Kölner Beiträge zur Sprachdidaktik

herausgegeben von
Hartmut Günther, Ursula Bredel & Michael Becker-Mrotzek

Reihe A

Daniela Neumann

Schwierigkeitsbeeinflussende Merkmale bei Aufgaben zum Hörverstehen im Fach Deutsch in der Sekundarstufe I

KöBeS (8) 2012
Gilles & Francke Verlag

Daniela Neumann

Schwierigkeitsbeeinflussende Merkmale bei Aufgaben zum Hörverstehen im Fach Deutsch in der Sekundarstufe I

Dissertation zur Erlangung des akademischen Grads Dr. phil.
im Fach Erziehungswissenschaften

---

eingereicht am 11. März 2011 an der Philosophischen Fakultät IV
der Humboldt-Universität zu Berlin von Daniela Neumann, 09.11.1975,
Gunzenhausen

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Philosophischen Fakultät IV
Prof. Dr. Ernst von Kardorff

Gutachter
1. Prof. Dr. Michael Becker-Mrotzek
2. Prof. Dr. Olaf Köller

Danksagung

Bedanken möchte ich mich an erster Stelle bei meinen beiden Betreuern Prof. Dr. Olaf Köller und Prof. Dr. Michael Becker-Mrotzek. Durch die Anbindung der Dissertation an die Arbeiten am Institut zur Qualitätsentwicklung (IQB) konnte diese Dissertation erst entstehen und ich bedanke mich ausdrücklich für die Möglichkeit, mit den IQB-Aufgaben und –Daten arbeiten zu dürfen. Beide Betreuer unterstützten und berieten mich in allen Phasen der Dissertation, gewährten mir stets ein offenes Ohr und ermutigten mich, die Arbeit zu Ende zu bringen. Danken möchte ich auch meinen Kolleginnen und Kollegen am IQB, die mich in vielfacher Weise motiviert und unterstützt haben. Besonders erwähnen möchte ich an dieser Stelle Ina Lindow, Janis Nalbadidacis und Christoph Schulte, deren Urteile im Rahmen der Ratings in dieser Arbeit unerlässlich waren und durch deren kritischen Blick und fachkundige Diskussion viele Merkmale in ihren Beschreibungen geschärft werden konnten. Besonders erwähnt soll auch Henrik Winkelmann sein, der auch in Momenten der Unsicherheit immer für mich da war und dessen Rat mir vielfach geholfen hat, die Arbeit zu optimieren. Viele Menschen haben Interesse an meiner Arbeit gezeigt und mich durch ihre Nachfragen und kritischen Gedanken nicht nur zur Weiterarbeit motiviert, sondern mir auch wertvolle Impulse und Anregungen gegeben. Dafür möchte ich mich an dieser Stelle bedanken. Ausdrücklich möchte ich auch meinen Eltern und meinem Bruder Tobias für die immerwährende Unterstützung danken. Danke, dass ihr an mich geglaubt habt. Zuletzt möchte ich mich bei Alexander Robitzsch bedanken. Er stand mir als verlässlicher Freund bei der Arbeit zur Seite und nahm sich viel Zeit, mich beratend zu unterstützen. Insbesondere in technischen Fragen war er ein sehr kompetenter und äußerst geduldiger Ansprechpartner. Ohne seinen motivierenden Beistand und ohne die vielen inspirierenden Diskussionen mit ihm hätte diese Arbeit nicht entstehen können.

Vielen herzlichen Dank!

Für Luise

Informationen über KöBeS – Kölner Beiträge zur Sprachdidaktik – finden Sie unter folgender Internet-Adresse: www.koebes.uni-koeln.de

ISBN 978-3-925348-94-5

Bibliografische Information der Deutschen Bibliothek:
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://www.ddb.de abrufbar.

Inhalt

I
Einleitung und bildungspolitische Einordnung
Seite 26

II
Theoretische Grundlagen
Seite 30

III
Darstellung der Untersuchung
Seite 126

IV
Darstellung der Ergebnisse
Seite 168

V
Zusammenfassung und Diskussion
Seite 256

I Einleitung und bildungspolitische Einordnung 26

II Theoretische Grundlagen 30

1. Rahmenbedingungen des Hörverstehens 31

1.1. Zuhören als Teilbereich mündlicher Kommunikation 32

1.1.1. Inhaltliche Merkmale 36
1.1.2. Lautliche und paralinguistische Merkmale 37
1.1.3. Lexikalische Merkmale 37
1.1.4. Syntaktische Merkmale 38

1.2. Text und Diskurs 39

2. Hörverstehen – eine psycholinguistische Perspektive 40

2.1. Verarbeitung eintreffender Informationen 41

2.1.1. Informationsverarbeitung 42
2.1.1.1. Theorie des Kurzzeitgedächtnisses 43
2.1.1.2. Baddeleys Modell des Arbeitsgedächtnisses 45
2.1.1.3. Capacity Theory of Comprehension 46
2.1.2. Speicherung von Wissen 48

2.2. Sprachverstehen 50

2.2.1. Bottom-up und Top-down Verarbeitungsprozesse 53
2.2.2. Verstehen durch Wahrnehmung, Parsing und Verwendung 54
2.2.3. Das Construction-Integration-Modell 55
2.2.4. Kognitive Prozessmodelle 57

3. Hörverstehen – Zusammenfassung und Ausblick 59

3.1. Rahmenbedingungen des Zuhörens 60
3.2. Modell des Zuhörens als Grundlage für diese Arbeit 64
3.3. Operationalisierung des Konstrukts „Hörverstehen“ in Testaufgaben 66
3.4. Ausblick: Ableitung von weiteren Merkmalen 68
3.5. Ausblick: Studien zu schwierigkeitsbeeinflussenden Merkmalen 69

4. Hörverstehen - Eine kompetenzdiagnostische Perspektive 72

4.1. Messung von Sprachkompetenz 72
4.2. Hörverstehen im Gemeinsamen Europäischen Referenzrahmen für Sprachen(GER) 76

4.3. Zuhören im Large-Scale-Assessment ... 79
4.3.1. Nationale Studien ... 79
4.3.2. Internationale Studien ... 84
4.4. Hörverstehen im Bildungswesen Deutschlands ... 86
4.4.1. Hörverstehen in den Bildungsstandards ... 86
4.4.2. Kompetenzstufenmodelle Zuhören ... 90
4.4.2.1. Das Kompetenzstufenmodell der Bildungsstandards ... 91
4.4.2.2. Exkurs: Luxemburg ... 93
4.4.3. Hörverstehen in den Rahmenplänen der Länder ... 95

5. Ausgewählte Merkmale und ihr Einfluss auf die Itemschwierigkeit 98

5.1. Stimulusmerkmale ... 100
5.1.1. Sprachliche Merkmale des Stimulus ... 100
5.1.2. Struktur der Stimuli ... 103
5.1.3. Thematische Merkmale ... 107
5.1.4. Präsentationsmerkmale ... 107
5.1.5. Subjektive Einschätzungen der Stimuli ... 110
5.2. Itemmerkmale ... 111
5.2.1. Itemformat ... 111
5.2.2. Multiple-Choice-Items ... 113
5.2.3. Zeitpunkt der Itembearbeitung ... 115
5.2.4. Überlappung der Item-Formulierungen mit dem Stimulus ... 116
5.2.5. Die zur Beantwortung eines Items notwendige Information (NI) ... 117
5.2.5.1. Art der NI ... 117
5.2.5.2. Position der NI ... 119
5.2.5.3. Auftretenshäufigkeit der NI ... 120
5.3. Merkmale der Testpersonen ... 120
5.3.1. Motivation und Interesse ... 121
5.3.2. Hintergrundwissen ... 121
5.3.3. Arbeitsgedächtnis ... 122
5.3.4. Sprachkenntnisse ... 123

III Darstellung der Untersuchung 126

1. Wissenschaftliche Fragestellungen 127

1.1. Forschungsfragen ... 127
1.2. Hypothesen ... 128
1.2.1. Hypothese 1 - Einzelmerkmale ... 128
1.2.2. Hypothese 2: Merkmalsgruppen ... 132
1.2.3. Hypothese 3: Globalurteil und Fragebogen ... 133
1.2.4. Hypothese 4: Personenmerkmale ... 133

2. Instrument und Datengrundlage 133
2.1. Übersicht über die identifizierten Variablen 138
2.2. Beschreibung der identifizierten Variablen 140
2.2.1. Stimulusmerkmale aus den IQB-Ratings 140
2.2.1.1. Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale 140
2.2.1.2. Merkmalsgruppe II: Präsentationsmerkmale 143
2.2.1.3. Merkmalsgruppe III: Inhaltlich-thematische Merkmale 144
2.2.1.4. Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte 145
2.2.1.5. Merkmalsgruppe V: Globalurteil 146
2.2.2. Stimulusmerkmale aus dem Lehrerfragebogen 146
2.2.3. Itemmerkmale aus den IQB-Ratings 147
2.2.3.1. Merkmalsgruppe I: Itemformat 147
2.2.3.2. Merkmalsgruppe II: Merkmale der Itempräsentation 148
2.2.3.3. Merkmalsgruppe III: Merkmale von MC-Items 148
2.2.3.4. Merkmalsgruppe IV: Kognitive Anforderungen der Items 150
2.2.3.5. Merkmalsgruppe V: Globalurteil 153
2.2.4. Personenmerkmale 153
2.2.4.1. Einschätzung der Aufgaben durch die Schüler 153
2.2.4.2. Arbeitsgedächtniskapazität 154
2.2.4.3. Sprachkenntnisse 154

3. Verwendete Forschungsmethoden 155
3.1. Theoretische Grundlagen 155
3.2. Deskriptive Itemanalysen 157
3.3. Item-Response-Theory (IRT) 157
3.3.1. Das Rasch-Modell 159
3.3.2. Zweiparametrische (2-)PL Modelle 160
3.4. Analyse von Verteilungen 160
3.5. Korrelationsanalysen 161
3.6. Varianzanalysen 161
3.7. Lineare Regressionsanalysen 162
3.8. Methoden zur Trennung von Item- und Stimuluseffekten 164
3.9. Dimensionsanalysen 164
3.10. Verfahren zur Erfassung der Beurteilerübereinstimmung 166

IV Darstellung der Ergebnisse 168

1. Explorative Dimensionsanalysen von Item- und Stimulusmerkmalen 169
1.1. Stimulusmerkmale 169

1.1.1. Korrelation der Stimulusmerkmale aus den IQB-Ratings mit der Schwierigkeit 169
1.1.1.1. Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale 170
1.1.1.2. Merkmalsgruppe II: Präsentationsmerkmale 171
1.1.1.3. Merkmalsgruppe III: Inhaltlich-thematische Merkmale 172
1.1.1.4. Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte 173
1.1.1.5. Merkmalsgruppe V: Globalurteil 173
1.1.2. Korrelation der Stimulusmerkmale aus den Lehrerfragebögen mit der Schwierigkeit 173
1.1.3. Faktorenanalysen der Stimulusmerkmale aus den IQB-Ratings 174
1.1.4. Faktorenanalysen der Stimulusmerkmale aus dem Lehrerfragebogen 176
1.2. Itemmerkmale 178
1.2.1. Korrelationsübersicht der Itemmerkmale 178
1.2.1.1. Merkmalsgruppe I: Itemformat 178
1.2.1.2. Merkmalsgruppe II: Merkmale der Itempräsentation 179
1.2.1.3. Merkmalsgruppe III: Merkmale von MC-Items 179
1.2.1.4. Merkmalsgruppe IV: Kognitive Anforderungen der Items 179
1.2.1.5. Merkmalsgruppe V: Globalurteil 179
1.2.2. Faktorenanalysen der Itemmerkmale 179
1.2.2.1. Allgemeine Itemmerkmale 179
1.2.2.2. Merkmale zur Beschreibung von MC-Items 180
1.2.2.3. Merkmale zur Beschreibung der NI 181
1.3. Personenmerkmale 181
1.3.1. Einschätzung der Stimuli durch die Schüler 181
1.3.2. Arbeitsgedächtnis 186
1.3.3. Sprachkenntnisse 187
1.4. Zusammenfassung der explorativen Dimensionsanalysen 189
1.4.1. Zusammenfassung Stimulusmerkmale 189
1.4.2. Zusammenfassung Itemmerkmale 192
1.4.3. Zusammenfassung Personenmerkmale 196

2. Zusammenhangsanalysen 197
2.1. Stimulusmerkmale 197
2.1.1. Mittelwertsvergleiche der Stimulusmerkmale aus den IQB-Ratings 197
2.1.1.1. Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale 198
2.1.1.2. Merkmalsgruppe II: Präsentationsmerkmale 205
2.1.1.3. Merkmalsgruppe III: Inhaltlich-thematische Merkmale 207
2.1.1.4. Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte 209
2.1.1.5. Merkmalsgruppe V: Globalurteil 214
2.1.2. Mittelwertsvergleiche der Stimulusmerkmale aus dem Lehrerfragebogen 214

2.2. Itemmerkmale 217
2.2.1. Mittelwertsvergleiche der Itemmerkmale 217
2.2.1.1. Merkmalsgruppe I: Itemformat 218
2.2.1.2. Merkmalsgruppe II: Merkmale der Itempräsentation 220
2.2.1.3. Merkmalsgruppe III: Merkmale von MC-Items 221
2.2.1.4. Merkmalsgruppe IV: Kognitive Anforderungen der Items 222
2.2.1.5. Merkmalsgruppe V: Globalurteil 229
2.3. Zusammenfassung der Zusammenhangsanalysen 229
2.3.1. Zusammenfassung Stimulusmerkmale 229
2.3.2. Zusammenfassung Itemmerkmale 237

3. Regressionsanalysen 244
3.1. Aufgabenschwierigkeiten und Stimulusmerkmale 244
3.1.1. Die am höchsten mit der Aufgabenschwierigkeit korrelierenden Merkmale 245
3.1.2. Faktorenanalytisch bestimmte Merkmalsgruppen 245
3.1.3. Thematisch gebildete Variablengruppen 246
3.1.4. Ausgewählte Stimulusmerkmale der vorhergehenden Analysen 247
3.2. Itemschwierigkeiten und Stimulusmerkmale 247
3.2.1. Die am höchsten mit der Itemschwierigkeit korrelierenden Merkmale 247
3.2.2. Faktorenanalytisch bestimmte Merkmalsgruppen 247
3.2.3. Thematisch gebildete Variablengruppen 248
3.2.4. Ausgewählte Stimulusmerkmale der vorhergehenden Analysen 249
3.3. Itemschwierigkeiten und Itemmerkmale 250
3.3.1. Die am höchsten mit der Itemschwierigkeit korrelierenden Merkmale 250
3.3.2. Faktorenanalytisch bestimmte Merkmalsgruppen 250
3.3.3. Thematisch gebildete Variablengruppen 251
3.3.4. Ausgewählte Itemmerkmale der vorhergehenden Analysen 252
3.4. Zusammenfassung der Regressionsanalysen 252
3.4.1. Zusammenfassung: Einfluss Stimulusmerkmale auf die Aufgabenschwierigkeit 252
3.4.2. Zusammenfassung: Einfluss Stimulusmerkmale auf die Itemschwierigkeit 253
3.4.3. Zusammenfassung: Einfluss Itemmerkmale auf die Itemschwierigkeit 254

V Zusammenfassung und Diskussion 256

Literatur 268
Weblinks 290
Verzeichnis der Anhänge 292

Verzeichnis der Abbildungen

Abbildung II-2.1.2.:
Modell eines semantischen Netzwerks (vgl. Anderson, 2001: 186) 50

Abbildung II-2.2.4.:
Informationsverarbeitungsmodell für MC-Items zum Leseverstehen
(Embretson & Wetzel, 1987: 17) 57

Abbildung II-3.2.:
Zuhörprozesse 66

Abbildung II-3.3.:
Am Zuhörprozess beteiligte Aspekte 67

Abbildung II-4.1.:
Modell der kommunikativen Sprachkompetenz (Bachman & Palmer, 1996) 75

Abbildung II-4.2.:
GER Schwierigkeitsbeeinflussende Merkmale 78

Abbildung II-4.3.1a.:
Merkmale der DESI-Leseverstehensaufgaben Deutsch (vgl. Klieme et al., 2003: 38ff) 82

Abbildung II-4.3.1b.:
DESI Testkonzept Hörverstehen (Klieme et al., 2003: 76) 83

Abbildung II-4.3.1c.:
Merkmale der DESI-Hörverstehensaufgaben Englisch (Klieme et al., 2003: 79) 84

Abbildung II-4.4.1.:
Kompetenzmodell der Bildungsstandards (KMK, 2004: 8) 87

Abbildung II-4.4.2.2.:
Schwierigkeitsbeeinflussende Merkmale Luxemburg
(vgl. Ministère de l'Education nationale et de la Formation professionnelle, 2008: 61ff) 94

Abbildung III-2.2.2.:
Lehrerfragebogen 146

Abbildung III-2.2.4.1.:
Einschätzung der Aufgaben durch die Schüler im Testheft 154

Verzeichnis der Tabellen

Tabelle II-1.2.:
Übersicht über „Text“ und „Diskurs“ ... 40

Tabelle II-4.1.:
Klassifikation von Teilbereichen sprachlicher Kompetenz ... 73

Tabelle II-4.4.1.:
In den IQB-Items realisierte Bildungsstandards ... 89

Tabelle II-4.4.2.1.:
Kompetenzstufenmodell Bildungsstandards ... 92

Tabelle II-4.4.3a.:
Vergleich BS 1.4. mit Lehrplanformulierungen ... 97

Tabelle II-4.4.3b.:
Vergleich Bildungsstandard – Lehrplan ... 97

Tabelle II-5.1.2.:
Relationstypen nach Anderson (2001) ... 103

Tabelle III-1.2.1a.:
Übersicht über den vermuteten Einfluss der Stimulusmerkmale ... 130

Tabelle III-1.2.1b.:
Übersicht über den vermuteten Einfluss der Itemmerkmale ... 132

Tabelle III-2a.:
Übersicht über die Hörverstehens-Aufgaben (Normierung) ... 137

Tabelle III-2b.:
Übersicht über die Hörverstehens-Aufgaben (2. Testtag, Ländervergleich) ... 137

Tabelle III-2.1a.:
Übersicht über alle untersuchten Stimulusmerkmale ... 138

Tabelle III-2.1b.:
Übersicht über alle untersuchten Itemmerkmale ... 139

Tabelle III-2.1c.:
Übersicht über alle untersuchten Personenmerkmale ... 139

Tabelle III-2.1.1.1a.:
Übersicht über die Einzelcodes der Variable STR ... 141

Tabelle III-2.1.1.1b.:
Übersicht über die Einzelcodes der Variable RHE ... 142

Tabelle III-2.2.1.3.:
Übersicht über die Einzelcodes der Variable TFU ... 144

Tabelle III-2.2.1.4.:
Übersicht über die Einzelcodes der Variable REL ... 145

Tabelle III-2.2.3.1a.:
Übersicht über die Einzelcodes der Variable IFK ... 148

Tabelle III-2.2.3.1b.:
Übersicht über die Einzelcodes der Variable IFA ... 148

Tabelle III-2.2.3.2.:
Übersicht über die Einzelcodes der Variable PIA 149

Tabelle III-2.2.3.3.:
Übersicht über die Einzelcodes der Variable PDI 149

Tabelle III-2.2.3.4a.:
Übersicht über die Einzelcodes der Variable BS 150

Tabelle III-2.2.3.4b.:
Übersicht über die Einzelcodes der Variable ANI 150

Tabelle III-2.2.3.4c.:
Übersicht über die Einzelcodes der Variable ARN 151

Tabelle III-2.2.3.4d.:
Übersicht über die Einzelcodes der Variable TCO 152

Tabelle III-2.2.3.4e.:
Übersicht über die Einzelcodes der Variable TNI 153

Tabelle IV-1.1.1:
Zusammenfassung Korrelation Merkmal mit der Aufgabenschwierigkeit 170

