1         1         0         1
12      0         0         1         0         0         0         0         1         0         1
13      0         0         1         0         0         0         1         0         1         0
14      0         0         1         0         0         1         0         0         0         0
15      0         0         1         0         0         1         0         1         1         1
16      0         1         0         0         0         0         0         0         1         0
17      0         1         0         0         0         0         1         0         0         0
18      0         1         0         0         0         0         1         1         1         1
19      0         1         0         0         0         1         0         0         1         1
20      1         0         0         0         0         0         1         1         1         1
21      1         0         0         0         1         0         0         0         1         1
22      1         0         0         0         1         0         0         1         0         1
23      1         0         0         0         1         0         1         0         0         0
24      1         0         0         0         1         0         1         1         1         0
25      1         0         0         0         1         1         0         0         0         0
26      1         0         0         1         0         0         0         0         1         1
27      1         0         0         1         0         0         1         0         1         0
28      1         0         0         1         0         1         0         0         0         1
29      1         0         0         1         0         1         0         1         1         0
30      1         0         0         1         0         1         0         1         1         1
31      1         0         1         0         0         0         0         0         1         0
32      1         0         1         0         0         0         0         0         1         1
33      1         0         1         0         0         0         1         0         0         0
34      1         0         1         0         0         0         1         1         0         1
35      1         0         1         0         0         1         0         1         0         0
36      1         1         0         0         0         0         0         1         0         0
37      1         1         0         0         0         0         0         1         0         1
38      1         1         0         0         0         0         1         1         1         0
39      1         1         0         0         0         1         0         0         0         0
40      1         1         0         0         0         1         0         0         0         1

```

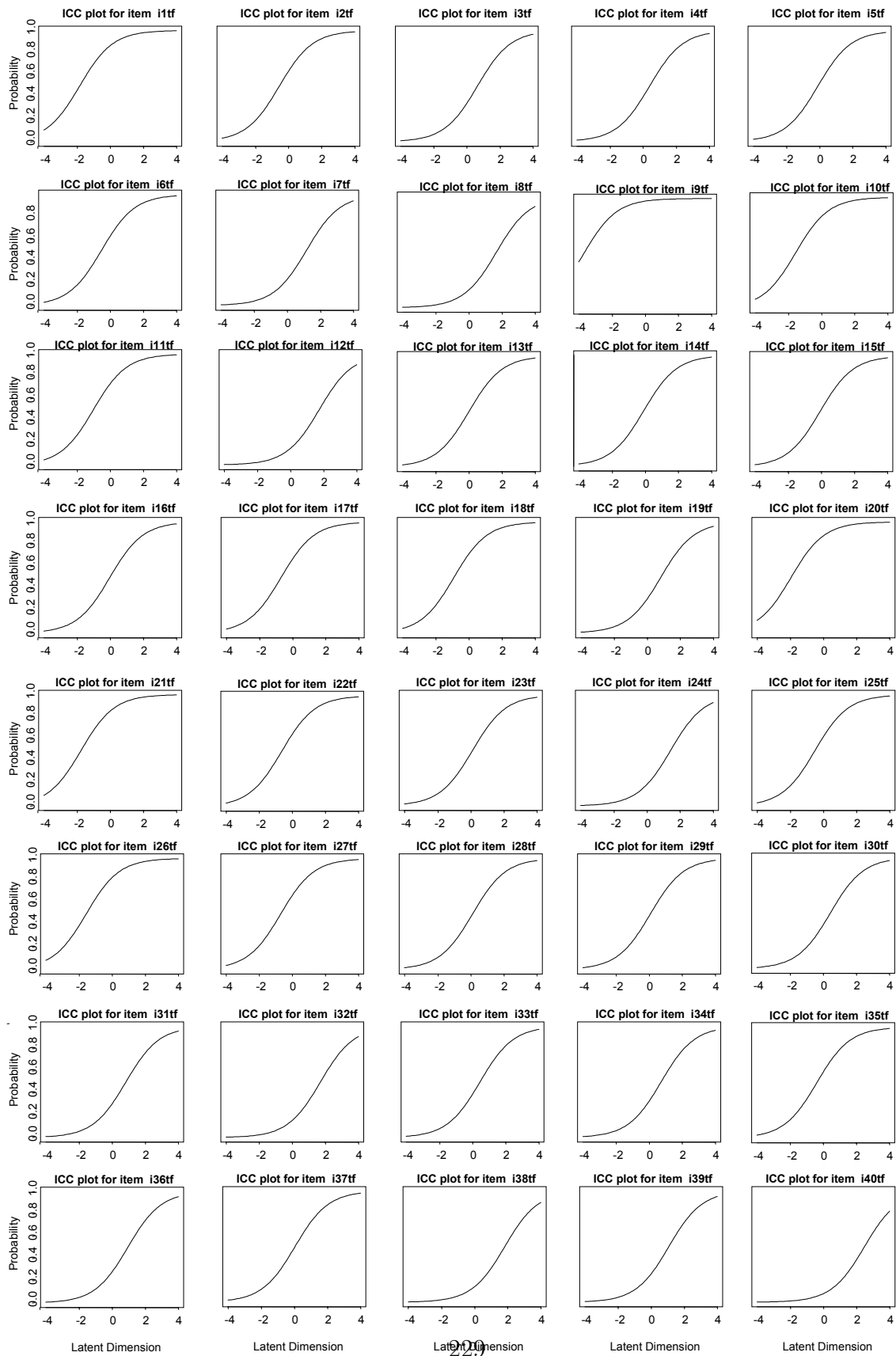


Figure A.17.

Item characteristic curves for all 40 FAT items (Study 1)

B

Study 2 – Additional Materials

Additional analyses for NST items with distinct surface patterns

Answer patterns on items labelled as “wrong track”-items (i.e., items with distinct, potentially distracting, surface patterns) were compared with answers on items that did not offer any particular wrong solution strategies. Figure B.2 summarizes these items. Test performance of test-takers following one or more of the “wrong tracks” on all other (non-WT) items was analyzed. If the wrong-track character of an item led able individuals on a wrong track, actually distracting them from answering an item correctly (that they would have answered correctly if it were not WT item), choosing WT solutions should be positively related to correct performance on all other items. In this case, test-takers getting in principle many items right would be more likely to get WT items wrong. That is, discrimination parameters of these items should be negative or very low. Such items would be not helpful for any productive measurement. The opposite case would be that choosing WT solutions would be negatively correlated with correct performance on all other items. In this case, test-takers getting in principle many items right would be likely to get WT items right as well. That is, discrimination parameters of these items should be at as high as for all other items. The latter should be the case if the test is robust against surface characteristics. Only if this is the case could WT items be used for productive measurement and would not need to be dropped from the test.