Tabelle IV-1.1.2.:
Zusammenfassung Korrelation Lehrerfragebogen mit der Aufgabenschwierigkeit 174

Tabelle IV-1.1.4c.:
Vergleich Merkmalsgruppen Aufgabenebene und Ratingebene 178

Tabelle IV-1.3.1a.:
Korrelationsübersicht auf Blockebene: Summenscore – Schülereinschätzung (Block H1 und H4) 183

Tabelle IV-1.3.1b.:
Korrelationsübersicht auf Blockebene: Summenscore – Schülereinschätzung (Block H8 und H10) 183

Tabelle IV-1.3.1c.:
Korrelationen der Summenscores mit den Schülereinschätzungen für Block H1 184

Tabelle IV-1.3.1d.:
Mittelwerte der Lehrer- und der Schülereinschätzungen zum Interessantheitsgrad der Stimuli 186

Tabelle IV-1.3.2.:
Korrelationsübersicht auf Aufgabenebene: Summenscores mit dem Test zur Arbeitsgedächtniskapazität 187

Tabelle IV-1.3.3a.:
Korrelation Angaben zum Sprachstand – Leistungsdaten Zuhören 188

Tabelle IV-1.3.3b.:
Korrelation Angaben zum Sprachstand – Leistungsdaten Zuhören: Aufgabenspezifisch 188

Tabelle IV-1.4.1a.:
Übersicht über den vermuteten Einfluss der Stimulusmerkmale 190

Tabelle IV-1.4.1b.:
Vergleich Merkmalsgruppen didaktisch/sprachwissenschaftlich und faktorenanalytisch 192

Tabelle IV-1.4.2a.:
Übersicht über den vermuteten und tatsächlichen Einfluss der Itemmerkmale 193

Tabelle IV-1.4.2b.:
Überblick über die Zugehörigkeit der Itemvariablen zu Merkmals- und Faktorengruppen 195

Tabelle IV-2.2.1.1a.:
Kreuztabelle – IFA/IFK 218

Tabelle IV-2.3.1a.:
Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale
Merkmalsgruppe I „Komplexität des Wortschatzes und sprachliche Merkmale“ 230

Tabelle IV-2.3.1b.:
Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale
Merkmalsgruppe II „Präsentationsmerkmale“ 232

Tabelle IV-2.3.1c.:
Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale
Merkmalsgruppe III „Inhaltlich-thematische Merkmale“ 234

Tabelle IV-2.3.1d.:
Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale
Merkmalsgruppe IV „Struktur der Stimuli und propositionale Dichte“ 235

Tabelle IV-2.3.1e.:
Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale
Merkmalsgruppe V „Globalurteil“ 237

Tabelle IV-2.3.2a.:
Zusammenfassung Zusammenhangsanalysen Itemmerkmale
Merkmalsgruppe I „Itemformat“ 238

Tabelle IV-2.3.2b.:
Zusammenfassung Zusammenhangsanalysen Itemmerkmale
Merkmalsgruppe II „Merkmale der Itempräsentation“ 238

Tabelle IV-2.3.2c.:
Zusammenfassung Zusammenhangsanalysen Itemmerkmale
Merkmalsgruppe III „Merkmale von MC-Items“ 239

Tabelle IV-2.3.2d.:
Zusammenfassung Zusammenhangsanalysen Itemmerkmale
Merkmalsgruppe IV „Kognitive Anforderungen der Items“ 240

Tabelle IV-2.3.2e.:
Zusammenfassung Zusammenhangsanalysen Itemmerkmale
Merkmalsgruppe V „Globalurteil“ 243

Tabelle IV-3.3.2a.:
Übersicht über die faktorenanalytischen Ergebnisse bei den Itemmerkmalen 250

Zusammenfassung

In dieser Arbeit wird der Einfluss ausgewählter Merkmale auf die Schwierigkeit von Items und Stimuli des IQB-Aufgabenpools für das Fach Deutsch in der Sekundarstufe I im Kompetenzbereich „Zuhören“ untersucht. Aufgrund linguistischer Überlegungen zur Eigenart mündlicher Kommunikation und den besonderen Umständen in Art und Produktion gesprochener Sprache sowie den psychologischen, psycholinguistischen und didaktischen Grundlagen des Hörverstehens ergeben sich Merkmale der Stimuli, der Items sowie Merkmale, die aus der Interaktion der Items mit dem entsprechenden Stimulus resultieren. Diese Merkmale werden im Rahmen korpusanalytischer Methoden untersucht und sind überwiegend auf lexikalischer und syntaktischer Ebene zu finden. Außerdem wird die Übereinstimmung der Einschätzung der Aufgabenentwickler bezüglich der Schwierigkeit von Items und Stimuli mit den empirischen Schwierigkeiten überprüft. Zu den Stimuli liegen ferner Einschätzungen von Lehrkräften auf der Grundlage einer Ratingskala vor. Zusätzlich werden Merkmale der Testpersonen in die Analysen miteinbezogen.

Die Analysen werden an Aufgaben durchgeführt, die auf der Grundlage der KMK-Bildungsstandards zur Überprüfung der Hörverstehenskompetenz im Fach Deutsch von Fachlehrkräften aus allen Bundesländern unter der Leitung des Instituts zur Qualitätsentwicklung im Bildungswesen entwickelt wurden. Die Aufgaben für den Kompetenzbereich „Zuhören“ operationalisieren Aspekte des Hörverstehens, wie sie von den Lehrplänen der Länder und den 2003 und 2004 von der Kultusministerkonferenz für alle Bundesländer verbindlich eingeführten Bildungsstandards für den Hauptschulabschluss (Jahrgangsstufe 9) und den Mittleren Schulabschluss (Jahrgangsstufe 10) gefordert werden.

Die auf die einzelnen Merkmale bezogenen Analysen ergeben einen Zusammenhang mit der Schwierigkeit für 23 Stimulusmerkmale, die überwiegend der Kategorie „Komplexität des Wortschatzes und sprachliche Mittel“ zuzuordnen sind, und für sieben Itemmerkmale, von denen die meisten in die Kategorie „Itemformat“ fallen. Untersucht wird ferner, ob bestimmte

Merkmalsgruppen einen höheren Einfluss auf die Schwierigkeit haben als einzelne Merkmale. In der Gruppe der am höchsten mit der Aufgabenschwierigkeit korrelierenden Merkmale wird für zwölf Stimulusmerkmale ein Anteil an der Gesamtvarianzaufklärung von 87.9%, für drei Itemmerkmale von 41.3% ermittelt. Eine Gruppierung der Merkmale aufgrund faktorenanalytischer Kriterien führt nicht zur Identifikation besonders vorhersagestarker Variablengruppen. Eine nach thematischen Überlegungen vorgenommene Gruppierung der Merkmale führt zu einer maximalen Varianzaufklärung im Bereich der Stimulusmerkmale von 63.2%.

Die aus den vorhergehenden Regressionsanalysen ausgewählten Merkmale eignen sich im Vergleich zu den anderen Gruppen am besten dazu, die Schwierigkeit vorherzusagen. Zwölf Stimulusmerkmale erklären zusammen 90.1%, vier Itemmerkmale erklären 45% der Gesamtvarianz. Die Merkmale, die einen Einfluss in den meisten Gruppen zeigen, waren „Worthäufigkeit“ und „Sprechgeschwindigkeit“. Das Globalurteil der Aufgabenentwickler ist wenig dazu geeignet, die empirische Schwierigkeit vorherzusagen. Die Prognose der Stimulusschwierigkeit mittels der durch den Lehrerfragebogen erhaltenen Einschätzungen funktioniert für die meisten Variablen hingegen sehr gut.

Zusätzlich wird untersucht, in welcher Weise personenbezogene Merkmale die Testleistung und damit die Schwierigkeit von Items und Stimuli beeinflussen. Beurteilt werden von den Schülern der Interessantheits- bzw. Bekanntheitsgrad der Stimuli sowie die Präsentationsqualität. Für keine dieser Einschätzungen kann ein systematischer signifikanter Zusammenhang mit den Leistungsdaten der Schüler nachgewiesen werden. Die Ergebnisse des Tests zur Arbeitsgedächtniskapazität sowie das Merkmal „Sprachkenntnisse“ korrelieren dagegen in allen Fällen signifikant ($p < 0.05$) mit den Leistungsdaten.

Abstract

In this work the influence of certain features on item difficulty and stimuli of listening comprehension exercises is investigated. The data were taken from the IQB (Institute for Educational Progress) data pool of listening comprehension tasks used for secondary education tests for German. Certain characteristics of the stimuli, the items and characteristics of item-stimuli- interaction result from linguistic considerations on the nature of oral communication and the particular circumstances of spoken language as well as the psychological, psycholinguistic and didactic principles of listening. These characteristics are investigated mainly with corpus-analytic methods focusing on lexical and syntactic features. In addition the test developers' judgement on the degree of difficulty of items and stimuli is compared with empirical evidence. The stimuli are also assessed by teachers based on a rating scale. Moreover, characteristics of the test subjects are included in the analysis.

All test items were developed by schoolteachers of all federal states in Germany on the basis of the national educational standards for assessing listening comprehension in German. The supervision and responsibility of setting the test is entrusted to the Institute for Educational Progress. These tasks operationalize aspects of listening comprehension, as required by the curricula of the German federal states and by the educational standards for secondary school (grade 9) and intermediate school (grade 10) that were established by the Standing Conference of the Ministers of Education and Cultural Affairs of the federal states in 2003 and 2004.

The analysis reveals that there is a correlation between the item difficulty and 23 stimulus characteristics, mainly in the category „complexity of vocabulary and linguistic resources". A correlation also exists between the item difficulty and seven item characteristics, most of which within the category "item formats". Furthermore it is examined whether certain combinations of characteristics have a higher impact on item difficulty than individual ones. In this respect twelve stimulus characteristics relate to the most highly degree of task difficulty, with a proportion of 87.9% on total variance explained, while three item characteristics

explain 41.3%. A grouping of characteristics according to factor-analytic criteria does not lead to the identification of particularly strong predictive groups of variables. However, a grouping according to thematic considerations leads to a maximum explanation of variance within the stimulus characteristics of 63.2%.

It has been proven that those characteristics selected from preceding regression analyses are best suitable for predicting item difficulty. Twelve stimulus characteristics together explain 90.1% of total variance, while four item characteristics explain 45%. The characteristics, which are present in most of these groupings, are „word frequency" and „mean speech rate". Interestingly, the global judgement of the test developers is not suited for predicting empirical difficulty, whereas the attained data from teacher questionnaires predict very well stimulus difficulty for most variables.

This paper also investigates the way in which personal characteristics of test subjects influence test performance and therefore the difficulty of items and stimuli. The students assessed how interesting and known the stimuli were to them, as well as to what extent the stimuli were difficult to understand. The analysis shows that for none of these estimates a systematic and significant connection to students' performance can be proven. Nevertheless the test results show that working memory capacity and „language skills" always correlate significantly ($p < 0.05$) with students' performance.

I

Einleitung und bildungspolitische Einordnung

I Einleitung und bildungspolitische Einordnung

Im Bereich des Spracherwerbs spielen Zuhören und Hörverstehen eine zentrale Rolle. Mündliche Kommunikation folgt eigenen Regeln und hat für das Leben in der Gemeinschaft eine wichtige Schlüsselfunktion. Trotz ihrer Bedeutung für das Sprachhandeln nimmt die Arbeit mit gesprochener Sprache sowie die Förderung und Überprüfung von Hörverstehen im schulischen Kontext bislang eine Randposition ein. Im Alltag stellt Zuhören zwar eine Schlüsselkompetenz dar, im Unterricht findet der Kompetenzbereich jedoch meist nur am Rand Beachtung. Dies mag damit zusammenhängen, dass Hörverstehen zu einem erheblichen Teil bereits vorschulisch erworben wird und traditionell der Auftrag des Sprachunterrichts viel stärker im Erwerb der Schriftsprache gesehen wird.

Merkmale, die einen Einfluss auf die Item[1]- und Stimulusschwierigkeit[2] haben, stellen seit einiger Zeit vor allem im fremdsprachlichen Bereich einen wichtigen Untersuchungsgegenstand dar (vgl. Best et al., 2005; Grotjahn, 2000; Freedle & Kostin, 1993a). Weshalb wird diesem Bereich in neuerer Zeit so viel Aufmerksamkeit gewidmet? Stiggins (2002) betont, dass Schulleistungstests und Leistungserhebungen nicht nur den Kompetenzstand der Schülerinnen und Schüler[3] messen dürfen, sondern immer auch dazu dienen müssen, weitere Lernprozesse zu fördern. Die genaue Kenntnis dessen, was Aufgaben[4] leichter oder schwieriger machen kann, also von schwierigkeitsbeeinflussenden[5] Merkmalen, ist für das professionelle Lehrerhandeln essentiell (vgl. Berkemeyer & Bos, 2009). Zum einen können Lehrkräfte dadurch reliable Aussagen über Schülerfähigkeiten auf der Grundlage der erzielten Testergebnisse machen. Verlässliche Informationen zum Lernstand der Schüler in Form kriterialer und ipsativer Rückmeldungen sind dabei die Grundlage für Intervention und Förderung der sprachlichen Kompetenzen im Rahmen des Unterrichts. (vgl. Grotjahn, 2000; Helmke, 2006) Zum anderen können Unterrichts- und Testmaterialien den Fähigkeiten der Zielgruppe entsprechend ausgewählt werden. Die Auswahl geeigneter Unterrichtsmaterialien und deren Modifikation im Rahmen von Binnendifferenzierung ist gerade unter den Anforderungen von Inklusion eine wichtige Fähigkeit, die Lehrkräfte für ihre Tätigkeit benötigen. (vgl. Pressemitteilung Hessisches Kultusministerium vom 02.12.2010) Studien in didaktisch-pädagogischen Kontexten ergaben jedoch beispielsweise für den Bereich des Leseverstehens, dass Lehrkräfte nicht immer über die notwendige diagnostische Kompetenz verfügen, die Schwierigkeit von Aufgaben richtig einzuschätzen (z. B. Artelt et al., 2009). Merkmale, welche einen Einfluss auf die Stimulus- und/oder Itemschwierigkeit haben, sind ihnen häufig nicht bekannt (vgl. BMBF, 2005: 9).

1 Unter Items werden in dieser Arbeit die einzelnen Fragen zu einem Stimulus verstanden.

2 Unter einem Stimulus wird in dieser Arbeit der auditive Input verstanden, den die Testpersonen hören und zu dem sie Fragen (die Items) beantworten sollen.

3 Der besseren Lesbarkeit wegen wird in dieser Arbeit stets die männliche Form verwendet. Gleichwohl sind immer Schülerinnen und Schüler, weibliche und männliche Personen gemeint.

4 Eine Aufgabe ist in dieser Arbeit die Gesamtheit von Stimulus und den dazugehörigen Items.

5 In dieser Arbeit wird bewusst von schwierigkeitsbeeinflussenden und nicht schwierigkeitsgenerierenden (vgl. Buse, 2008; Leucht et al. oder Köller et al., 2009) oder schwierigkeitsbestimmenden (vgl. Schweitzer, 2007; Neumann, 2007 oder Helmke et al. 2004) Merkmalen gesprochen, da die Schwierigkeit einer Aufgabe von einer Vielzahl Faktoren abhängt und die einzelnen Merkmale die Schwierigkeit im besten Fall beeinflussen, jedoch weniger generierend oder bestimmend wirken.

Im Bereich der empirischen Bildungsforschung ist die Kenntnis von schwierigkeitsbeeinflussenden Merkmalen für die theoriegeleitete Aufgabenkonstruktion von Bedeutung. Durch die Kenntnis von Stimulus- und Itemmerkmalen können zunächst optimal geeignete Stimuli für die Aufgabengenerierung gefunden werden. Im Rahmen der Itementwicklung können dann die Items gezielt in ihrer Schwierigkeit für bestimmte Zielgruppen angepasst werden und der Prozess der Itemrevision wird erleichtert. (vgl. Grotjahn, 2000: 7) Durch die Kenntnis schwierigkeitsbeeinflussender Merkmale wird nicht nur der Prozess der Aufgabenentwicklung effizient und professionell, das Wissen über derartige Prädiktoren ist auch für die Beschreibung der einzelnen Stufen in einem Kompetenzstufenmodell von Bedeutung. (vgl. Nold & Rossa, 2007; Willenberg, 2007) Perspektivisch könnten unter Umständen sogar Items gezielt gemäß den unterschiedlichen Kompetenzstufen manipuliert werden. Retrospektiv hilft Wissen über Prädiktoren für die Itemschwierigkeit auf Stimulus- und Itemebene, unerwartet leichte oder schwierige Items auch fachdidaktisch besser interpretieren zu können und die in den Daten auftretende Varianz theoriegeleitet zu erklären. (vgl. Freedle & Kostin, 1993a: 166; Buck & Tatsuoka, 1998: 120) Auf diese Weise kann auch das zu überprüfende Konstrukt genauer definiert und validiert werden. (Embretson, 1983)

Bildungspolitisch sind die Testaufgaben zu den Bildungsstandards ein Resultat der empirischen Wende. Bereits im Oktober 1997 beschloss die *Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland* (KMK) auf ihrer 280. Plenarsitzung in Konstanz länderübergreifende Vergleichsuntersuchungen zum Lern- und Leistungsstand von Schülern. Um die Gleichwertigkeit der schulischen Ausbildung, die Vergleichbarkeit der Schulabschlüsse und die Durchlässigkeit des Bildungssystems in Deutschland zu gewährleisten, hielt sie Maßnahmen zur Sicherung der Qualität schulischer Bildung für unabdingbar. (http://www.kmk.org/schul/pisa/konstanz.htm) Die Notwendigkeit für länderbezogene Qualitätssicherungsmaßnahmen wurde nach den Ergebnissen der großen, internationalen Schulleistungsstudien PISA (*Programme for International Student Assessment*, vgl. Deutsches PISA-Konsortium, 2001, 2002, 2004, 2005), TIMSS (*Third International Mathematics and Science Study*, vgl. Baumert et al., 1997; Baumert et al., 2000a, 2000b) und IGLU (*Internationale Grundschul-Lese-Untersuchung*, vgl. Bos et al., 2003) noch verschärft. Die KMK sah dringenden Handlungsbedarf, um die offensichtlich gewordenen Schwierigkeiten des deutschen Bildungssystems zu beheben. Zunächst verabschiedete sie also 2003 und 2004 verbindliche länderübergreifende Bildungsstandards, um mit ihnen die Vergleichbarkeit schulischer Abschlüsse und die Durchlässigkeit des Bildungssystems zu sichern. In diesem Rahmen wurde 2004 auch das *Institut zur Qualitätsentwicklung im Bildungswesen* (IQB) gegründet. Das IQB ist eine wissenschaftliche Einrichtung der 16 Länder an der Humboldt-Universität zu Berlin. Als Folge der 2006 verabschiedeten Plöner Beschlüsse leitete die KMK dann eine Bildungsreform im Sinne einer Gesamtstrategie zur Qualitätssicherung im Primarbereich und in der Sekundarstufe I ein.

Die Bildungsstandards in der Sekundarstufe I wurden in den Jahren 2003 und 2004 verbindlich in allen Ländern für den Hauptschulabschluss (HSA) und den Mittleren Schulabschluss (MSA) eingeführt. Die Bildungsstandards beschreiben erwartete Lernergebnisse, indem sie allgemeine Bildungsziele aufgreifen und Kompetenzen benennen, die Lernende bis zum Ende ihrer Schullaufbahn an zentralen Inhalten erworben haben sollten. Dabei konzentrieren sich

die Bildungsstandards auf die Kernbereiche eines Faches und formulieren fachliche und fachübergreifende Basisqualifikationen. (KMK, 2004) Das IQB illustriert einerseits die Bildungsstandards mit Unterrichtsaufgaben und macht sie andererseits mit Testaufgaben überprüfbar mit dem Ziel, fundierte Ergebnisse der Stärken und Schwächen der Lernenden in den zentralen Kompetenzbereichen zu bekommen. Die vom IQB entwickelten Testaufgaben wurden in repräsentativen Studien pilotiert und normiert. Mit diesen Aufgaben sollen in sechsjährigen Abständen flächendeckende Ländervergleiche durchgeführt werden, um den Lernstand der Schüler zu erheben. Ziel ist eine Sicherung und Verbesserung von Unterrichtsqualität, wobei die stete Überprüfung und gegebenenfalls Verbesserung der Bildungsstandards durch empirische Erhebungen unabdingbar ist.

Für den Deutschunterricht legen die Standards fest, welche Sach- und Methodenkompetenz die Schüler mit dem Erreichen der 9. (HSA) bzw. 10. Jahrgangsstufe (MSA) erworben haben sollten. Übergeordnetes Ziel ist dabei Orientierungs- und Handlungswissen in Sprache, Literatur und Medien sowie einer entsprechenden Verstehens- und Verständigungskompetenz zu vermitteln. Diese Fähigkeiten und Fertigkeiten werden für die Kompetenzbereiche „Mit Texten und Medien umgehen“, „Sprache und Sprachgebrauch untersuchen“, „Schreiben“, „Richtig Schreiben“ sowie „Sprechen und Zuhören“ beschrieben. Der Kompetenzbereich „Sprechen und Zuhören“ wird von den KMK-Bildungsstandards nur ungenau definiert. Zu den Aufgaben des IQB gehört es jedoch, auch für diesen Kompetenzbereich Testaufgaben zu entwickeln, um die Fähigkeiten der Schüler zukünftig entsprechend testen und vergleichen zu können. Die vorliegende Arbeit konzentriert sich auf den Bereich „Zuhören“, da zum Teilbereich „Sprechen“ bislang keine Aufgaben entwickelt wurden.

Dieser Arbeit liegt die Annahme zugrunde, dass die Merkmale gesprochener Sprache und die Verarbeitung von gesprochenen Informationen sich von geschriebener Sprache und deren Verarbeitung unterscheiden. Im Folgenden soll deshalb untersucht werden, welchen spezifischen Bedingungen Hörverstehen unterliegt und welche Merkmale der Stimuli und der Aufgaben Hörverständnis erschweren bzw. erleichtern. Aus diesem Grund wird zunächst die Eigenart von gesprochener Sprache und mündlicher Kommunikation dargestellt um zu zeigen, vor welchen Herausforderungen die Kompetenzdiagnostik im Bereich des Hörverstehens in der deutschen Sprache steht. Es wird dann der Stand der Forschung zur Hörverstehensdiagnostik in den Bereichen der Psychologie, der Psycholinguistik und der Deutschdidaktik vorgestellt, der für die Testaufgabenentwicklung zu den Bildungsstandards im Rahmen der Arbeiten am IQB Beachtung fand. Die Analysen wurden an Aufgaben durchgeführt, die zur Überprüfung der Zuhörkompetenz im Fach Deutsch am IQB entwickelt wurden. Die Aufgaben für den Kompetenzbereich „Zuhören“ operationalisieren Aspekte des Hörverstehens, wie sie von den Bildungsstandards für den Hauptschulabschluss (HSA) (Jahrgangsstufe 9) und den Mittleren Schulabschluss (MSA) (Jahrgangsstufe 10) gefordert werden.

II

Theoretische Grundlagen

II Theoretische Grundlagen

Der Teil „Theoretische Grundlagen" informiert über die Hintergründe, welche bei der Erhebung von Leistungen im Bereich Hörverstehen eine Rolle spielen. Zunächst werden dafür in Kapitel 1. die Rahmenbedingungen des Hörverstehens dargestellt. Zuhören ist streng genommen ein Teilbereich mündlicher Kommunikation (Kapitel 1.1.), die währenddessen entstehenden Produkte der Sprecher und Zuhörer sind Diskurse. Da bei den IQB-Sprachtests jedoch Zuhörleistungen erbracht werden müssen, die nicht im Rahmen mündlicher Kommunikation stattfinden, sondern sich auf das Verstehen von Texten (z. B. verskriptete Radiobeiträge) beziehen, werden beide Stimulusgrundlagen, Diskurse und Texte, vorgestellt (Kapitel 1.2.). In Kapitel 2. werden die psycholinguistischen Grundlagen des Hörverstehens beschrieben. Dabei wird zunächst in Kapitel 2.1. die Verarbeitung eintreffender Informationen thematisiert und anschließend werden in Kapitel 2.2. unterschiedliche Theorien zum Sprachverstehen dargestellt. Kapitel 3. gibt dann einen Überblick über das dieser Arbeit zugrundeliegende Konstrukt zum Hörverstehen. Eine kompetenzdiagnostische Perspektive zum Hörverstehen wird in Kapitel 4. eröffnet. Bevor ein Überblick darüber gegeben wird, welche Rolle Zuhörkompetenz im Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER) (Kapitel 4.2.), in nationalen und internationalen Schulleistungsstudien (Kapitel 4.3.) und im Bildungswesen Deutschlands (Kapitel 4.4.) spielt, wird beschrieben, inwieweit das Konstrukt „Hörverstehen" überhaupt messbar ist (Kapitel 4.1.). Abschließend wird in Kapitel 5. ein Überblick über unterschiedliche Studien gegeben, die den Einfluss ausgewählter Item- und Stimulusmerkmale sowie Merkmale von Testpersonen auf die Itemschwierigkeit untersuchen.

1. Rahmenbedingungen des Hörverstehens

Mit dem ersten Kapitel „Rahmenbedingungen des Hörverstehens" wird die besondere Qualität des Hörverstehens im Gegensatz zum Leseverstehen verdeutlicht. Dazu wird Hörverstehen als Teil mündlicher Kommunikation beschrieben und die daraus resultierenden Merkmale gesprochener Sprache werden dargestellt (Kapitel 1.1.). In einem zweiten Teil werden die Begriffe „Text" und „Diskurs" eingeführt, um eine sprachlich präzise Einordnung der Stimuli zu ermöglichen (Kapitel 1.2.).

Obwohl gesprochene Sprache im täglichen Leben präsenter als die geschriebene Sprache ist und Hörverstehen eine Schlüsselposition im Rahmen von Spracherwerb und mündlicher Kommunikation zufällt, spielt die gezielte Förderung des Hörverstehens im schulischen Bereich eine untergeordnete Rolle. Begründet werden kann dies mit den Besonderheiten gesprochener Sprache, die sich aus der Rolle des Zuhörens im Rahmen mündlicher Kommunikation ergeben (vgl. Fiehler et al., 2004: 56). Zu nennen sind hier vor allem die Flüchtigkeit und die Zeitlichkeit gesprochener Sprache, die eine sehr schnelle Verarbeitung des Gesagten erfordern. Von Bedeutung sind aber auch die Anzahl und die Größe der an den Gesprächen beteiligten Parteien sowie ihre Kopräsenz, die in einer gemeinsamen Situation für die Gesprächsparteien resultiert. Aufgrund dieser kopräsenten Gesprächssituationen nehmen sich

die Gesprächspartner wechselseitig wahr. Diese Rahmenbedingungen gesprochener Sprache führen zu distinkten Merkmalen, von denen die Kompetenz „Zuhören" wesentlich beeinflusst wird. Die Merkmale zeigen sich bei der inhaltlichen und lautlichen Betrachtung gesprochener Sprache, aber auch im paralinguistischen, lexikalischen und syntaktischen Bereich.

Ferner ist für die genaue Beschreibung der in dieser Arbeit verwendeten Stimuli die Unterscheidung der Begriffe „Text" und „Diskurs" von Bedeutung. In Anlehnung an die Funktionale Pragmatik werden mit dem Begriff „Text" die Stimuli bezeichnet, die auf schriftlichen Grundlagen basieren. Erst nachträglich verschriftlichte Stimuli, die im Sinne einer unmittelbaren Kommunikationssituation bei gleichzeitiger Anwesenheit aller Gesprächsbeteiligten entstanden, werden als „Diskurse" bezeichnet.

1.1. Zuhören als Teilbereich mündlicher Kommunikation

Sowohl individual- wie auch menschheitsgeschichtlich entwickelte sich Schriftlichkeit erst nach der Mündlichkeit und viele Sprachen weisen bis heute kein Schriftsystem auf. Vor ca. 200 000 – 40 000 Jahren entwickelten Hominiden Formen der lautsprachlichen Verständigung, die Entstehung von Schriftsystemen schätzt man auf die Zeit um ca. 5000 - 3000 v. Chr. Lautsprachliche Verständigung mittels Sprechen, Zeigen, Hören und Sehen war also lange Zeit die einzige Möglichkeit zur Kommunikation. (Becker-Mrotzek, 2003: 70) Bis hin zur vorindustriellen Gesellschaft wurde die Schriftsprache nur von einem kleinen Teil der Bevölkerung beherrscht. Erst die allgemeine Schulpflicht in Deutschland seit dem 19. Jahrhundert trug dazu bei, dass die Mehrheit der Bevölkerung lesen und schreiben lernte. (Fiehler, 2009) Heute wird der Anteil der illiteraten Erwachsenen noch immer auf weltweit 774 Millionen geschätzt. (EFA Global Monitoring Report Team, 2008) In Deutschland wird derzeit von ungefähr vier Millionen funktionellen Analphabeten ausgegangen, Menschen also mit erheblichen Schwächen im Lesen und Schreiben. (Pressemitteilung BMBF, 2007) Obwohl in der täglichen Kommunikation gesprochene Sprache zur geschriebenen Sprache in einem Verhältnis von 95% zu 5% steht und Hörverstehen die dominierende Kompetenz und Voraussetzung für andere Sprachtätigkeiten ist, wächst dennoch die gesellschaftliche Bedeutung von Schriftlichkeit. (Thaler, 2007) Auch der überwiegende Teil der schulischen Instruktionszeit entfällt auf die Kompetenzbereiche „Schreiben", „Lesen" und „Sprechen". Dem Kompetenzbereich „Zuhören" wird im schulischen Bereich bislang kaum Aufmerksamkeit gewidmet (vgl. Kapitel 4.4.3. *Hörverstehen in den Rahmenplänen der Länder*). Ohne Zuhörkompetenzen ist eine Beteiligung an der Unterrichtskommunikation (und sei es auch nur als Zuhörer) nicht möglich. Gründe für die Vernachlässigung des Kompetenzbereichs „Zuhören" im Unterricht liegen unter anderem in der schriftsprachlich orientierten Unterrichtstradition sowie in den besonderen Merkmalen gesprochener Sprache und mündlicher Kommunikation.

Im Folgenden soll kurz die Rolle des Zuhörens im Rahmen mündlicher Kommunikation verdeutlicht werden, um in diesem Zusammenhang auch auf die Besonderheiten gesprochener Sprache einzugehen. Dabei wird der Begriff „Zuhören" wie folgt vom Begriff „Hören" abgegrenzt: „Zuhören" umfasst die Selektion, Organisation und Integration von akustisch vermittelter Information sprachlicher oder nichtsprachlicher Art. Im Gegensatz zum „Hören", werden Zuhörprozesse bewusst vom Zuhörer durch die Intention zu selektieren gesteuert. Während

des Zuhörens wird neue Information in existierende kognitive Strukturen integriert. Dieser Vorgang setzt die kognitive und motivationale Aktivität des Rezipienten voraus und erfordert den Einsatz energetischer Aktiviertheit. (Imhof, 2003: 15ff) Rost fasst dies folgendermaßen zusammen: „Listening is primarily a cognitive activity, involving the activation and modification of concepts in the listener's mind." (Rost, 2002: 62)

Nach Becker-Mrotzek wird unter mündlicher Kommunikation „prototypisch die aktuale Verständigung zwischen mindestens zwei Aktanten verstanden (…), die sich in einem gemeinsamen Sprechzeit-Raum befinden, sodass die Kommunikation auf der Grundlage akustischer und visueller Wahrnehmungen verbal und nonverbal (körperlich) erfolgen kann." (Becker- Mrotzek, 2009: 79) Verständigung erfolgt multimodal, d. h. sie umfasst in der Regel gleichzeitig mehrere Teilbereiche der mündlichen Kommunikation, und zwar nonverbale, wahrnehmungs- und inferenzgestützte und verbale Kommunikation. (Rost, 2002) Jeder Gesprächsbeitrag hat innerhalb eines Gesprächs eine bestimmte Funktion und ein Gespräch kann als ein Komplex von miteinander vernetzten und interagierenden Aufgaben gesehen werden. Die Gesprächsbeteiligten erfüllen diese Aufgaben gemeinsam, indem sie meist unbewusst prosodisch, nonverbal und/oder verbal eine Beziehung zueinander herstellen und sich gegenseitig beeinflussen und steuern. Während eines Gesprächs bauen die jeweiligen Zuhörer ständig Projektionen auf. Diese Projektionen sind inhaltliche aber auch grammatische Vermutungen über die weitere Gestalt der Sätze, beispielsweise über die Valenz bestimmter Verben oder die Erwartung einer Vergleichsgröße bei einem Adjektiv im Komparativ. Die jeweiligen Sprecher können mit dem tatsächlichen Sprachbeitrag diese Projektionen erfüllen oder sie durch einen Satzabbruch o. ä. enttäuschen. (vgl. Hennig, 2006) Außer im Falle eines Selbstgesprächs ist ein Gespräch als geordnete Abfolge von Gesprächsbeiträgen also stets das kollektive Produkt aller Beteiligten, da die Gesprächsbeteiligten durch ihre Beiträge ständig auf allen Ebenen des Handelns interagieren. (Fiehler, 2005)

Die beschriebene Interaktivität, der zum Teil sehr schnelle Wechsel der Sprecherrollen und die Verankerung des Sprachbeitrags in der laufenden Interaktion haben eine geringe Verarbeitungszeit des Gesagten und auch eine geringe Vorausplanbarkeit von weiteren Beiträgen für Sprecher und Zuhörer zur Folge. (Fiehler, 2005) Mündliche Kommunikation erfordert von den Beteiligten also hohe Aufmerksamkeit, da sie von vielen Faktoren bestimmt wird, die ständig überprüft werden müssen: Jeder zukünftige Redebeitrag wird durch das eben Gesagte und die Rückmeldungen des Zuhörers beeinflusst. Dabei muss unterschieden werden zwischen der Äußerungsinformation durch den Sprecher und der Kontextinformation, die vom Zuhörer bereitgestellt wird. Die Kontextinformationen stammen aus bereits Gesagtem, der gemeinsamen Situation und dem Weltwissen des Zuhörers.

Die wissenschaftliche Sprachreflexion beschäftigte sich lange Zeit fast ausschließlich mit geschriebenen Texten und entwickelte ein linguistisches Kategoriensystem für die Analyse von Schriftsprache (z. B. Duden, 2005; Engel, 1988; Helbig & Buscha, 2001), das es in vergleichbarer Form für die gesprochene Sprache nur in Grundzügen gibt (z. B. Hennig, 2006; Ágel & Hennig, 2007). Ein Grund dafür ist, dass das gesellschaftliche Sprachbewusstsein schriftsprachlich dominiert ist. Auch die Unsichtbarkeit und mangelnde Dauerhaftigkeit mündlicher Kommuni-

kation ließen Untersuchungen lange Zeit nur sehr eingeschränkt zu. Erst reproduzierbare Aufnahmen und Transkriptionen seit den 60er Jahren des 20. Jahrhunderts machten Forschung in einer hinreichenden Detailtiefe möglich. (vgl. Fiehler, 2009; Deppermann et al. 2006)

Unter gesprochener Sprache werden „die verbalsprachlichen Anteile der mündlichen Kommunikation einschließlich aller bedeutungstragenden stimmlichen und prosodischen Erscheinungen" (Fiehler, 2009: 33) verstanden. Im Vergleich zur geschriebenen Sprache zeichnet sich gesprochene Sprache durch verschiedene Besonderheiten aus, wie auf Kontextinformationen basierende strukturelle Mittel. Auf diese Merkmale wird im Folgenden in Anlehnung an Fiehler et al. (2004) genauer eingegangen. Fiehler et al. (2004: 56) benennen elf Merkmale gesprochener Sprache bzw. Rahmenbedingungen auf gesprochene Sprache:

1) Kurzlebigkeit/Flüchtigkeit
2) Zeitlichkeit
3) Anzahl und Größe der Parteien
4) Kopräsenz der Parteien und Gemeinsamkeit der Situation
5) Wechselseitigkeit der Wahrnehmung
6) Multimodalität der Verständigung
7) Interaktivität
8) Bezugspunkt der Kommunikation
9) Institutionalität
10) Verteilung der Verbalisierungs- und Thematisierungsrechte
11) Vorformuliertheit von Beiträgen

Gesprochene Sprache ist eine kontinuierliche Modulation von Schallwellen, die durch Pausen unterbrochen wird und flüchtig und schnell vergänglich ist. (Klein, 1985) Mündliche Beiträge können im Unterschied zu schriftsprachlichen Texten nicht beliebig oft gelesen oder überarbeitet werden und die Ausführungen sind aufgrund ihrer Flüchtigkeit nicht mehr zurückzunehmen. Deshalb kommt es häufig zu nachträglichen Äußerungsbearbeitungen, wie der Korrektur von Fehlern (phonologische Vertauschungen, falsche Wortwahl etc.), Abbrüchen, Paraphrasen/Reformulierungen und Neuansätzen, aber auch zur Bearbeitung von Wortfindungsschwierigkeiten und anderen Formulierungsproblemen. Die beschriebenen Prozesse dienen sowohl der Schaffung von Eindeutigkeit als auch der Präzisierung, Spezifizierung, der inhaltlichen Abschwächung/Distanzierung oder der Selbstkorrektur (Reparatur). Um dem Gesprächspartner anzuzeigen, dass eine Äußerung nachträglich bearbeitet werden soll, wird zu entsprechenden kommunikativen Verfahren und sprachlichen Mitteln gegriffen, insbesondere prosodischen Elementen. Der Zuhörer muss diese sprachlichen Mittel und kommunikativen Verfahren erkennen und deuten können und verfügt seinerseits über Mittel und Verfahren, um dies auszudrücken. Für die Kommunikation sind Kurzlebigkeit und Flüchtigkeit auch deshalb problematisch, da die Äußerungen im Gespräch nicht fixiert vorliegen und die Gesprächspartner auf ihre Gedächtnisleistung angewiesen sind. Um das gegenseitige Verständnis so weit wie möglich zu erleichtern und zu sichern, können Struktur und Funktion von Äußerungen z. B. durch Redundanz und Metakommunikation transparent gestaltet werden. (vgl. Fiehler 2005) Die Flüchtigkeit ist auch Ursache dafür, dass gesprochene Sprache weniger leicht in

empirische Tests umzusetzen ist als geschriebene Sprache. Da weiterhin Lesekompetenzen als zentrale Voraussetzungen für das selbst regulierte Lernen mit Texten gelten, fiel bisher der Überprüfung des Leseverständnisses, beispielsweise in der PISA-Studie, eine größere Rolle zu. (vgl. Deutsches PISA-Konsortium, 2001, 2002, 2004, 2005)

Mündliche Verständigung erstreckt sich über die Zeit hinweg. Der Zuhörer ist auf die Sprechgeschwindigkeit des Sprechers angewiesen. Er muss die Botschaft unmittelbar in der vom Sprecher artikulierten Geschwindigkeit verarbeiten, ohne sich ihr später nochmals zuwenden zu können. (Buck, 2001: 6) Dabei spielt die Aufnahmekapazität des Arbeitsgedächtnisses eine wichtige Rolle (vgl. Kapitel 2.1.2. *Speicherung von Wissen*). Die zu übermittelnden Informationen müssen vom Sprecher und vom Zuhörer in Einheiten portioniert werden. Dem Gesprächspartner muss signalisiert werden, wo diese Einheiten beginnen und enden, von welchem Typ sie sind und welche Beziehung zwischen den einzelnen vorhergehenden und nachfolgenden Einheiten bzw. Einheiten über- bzw. untergeordneten Formats besteht. Dies geschieht durch Sprechpausen und Intonation. (vgl. Fiehler, 2005: 1191)

Bei der Beschreibung von mündlicher Kommunikation spielt auch der Interaktionsgrad zwischen den Gesprächsbeteiligten eine Rolle. (Buck, 2001: 12) Im Unterricht sprechen meist der Lehrer und ein Schüler oder auch zwei Schüler miteinander. Obwohl der Rest der Klasse diesen Gesprächen zuhört, ist ihre Rolle doch nicht passiv, da sie sich jederzeit in das Gespräch einschalten könnten. Im Fall von Rundfunk- oder Fernsehübertragungen sind die Zuhörer hingegen nicht in der Situation des kommunikativen Geschehens gegenwärtig und haben damit keine Möglichkeit, sich in irgendeiner Form an der Kommunikation zu beteiligen. (Fiehler, 2005: 1191f) Auch die Schüler in der Testsituation können sich nicht an der von Band vorgespielten Kommunikation der Hörverstehensaufgaben beteiligen und haben keine Möglichkeit, in diese Gespräche einzugreifen oder nachzufragen. In der Testsituation stellen die Schüler also eine reine Zuhörergruppe dar.