For each participant, two scores were computed. First, a count variable was computed that describes how many of the attractive wrong solutions were chosen. This score is an indicator of the distractibility of a test-takers by the distinctive patterns in a subset of

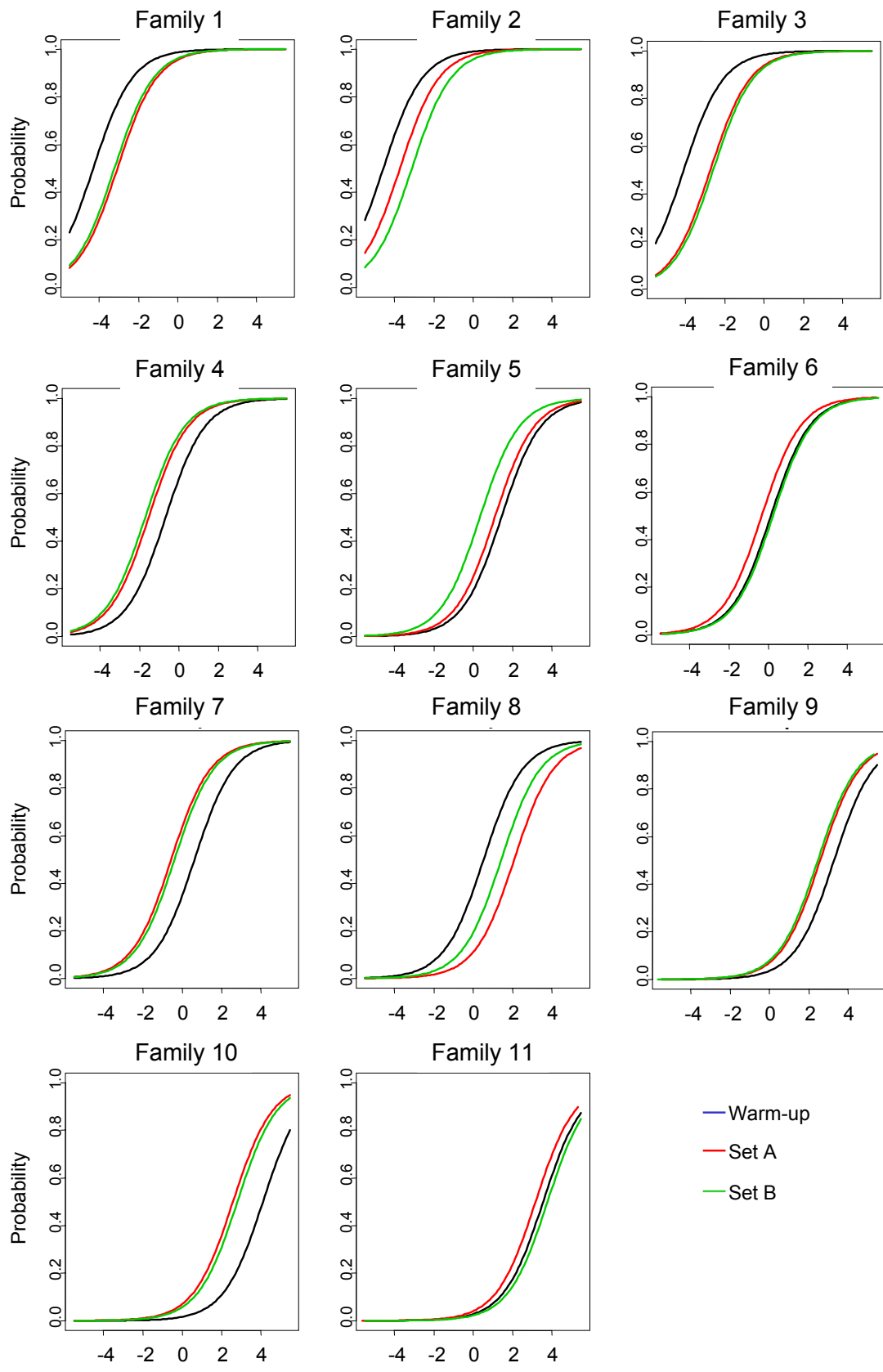


Figure B.1.
Item characteristic curves for all 33 NST items (Study 2)

Correct series	CS2	13	10	5	6	11	8
Wrong rule	Fib			5	6	11	17
Correct series	CS2	9	16	16	14	12	8
Wrong rule	-2			16	14	12	10
Correct series	CS1(Fib)	3	7	1	8	9	8
Wrong rule	Fib		7	1	8	9	17
Correct series	CS1(Fib)	6	4	1	5	6	2
Wrong rule	Fib		4	1	5	6	11
Correct series	Fib – CS1	5	12	14	21	32	48
Wrong rule	+7, Hierarchy	5	12	14	21	32	39

Figure B.2.
 “Wrong-track” items (Study 2)

the elements of a series. Scores could be between zero (none of the possible attractive wrong solutions) to 5 (all of the possible attractive wrong solutions were chosen). Second, the number of correct answers on all WT items was used as an indicator of the ability to induce and apply the correct solution principles from a given number series, regardless of distracting surface structures in an item. Scores could be between zero (all WT items were answered incorrectly) to 5 (all WT items were answered correctly). The difference between these two scores captures wrong responses on the five WT items that did not match the expected wrong responses when a test-taker followed the suggested incorrect rule(s). Correlations of these scores with total test scores, the score on all non-WT items, general cognitive ability, and scholastic performance are given in Table B.1. Individuals with a tendency to choose such a “wrong track” solution had also lower solution probabilities on all other items that did not offer such a “wrong track”. No significant correlations with CFT scores or math grades were found. All WT items had sufficiently high item-total correlations (average IT-correlations: $\bar{r}_{it} = .315$ for WT items compared to $\bar{r}_{it} = .372$ for non-WT-items). Based on these analyses, no obvious differences between WT and not-WT items were identified.

Table B.1.

Correlations of responses on WT items with responses of non-WT items, general cognitive ability, and scholastic performance (Study 2)

	WT	WT correct	non-WT correct	CFT20-R	Maths
WT	1	-.476**	-.203**	-.114	.089
WT correct		1	.628**	.351**	-.160**
non-WT correct			1	.483**	-.260**
CFT20-R				1	-.283**
Mathematics					1

Note. * : $p < .05$, ** : $p < .01$. WT: number of attractive wrong solutions chosen; WT correct: number of WT items that were solved correctly; non-WT correct: number of correct answers on all non-WT items

Table B.2.

Results for separate LLTM models for Russian and German test takers (Study 2)

Fixed Effects	GER		RUS	
	Est	SE	Est	SE
Intercept	4.69	0.21	5.70	0.28
Const	3.04	0.22	2.64	0.20
CS1	-1.88	0.25	-1.67	0.25
CS2	-3.10	0.16	-2.79	0.21
Fib	-1.48	0.13	-1.48	0.15
Add	-4.74	0.24	-4.27	0.24
Sub	-5.44	0.27	-5.20	0.28
Comp	-1.71	0.27	-2.73	0.28

Note. facet parameters are logits; smaller values indicate larger difficulties for the respective facets; all parameters are statistically significant with $p < .001$

Results for separate analyses of data from Russian and German students

In order to assure that facet-level results reported for the NST are not distorted by potential cross-cultural bias, LLTM models were re-run for the two separate subsamples as well and compared for consistency. Results are summarized in Table B.2. As the results stem from two models estimated separately the absolute values of the parameters cannot be directly compared. In order to see if there are major differences across countries, the relative magnitudes and order of parameters across both samples is investigated. Results show that there is a high consistency across samples no major differences in terms of the implications and conclusions drawn from analyses of the full sample. Close correspondence in the relative order of different facet parameters could be found. Only one parameter seems to show a larger difference, the complex combination principle of rules (“Comp”). While it is important that this difference is investigated further in future studies, for the current study, there is no indication that conclusions regarding the research questions are distorted by the heterogeneity of the sample and by the difference in the parameter estimate for this radical. However, these findings need to be cross-validated and further research should be conducted to understand why this radical might work differently across cultures. Future studies should apply more complex statistical modeling, such as the DIF and DFF models used in Study 3, to further elaborate on the cross-cultural validity of the NST.