Gesprochene Sprache unterscheidet sich von der geschriebenen durch den Grad der Situationsgebundenheit. In der Regel fallen im Gespräch Sprechzeit und –ort sowie Hörzeit und -ort zusammen. (Klein, 1985: 15) Die gemeinsame Gesprächsumgebung stellt den Rahmen für die Wahrnehmungen und Handlungen der anwesenden Gesprächsbeteiligten dar. Während des Gesprächs nehmen die Gesprächspartner anwesende Personen sowie fokussierbare Objekte ihrer Umgebung wahr und sie nehmen auch in einer gemeinsamen Gesprächssituation unterschiedliche räumliche und mentale Positionen ein. Durch diese Wechselseitigkeit der Wahrnehmung wird Wissen vermittelt, das die Anpassung der Gesprächsbeiträge an vorher Gesagtes oder an die Gesprächsparteien ermöglicht. Dabei ist jedes Verhalten sozial mehrdeutig und muss vom Gesprächspartner interpretiert werden. (Hennig, 2006: 91)

Die gemeinsame Situation liefert Bezugspunkte für deiktische Ausdrücke, die zur Referenz auf den Sprecherstandpunkt benutzt werden. Aufgrund der Situationsgebundenheit sind für die gesprochene Sprache Redekonstellation und das Verhältnis der Sprechenden zueinander (z. B. Alter, Rang, Geschlecht, Dialektraum, Einstellung und Verhalten) wichtiger als für die geschriebene Sprache. (Fiehler, 2005: 1192) Die richtige Interpretation von Deixis (also dem

Verweis mit sprachlichen Ausdrücken auf Ort, Personen oder Zeit der Gesprächssituation oder auf eine andere Stelle im gleichen Text), Redekonstellation und dem Verhältnis der Sprechenden ist die Grundlage für das gegenseitige Verständnis der Gesprächsbeteiligten. Im Falle der Testsituation könnten genau diese Aspekte Verständnis erschwerend wirken, da die Schüler nicht an der Gesprächssituation beteiligt sind und z. B. die Bezugspunkte für deiktische Ausdrücke erahnen müssen. Gerade bei modernen, technikbasierten Formen der Kommunikation, wie einer telefonischen Konferenzschaltung, ist Kopräsenz der Parteien nicht immer gegeben. Durch den Einsatz technischer Geräte kann dennoch eine reduzierte, gemeinsame Situation herbeigeführt werden. Meist ist hier jedoch die Wechselseitigkeit der Wahrnehmung eingeschränkt und erschwert dadurch die Kommunikation. (Fiehler, 2005: 1192f)

Die dargestellten Rahmenbedingungen mündlicher Kommunikation führen dazu, dass gesprochene Sprache distinkte Besonderheiten aufweist, die sich von geschriebener Sprache unterscheiden. Die Schallwellen der gesprochenen Sprache werden durch Tonhöhe, Lautstärke, Klangfarbe sowie durch Sprechgeschwindigkeit und Pausen moduliert. Diese prosodischen Merkmale existieren für die geschriebene Sprache nicht. (Klein, 1985) Der Zuhörer ist abhängig von der Präsentationsgeschwindigkeit und kann sich nicht in seinem eigenen Tempo mit dem Stimulus auseinandersetzen. Häufig werden zwei bis drei Wörter pro Sekunde bzw. drei bis fünf Silben pro Sekunde geäußert. (vgl. Rost, 2002: 30) Da deshalb die Sprecher ihre Botschaften in Echtzeit planen und organisieren, liegt der Fokus der Aufmerksamkeit meist auf der gedanklichen Stimmigkeit. Aus diesem Grund weist gesprochene Sprache häufig ungrammatische Formen oder phonologische Irregularitäten auf. Es handelt sich dabei z. B. um phonologische Vertauschungen, falsche Wortwahl, Wortfindungsschwierigkeiten oder syntaktische Abweichungen. Diese Fehler werden häufig nicht bemerkt, da die Zuhörer im Wesentlichen versuchen, den Inhalt der Äußerung zu erfassen und deshalb weniger auf derartige Irregularitäten achten. (vgl. Richards, 1983)

1.1.1. Inhaltliche Merkmale

Inhaltlich zeichnet sich gesprochene Sprache dadurch aus, dass das Gesagte nicht widerrufen werden kann, es situationsbezogen und kontextsensitiv ist. Gesprächsinhalte werden meist kooperativ konstruiert, da mehrere Gesprächsparteien (z. B. Sprecher und Zuhörer) an der Kommunikation teilhaben. Die Gesprächspartner gehen aufeinander ein, indem sie beispielsweise bei gemeinsamem Wissen über ein Thema schneller sprechen, Informationen lediglich andeuten oder redundante Begriffe weniger deutlich aussprechen. (Buck, 2001) Je nach Gesprächsform sind eine größere Formelhaftigkeit, eine stärkere Bildhaftigkeit des Gesagten sowie ein höherer Anteil an Bewertungen und Intensivierungen zu vermerken. Im Unterschied zur schriftlichen Sprache sind mit gesprochener Sprache lautliche Nuancierung und damit der Ausdruck von Emotionen möglich. Durch nonverbale Signale wird zusätzlich der Grad an Formalität und das Verhältnis der Gesprächsbeteiligten zueinander ausgedrückt. (Richards, 1983) In der geschriebenen Sprache wird die Beziehung zwischen Propositionen in der Regel durch syntaktische Mittel ausgedrückt. In der gesprochenen Sprache kann es jedoch vorkommen, dass die Propositionen unverbunden nebeneinander stehen und nur durch den Kontext miteinander verbunden werden. (Buck, 2001)

1.1.2. Lautliche und paralinguistische Merkmale

Beim Erkennen gesprochener Sprache müssen einzelne Phoneme aus dem Lautstrom ausgegrenzt werden. Unter einem „Phonem" versteht man die kleinste sprachliche Einheit, deren Veränderung i. d. R. zu einer Bedeutungsänderung führt. Die von einem Sprecher konkret produzierten Exemplare eines Phonems nennt man „Laute". Die einzelnen Sprecher unterscheiden sich bei der Produktion derselben Phoneme, wenn sie beispielsweise mit Akzent oder Dialekt sprechen. Um Bedeutungen so effizient wie möglich zu übermitteln, werden unwichtige Begriffe von den Sprechern bei der Artikulation z. T. vernachlässigt und verzerrt oder abgeschnitten ausgesprochen. Wichtige Begriffe werden dagegen sorgfältig artikuliert und besonders betont. (Buck, 2001: 4f)

Typisch für die gesprochene Sprache sind: Wegfall von unbetonten Vokalen und Konsonanten, Assimilation, Vereinfachung von Konsonantensequenzen, Verschmelzungen und Abschwächungen. Eine Schwierigkeit des Zuhörens ist es also, Lauten die entsprechenden Phoneme zuzuordnen. Diese lautlichen Besonderheiten betreffen jedoch nicht nur einzelne Wörter. Da der Sprachfluss kontinuierlich verläuft, sind die Phoneme nicht in gleicher Weise voneinander getrennt, wie die Grapheme/Wörter in der geschriebenen Sprache. Dies kann dazu führen, dass Wortgrenzen ggf. nicht erkannt werden und Verständnisschwierigkeiten auftreten. (Anderson, 2001: 57f) Auch die Koartikulation der Phoneme kann bei der Sprachwahrnehmung zu Schwierigkeiten führen. Während der Vokaltrakt einen Laut produziert, verändert er sich bereits in seiner Form in die Richtung des als nächstes zu produzierenden Lauts. Bei der Artikulation der Phoneme tritt also meist eine gewisse Überlappung auf und das Klangmuster, das für ein bestimmtes Phonem produziert wird, wird durch die Phoneme in der Umgebung dieses Phonems mitbestimmt. (Imhof, 2003: 25f)

Zusätzlich zur lautlichen Wahrnehmung des Sprachflusses muss der Zuhörer auch paralinguistische Merkmale wie Prosodie, Lautstärke oder Tempo verarbeiten. Im Kommunikationsprozess dienen die prosodischen Elemente (z. B. Stimmsenkung oder -erhöhung, gefüllte oder ungefüllte Pausen) der Segmentierung und der Kontextualisierung des Gesprochenen. Sie leisten einen wichtigen Beitrag im Bereich der Organisation des Beitragswechsels und der Gesprächsorganisation, indem mit ihnen das Ende von Äußerungen, die Relevanz eines Beitrags oder die Gesprächsmodalität (z. B. scherzhaft) angezeigt werden kann. Zusätzlich erfüllen prosodische Gestaltungsmittel personenbezogene Funktionen wie das Anzeigen von Emotionen oder die Intensität der Beteiligung am Gespräch. (Fiehler, 2005: 1205ff) Auch Mehrdeutigkeiten können bei gesprochener Sprache häufig erst durch entsprechende Intonation oder den Kontext aufgelöst werden. (Imhof, 2003: 30ff)

1.1.3. Lexikalische Merkmale

Lexikalische Besonderheiten der gesprochenen Sprache sind beispielsweise deiktische und anaphorische Ausdrücke. Deiktische Ausdrücke, wie „ich", „du", „hier", „jetzt", etc. stehen nicht lexikalisch fest, sondern ergeben sich situativ und können sich mit jedem Sprecherwechsel ändern. Aus diesem Grund ist für alle deiktischen Ausdrücke die Beziehung zur Sprechsituation notwendig. Erst in der konkreten Sprechhandlung wird deutlich, worauf sich die Ausdrücke beziehen. In schriftlichen Texten müssen deiktische Ausdrücke aufgrund ihrer Situations-

gebundenheit sprachlich expliziert werden. (vgl. Graefen & Liedke, 2008: 94ff) Anaphorische Ausdrücke, wie „er", „sie", „es", „dann", „dahinter", „vorher" etc., basieren auf Vorgängerinformation, indem sie vorher Genanntes wieder aufnehmen. (ebd., 20)

Gesprochene Sprache zeichnet sich außerdem durch einen höheren Anteil an Funktionswörter im Gegensatz zu Inhaltswörtern sowie durch häufig gebrauchte Wörter des Wortschatzes aus. (Rost, 2002: 31) Funktionswörter sind Wörter, die nur innerhalb des Systems Sprache bedeutungstragend sind, wie Präpositionen oder Konjunktionen. Inhaltswörter sind bedeutungstragende Wörter wie Nomen, Verben, Adjektive und Adverbien.

1.1.4. Syntaktische Merkmale

Während die Organisationseinheit der geschriebenen Sprache der Satz ist, ist sie für die gesprochene Sprache die Äußerung. Äußerungen sind meist einfache Beifügungen oder Zusätze. Sowohl in der Planung als auch in der Ausführung der gesprochenen Sprache scheinen sie die zentrale Komponente zu sein und werden meist nur durch Konjunktionen miteinander verbunden. (Richards, 1983)

Hennig (2006) nennt als syntaktische Einheiten gesprochener Sprache unter anderem Ellipse (syntaktisch notwendige Satzteile fehlen), Anakoluth (Satzabbrüche) und Nähezeichen. Besteht geteiltes Wissen über Struktur und Komponenten einer Handlung, so bieten sich Ellipsen zur Redundanzvermeidung an. Ellipsen werden vom Sprecher häufig dann eingesetzt, wenn an die noch präsente syntaktische Struktur der Vorgängeräußerung angeknüpft werden kann. Verbalisiert werden muss in beiden Fällen nur die Information, die dem Zuhörer neu ist. Auch Interjektionen, z. B. Aufforderungen, Warnungen, Ausrufe u. ä. können als sprachliche Handlung häufig kommunikativ vollständig ausgeführt werden, ohne dass die verwendeten sprachlichen Mittel der Satzform unterliegen. (Fiehler, 2005) Dabei werden Ellipsen als Konstruktionen verstanden, die „grammatisch unvollständig, aber kommunikativ vollständig" (Hennig, 2006: 198) sind, d. h. obwohl aus grammatischer Perspektive Teile fehlen, sind Ellipsen dennoch durch den sprachlichen Kontext verstehbar. Im Laufe der Versprachlichung kann jedoch auch der Fall eintreten, dass Zuhörer-Projektionen nicht erfüllt und begonnene syntaktische Konstruktionen nicht oder anders zu Ende geführt werden. Der Abbruch einer Äußerung, die damit als nicht beendetes Fragment stehen bleibt, ist das geläufigste Unterscheidungsmerkmal von Anakoluthen zu Ellipsen. Alle sprachlichen Ausdrücke, die nicht als Satz, als Anakoluth oder Ellipse qualifizieren, werden als „Nähezeichen" bezeichnet. Dazu gehören beispielsweise Responsive, Engführungssignale, Rederechtssignale, Zögerungssignale sowie die Operatoren in Operator-Skopus-Strukturen. Sie bauen keine syntaktischen Projektionen auf. (vgl. Hennig, 2006: 101)

Ergänzend benennt Fiehler (vgl. 2009: 48ff) weitere spezifische syntaktische Strukturen, die entweder ausschließlich oder überwiegend in der gesprochenen Sprache auftauchen: Referenz- Aussage-Strukturen und Verberststellung. Referenz-Aussage-Strukturen weisen ein Referenzobjekt auf und eine Einheit, mit der eine Aussage über das Referenzobjekt gemacht wird. Diese Einheit enthält häufig ein Element, das auf das Referenzobjekt zurückweist, wie beispielsweise im Satz: „Das Wetter, das ist heute ganz ausgezeichnet." Beim Merkmal

„Verberststellung“ steht in Aussagesätzen ein Prädikat an erster Stelle im Satz: z. B. „Sind interessante Themen drin“.

1.2. Text und Diskurs

Bei den IQB-Stimuli handelt es sich nur teilweise um Diskurse, um Sprachbeiträge, die während der mündlichen Kommunikation gemeinsam von den Gesprächsbeteiligten erstellt werden. Ein Teil der Stimuli sind auch Texte. Die Begriffe „Diskurs“ und „Text“ werden im Folgenden kurz vorgestellt: Es gibt zahlreiche Versuche sprachliche Äußerungen entsprechend ihrer Konzeption zu klassifizieren. Im angloamerikanischen Raum wird der Begriff „discourse“ für alles Gesprochene und Geschriebene verwendet. Die beiden Modalitäten werden durch die Ergänzungen „spoken discourse“ (gesprochener Diskurs) und „written discourse“ (geschriebener Diskurs) voneinander unterschieden (vgl. Graefen & Liedke, 2008: 250). Ochs (1979 zitiert nach Buck, 2001: 9) unterscheidet „spoken discourse“ weiter in „unplanned discourse“, der spontan und ungeplant produziert wird, und „planned discourse“, der häufig auch schriftlich vorbereitet wurde.

In der deutschsprachigen Linguistik spielt der Begriff „Text“ aufgrund der wissenschaftstheoretischen Tradition eine wichtigere Rolle, aber auch hier wird zwischen „geschriebenen“ und „gesprochenen Texten“ differenziert (vgl. Graefen & Liedke, 2008: 250). Tannen (1982, 1985) unterscheidet beispielsweise zwischen „oral texts“ und „literate texts“, wobei „oral texts“ mehr Merkmale gesprochener Sprache tragen und „literate texts“ mehr Kennzeichen geschriebener Sprache, insbesondere darstellender Prosa, aufweisen. Koch und Oesterreicher (1985) sowie Ágel und Hennig (Hennig, 2006) beschreiben sprachliche Produktionen bezüglich der Wahl des sprachlichen Registers auf einem Kontinuum zwischen Medium und Konzeption. Dabei schlagen Koch und Oesterreicher vor, Produktionen in der Sprache der Nähe als „Diskurs“ zu bezeichnen und Produktionen in der Sprache der Distanz als „Text“. (Koch & Oesterreicher, 1985: 22)

In der Funktionalen Pragmatik wird unter dem Begriff „Text“ das sprachliche Handeln bei Überlieferungen verstanden. Nach Ehlich (1983) bezeichnet der Begriff „Text“ Sprechhandlungen, die aus ihrer primären Situation herausgelöst wurden und über Raum und Zeit hinweg transportiert werden, um für weitere Sprechsituationen gespeichert zu werden. Ein wichtiges Kennzeichen von Text ist also, dass er zum Zweck der Überdauerung von Sprache produziert wurde. Die Gesprächspartner sind nicht zur gleichen Zeit am selben Ort, was eine Zerdehnung der ursprünglichen Gesprächssituation zur Folge hat. Die Äußerungen des Sprechers müssen fixiert werden, damit eine spätere Rezeption möglich wird.

Mit dem Begriff „Diskurs“ wird in der Funktionalen Pragmatik eine unmittelbare Kommunikationssituation bezeichnet, bei der alle Gesprächspartner in Echtzeit miteinander kommunizieren. Nach Becker-Mrotzek und Meier bezeichnen „Diskurse (…) Folgen von Sprechhandlungen mehrerer Beteiligter in einer unmittelbaren, gemeinsamen Sprechsituation, die entweder in direkter Interaktion, d. h. face-to-face, oder medial vermittelt hergestellt ist.“ (Becker- Mrotzek & Meier, 2002: 20) Becker-Mrotzek verwendet den Begriff „Diskurs“ als Oberbegriff für diverse mündliche Kommunikationsformen, wobei die Ausdrücke Gespräch bzw. Konversation mit

ähnlicher Bedeutung gebraucht werden. Entscheidend bei Diskursen ist, dass sie primär auf einer gleichzeitigen Anwesenheit von Sprecher und Zuhörer in einem gleichzeitigen Sprechzeit-Raum basieren. Die gleichzeitige Anwesenheit kann dabei auch medial z. B. durch ein Telefongespräch hergestellt sein. Durch die Produktion eines Gesprächsbeitrages in einem gleichzeitigen Sprechzeit-Raum, tragen Diskurse i. d. R. mehr Merkmale gesprochener Sprache als vertonte Texte.