Study 3 – Additional Materials

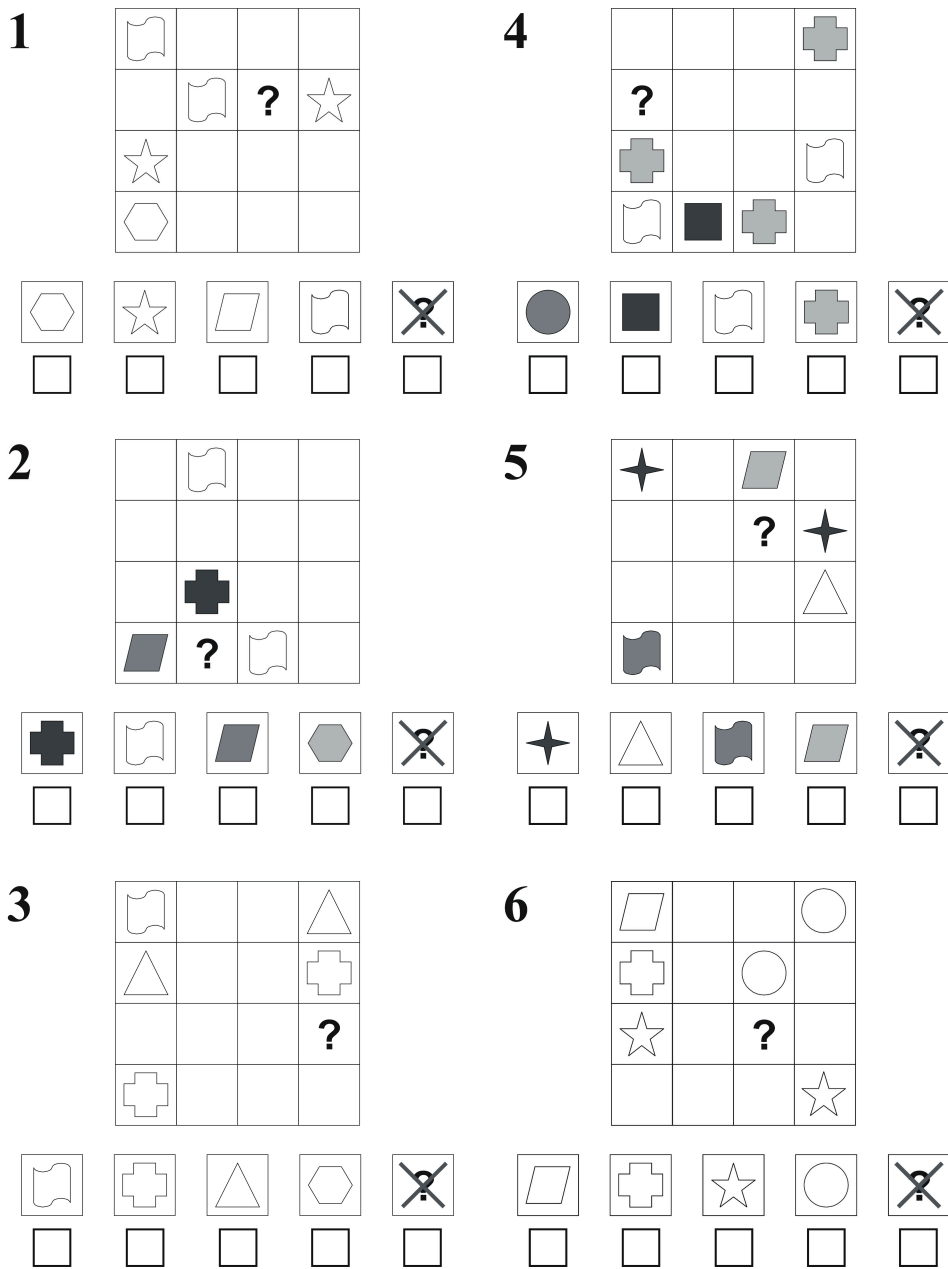


Figure C.1.
LST: Items 1-6

7

		?	

				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

8

			?

				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

9

		?	

				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

10

			?

				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

11

?			

				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

12

			?

				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure C.2.
LST: Items 7-8

13

			⬛
		●	
?			☁
▱		⬛	

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

14

			+
	+		
		?	
		▱	

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

15

▧			
		?	
	▧		
		✦	

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

16

☁			■
	☁	✦	
		?	
	☆		

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

17

☆	✦		
	☆		
?		☁	✦
	○		☆
✦			

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

18

▲		☆	■	?
☆		■	⬡	▲

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
--------------------------	--------------------------	--------------------------	--------------------------	--------------------------	--------------------------

Figure C.3.
LST: Items 13-18

19

	□		☆	
	▱			☆
		□	+	
?				☆

□	☆	+	▱	☆	✗
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

20

	+			
	☆			△
	△			
☆		+	?	
		▱		

+	△	⬡	☆	▱	✗
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

21

	?			⬡
○	▲			
			▲	+
	⬡			
	+	○		

■	○	+	▲	⬡	✗
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

22

⬡		☆		▱
	▱			
				⬡
	☆			
▱		?	+	

⬡	☆	▱	+	△	✗
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

23

	☆			▱
▱				?
			☆	○
☆		▱		
				☆

☆	▱	☆	⬡	○	✗
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

24

				⬡
			⬡	+
				●
				▲
▲	●	+	?	

☆	▲	●	+	⬡	✗
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure C.4.
LST: Items 19-24

25

				■
		✦		
	?	■	☆	
◌	✦			
				☆

✦ ☆ ◌ ■ ● ✕

26

+			○	
		◌		
	?		○	
◌			◌	
		+	▱	

◌ + ▱ ○ ◌ ✕

27

				●
▱			◌	
			★	
	▱		?	

● ▱ ■ ◌ ★ ✕

28

				✦
▨	✦		?	+
+		▨		
✦		+		
◌				

◌ ✦ ◌ ▨ + ✕

29

	✦			+
		+		
			✦	
+				
◌	?	◌		◌

◌ ✦ ◌ ◌ + ✕

30

	✦			
			◌	?
	◌	▴		
▴				
★		◌	◌	

★ ◌ ◌ ▴ ✦ ✕

Figure C.5.
LST: Items 25-30

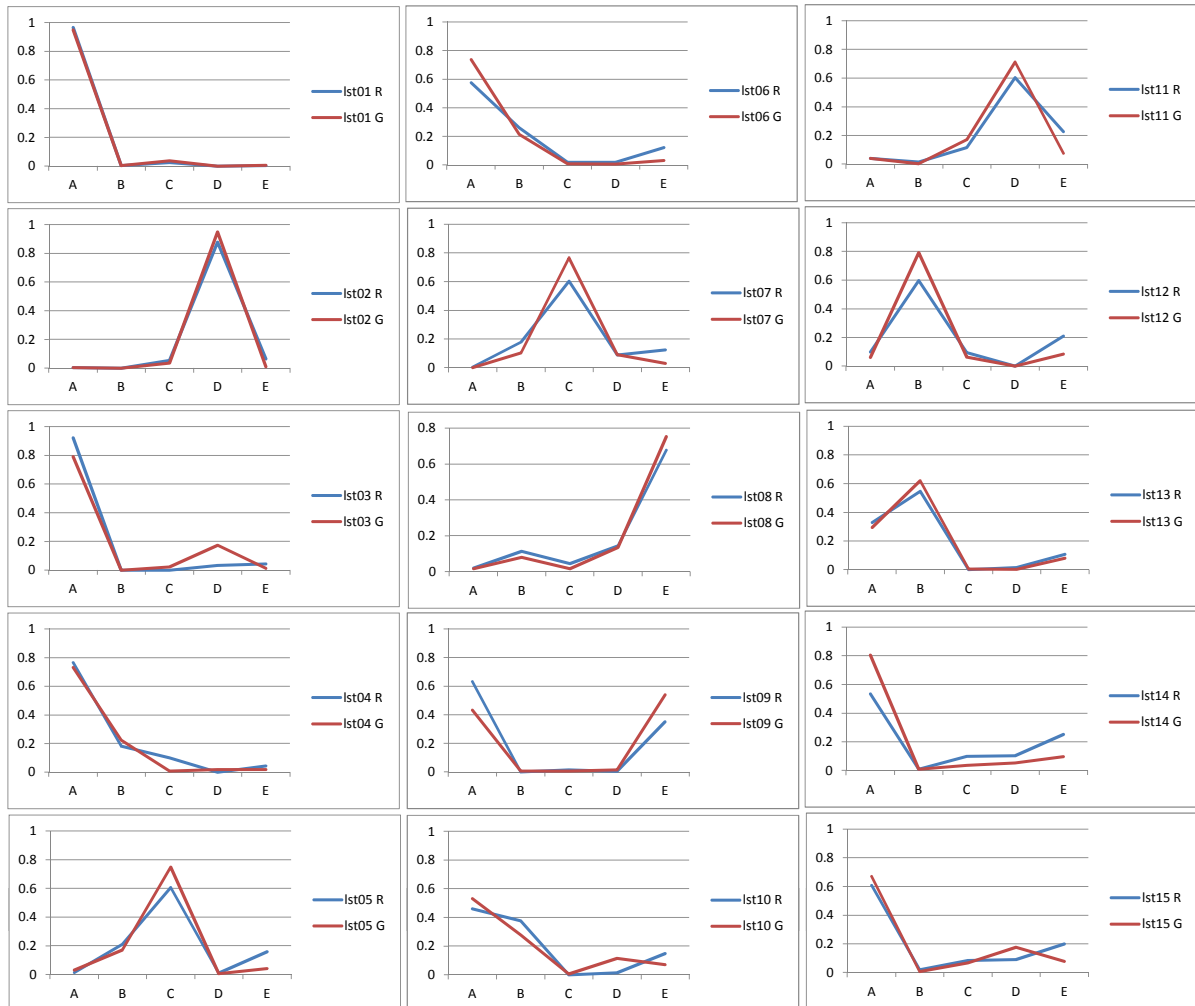


Figure C.6.