In der Funktionalen Pragmatik können mündliche Texte und mündliche Diskurse, aber auch schriftliche Texte und schriftliche Diskurse auftreten (vgl. Graefen & Liedke, 2008: 250f) Tabelle II-1.2. gibt einen Überblick über die Auftretensmöglichkeiten:

Tabelle II-1.2.: Übersicht über „Text" und „Diskurs"

	Medium Mündlichkeit oral - auditiv	**Medium Schriftlichkeit graphisch - visuell**
zerdehnte Kommunikationssituation (Sender – Empfänger über Zeit/Raum hinweg getrennt)	mündlicher Text (z. B. vorgelesenes Märchen)	schriftlicher Text (z. B. Brief)
unmittelbare Kommunikationssituation (Sender - Empfänger in direktem Kontakt)	mündlicher Diskurs (z. B. Gespräch)	schriftlicher Diskurs (z. B. Internet-Chat)

In Anlehnung an die Benennungskonventionen der Funktionalen Pragmatik werden in dieser Arbeit Stimuli als „Texte" bezeichnet, die auf einer schriftlichen Grundlage basieren und erst nachträglich vertont wurden. Stimuli, die unter einer gleichzeitigen Anwesenheit der Sprecher und Zuhörer entstehen, werden als „Diskurse" bezeichnet, auch wenn sie nachträglich verschriftlicht wurden.

2. Hörverstehen – eine psycholinguistische Perspektive

Merkmale, die einen Einfluss auf die Schwierigkeit von Stimuli und Items haben, ergeben sich jedoch nicht nur aus den Besonderheiten der gesprochenen Sprache, sondern resultieren auch aus den kognitiven Prozessen der Informationsverarbeitung und der Speicherung von Wissen, die im Rahmen des Zuhörens ablaufen (Kapitel 2.1.). Aus diesem Grund werden diese Prozesse im folgenden Kapitel genauer beschrieben. Dabei wird auch darauf eingegangen, welche Anforderungen sie an den Zuhörer stellen und welche typischen Schwierigkeiten auftreten können. Zunächst spielen Wahrnehmung und Aufmerksamkeit hier eine wichtige Rolle, denn Reize, denen nicht unmittelbar Aufmerksamkeit zugewendet wird, werden nicht weiter verarbeitet.

Zur Informationsverarbeitung im Allgemeinen gibt es verschiedene Theorien, von denen drei für diese Arbeit relevante vorgestellt werden (Kapitel 2.1.1.): Die „Theorie des Kurzzeitgedächtnisses" (Kapitel 2.1.1.1.) legt ihren Schwerpunkt auf die eingeschränkte Speicherkapazität des Kurzzeitgedächtnisses. Eintreffende Reize werden nacheinander in verschiedenen Gedächtnissystemen verarbeitet. Eine parallele Bearbeitung ist aufgrund der begrenzten

Speicherkapazität der Systeme nur sehr eingeschränkt möglich. Überlastung der Systeme führt zu einem sog. „Flaschenhals", bei dem nicht mehr alle eintreffenden Informationen verarbeitet werden können. Auch die zweite Theorie, Baddeleys Modell des Arbeitsgedächtnisses (Baddeley, 2002) (Kapitel 2.1.1.2.), fokussiert auf dem Kurzzeitgedächtnis, das nach Baddeley ein Arbeitsgedächtnis mit mehreren Hilfssystemen enthält. Diese sind in ihrer Aufnahmekapazität beschränkt. Unterschiede in der Informationsverarbeitungsfähigkeit werden mit der unterschiedlichen Leistungsfähigkeit des Arbeitsgedächtnisses erklärt. Dieses Konzept wurde von Just und Carpenter (1992) in ihrer „Capacity Theory of Comprehension" aufgegriffen (Kapitel 2.1.1.3.). Sie beschreibt die Verarbeitung und Speicherung von Informationen abhängig von der Arbeitsgedächtniskapazität.

Die Speicherung der eingetroffenen Informationen ist erst möglich, wenn ihnen Bedeutung zugeschrieben wurde und sie in Sinnzusammenhänge gebracht wurden. Entsprechend gibt es unterschiedliche Theorien, in welcher Form Informationen sinnhaft gespeichert werden. So wird angenommen, dass die Speicherung von Informationen in Form von mentalen oder propositionalen Repräsentationen erfolgt. Andere Theorien gehen hingegen davon aus, dass die Informationen im Rahmen von Prototypen oder Schemata bzw. Skripten gespeichert werden. Die Vernetzung des gespeicherten Wissens wird als semantisches Netzwerk gesehen.

In einem nächsten Schritt wird genauer auf verschiedene Theorien zu den spezifischen Prozessen des Sprachverstehens eingegangen (Kapitel 2.2.). Bei der Theorie, es handele sich bei der Sprachverarbeitung um bottom-up und top-down Prozesse, steht im Vordergrund, dass Sprachverarbeitung auf unterschiedlichen kognitiven Ebenen angesiedelt ist (Kapitel 2.2.1.). Die involvierten Prozesse laufen nach neueren Erkenntnissen gleichzeitig und interagierend ab. Anderson (1995) geht hingegen davon aus, dass Sprachverarbeitung in unterschiedlichen Phasen, nämlich Wahrnehmung, Parsing und Anwendung, abläuft (Kapitel 2.2.2.). Die Dynamik des Verstehensprozesses und die Bedeutung des Kontextes sowie des Vorwissens des Zuhörers/Lesers werden besonders im Modell von Kintsch (1998) herausgestellt (Kapitel 2.2.3.). Zentral darin ist die Neuaufnahme von Propositionen und ihre Verknüpfung mit bereits vorhandenen Propositionen.

Ergänzend zu diesen Theorien der Sprachverarbeitung werden zwei weitere kognitive Prozessmodelle vorgestellt, die sich explizit auf das Lösen von Multiple-Choice-Items beziehen (Kapitel 2.2.4.). Beide Modelle fokussieren weniger auf den Prozessen der Sprachverarbeitung als auf den bei der Itembeantwortung involvierten Prozessen. Das Modell von Embretson und Wetzel (1987) beschreibt die für die Itemschwierigkeit relevanten kognitiven Charakteristika von Items, während das Modell von Sheehan und Ginther (2001) den Schwerpunkt auf die bei der Bearbeitung von MC-Items ablaufenden Prozesse legt.

2.1. Verarbeitung eintreffender Informationen

Bevor eintreffende Informationen verarbeitet werden können, müssen sie vom Zuhörer wahrgenommen werden. Der Begriff „Wahrnehmung" bezeichnet die „Interpretation und Bedeutungszuschreibung der durch die verschiedenen Sinne aufgenommenen Informationen." (Gerstenmaier, 1995: 263) Zur Wahrnehmung gehört das einfache Registrieren sensorischer

Informationen aber auch die Fähigkeit zur Mustererkennung, durch welche die Gliederung des Sprachflusses erkannt wird. (Cassells & Green, 1995: 69)

Wahrgenommene Informationen, denen nicht unmittelbar Aufmerksamkeit zugewendet wird, gehen verloren. Diese Reduktionsstrategie wird angewendet, da das Gehirn nur begrenzt über Verarbeitungskapazität verfügt. Unter „Aufmerksamkeit" versteht man die „Zuteilung kognitiver Ressourcen zu laufenden Prozessen." (Anderson, 2001: 460) Aufmerksamkeit ist stark an das Bewusstsein gebunden und wie dieses ein einheitliches System. Sie kann unterschieden werden in auditive und visuelle Aufmerksamkeit sowie selektive und automatisch ablaufende Prozesse. (Reddy, 1995: 93) Durch selektive Aufmerksamkeit werden Reize ausgefiltert, damit die Aufmerksamkeit in eine bestimmte Richtung gelenkt werden kann. „Selective attention is the process of focusing processing resources (electrical activity in the brain) onto one idea, and allowing the processing of other ideas or thoughts to terminate." (Rost, 2002: 15)

Welche Faktoren bewirken, dass einem Reiz Aufmerksamkeit zugewendet wird? Aufmerksamkeitszuwendung erfolgt häufig unbewusst, sie wird aber von Signalfaktoren und von motivationalen Faktoren beeinflusst: Signalfaktoren beziehen sich auf die physikalische Anordnung des Versuchssettings, wie die Intensität oder die Dauer eines dargebotenen Reizes. Motivationale Faktoren betreffen die jeweilige Versuchsperson und können beispielsweise eine Leistungsrückmeldung oder die Anwesenheit anderer im Raum sein. (Hayes, 1995: 19f) Nach Heckhausen (1989) beeinflusst die Motivation einer Person auch die dazugehörigen kognitiven Prozesse. In Bezug auf die Prozesse des Zuhörens bedeutet das, dass sich der Umfang und die Qualität der Informationsverarbeitung bei motivierten und weniger motivierten Zuhörern unterscheiden.

2.1.1. Informationsverarbeitung

Da eine Gliederung in diskrete Einheiten bei der gesprochenen Sprache nicht in gleicher Weise wie bei Schrift vorliegt, sind auch bei der Reizverarbeitung unterschiedliche Prozesse zu beobachten. Dabei muss zwischen der akustischen und der auditorischen Verarbeitung von Information unterschieden werden. Akustische Verarbeitung bezeichnet die Wahrnehmung und Verarbeitung eines Lautstroms aufgrund physikalischer Kriterien wie Tonhöhe, Lautstärke, Rhythmus etc. und beinhaltet Prozesse der Registrierung und einfachen Kategorisierung. Die differenzierte Analyse sowie die kognitive Verarbeitung einer akustisch vermittelten Botschaft mithilfe kognitiver Operationen erfolgt während der auditorischen Verarbeitung. Akustische und auditorische Prozesse werden unter dem Begriff „auditiv" zusammengefasst. (Imhof, 2003: 12ff)

Im Gegensatz zu optischen Reizen, die auf der Retina so abgebildet werden, dass Informationen über den entsprechenden Ort erhalten bleiben, summieren und vermischen sich akustische Reize. Der entstehende Lautstrom enthält keine zusätzlichen Informationen über die Art der Schallwellen und wird zunächst zeitlich gegliedert aufgenommen. Entsprechend erfolgt auch die Selektion und Strukturierung des Inputs erst nach der Reizaufnahme und einzelne Geräuschquellen oder der Ort der Geräuschquellen werden nachträglich ermittelt. Sprachwahrnehmung erfolgt durch das Erfassen von Frequenzen, Dauer und Amplituden des Laut-

flusses. Da der Lautfluss redundant ist, wird eine selektive Erfassung möglich. (Buck, 2001: 18) Stimmen variieren von Geräuschen systematisch in der Frequenz und in der Amplitude und können deshalb vom Zuhörer unterschieden werden (vgl. Imhof, 2003: 74f). Auch die Trennung mehrerer Stimmen, die sich in Tonhöhe und Stimmlage voneinander unterscheiden, bereitet dem Zuhörer i. d. R. kaum Schwierigkeiten. Handelt es sich dabei jedoch um sehr ähnliche Stimmen, z. B. die Stimmen mehrerer gleichaltriger Mädchen, so wird die Differenzierung schwieriger.

Für die nachträgliche Bearbeitung des Lautstroms sind Konzentration und eine präzise Fokussierung der akustischen Wahrnehmung notwendig. (Imhof, 2003: 23ff) Schlechte Tonqualität, bei der viele Störgeräusche aus dem Lautstrom herausgefiltert werden müssen, führt deshalb also sehr viel schneller zu Ermüdungserscheinungen beim Zuhörer als schlechte Druckqualität den Leser ermüden würde, der zwischendurch eine Lesepause einlegen kann. Gerade wenn starke Hintergrundgeräusche aus dem Lautstrom herausgefiltert werden müssen, nehmen Faktoren wie Prosodie, Wortlänge, der lexikalische Status eines Wortes, die relative Auftretenshäufigkeit eines Wortes etc. an Bedeutung für die Lautdiskrimination zu.

Um die Prozesse der auditiven Verarbeitung besser verstehen und erklären zu können, wurden seit den 60er Jahren verschiedene Gedächtnismodelle entwickelt. Dabei besteht relative Einigkeit darüber, dass Gedächtnis der „gesamte Prozess des Enkodierens, Speicherns (über einen gewissen Zeitraum) und Abrufens (zum gegebenen Zeitpunkt infolge bestimmter Hinweis- und Abrufreize) von Informationen." (Gerstenmaier, 1995: 257) ist. Rost erweitert diese Definition in Bezug auf das Hörverstehen, indem er darauf hinweist, dass auch bereits gespeicherte Gedächtnisinhalte eine wichtige Rolle im Zuhörprozess spielen: „When we refer to memory in listening, we mean both the process of activation relevant memories to assist in comprehension and the process of forming or updating memories during comprehension." (Rost, 2002: 69)

2.1.1.1. Theorie des Kurzzeitgedächtnisses

Informationen, denen Aufmerksamkeit zugewendet wird, können nach Durchlaufen eines sensorischen Gedächtnisspeichers in ein zwischengeschaltetes Kurzzeitgedächtnis überführt werden. Das Kurzzeitgedächtnis wird häufig auch als Arbeitsgedächtnis bezeichnet, da in ihm die Verarbeitung der eintreffenden Informationen stattfindet. Der sensorische Speicher besteht aus dem ikonischen Gedächtnis und dem echoischen Gedächtnis. Visuelle Stimuli werden ca. ½ Sekunde lang im ikonischen Gedächtnis gespeichert, auditive Stimuli verweilen etwa ebenso lange im echoischen Gedächtnis. Nach der Verarbeitung im Kurzzeitgedächtnis können die Informationen in ein relativ andauerndes Langzeitgedächtnis übertragen werden. (Cassells, 1995: 156f)

Reize werden aufeinanderfolgend in diesen verschiedenen Systemen verarbeitet, wobei eine parallele Bearbeitung nur sehr eingeschränkt möglich ist. Wenn bei der Verarbeitung einer bestimmten Aufgabe für die einzelnen Prozesse zu viele Ressourcen benötigt werden, kann der gesamte Verarbeitungsablauf zusammenbrechen. Vor allem junge Lernende oder Fremdsprachenlerner laufen Gefahr, aufgrund mangelnder Teilfähigkeiten und Konzepte bei der

Informationsverarbeitung zu scheitern. (Klein-Braley, 1994: 162) Der Moment, in dem parallele Verarbeitung von Informationen unmöglich wird - ein Engpass in den informationsverarbeitenden Bahnen also dazu führt, dass nicht mehr alle eintreffenden Reize verarbeitet werden - wird als „Flaschenhals" bezeichnet. (Hayes, 1995: 23) Aus den weiterhin eintreffenden Reizen werden nur die wichtigsten Informationen herausgefiltert, welchen nach wie vor Aufmerksamkeit zugewendet werden soll. Dabei ist ungeklärt, zu welchem Zeitpunkt der Informationsverarbeitung der Flaschenhals auftritt. (Anderson, 2001: 75f) Der Flaschenhals wird häufig auch durch die zentrale Kognition bei der Koordination vieler Aufgaben gebildet. Wenn durch Übung die meisten Erfordernisse an die zentrale Kognition überflüssig werden, tritt Automatisiertheit auf. Eine Tätigkeit kann dann ohne bewusste, fokussierte Aufmerksamkeit vollzogen werden. In diesem Fall besteht die Ausführung der Aufgabe überwiegend aus diversen perzeptuellen und motorischen Systemen und das Konfliktpotential innerhalb eines Systems ist weitgehend aufgehoben. (ebd., 100f)

Im Gegensatz zum Langzeitgedächtnis weist das Kurzzeitgedächtnis nur begrenzte Speicherkapazität auf. Informationen gehen verloren, wenn sie nicht innerhalb von 60-90 Sekunden im Kurzzeitgedächtnis verarbeitet und weitergereicht werden. (Rost, 2002: 145) Da im Kurzzeitgedächtnis ältere Informationen ständig von neu eintreffenden Informationen verdrängt werden, eignet es sich nicht zur dauerhaften Speicherung von Elementen. Die Überführung von Informationen ins Langzeitgedächtnis wird durch das Ausmaß des Memorierens aber auch durch die Tiefe der Verarbeitung beeinflusst. Neuere Theorien vermuten, dass Informationen auch direkt von den sensorischen Gedächtnissystemen ins Langzeitgedächtnis gelangen können, wenn die Informationsverarbeitung in einer tiefen und bedeutungshaltigen Art und Weise erfolgt. Das Langzeitgedächtnis kann viele Informationen für längere Zeit aufnehmen, Spurenzerfall oder Interferenzen können jedoch zum Vergessen bestimmter Informationen führen. (Anderson, 2001: 174ff)

In der Regel liegt der Fokus selektiver Aufmerksamkeit innerhalb einer Modalität (visuell oder akustisch) auf einem Stimulus und ignoriert andere Stimuli weitgehend. Es wird immer nur ein Sachverhalt zu einem bestimmten Zeitpunkt verarbeitet. Werden jedoch Reize in zwei unterschiedlichen Modalitäten angeboten und sprechen damit verschiedene sensorische Systeme an, dann ist es möglich, bereits während der Arbeit an einer visuellen Aufgabe einem auditiven Stimulus Aufmerksamkeit zu widmen. Ein zentraler Flaschenhals tritt jedoch auf, wenn Denkarbeit für beide Aufgaben zu leisten ist. Dann muss die Arbeit an einer Aufgabe solange unterbrochen werden, bis die Arbeit an der anderen Aufgabe abgeschlossen ist. (Anderson, 2001: 98ff) Solange die Modalitäten von Doppelaufgaben getrennt sind (Versuchspersonen sollen gleichzeitig etwas nachsprechen und etwas anderes abschreiben), werden diese Aufgaben gut bewältigt. Wenn die beiden Aufgaben jedoch in Bezug auf Input oder Output miteinander interferieren (wenn z. B. sowohl das Nachzusprechende als auch das zu Schreibende auditiv präsentiert wird), sinkt die Leistung. Dem entsprechen auch die Studien von Schneider und Shiffrin (1977), die von zwei unterschiedlichen Informationsverarbeitungsmodi ausgehen: der seriellen und der parallelen Informationsverarbeitung. Bei der seriellen Informationsverarbeitung werden eintreffende Reize nacheinander verarbeitet. Die Aufmerksamkeit richtet sich flexibel auf die eintreffenden Reize, ist aber in ihrer Kapazität

begrenzt. Bei der parallelen Informationsverarbeitung werden mehrere Aufgaben gleichzeitig und parallel automatisch nebeneinander ausgeführt. Zwar ist die Aufmerksamkeitskapazität unbegrenzt, die automatisierten Abläufe lassen sich jedoch nur schwierig modifizieren und eintreffende Reize werden meist sowohl seriell als auch parallel verarbeitet.

2.1.1.2. Baddeleys Modell des Arbeitsgedächtnisses

Die Theorie des Arbeitsgedächtnisses, als einem präziseren Modell des Kurzzeitgedächtnisses, stammt von Baddeley und Hitch (1974). In den ursprünglichen Modellen des Kurzzeitgedächtnisses wurde davon ausgegangen, dass in diesem einheitlichen System nur wenige Aufgaben gleichzeitig ausgeführt werden können. Baddeley fand jedoch heraus, dass insbesondere die Ausführung mehrerer Aufgaben unterschiedlichen Typs häufig gut möglich ist. Nach Baddeley ist das Kurzzeitgedächtnis demzufolge kein einheitliches System sondern kann in mehrere Komponenten unterteilt werden.

Er nimmt an, dass im Kurzzeitgedächtnis, einem Arbeitsgedächtnis mit mehreren Hilfssystemen, ein räumlich-visueller Notizblock und eine artikulatorische Schleife zur Aufrechterhaltung von Informationen existieren. Beide Hilfssysteme sind in ihrer Aufnahmekapazität begrenzt. Im räumlich-visuellen Notizblock werden visuelles und räumliches Informationsmaterial gespeichert und dort zum unmittelbaren Abruf bereit gehalten. In der artikulatorischen Schleife wird verbales Material gehalten, das durch verbales Üben phonemisch verarbeitet wird. Beide Hilfssysteme werden durch eine zentrale Verarbeitungseinheit kontrolliert und dirigiert. Sie speist Informationen in die beiden Hilfssysteme ein, ruft sie daraus ab oder übersetzt sie von einem in das andere System. Ihre wesentlichen Funktionen sind, eine Verbindung zum Langzeitgedächtnis herzustellen sowie Aufmerksamkeit zu fokussieren (Baddeley, 2002). In den Hilfssystemen haben die Informationen im Gegensatz zum Kurzzeitgedächtnis keine Verweildauer, sondern bleiben für den Übertritt ins Langzeitgedächtnis verfügbar. (Anderson, 2001: 178ff)

Akustische und visuell vermittelte Botschaften können demnach getrennt verarbeitet werden, eine parallele Verarbeitung beschleunigt die Informationsaufnahme jedoch durch Synergieeffekte. Von beiden Systemen aus ist der Zugriff auf das semantische Gedächtnis möglich. Da beim Zuhören nur die Inhalte der artikulatorischen Schleife von zentraler Bedeutung sind, erfordert die Vermittlung von gesprochenen Informationen mehr Zeit als von geschriebenen. Außerdem ist die Verarbeitung auditiven Inputs störanfälliger als die Verarbeitung von Schrift. Tritt eine Störung ein, während Informationen in der artikulatorischen Schleife gehalten werden, so ist die Gefahr besonders groß, diese Informationen zu vergessen. Ähnlich störend kann ein für den Zuhörer unangemessenes Sprechtempo wirken. Zu schnell eintreffende Informationen können nicht von der artikulatorischen Schleife aufgenommen werden und gehen verloren. (Imhof, 2003: 28ff)

Das Konstrukt der phonologischen Schleife diente Baddeley dazu, vor allem drei Effekte zu erklären: den Effekt der phonologischen Ähnlichkeit, den Wortlängeneffekt und den irrelevant- speech-Effekt. Der Effekt phonologischer Ähnlichkeit beschreibt, dass ähnlich klingende Informationen schlechter memoriert werden können als unähnliche. Der Wortlängenef-

fekt beschreibt, dass kurze Wörter besser behalten werden können als sehr lange Wörter. Ein Schlüsselbegriff für das Verständnis des Modells ist dabei die „Gedächtnisspanne". Sie bezeichnet die Anzahl der Informationen, die ohne weitere Verarbeitung für eine kurze Zeit im Gedächtnis behalten und danach wiedergegeben werden können. Die Gedächtnisspanne ist insbesondere für das Hörverstehen relevant, da die gesprochenen Informationen flüchtig sind. Ebbinghaus (Cassells, 1995: 160) geht beispielsweise davon aus, dass sieben plus/minus zwei Elemente als Gedächtnisspanne behalten werden können. Chafe (1985) spricht von sieben Wörtern oder einer Zeitdauer von zwei Sekunden. Auch Baddeley (1986) geht von einer Zeitdauer von zwei Sekunden aus und hält als weiteres bestimmendes Merkmal in Bezug auf den Umfang der Gedächtnisspanne die Geschwindigkeit, mit der Informationen memoriert werden können. Der irrelevant-speech-Effekt beschreibt ein Sinken der Leistung bei einer verbalen Gedächtnisaufgabe, wenn Sprache oder Töne mit wechselnder Frequenz als Hintergrundgeräusch zu hören ist.

Im Jahr 2000 ergänzte Baddeley sein Modell durch den episodischen Puffer, da bestimmte Effekte durch das ursprüngliche Modell nicht mehr erklärt werden konnten (Baddeley, 2002). Der episodische Puffer kann als eigenständige Komponente betrachtet werden, die ebenfalls durch die zentrale Exekutive gesteuert wird. Er ist ein multimodaler Speicher mit begrenzter Kapazität, der sowohl visuelle als auch phonologische Informationen gebündelt in Form von „Episoden" speichern kann (Baddeley, 2002). Durch seine multidimensionale Kodierung kann er die Informationen der Subsysteme integrieren. Dies erleichtert es der zentralen Exekutive, diese Informationen zu koordinieren. Konkret können also zu Episoden gebündelte Informationen leichter behalten werden und die Gedächtnisspanne steigt dadurch an.

Baddeleys Modell erklärt zwar die Arbeitsweisen der Hilfssysteme räumlich-visueller Notizblock, artikulatorische Schleife und episodischer Puffer, geht jedoch kaum auf die Arbeitsweise der zentralen Exekutive oder die zwischen den einzelnen Systemen ablaufenden Prozesse ein. Auch die Verarbeitung anderer als auditiver und visuell-räumlicher Reize bleibt unbeachtet.

Unterschiede in der Informationsverarbeitungsfähigkeit von Individuen wurden bereits früh auf die Leistungsfähigkeit des Arbeitsgedächtnisses zurückgeführt (z. B. Adams & Collins, 1979). Eine wichtige Theorie dazu ist die „Capacity Theory of Comprehension" von Just und Carpenter (1992). Sie beschreiben darin die Informationsspeicherung und Verarbeitungsprozesse beim Sprachverstehen in Abhängigkeit von der Arbeitsgedächtniskapazität. Da auch die Arbeitsgedächtniskapazität im Rahmen des Ländervergleichs untersucht wurde und die Ergebnisse als Personenmerkmale in dieser Arbeit berücksichtigt werden, wird die „Capacity Theory of Comprehension" kurz vorgestellt.

2.1.1.3. Capacity Theory of Comprehension

Die Kapazität des Arbeitsgedächtnisses, d. h. die maximale Menge an im Arbeitsgedächtnis verfügbarer Aktivierung für die Speicherung und Verarbeitung von Informationen, ist begrenzt und individuelle Differenzen darin beeinflussen das Sprachverstehen. Dem Arbeitsgedächtnis kommt bei Just und Carpenter (1992) eine Schlüsselfunktion in der Sprachverarbeitung zu, da in ihm die (Zwischen-)Ergebnisse gespeichert werden, die bei der Verarbeitung des Laut-

stroms entstehen, wie z. B. das Textthema, die wichtigsten Propositionen der vorhergehenden Sätze, etc. Zusätzlich dazu kann das Arbeitsgedächtnis aber auch als der Ort betrachtet werden, an dem die operationalen Ressourcen liegen, mit denen diese (Zwischen-) Ergebnisse generiert werden, wie z. B. Vergleiche ziehen, logische Operationen, etc. Die Vorstellung, das Arbeitsgedächtnis enthalte beides, Sprachverständnis ermöglichende Prozesse und Ressourcen, entspricht nicht der Vorstellung von Baddeley, der von einer artikulatorischen Schleife und einer zentralen Exekutive ausgeht, die getrennt Speicher- und Koordinierungsfunktion besitzen.

Die vom Arbeitsgedächtnis bereitgestellte Aktivierung wird für die Speicherung von Informationen bzw. der Bereithaltung sogenannter „repräsentationaler Elemente" benötigt. Es wird davon ausgegangen, dass Wörter, Sätze, Propositionen oder Objekte der externalen Welt durch repräsentationale Elemente dargestellt werden, die sich auf bestimmten Aktivierungsniveaus befinden. Diese verschiedenen Aktivierungsniveaus resultieren aus der Enkodierung von Textinformationen beim Lesen oder aus dem Abruf von Informationen aus dem Langzeitgedächtnis. Die Verarbeitung eines repräsentationalen Elements ist abhängig vom jeweiligen Aktivierungsniveau. Nur wenn dieses über einem bestimmten Schwellenwert liegt, kann das Element im Arbeitsgedächtnis verarbeitet werden. (Just & Carpenter, 1992) Wenn durch den Bedarf an Aktivierungsenergie die Kapazitätsgrenze des Arbeitsgedächtnisses erreicht wird, werden alte Elemente von neu eintreffenden verdrängt, um die Verarbeitungskapazität so wieder zu erhöhen. Konkret bedeutet das, dass beispielsweise komplexe Repräsentationen, die am Satzanfang gebildet wurden u. U. am Satzende nicht mehr präsent sind, da sie aus dem Arbeitsgedächtnis entfernt wurden.

Aktivierung wird jedoch auch für die eigentlichen Verarbeitungsprozesse benötigt, die nach Just und Carpenter (1992) über ein Produktionssystem laufen. Darin wird Wissen dargestellt und deklaratives Wissen aus einem Datenspeicher und prozedurales Wissen aus einem Produktionenspeicher interagieren miteinander. Unter „Produktionen" werden in diesem Zusammenhang Regeln verstanden, die aus einer Vorbedingung und einer Aktion bestehen. Durch einen Abgleich mit dem aktuellen Inhalt des Datenspeichers wird geprüft, ob die jeweiligen Vorbedingungen erfüllt sind. Ist dies gewährleistet, wird die zur Vorbedingung gehörige Aktion ausgeführt. Nach Just und Carpenter (ebd.) verbreitet sich Aktivierung beim Sprachverstehen also über die Produktionsregeln über die repräsentationalen Elemente hinweg. Die Produktionsregeln können sowohl wiederholt ablaufen, wodurch sich das Aktivierungsniveau des Zielelements erhöht, als auch simultan ablaufen und dabei ihre jeweiligen Teilprodukte generieren. Ein Beispiel soll dies verdeutlichen: Beim Lesen eines Satzes werden Erwartungen aufgebaut, die die Valenz des Verbs betreffen. Es werden jedoch quasi zeitgleich auch syntaktische, semantische und pragmatische Eigenschaften des Satzes ermittelt und verarbeitet. Übersteigt die Anzahl an Prozessen die Kapazität des Arbeitsgedächtnisses, so wird die Aktivierung einzelner Elemente reduziert, um die Verarbeitungsprozesse dennoch aufrecht zu erhalten. Durch die Reduktion der Aktivierung erhöht sich die notwendige Anzahl an Verarbeitungszyklen und die Verarbeitung wird insgesamt verlangsamt.

Just und Carpenter (ebd.) machen die Geschwindigkeit, mit der Sprache verarbeitet wird, und die Behaltensleistung von Teilergebnissen also abhängig von der Kapazität des Arbeitsgedächtnisses. Aspekte wie Lesermotivation, -ziele oder Vorwissen bleiben bei dieser Theorie unberücksichtigt. Leser mit einer hohen Kapazität des Arbeitsgedächtnisses können parallel aktivierte, mehrfache Bedeutungen eines Wortes länger im Arbeitsgedächtnis aktiviert halten als Leser mit einer geringen Kapazität. Diese Leser entscheiden sich relativ früh für eine der Wortbedeutungen. Handelt es sich dabei um eine falsche Entscheidung, muss vergleichsweise viel Energie aufgewendet werden, um den Fehler zu korrigieren. (vgl. Imhof, 2003)

Kintsch (1998: 238ff) kritisiert die „Capacity Theory of Comprehension". Er weist darauf hin, dass die bessere Verstehensleistung der Personen mit größerem Arbeitsgedächtnis auch auf besseren Lesefertigkeiten beruhen könnte. Die differenziellen Effekte im Leseverstehen könnten ebenso mit Unterschieden im Vorwissen und in den Fertigkeiten der Testpersonen begründet sein.

2.1.2. Speicherung von Wissen

Da das Arbeitsgedächtnis nur über begrenzte Kapazität verfügt, kann nicht der exakte Wortlaut einer eintreffenden Botschaft gespeichert werden, sondern die im Gehirn eintreffenden Informationen müssen zunächst dekodiert werden. Erst wenn ihnen Bedeutungs- und Sinnzusammenhänge zugeschrieben werden, ist die Speicherung, Nutzung und Abrufbarkeit dieser Informationen möglich. (Hayes, 1995: 14) Zur Speicherung von Informationen im Gedächtnis gibt es mehrere Theorien, von denen fünf kurz vorgestellt werden. Einige Theorien besagen, dass Informationen in Form von mentalen (1) oder propositionalen Repräsentationen (2) gespeichert werden. Andere Theorien verstehen die Speicherung von Informationen in Form von Prototypen (3) oder Schemata bzw. Skripten (4). Die Vernetzung der gespeicherten Informationen wird hingegen häufig in Form von semantischen Netzwerken (5) gesehen. Aus jeder dieser Theorien resultieren Merkmale, die einen Effekt auf die Schwierigkeit von Items und Stimuli haben könnten.

(1) Informationen werden in Form von mentalen Repräsentationen gespeichert, die mit zusätzlich eintreffenden Informationen aktualisiert und damit weiter ausdifferenziert und komplexer werden. Mentale Repräsentationen sind interne kognitive Darstellungen, die externe Sachverhalte abbilden. Es wird davon ausgegangen, dass verschiedene Lerner beim Verstehen von Stimuli aufgrund von Stimulusinformationen und Weltwissen zu unterschiedlichen mentalen Repräsentationen gelangen. Mentale Repräsentationen sind die Grundlage für mentale Modelle. Dabei handelt es sich um ganzheitlich gebildete Repräsentationen, die den Gegenstand eines Stimulus darstellen. (Kürschner & Schnotz, 2008: 140) Dieses Konzept wird auch von Kintsch (1998) in seinem „Construction-Integration-Modell" aufgenommen. (vgl. Kapitel 2.2.3. *Das Construction-Integration-Modell*)

(2) Andere Forschungsansätze (vgl. Anderson, 2001: 147) gehen davon aus, dass Informationen in Form von propositionalen Repräsentationen gespeichert werden. Unter einer „Proposition" wird eine semantische Struktureinheit verstanden, die den vom Satz ausgedrückten Sachverhalt beinhaltet. (Brinker, 2005: 27) Eine Proposition kann in der Regel durch unterschiedliche

Sätze realisiert werden und ein Satz kann mehrere Propositionen enthalten, welche zusammen die Satzbedeutung ausmachen. Propositionen enthalten ein Prädikat und verschiedene Argumente. (Kintsch & van Dijk, 1978) Propositionen entstehen durch Abstraktion, indem auf die Wahrnehmung bezogene Details gelöscht werden. Im Gedächtnis gespeichert werden nur die wichtigen Beziehungen zwischen den Inhaltselementen. Diese wechselseitigen Beziehungen können durch Netzwerke gezeigt werden. Je mehr Propositionen in einem Satz ausgedrückt werden, desto mehr Zeit ist zum Verstehen notwendig. (Anderson, 2001)

(3) Die Prototypentheorie besagt, dass zu speichernde Informationen in Form von idealtypischen Prototypen in einem hierarchischen Klassifikationssystem gespeichert werden. Die Prototypen stehen stellvertretend für eine ganze Kategorie, wobei es graduelle Unterschiede in der Zugehörigkeit zu einer Kategorie gibt. (vgl. Graefen & Liedke, 2008: 72) Die Begriffe werden aufgrund von Familienähnlichkeiten einer Kategorie zugeordnet. Die Kategorien überschneiden sich häufig und überlappende Bedeutungen können als Erklärung für Polysemie gesehen werden. Begriffe aus Kategorien ähnlicher Niveaus können schneller abgerufen werden als Begriffe, die von unterschiedlichen Kategorieebenen stammen. Auch Begriffe, die als besonders typische, repräsentative Vertreter ihrer Klasse gelten, können schneller verifiziert werden, z. B. wird ein Rotkehlchen schneller als ein Vogel identifiziert als ein Pinguin. (Bärenfänger, 2002) Allerdings haben lexikalische Kategorien häufig unscharfe Grenzen und die dazugehörigen Begriffe (z. B. Präpositionen oder kulturelle, abstrakte Begriffe) können nicht immer eindeutig einer Kategorie zugeordnet werden. Eine Variation der Prototypentheorie stellt Rosch (1975) vor. Demnach sind Begriffe aus der täglichen Objekt- und Erfahrungswelt statt über Merkmale und Ähnlichkeiten über ihre Einsatz- und Handlungsmöglichkeiten miteinander verbunden.

(4) Auch Vorwissen spielt bei der Verarbeitung und Speicherung neuer Informationen eine Rolle. Die Rolle des Vorwissens im Hörverstehensprozess findet besonders in der Skript- und Schema-Theorie (Schank & Abelson, 1977; Adams & Collins, 1979) Beachtung. Ein Skript enthält Hintergrundwissen zu Ereignissen oder bestimmten Situationen (z. B. Was ist ein Supermarkt?). Dieses ist notwendig, um eine Botschaft zu verstehen, wird aber nicht explizit vom Sprecher kommuniziert. Skripte kommen beim Reflektieren über Ereignisse zum Einsatz und helfen, Leerstellen durch Wissen über ein prototypisches Ereignis zu füllen. (Anderson, 2001: 156ff) Schemata sind mit einem Drehbuch zu einer bestimmten Situation vergleichbar (Welche Einzeltätigkeiten fallen bei einem Einkauf im Supermarkt an?). Schemata bilden Regelhaftigkeiten einzelner Kategorien ab und dienen dazu, Handlungen zu steuern sowie die Beziehungen einzelner Ereignisse zueinander zu verstehen. (Rost, 2002: 62ff) Die grundsätzliche Annahme der Schema-Theorie ist, dass ein Text in sich selbst keine Bedeutung trägt. Er liefert lediglich Gebrauchsanweisungen für den Leser, wie auf der Grundlage des eigenen Vorwissens Bedeutungen zu entnehmen sind. Ziel der Schema-Theorie ist es, zu erklären, wie das Wissen des Lesers mit den Informationen des Textes interagiert und diese formt. Sie gibt auch Aufschluss darüber, wie das Wissen organisiert sein muss, um Interaktion mit dem Text zu ermöglichen. Ausgegangen wird von einer Hierarchie von Schemata, die in Subschemata eingebettet sind und mit zunehmender Hierarchietiefe zahlreicher werden, in ihrer Breite jedoch abnehmen. Die Schemata können nicht nur inhaltliche sondern auch strukturelle

Informationen beinhalten. Fehlen relevante Skripte oder Schemata, beispielsweise aufgrund unterschiedlicher kultureller Erfahrungen, wird die Kommunikation dadurch beeinträchtigt und es treten Missverständnisse auf. (Klein-Braley, 1994: 161)

(5) Die Vernetzung der Informationen erfolgt in semantischen Netzwerken. Das assoziationistische Modell der Begriffsbildung (vgl. Banyard & Hayes, 1995: 133) nimmt an, dass neu eintreffende Informationen über Assoziationen miteinander verbunden werden. Zu ähnlichen Informationen werden gleiche Begriffe gebildet, die in einem semantischen Netzwerk gespeichert werden. Auf Reize, die Gemeinsamkeiten mit bekannten Reizen aufweisen, wird in ähnlicher Weise reagiert. Werden zur Sprachverarbeitung zusätzliche Informationen benötigt, werden die Begriffe mit den entsprechenden Merkmalen eines übergeordneten Konzepts abgerufen. Problematisch bei der Theorie der semantischen Netzwerke ist, dass der graduelle Charakter kategorialer Informationen durch sie nicht erfasst wird. Aktivation breitet sich nach der Netzwerktheorie entlang der Pfade eines Netzwerkes aus. In Abbildung II-2.1.2. ist dies am Beispiel des Begriffs „Hund" erkennbar. Das Konzept „Hund" ist im Netzwerk mit dem Konzept „Knochen" verbunden. Wird das Konzept „Hund" aktiviert, breitet sich die Aktivation unbewusst auch auf die umliegenden Konzepte aus. Dabei ergeben sich folgende Zusammenhänge: Je stärker ein bestimmtes Material aktiviert wird, desto schneller kann es abgerufen werden. Das Ausmaß mit dem ein Gedächtnisinhalt aktiviert wird, hängt von der Stärke dieses Gedächtnisinhalts ab. (Anderson, 2001) Allerdings wird bezweifelt, ob eine assoziative Verbindung zwischen den Reizen allein genügt, um entsprechende Begriffe zu bilden.