Category frequencies for all LST: Items, comparison of Russian and German samples (Study 3)

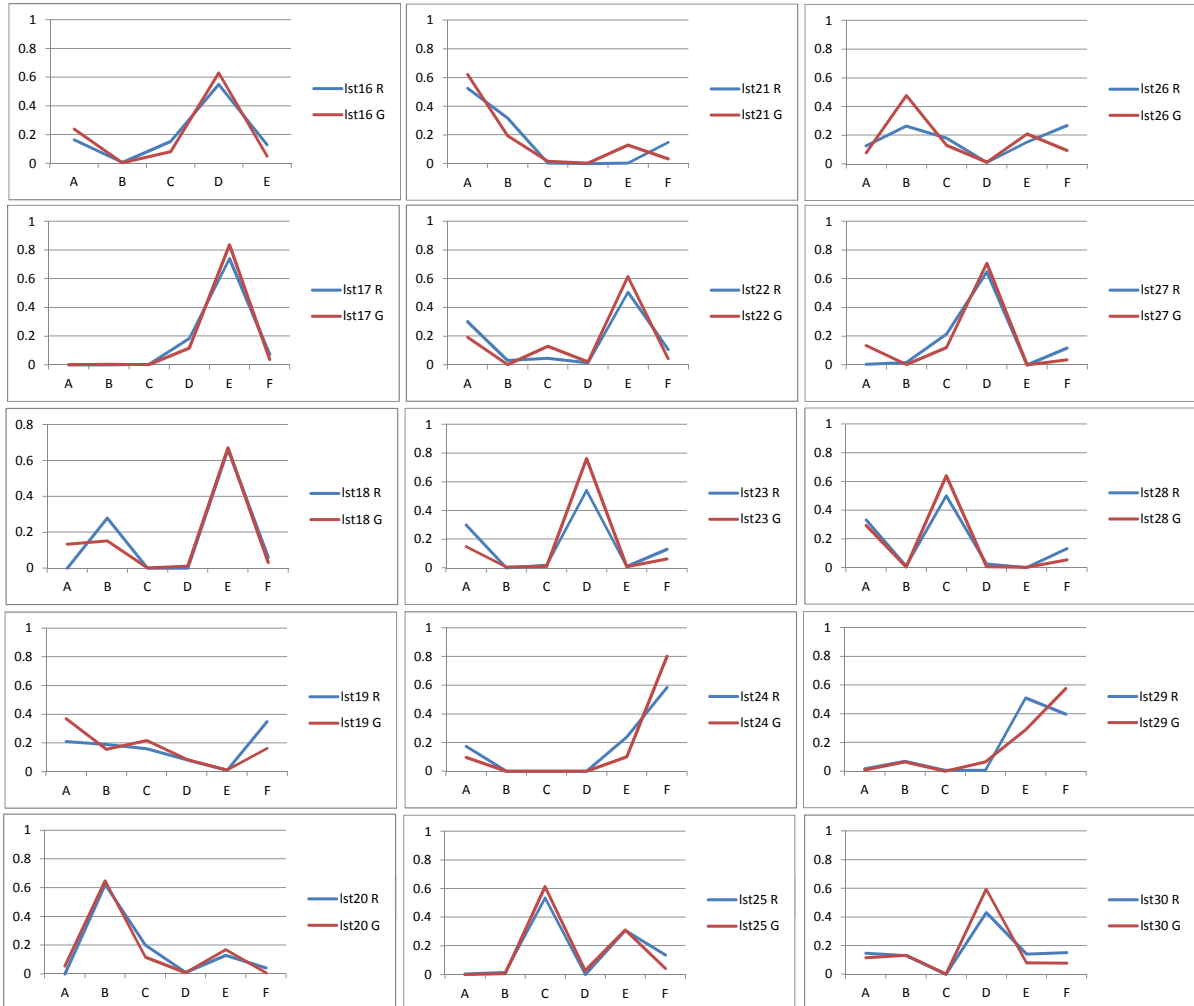


Figure C.7.
 Category frequencies for all LST: Items, comparison of Russian and German samples
 (cont'd)

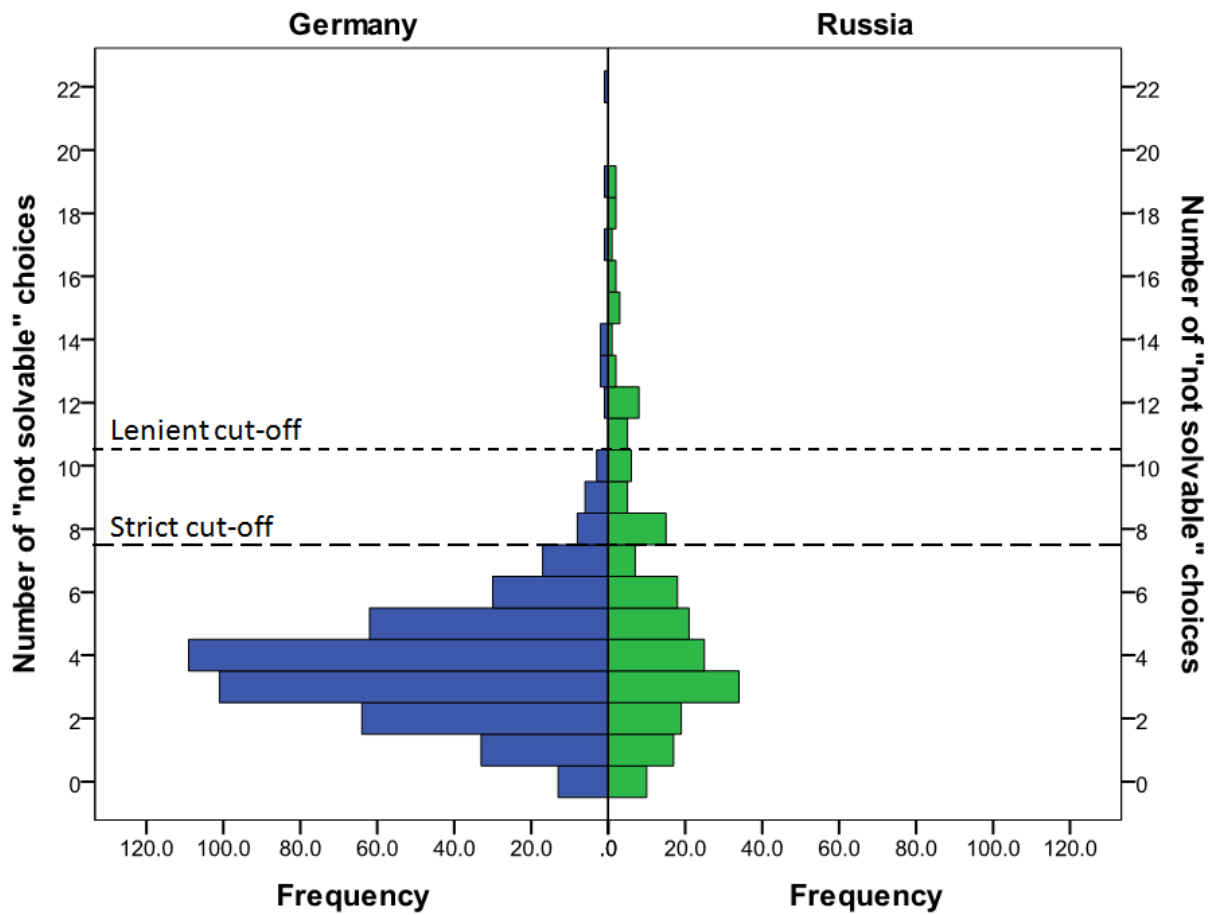


Figure C.8.
Frequencies for "not solvable" choices among Russian and German test-takers; the two dotted lines indicate the two cut-off values used when creating the subsamples.

cDIF uniform Logistic Regression

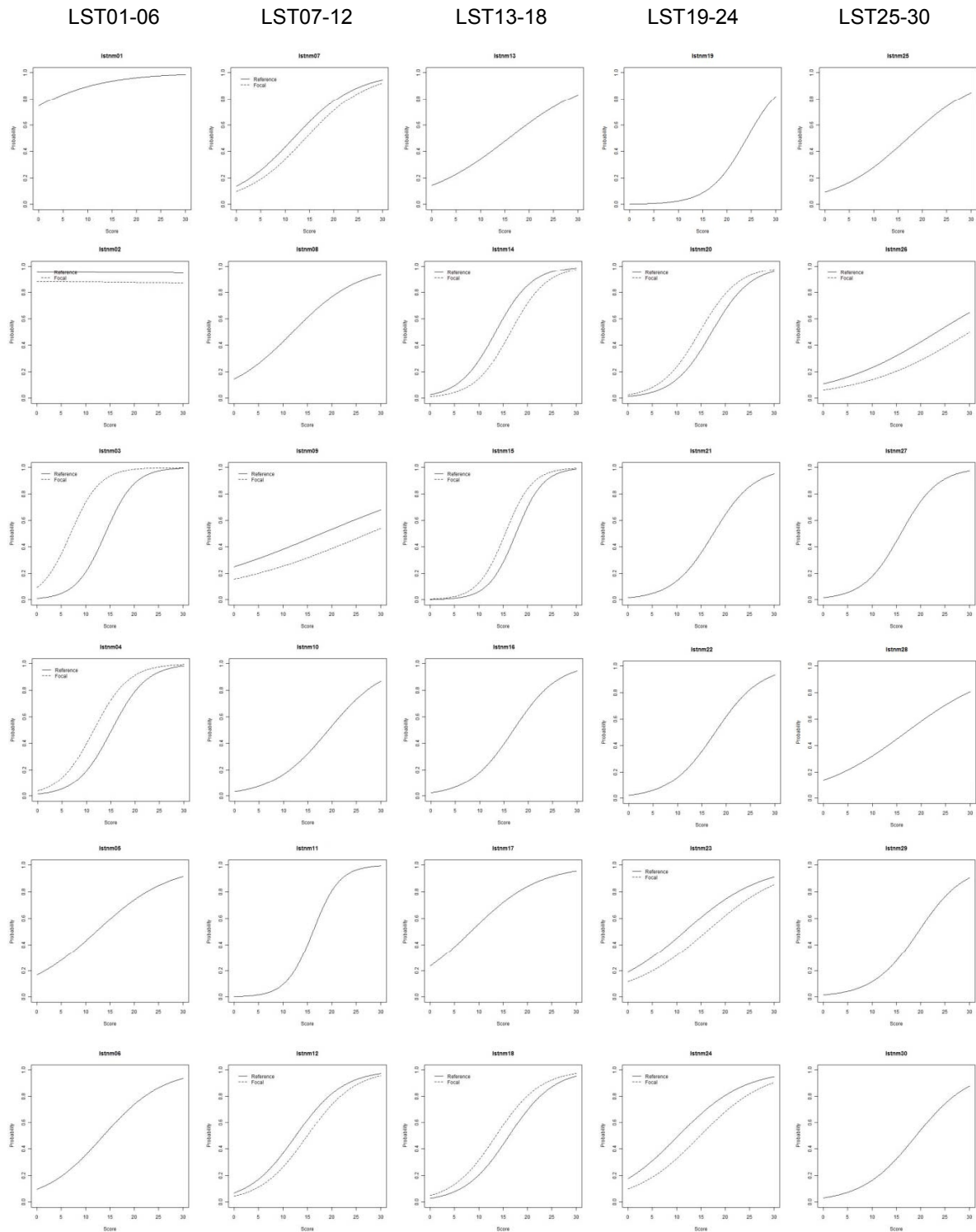


Figure C.9.
Item characteristic curves for uniform country-DIF based on the logistic regression model (Study 3)

cDIF nonuniform Logistic Regression

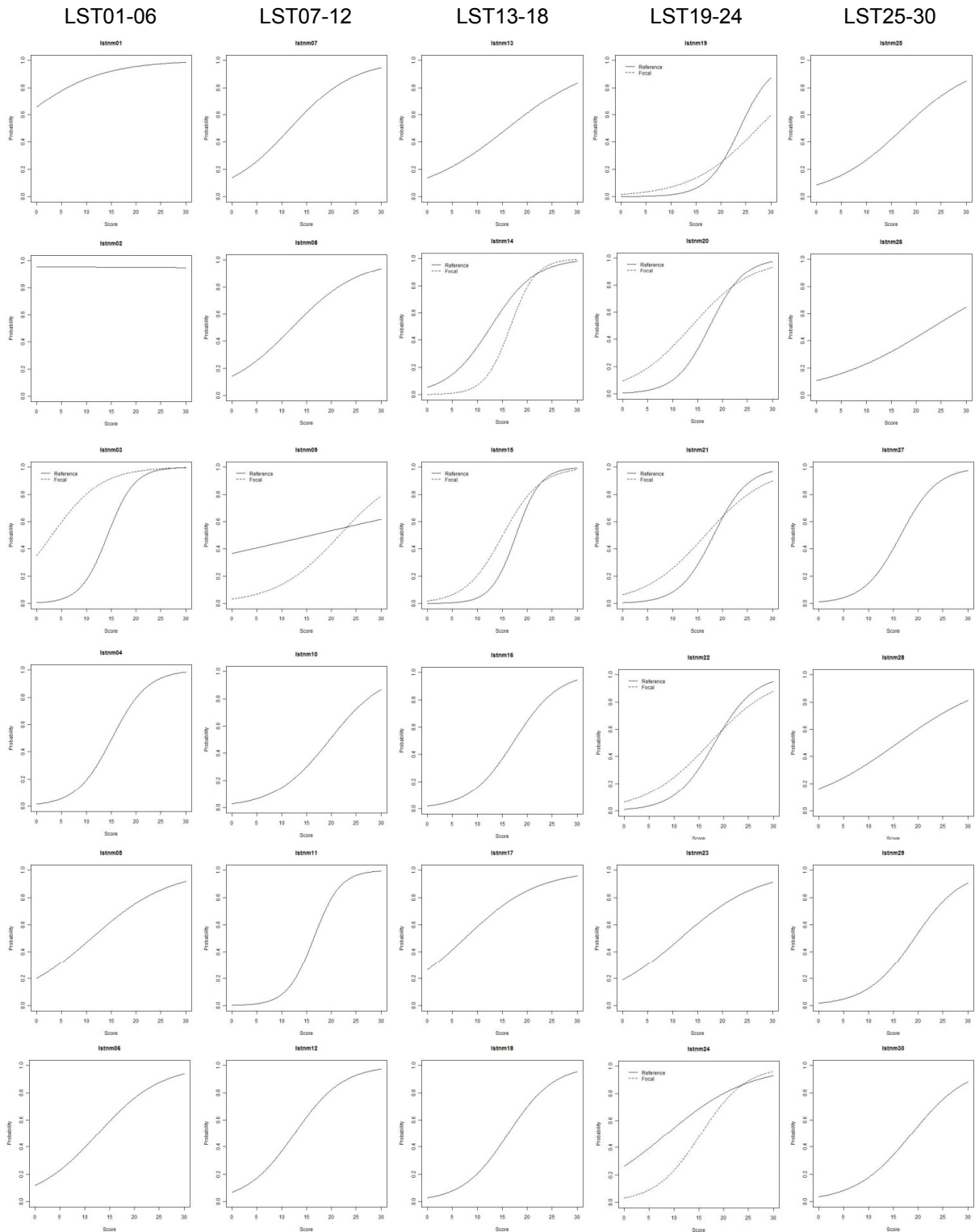


Figure C.10.