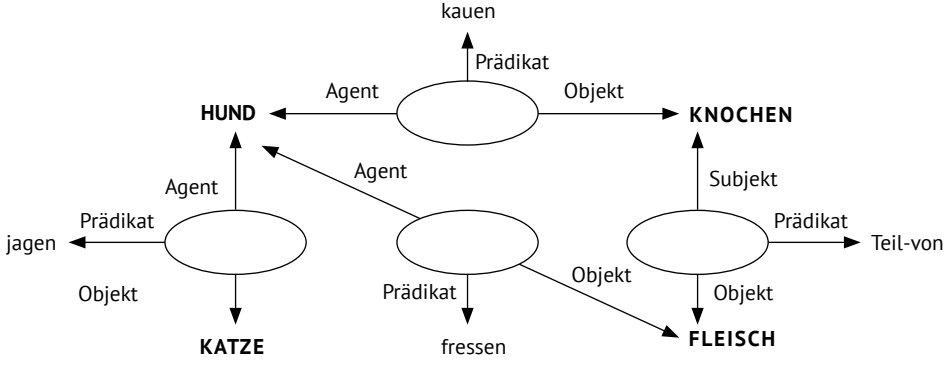


Abbildung II-2.1.2.: Modell eines semantischen Netzwerks (vgl. Anderson, 2001: 186)

2.2. Sprachverstehen

Grundannahme der Psycholinguistik ist, dass die menschliche Sprachfähigkeit auf Wissen basiert, das bei jedem kompetenten Mitglied der Sprachgemeinschaft verfügbar ist. Dieses Wissen kann in die Bereiche mentales Lexikon und mentale Grammatik unterschieden werden. Im mentalen Lexikon finden sich Informationen zu einzelnen sprachlichen Einheiten, wie die Bedeutungsebene der Einheiten, ihre syntaktischen Eigenschaften, Lautwissen und graphemisches Wissen über eine Einheit. Regeln zu Kombinationsmöglichkeiten der Einheiten stammen aus dem Bereich der mentalen Grammatik. (Anderson, 2001: 355)

Es wird davon ausgegangen, dass Sprachverarbeitung ein iterativer Prozess mit diversen „Prüfschleifen“ ist, durch den Verständnis erzeugt und gewährleistet werden soll. In einem durchschnittlichen Sprachfluss treffen ca. 125-175 Wörter gesprochener Sprache pro Minute auf den Zuhörer (Imhof, 2003: 81). Aufgrund einer durchschnittlichen Sprechgeschwindigkeit und einer damit verbundenen mittleren Silbendauer von etwa 185 ms (Pfitzinger, 2001: 155) muss damit auch die Worterkennung sehr schnell erfolgen.

Aus jedem Wort des Sprachflusses wird zunächst sofort so viel Information wie möglich extrahiert (Prinzip der unmittelbaren Verarbeitung). Worterkennung erfolgt durch das Zusammenspiel von wahrgenommenem Klang und dem Wissen über die Wahrscheinlichkeit, dass ein Wort in einem bestimmten Zusammenhang erscheinen kann. Die Verarbeitung von Sprache erfolgt sequenziell und die Worterkennung dient dazu, den Beginn des jeweils nachfolgenden Wortes zu identifizieren. Ein Wort wird genau dann erkannt, wenn alle anderen Möglichkeiten für dessen Bedeutung vom Zuhörer ausgeschlossen wurden. Dabei werden Wörter häufig schon erkannt, bevor sie vollständig ausgesprochen wurden, denn ein Zuhörer kann ein Wort schneller erfassen als ein Sprecher es formulieren kann. Da Sprache teilweise retrospektiv verarbeitet wird, kann der dadurch entstehende zeitliche Vorsprung dafür genutzt werden, unerkannte Wörter solange in einer artikulatorischen Schleife (vgl. Kapitel 2.1.1.2. *Baddeleys Modell des Arbeitsgedächtnisses*) zu halten, bis weitere Hinweise verarbeitet werden. Selbst wenn nicht alle Wörter eines Sprachflusses erkannt werden, ist aufgrund der Redundanz von Sprache und durch Inferenzen seitens des Zuhörers Sprachverständnis möglich. Schwierigkeiten bei der Worterkennung sind häufig eine fehlerhafte Identifikation der Wortgrenze, zum Beispiel durch die variierende Aussprache des Lautstroms und mangelndes Wissen über die Wortbedeutung. Während des Zuhörens berufen sich Sprachnutzer auf gespeicherte Prototypen von Lauten, die als Basis für die Interpretation von allophonischen Varianten dienen. Diese werden durch Assimilation (die Angleichung der Artikulation eines Lautes an einen vorangehenden Laut), Reduktion (Abschwächung von Vokalen oder Verlust von Lauten) und Elision (Wegfall eines Lautes) erzeugt. (vgl. Imhof, 2003: 81ff)

Beim Sprachverstehen wird der Sprachfluss dann in bedeutungshaltige Einheiten, sogenannte „Phrasen“, gegliedert, die interpretiert werden. (Imhof, 2003: 81ff) Als Indiz für die Theorie der Phrasenstruktur der Sprache gilt, dass bei der Generierung von Sätzen und Äußerungen zwischen den Phrasen in der Regel eine kurze Sprechpause erfolgt. Beispielsweise werden auch Versprecher mit Laut-Wort-Vertauschungen in der Regel durch Wiederholen der gesamten Phrase korrigiert. (Anderson, 2001: 362ff) Die Bedeutung eines Satzes bzw. einer Äußerung wird also Phrase für Phrase verarbeitet und ihre Interpretationen werden kombiniert (Rost, 2002). Der Zugriff auf die exakte Phrasenformulierung wird nur aufrecht gehalten, solange die Bedeutung verarbeitet wird. Am Ende jeder Phrase wird zusätzliche Integrationszeit aufgewendet, um alle Informationen zu verarbeiten. Dabei ist entscheidend, ob die eintreffenden Informationen dem Zuhörer vertraut oder fremd sind, denn insbesondere für das Verständnis mehrdeutiger Stimuli wird Wissen über den sozialen Kontext sowie Allgemeinwissen aktiviert. (Anderson, 2001: 392ff)

Aus pragmatischer Perspektive bedeutet Verständnis die richtige Interpretation aller eintreffenden Informationen (des Lautstroms und der Kontextinformationen) und nicht nur die Aufnahme bzw. Dekodierung des Gesagten. Indem er die eintreffende Botschaft rekonstruiert und unter Berücksichtigung seiner eigenen Zuhörziele, Erwartungen, Weltwissen, etc. verarbeitet, ist der Zuhörer für einen Großteil der gelungenen Kommunikation verantwortlich. Zuhören beinhaltet daher immer auch die Absicht, den Kommunikationsprozess zu vervollständigen und dem Sprecher beim Ausdrücken bestimmter Inhalte entgegenzukommen, kurz: „From a pragmatic perspective, listening is an intention to complete a communication process." (Rost, 2002: 40)

Die inhaltliche Bedeutung scheint für das Verständnis eines Satzes bzw. einer Äußerung also wichtiger als seine grammatikalische Struktur und es werden überwiegend semantische Strategien eingesetzt, mit deren Hilfe die Bedeutungen der Wörter in eine sinnvolle Beziehung gebracht werden. (Anderson, 2001: 395ff) O'Malley und Chamot fassen dies wie folgt zusammen: "Listening comprehension is an active and conscious process in which the listener constructs meaning by using cues from contextual information and existing knowledge, while relying upon multiple strategic resources to fulfill the task requirement." (O'Malley et al., 1989: 420)

Der eintreffende Sprachfluss wird aber auch mit einem grammatischen Sprachmodell abgeglichen und die eintreffenden Informationen werden in bereits vertraute Konzepte des Zuhörers integriert. Grammatische Hinweise, wie die Wortreihenfolge, Kongruenz von Subjekt und Verb, Kongruenz der Pro-Formen oder Flexionsformen, steuern die inhaltliche Interpretation beim Verstehen. (Rost, 2002: 26) Die Auswahl erfolgt meist nach syntaktischen und semantischen Kriterien, wie Anhaltspunkte, die sich aus Genus oder Numerus ergeben aber auch Pronomina, die auf Dinge oder Sachverhalte in derselben grammatischen Rolle (z. B. Subjekte vs. Objekte) referieren oder Weltwissen. (Just & Carpenter, 1987)

Viele Sätze/Äußerungen erlauben mehrere Interpretationen, weil in ihnen mehrdeutige Wörter oder mehrdeutige syntaktische Konstruktionen vorkommen. Syntaktische Mehrdeutigkeiten werden häufig am Ende des Satzes bzw. der Äußerung aufgelöst, lexikalische Mehrdeutigkeiten werden durch den Kontext geklärt. Sie stellen erhöhte Ansprüche an die Verarbeitung, da sie mit dem Prinzip der unmittelbaren Verarbeitung kollidieren. Einerseits wird versucht, jedes Wort und jede Phrase sofort zu interpretieren, andererseits lösen sich die Mehrdeutigkeiten erst mit dem Ende des Satzes oder der Äußerung auf. Bei Mehrdeutigkeiten wird eine Interpretation gewählt, die unter Umständen revidiert werden muss, wenn sie mit dem weiteren Verlauf des Satzes/der Äußerung nicht übereinstimmt. (Anderson, 2001: 400ff)

Häufig ist für das Sprachverständnis auch die Bildung von Inferenzen notwendig. Da der Zuhörer die vom Sprecher intendierte Bedeutung nicht kennt, muss er das tatsächlich Gesagte interpretieren und davon auf die mutmaßliche Bedeutung schließen. Dabei sind Sprachverwendungsstrategien relevant, aber auch Problemlösungsstrategien, frühere Erfahrungen und Weltwissen. Schlussfolgern oder inferieren bedeutet, diese Informationen für das weitere Textverständnis bereitzustellen. Die notwendigen Inferenzen können sehr vielseitig sein und beispielsweise das Erschließen eines Schemas (z. B. Restaurant), einer kategorialen Klasse (Apfel ist ein Mitglied der Klasse Obst) oder eines logischen Schlusses (x + y führen zu z) erfordern. (Rost, 2002: 64ff)

Im Einzelnen wird demnach für die linguistische Analyse des Sprachstroms graphemisches, phonologisches, morphologisches und lexikalisches Wissen sowie Wissen über Syntax und Semantik benötigt. Außerdem spielen beim Sprachverstehen Allgemeinwissen, Kenntnisse über den sozialen Kontext, die intendierte Sprechhandlung sowie Kenntnisse über den Gebrauch von Sprache (pragmatisches Wissen) und Strategiewissen eine Rolle. (Hartland, 1995: 203) Um einen Text bzw. Diskurs zu verstehen, genügt es folglich nicht, lediglich ein Wort im mentalen Lexikon wiederzufinden. Für das Verständnis einer Äußerung ist es notwendig, Beziehungen zu anderen Satzteilen aber auch zum Sprecher, zum Zuhörer, zum Ort und zur Zeit herzustellen. Dieses Herstellen von Beziehungen kann durch unterschiedliche Theorien erklärt werden, von denen die wichtigsten kurz vorgestellt werden.

2.2.1. Bottom-up und Top-down Verarbeitungsprozesse

In datenbasierten Theorien, sogenannten bottom-up Theorien, wird davon ausgegangen, dass die akustischen Informationen zunächst in Phoneme und dann in Wörter gegliedert werden. Auf einer syntaktischen Ebene werden dann semantische Informationen miteinbezogen, um die basale linguistische Bedeutung des Lautflusses auszumachen. In Anbetracht der kommunikativen Situation wird das Gesagte zuletzt interpretiert, um die Absicht des Sprechers zu erfassen. Die Verarbeitung eines auditiven Stimulus wird also als eine Abfolge von Prozessen auf verschiedenen Ebenen gesehen, bei denen stets das Ergebnis Arbeitsgrundlage für die nächste Ebene ist. Dagegen wird in wissensbasierten Theorien, sogenannte top-down Theorien angenommen, dass die Bedeutung eines Wortes oder einer akustischen Botschaft verstanden werden kann, ohne alle Einzelheiten dekodiert und analysiert zu haben. Vorgänge und Wissen auf höheren kognitiven Ebenen, wie Weltwissen und Erwartungen, beeinflussen den Verstehensprozess. (Cassells & Green, 1995: 77)

Für das Rezipieren von Stimuli nehmen die Erwartungshaltung des Lesers/Zuhörers und die Voraktivierung von Begriffen eine zentrale Rolle ein, denn die Stimulusinformationen interagieren mit dem Vorwissen des Lesers/Zuhörers. Bereits nach Aufnahme der ersten Informationen, die das Thema erkennbar machen, werden Wörter oder Sätze/Äußerungen aus diesem Themenbereich assoziiert und die Erwartung dieser Informationen wird beim Weiterlesen/Weiterhören bestätigt oder modifiziert. Verstehen kann deshalb als Wechselspiel von aufsteigenden bottom-up und absteigenden top-down Verarbeitungsprozessen gesehen werden. (vgl. Kürschner & Schnotz, 2008; van Dijk & Kintsch, 1983) Neuere psychologische Erkenntnisse legen nahe, dass es sich bei der Sprachverarbeitung nicht um streng nacheinander ablaufende, sondern vielmehr um parallel, sich abwechselnde und miteinander interagierende Prozesse handelt, bei denen hierarchieniedrige und –hohe Prozessebenen gleichzeitig oder in zeitlicher Überlappung aktiviert sein können. Höhere Prozesse können bereits einsetzen, bevor niedrigere abgeschlossen sind. (Kürschner & Schnotz 2008: 139). Ferner scheint sich das Gleichgewicht zwischen top-down und bottom-up Prozessen je nach Art der eintreffenden Botschaft zu verändern. Normalerweise werden die eintreffenden Informationen aktiv für die Wahrnehmung genutzt und die ablaufenden Prozesse erfolgen eher bottom-up. Steht jedoch nicht ausreichend Informationsmaterial zur Verfügung, erfolgt die Verarbeitung eher top-down und es wird versucht, die dargebotenen Reize mithilfe von Weltwissen und Erwartungen zu identifizieren und zu interpretieren. (Cassells & Green, 1995: 78)

Auch die Schema-Theorie (vgl. Kapitel 2.1.2. *Speicherung von Wissen*) geht davon aus, dass ein Stimulus simultan auf zwei Ebenen durch bottom-up und top-down Prozesse in allen Bereichen (z. B. syntaktisch, semantisch, interpretativ) verarbeitet wird. Bottom-up Informationen werden vom Stimulus geliefert. Zunächst werden bottom-up Prozesse durch die eintreffende Nachricht aktiviert (z. B. Dekodierung der Nachricht) und das am besten passende bottom-level Schema wird ausgewählt. Bei der top-down Verarbeitung wird Hintergrundwissen zum Verständnis der Nachricht aktiviert und Kontext oder allgemeines Weltwissen steuern die Wahrnehmung. Allgemeines Wissen bestimmt auf einer hohen Ebene, wie Wahrnehmungseinheiten auf einer niedrigeren Ebene interpretiert werden müssen. Hypothesen werden basierend auf allgemeinen Schemata gebildet und die eintreffende Nachricht wird auf Informationen geprüft, welche in diese Schemata passen. (Buck & Tatsuoka, 1998)

O'Malley et al. (1989) fanden bei einer Studie zu Hörverstehensstrategien im Fremdsprachbereich heraus, dass effektive Zuhörer sowohl von top-down als auch von bottom-up Strategien Gebrauch machten, wohingegen ineffektive Zuhörer sich stärker auf bottom-up Strategien verließen und eher versuchten die Bedeutung einzelner Wörter zu entschlüsseln und aus dem Kontext zu erschließen. Diese Ergebnisse konnten auch von Bacon (1992) bestätigt werden. Je anspruchsvoller der Input für die Zuhörer wird, desto stärker verlassen sie sich auf bottom-up Strategien. (z. B. Wolff, 1999; Rubin, 1994)

2.2.2. Verstehen durch Wahrnehmung, Parsing und Verwendung

Nach Anderson (1995) besteht Sprachverarbeitung der Muttersprache aus drei wesentlichen Phasen „Perception" (Wahrnehmung), „Parsing" (Parsing) und „Utilisation" (Verwendung). Zunächst wird die akustische Mitteilung im Rahmen wahrnehmungsbezogener Prozesse enkodiert. Daran schließt sich die syntaktische und semantische Analyse an, das Parsing. Dabei werden Propositionen ermittelt, indem erkannten Wörtern grammatische Kategorien (z. B. Inhaltswörter oder Funktionswörter) zugewiesen und die strukturellen und semantischen Beziehungen zwischen ihnen festgestellt werden. Dieser Vorgang kann prospektiv und retrospektiv erfolgen. Die Wörter werden dann in mentale Repräsentationen überführt, z. B. propositionale Netzwerke (vgl. Kapitel 2.1.2. *Speicherung von Wissen*), und die Bedeutung eines Satzes/einer Äußerung wird erschlossen. Durch Parsing wird es also möglich, ein propositionales Modell des Lautflusses zu erstellen, das die Referenten und ihre Beziehung zueinander repräsentiert. (Anderson, 2001: 389) Sobald der Zuhörer den Lautfluss entschlüsselt hat, werden die gewonnenen Informationen inhaltlich, nicht formal, im Langzeitgedächtnis gespeichert und das Abbild der Nachricht im Kurzzeitgedächtnis wird gelöscht. Gespeichert wird nur der Propositionsgehalt einer Botschaft, nicht gespeichert werden die verwendeten Wörter oder die grammatischen Strukturen.

Wahrnehmung, Parsing und Verwendung laufen überwiegend nacheinander ab, können sich zum Teil aber auch überschneiden. Aufgrund der Redundanz der Sprache ist es für das Verstehen nicht notwendig, dass der Abgleich mit dem grammatischen Sprachmodell vollständig erfolgt. Um die Verarbeitungskapazität des Gehirns optimal zu nutzen, wendet sich der Zuhörer zunächst kommunikativen Aspekten des Sprachflusses zu, bevor die grammatischen Einzelheiten einer Botschaft beachtet werden. (Rost, 2002: 26)

2.2.3. Das Construction-Integration-Modell

Das „Construction-Integration-Modell“ (Kintsch, 1998) erklärt unter Berücksichtigung der begrenzten Kapazität des Arbeitsgedächtnisses die Prozesse beim Entnehmen und Verarbeiten von Textinformation und integriert dabei auch Vorwissen, Ressourcen und Ziele/Motivationen des Textrezipienten. Das Modell macht auf die Wechselwirkung von der Neuaufnahme von Propositionen und ihrer Aktivierung bzw. Verknüpfung mit bereits vorhandenen Propositionen aufmerksam und betont dadurch vor allem die Dynamik des Verstehensprozesses und die Bedeutung des Kontextes und Vorwissens. Kintsch (1998) unterscheidet zwischen der mentalen Repräsentation der Textbasis, welche die semantisch abgeleiteten Relationen enthält, und dem Situationsmodell mit vom Wissen des Lesers hergeleiteten Relationen. Die Textbasis orientiert sich stark an den Oberflächen- und Strukturmerkmalen eines Textes. Tieferes Verständnis des Textes entsteht aber erst mit dem Aufbau eines angemessenen Situationsmodells.