 Item characteristic curves for non-uniform country-DIF based on the logistic regression model (Study 3)

cDIF 1PL Lord

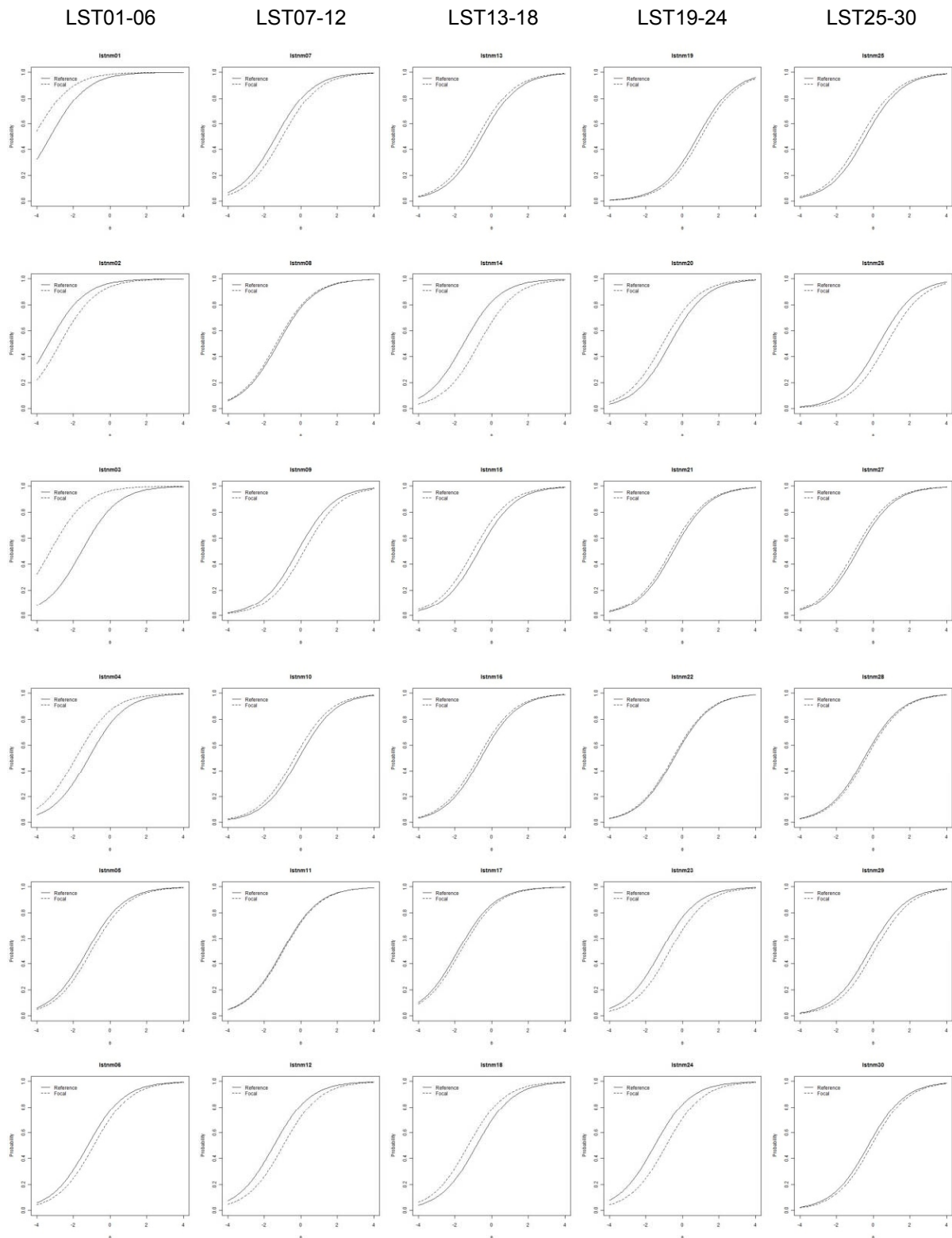


Figure C.11.
Item characteristic curves for uniform country-DIF based on Lord's approach (Study 3)

sDIF uniform Logistic Regression

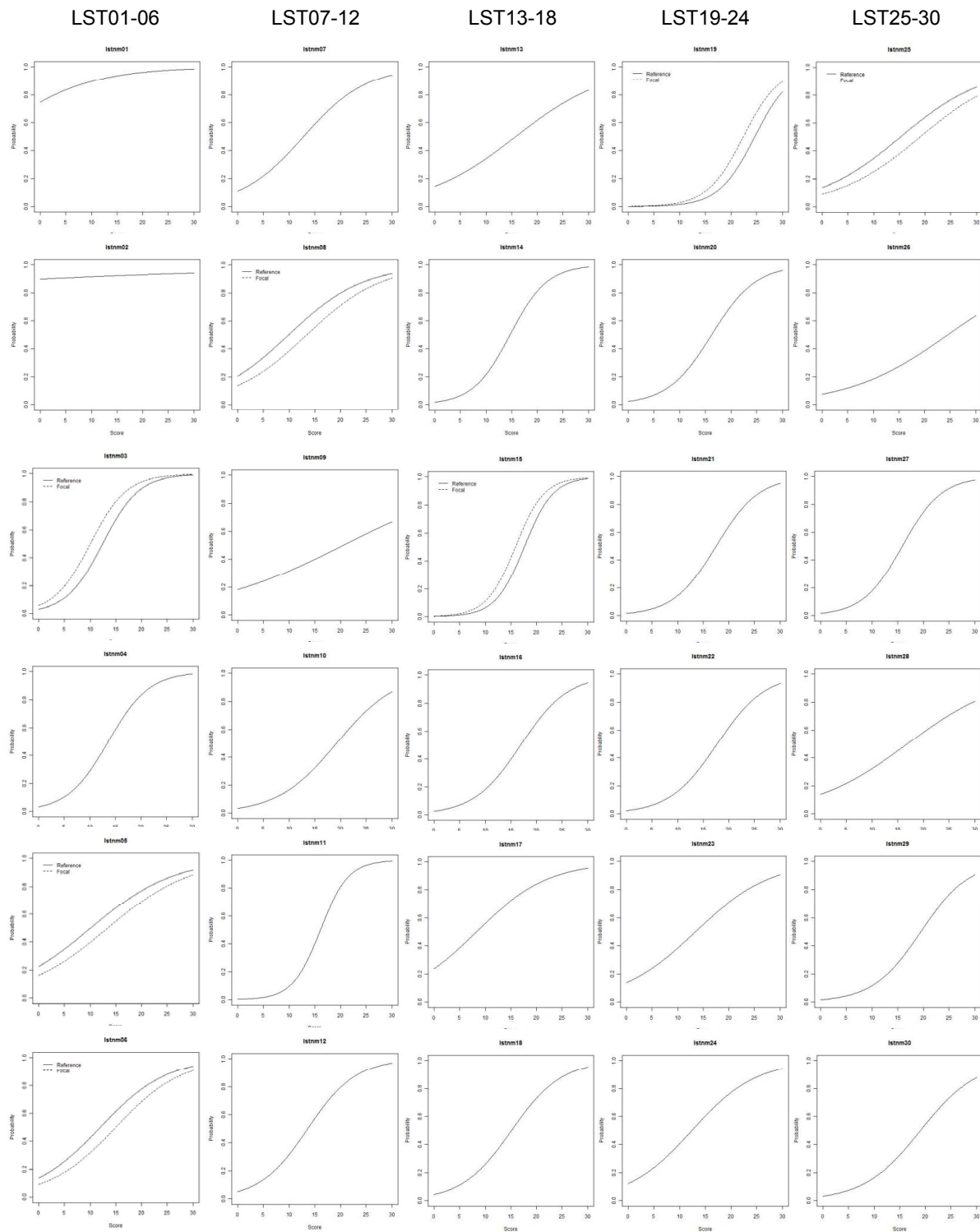


Figure C.12.

Item characteristic curves for uniform pre-knowledge-DIF based on the logistic regression model (Study 3)

sDIF nonuniform Logistic Regression

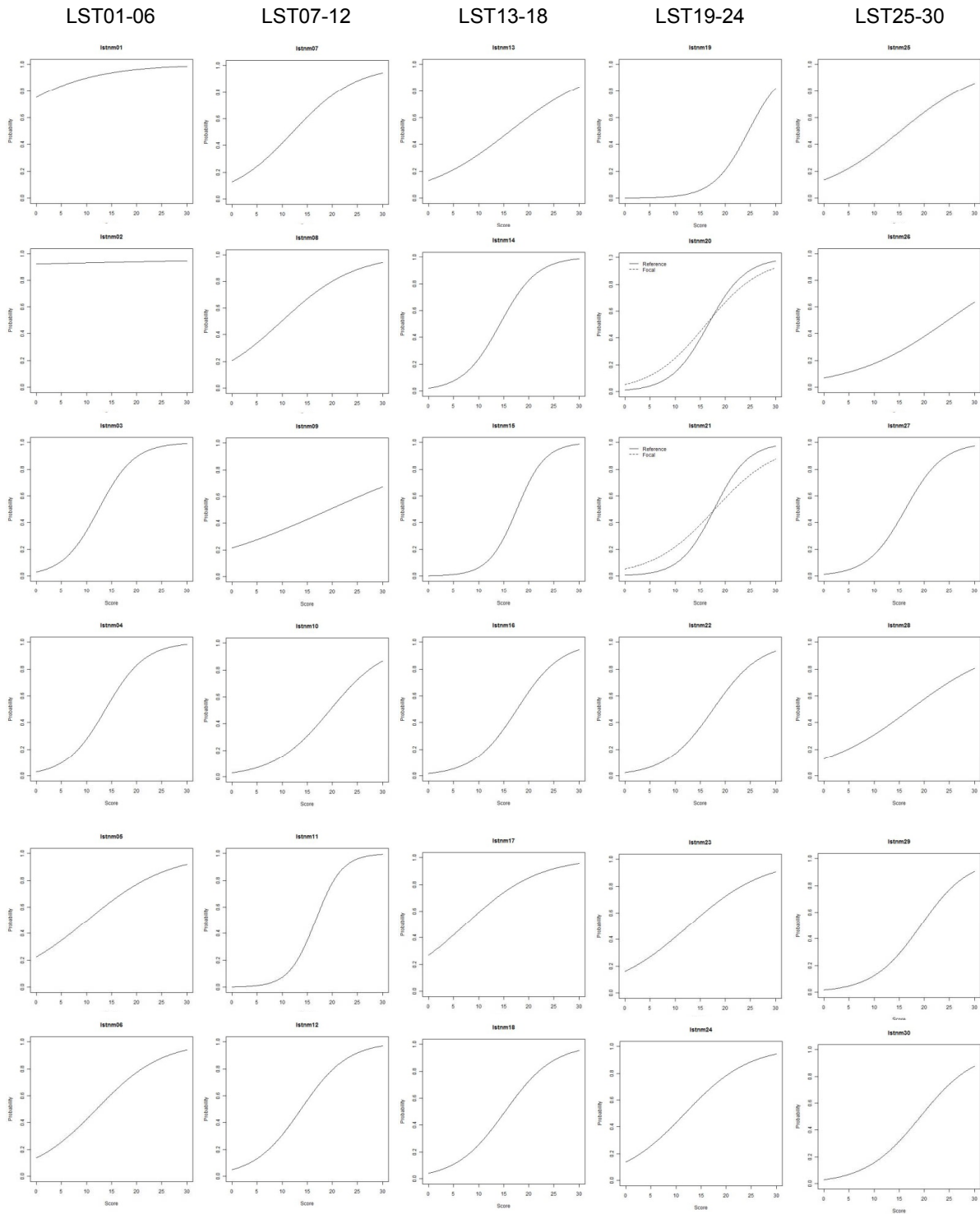


Figure C.13.
 Item characteristic curves for non-uniform pre-knowledge-DIF based on the logistic regression model (Study 3)

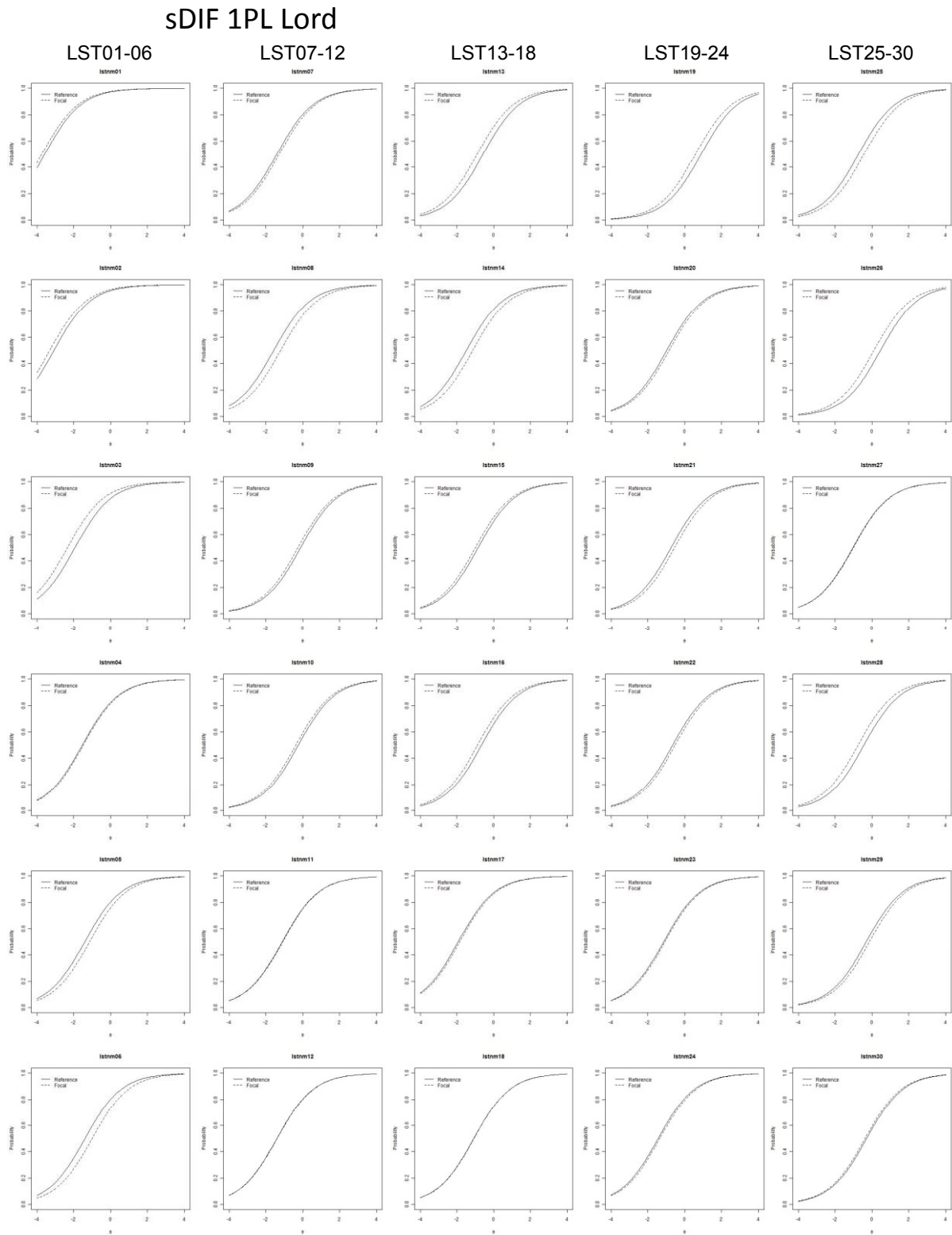


Figure C.14.
Item characteristic curves for uniform pre-knowledge-DIF based on Lord's approach (Study 3)

Table C.1.

Sum-normed RM item difficulty parameters for full Russian sample ($N = 201$) and two subsamples of all Russian participants ($N_1 = 177$; $N_2 = 151$) that chose the response category “not solvable” less than 11/ less than 8 times; Item Parameters for item 1 are excluded here because all subjects in the two subsamples solved this item correctly

Item	Total sample	Subsample a (<11 times not solvable)	Subsample b (<8 times not solvable)	Difference a	Difference b	Absolute difference a	Absolute difference b
N	201	177	151				
LST2	-1.70	-1.64	-1.63	-0.05	-0.07	0.05	0.07
LST3	-2.36	-2.59	-2.78	0.23	0.42	0.23	0.42
LST4	-0.91	-0.89	-1.01	-0.02	0.10	0.02	0.10
LST5	-0.06	-0.02	-0.11	-0.04	0.05	0.04	0.05
LST6	0.06	0.06	0.02	0.00	0.04	0.00	0.04
LST7	-0.06	-0.13	-0.11	0.07	0.05	0.07	0.05
LST8	-0.37	-0.08	0.08	-0.29	-0.45	0.29	0.45
LST9	1.12	1.13	1.29	-0.01	-0.17	0.01	0.17
LST10	0.59	0.59	0.53	0.00	0.06	0.00	0.06
LST11	-0.01	-0.18	-0.24	0.17	0.22	0.17	0.22
LST12	0.03	-0.18	-0.27	0.22	0.30	0.22	0.30
LST13	0.19	0.26	0.32	-0.07	-0.13	0.07	0.13
LST14	0.28	0.11	0.05	0.17	0.23	0.17	0.23
LST15	-0.01	-0.13	-0.24	0.12	0.22	0.12	0.22
LST16	0.19	0.21	0.23	-0.02	-0.04	0.02	0.04
LST17	-0.71	-0.79	-0.71	0.08	0.00	0.08	0.00
LST18	-0.30	-0.30	-0.34	0.00	0.04	0.00	0.04
LST19	1.94	1.89	1.86	0.05	0.08	0.05	0.08
LST20	-0.08	-0.08	-0.04	-0.01	-0.04	0.01	0.04
LST21	0.30	0.34	0.26	-0.04	0.04	0.04	0.04
LST22	0.41	0.44	0.41	-0.03	0.00	0.03	0.00
LST23	0.26	0.31	0.26	-0.06	-0.01	0.06	0.01
LST24	0.06	0.21	0.32	-0.16	-0.27	0.16	0.27
LST25	0.30	0.34	0.29	-0.04	0.01	0.04	0.01
LST26	1.66	1.69	1.74	-0.03	-0.08	0.03	0.08
LST27	-0.08	-0.18	-0.20	0.10	0.12	0.10	0.12
LST28	0.55	0.54	0.59	0.01	-0.04	0.01	0.04
LST29	0.98	1.21	1.32	-0.23	-0.34	0.23	0.34
LST30	0.84	0.89	0.86	-0.05	-0.02	0.05	0.02

Correl.	1	.992	.982	Average absolute differences	0.081	0.126
	.992	1	.997			
	.982	.997	1			

Note. Subsample a is based on the lenient cut-off (11 items), subsample b is based on the strict cut-off (8 items); the columns “Difference a” and ‘Difference b” show differences in sum-normed item difficulty parameters between each of the subsamples and the total sample.

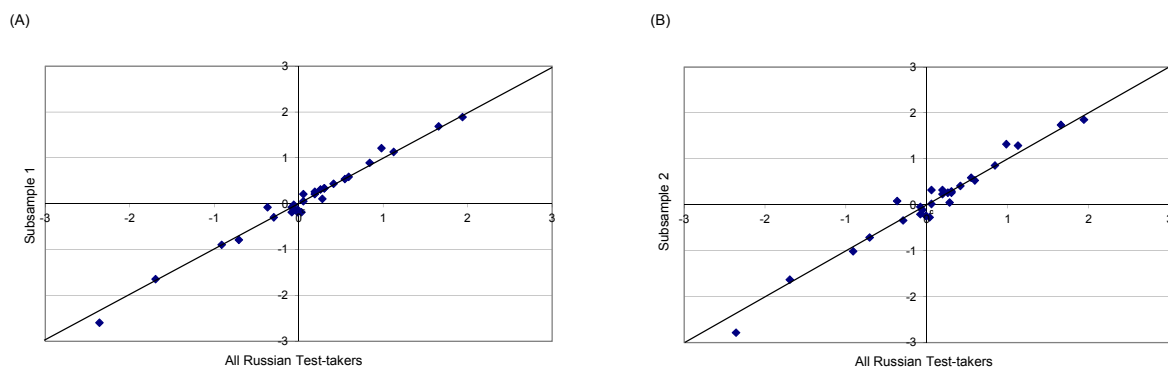


Figure C.15.

Alignment sum-normed RM item difficulty parameters for full Russian sample ($N = 201$) and two subsamples of all Russian participants ($N_1 = 177$; $N_2 = 151$) that chose the response category “not solvable” less than 11/ less than 8 times; A) lenient cut-off, B) strict cut-off value

Table C.2.

Frequencies of A, B, C DIF for items that (not) allow for a reduction of considerable response alternatives (Study 3)

ETS	Reduction of response alternatives		
	> 2	2	total
A	13	3	16
B	4	3	7
C	2	5	7
total	19	11	30

Table C.3.

Frequencies of A, B, C DIF for items that (not) allow for the application of an easy falsification strategy (Study 3)

ETS	Easy falsification		
	not possible	possible	total
A	14	2	16
B	5	2	7
C	2	5	7
total	21	9	30