Das „Construction-Integration-Modell“ (Kintsch, 1998) basiert auf Forschung zum Leseverstehen. Der Leser baut mit seinem Vorwissen neue Wissensstrukturen auf und beim Verstehen eines Textes entsteht durch die Integration von Textinformationen und Vorwissen ein mentales Modell. Van Dijk und Kintsch (1983) nehmen an, dass die Prozesse des Sprachverstehens auf verschiedenen Ebenen ablaufen: Auf der untersten Ebene finden basale Wahrnehmungsprozesse statt, bei denen auditive Reize erfasst und verarbeitet werden müssen. Diese basalen Prozesse werden im Modell außer Acht gelassen. Auf diesen Wahrnehmungsprozessen bauen Prozesse der Buchstaben- und Worterkennung auf. Im natürlichen Textverstehensprozess Wörter werden nicht isoliert, sondern im sprachlichen Kontext verarbeitet. Deshalb wird die Wortidentifikation zu einem beträchtlichen Ausmaß vom Kontext mitbestimmt und die vorliegenden Informationen werden mit dem eigenen mentalen Lexikon abgeglichen. Die Identifikation einzelner Wörter findet durch die Bildung einer Oberflächenstruktur auf der Oberfläche statt. Sie genügt jedoch nicht zum Verständnis der Satzbedeutung.

Wortfolgen müssen basierend auf ihren semantischen Relationen aufeinander bezogen werden. So entsteht zunächst eine Repräsentation der Textoberfläche (Textbasis), die alle sprachlichen Details sowie alle lexikalischen und syntaktischen Informationen des Textes enthält. Im Anschluss daran führen semantische Verarbeitungsprozesse dazu, dass die Wortfolgen zu Propositionen zusammengefasst werden. Diese stellen im „Construction-Integration-Modell“ die sprachlichen Grundeinheiten dar. Propositionen repräsentieren die Tiefenstruktur eines Satzes bzw. eines Textes und drücken in einer Prädikat-Argument-Struktur Handlungs- und Zustandsrelationen aus. (Kintsch, 1998: 38) Mehrere einfache Propositionen können sich zu einer komplexeren Proposition zusammensetzen. Die Textbasis stellt die Gesamtmenge der in einem Text enthaltenen Propositionen dar. Neben der propositionalen Struktur beinhaltet sie auch die Oberflächen-Struktur des Textes, d. h. die konkreten Wörter und Sätze. Struktur-Schemata mit Informationen über die zu erwartende Textstruktur werden aktiviert. Aus der Textbasis wird ein propositionales Netzwerk gebildet. Die Verbindungsstärken der Propositionen im Netzwerk können variieren und eingebettete, untergeordnete Verbindungen können z. B. stärker als direkte Verbindungen zwischen Propositionen sein oder umgekehrt. (Kintsch, 1998) Das propositionale Netzwerk ist teilweise ungeordnet und inkohärent, weshalb Assoziationen und erste Inferenzen notwendig werden um Widersprüche aufzulösen und Wichtiges

von Unwichtigem zu trennen. Neu eintreffende Informationen im Arbeitsgedächtnis aktivieren benachbarte Knoten im Netzwerk mit zur jeweiligen Verbindungsstärke proportionaler Wahrscheinlichkeit.

Erst auf der höchsten Verarbeitungsebene wird versucht, die Informationen aus den Sätzen zu integrieren und die Bedeutung des ganzen Textes zu erfassen. Van Dijk und Kintsch (1983) nehmen an, dass auf der hierarchiehohen Prozessebene ein mentales Situationsmodell entwickelt wird, das „Construction-Integration-Modell". Bei der Textverarbeitung werden zunächst sequenziell Wort für Wort sowie Satz für Satz verarbeitet, wobei jedes Textsegment umgehend in den restlichen, im Arbeitsgedächtnis gehaltenen Text integriert wird. Eine Proposition wird nach der anderen verarbeitet, wobei neue Propositionen einer zusammenhängenden Struktur von Propositionen hinzugefügt werden. Neue Propositionen werden also in das bestehende Netzwerk integriert, das auf diese Weise konstruiert wird, was zum Namen des Modells führte. Diese unmittelbar verarbeiteten Propositionen werden neben der jeweils aktuellsten und für die weitere Verarbeitung sehr zentralen Propositionen aktiv im Arbeitsgedächtnis gehalten. Dort können nach Kintsch und van Dijk eine durchschnittliche Menge von vier Propositionen aktiv gehalten werden, wobei sogenannte Makropropositionen gebildet werden können, die den Textinhalt zusammenfassen. Die Auswahl der im Arbeitsgedächtnis bleibenden Propositionen, erfolgt auf der Grundlage einer Kombination aus zeitlicher Nähe und Wichtigkeit der Information. (Anderson, 2001: 420)

Die jeweils bis dahin vorliegende Repräsentation des Textes wird ins Langzeitgedächtnis übertragen. Aber auch die Informationen im Langzeitgedächtnis sind für die weitere Verarbeitung noch nutzbar. Sie werden abgerufen, wenn in den weiteren Sätzen bestimmte Stichwörter als Hinweisreize vorkommen. (Anderson, 2001: 420) So wird die Textrepräsentation zyklisch aktualisiert und korrigiert, bis die entstehenden Textrepräsentationen eine kohärente Struktur besitzen und keine einfachen Sequenzen von Sätzen mehr sind. Durch frühere Textpassagen, Allgemeinwissen und persönliche Erfahrungen werden Sinnzusammenhänge hergestellt. (Kintsch, 1998) Dabei ist das Ausmaß dieser Sinnzusammenhänge einerseits vom jeweiligen Text abhängig, da die zugehörige Textbasis unterschiedlich kohärent und vollständig sein kann. Andererseits bestimmt auch der Leser mit seinen verfügbaren Ressourcen und seiner Motivation das jeweilige Ausmaß.

Ältere werden mit neueren Propositionen im Arbeitsgedächtnis durch das Überlappen von Ausdrücken verknüpft. Dabei bereiten Propositionen, die über Satzgrenzen hinweg verknüpft werden müssen, eher Schwierigkeiten beim Verstehen, denn in diesen Fällen müssen sogenannte „Überbrückungsinferenzen" (Haviland & Clark, 1974 zitiert nach Anderson, 2001: 419) gebildet werden. Dadurch werden schlussfolgernd weitere Propositionen integriert, welche sonst unzusammenhängende Ausdrücke miteinander verbinden. Unter Umständen kann eine neue Proposition nicht mit einer älteren verbunden werden, weil sich diese nicht mehr im Arbeitsgedächtnis befindet. In diesem Fall müssen frühere Propositionen aus dem Langzeitgedächtnis reaktiviert werden, um diejenige Proposition zu finden, die eine Argumentüberlappung mit der neuen Proposition aufweist. Je weiter der zugehörige Referenzausdruck im Text zurückliegt, desto mehr Zeit wird also für die Verarbeitung eines referentiellen Ausdrucks

benötigt. (vgl. Anderson, 2001: 418ff) Diese mentale Konstruktion, die die Textbasis mit relevanten Wissensaspekten des Lesers/Zuhörers verbindet, wird von Kintsch als Situationsmodell bezeichnet. Situationsmodelle unterschiedlicher Personen zum gleichen Text können aufgrund unterschiedlichen Vorwissens stark voneinander differieren.

2.2.4. Kognitive Prozessmodelle

Da ca. 30% der IQB-Items dem Format Multiple-Choice (MC) angehören, werden im Folgenden noch zwei kognitive Prozessmodelle vorgestellt, die für das Lösen von Multiple-Choice-Items zum Leseverstehen entwickelt wurden. Ziel dieser Modelle war es, aus ihnen manipulierbare Aufgabenmerkmale abzuleiten, welche die Itemschwierigkeit beeinflussen können. Im Gegensatz zu den vorher beschriebenen Modellen legen beide Modelle ihren Schwerpunkt weniger auf Prozesse die beim Verständnis der Stimuli auftreten, sondern beziehen sich stärker auf die bei der Beantwortung der Items involvierten Prozesse. Das erste Modell stammt von Embretson und Wetzel (1987) und soll die kognitiven Charakteristika von Items beschreiben, welche sich auf die Itemschwierigkeit auswirken. In dem Modell werden die Prozesse der „Textrepräsentation" (Text Representation) und der „Entscheidung für eine Antwortalternative" (Response Decision) voneinander unterschieden (vgl. Abbildung II- 2.2.4). Es geht in dem Modell also um Merkmale, die eher einen Einfluss auf das Textverständnis haben und solche, die sich eher auf den Antwortprozess beziehen.

Abbildung II-2.2.4. gibt einen Überblick über die Teilbereiche des Modells:

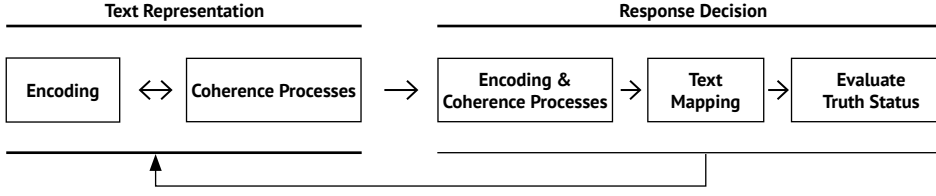


Abbildung II-2.2.4.: Informationsverarbeitungsmodell für MC-Items zum Leseverstehen (Embretson/Wetzel, 1987: 177)

Der Textrepräsentationsprozess mit den Teilbereichen lexikalisches „Enkodieren" (Encoding) und „Kohärenzprozesse" (Coherence Processes) bezieht sich auf das Textverstehen. Unter „lexikalischem Enkodieren" versteht man die Überführung des Stimulus in eine sinnvolle Repräsentation des Textes. Die Schwierigkeit dieses Prozesses ist abhängig von der Vertrautheit des Lesers mit den Wörtern des Textes. Aus diesem Grund spielen Worthäufigkeitsangaben hier eine wichtige Rolle. Zur Beschreibung des Kohärenzprozesses lehnen sich Embretson und Wetzel (1987) an die Modellvorstellung von Kintsch und van Dijk (1978) an (vgl. Kapitel 2.2.3. Das Construction-Integration-Modell) und gehen davon aus, dass die Propositionen aus einem Text über mehrere Verarbeitungszyklen zu einer kohärenten Textrepräsentation verbunden werden. Dabei nehmen die Autoren an, dass die Anzahl der Propositionen bei konstanter Wortzahl die Schwierigkeit der Kohärenzprozesse bestimmt. Der Verarbeitungsaufwand erhöht

sich, je mehr Propositionen in das Netzwerk integriert werden müssen, da viele Propositionen gleichzeitig aktiv gehalten werden müssen und u. U. das Arbeitsgedächtnis überlastet werden kann. Die propositionale Dichte eines Textes kann das Textverständnis also erleichtern oder erschweren.

Der „Entscheidungsprozess" wird in die drei Bereiche „Enkodierungs- und Kohärenzprozesse" (Encoding & Coherence Processes), „Textmapping" (Textmapping) und „Wahrheitsgehalt abschätzen" (Evaluate Truth Status) unterteilt. Nach der Enkodierung und der Verarbeitung der Antwortalternativen liegt ihre Bedeutung kohärent vor. Während des Textmappings werden die Propositionen der Antwortalternativen mit denen des Textes abgeglichen. Es müssen die Propositionen gefunden werden, mit deren Hilfe die Antwortalternativen bestätigt oder verworfen werden können. Bei Propositionen, die sich in einem einzigen kurzen Satz befinden, dürfte dieser Vorgang unproblematisch sein. Sind die Propositionen jedoch über den Text verstreut oder bezieht sich ein Item auf die Kernaussage des Textes, ist der Abgleich mit dem Stimulus schwieriger. Embretson und Wetzel (1987) gehen davon aus, dass die Schwierigkeit steigt, je länger ein Textabschnitt ist, in dem relevante Propositionen vorkommen. Der Wahrheitsgehalt der Antwortalternativen wird in einem zweistufigen Prozess durch Falsifikation und Konfirmation abgeschätzt. Zunächst ist es im Rahmen der „Falsifikation" von Bedeutung, so viele unzutreffende Antwortoptionen wie möglich zu identifizieren. Am einfachsten ist dies in Fällen, in denen die Propositionen aus dem Stimulus den Propositionen in den Distraktoren explizit widersprechen. Im Rahmen der „Konfirmation" werden dann die nicht falsifizierten Antwortalternativen mit dem Text abgeglichen. So soll herausgefunden werden, ob aufgrund der Textinformationen eine der möglichen Antwortalternativen als korrekt bestätigt werden kann. Je mehr Propositionen anhand der Textgrundlage falsifiziert werden können, desto einfacher ist ein Item. Wenn die Itemalternativen viele mit dem Text abzugleichende Propositionen enthalten, erhöht dies die Itemschwierigkeit.

Im Gegensatz zum Modell von Embretson und Wetzel (1987) beschreibt das Modell von Sheehan und Ginther (2001) die Prozesse, die bei der Bearbeitung von Multiple-Choice-Items ablaufen. Die Prozesse beim Verstehen des Textes spielen in diesem Modell keine Rolle. Dabei liegt der Schwerpunkt auf der Tiefe, mit welcher ein Thema oder eine Frage im Text behandelt wird. Eine stark elaborierte Information bleibt besser in der entsprechenden mentalen Repräsentation enthalten. Erscheint eine derartige Information im Attraktor (der richtigen Antwortalternative im MC-Item), wird das Item dadurch leichter. Taucht sie in den Distraktoren (den falschen Antwortalternativen) auf, wird die richtige Beantwortung des Items erschwert. In dem Modell wird davon ausgegangen, dass der Attraktor und die Distraktoren auf einem bestimmten Aktivierungsniveau liegen. Da stärker aktivierte Optionen mit größerer Wahrscheinlichkeit ausgewählt werden, besteht zwischen dem Aktivierungsniveau und der Itemschwierigkeit ein Zusammenhang. Die Schwierigkeit eines Items ist demzufolge abhängig von der Aktiviertheit der Antwortoptionen. Ist der Attraktor stärker aktiviert als die Distraktoren, ist das Item eher leicht, im umgekehrten Fall handelt es sich um ein eher schwieriges Item.

Im Modell werden drei Typen von Effekten unterschieden, die das Aktivierungsniveau der Antwortoptionen beeinflussen können: Lokalisations-, Korrespondenz- und Elaborationseffekte.

Bei den Lokalisationseffekten spielt das Modell von Kintsch (vgl. Kapitel 2.2.3. *Das Construction-Integration-Modell*) eine Rolle: Beim Aufbau mentaler Repräsentationen bewegen sich die Rezipienten i. d. R. sehr nahe an der Textbasis. Folglich werden benachbarte Informationen wahrscheinlich auch in der entsprechenden mentalen Repräsentation eng beieinander stehen. Finden sich im Text benachbarte Informationen auch in einer ähnlichen Kombination in einer Antwortoption im Item wieder, so wirkt diese Antwortoption besonders attraktiv. Im Fall des Attraktors, wird das Item so leichter, im Fall von Distraktoren wirkt diese erhöhte Attraktivität erschwerend. Dabei können auch die Vorerwartungen der Rezipienten über die räumliche Position von Informationen im Text eine Rolle spielen. Bei der lexikalischen Überlappung zwischen dem Stimulus und dem Item (Formulierungen im Stamm, im Attraktor sowie in den Distraktoren) und der semantischen Ähnlichkeit zwischen ihnen (Korrespondenzeffekte) gehen Sheehan und Ginther (2001), wie auch R. C. Anderson (1972) und Embretson und Wetzel (1987), von mehreren Stufen aus, die in der Schwierigkeit zunehmen: wort-wörtliche Entsprechung, paraphrasierte Entsprechung, einfaches Inferieren, komplexeres Inferieren und Vor- oder Fachwissen erforderlich. Der dritte Typ ist die Elaboration, nach Strube et al. (1996) der aktive Prozess der Anreicherung beim Verstehen oder Speichern von Informationen. Um zu einer umfassenderen und differenzierteren Wissensrepräsentation zu gelangen, werden neue Informationen nicht nur reproduziert, sondern auch mit zusätzlichen Informationen verknüpft. Dies führt auch zu einer besseren Behaltensleistung. (vgl. Strube et al., 1996)

3. Hörverstehen – Zusammenfassung und Ausblick

Das dritte Kapitel stellt im Gesamtzusammenhang der Arbeit eine Zusammenfassung des in den vorhergehenden Kapiteln Gesagten und einen Ausblick des noch Folgenden dar. Zunächst werden die Rahmenbedingungen des Zuhörens noch einmal kurz skizziert. Darauf aufbauend wird das Modell des Zuhörens vorgestellt, wie es die Grundlage für diese Arbeit liefert. Dabei wird auch diskutiert, inwieweit die herangezogenen Theorien (z. B. das „Construction- Integration-Modell“ oder die „Capacity Theory of Comprehension“) auf Hörverstehen übertragbar sind, da sie auf Forschung zum Leseverstehen basieren. Bislang wird davon ausgegangen, dass sich Lesen und Zuhören im Wesentlichen im Rahmen der Wahrnehmung von Informationen bzw. deren Aufnahme unterscheiden und der Informationsverarbeitung hingegen weitgehend ähnliche Prozesse zugrunde liegen. Aus diesem Grund erscheint nicht nur die Beachtung von Studien zum Leseverstehen sinnvoll, sondern auch die Prüfung von Merkmalen, die einen Einfluss auf die Schwierigkeit von Leseverstehensaufgaben haben. Unter Einbezug unterschiedlicher Modelle kommunikativer Kompetenz wird zuletzt dargestellt, wie das Konstrukt „Zuhören“ in den IQB-Testaufgaben operationalisiert wurde.

Schwierigkeitsbeeinflussende Merkmale werden jedoch nicht nur aus den Rahmenbedingungen gesprochener Sprache und des Hörverstehens sowie den psycholinguistischen Annahmen der Informationsverarbeitung abgeleitet. Auch Vorarbeiten im Rahmen nationaler und internationaler Studien sowie relevante Dokumente wie der Gemeinsame Europäische Referenzrahmen für Sprachen oder die Lehrpläne der Länder fanden zur Ableitung von Merkmalen Beachtung.

Aus Studien zum Lese- und Hörverstehen in der Erst- sowie Zweit- und Fremdsprache werden weitere Merkmale identifiziert, die an den IQB-Aufgaben untersucht werden. Diese Merkmale lassen sich grob in Merkmale gliedern, die die Stimuli und die Items betreffen. Einige Merkmale beruhen auf einer Interaktion der Items mit den Stimuli. Auch personenbezogene Merkmale werden in die Analysen mit einbezogen.

3.1. Rahmenbedingungen des Zuhörens

Obwohl mündliche Kommunikation menschheitsgeschichtlich älter als schriftliche Kommunikation ist, wird der gesprochenen Sprache aufgrund ihrer Flüchtigkeit weniger gesellschaftliche Bedeutung beigemessen. Die besonderen Umstände in der Art (z. B. Übertragung durch Schallwellen) und Produktion (z. B. in Echtzeit durch mehrere Gesprächspartner) gesprochener Sprache resultieren in bestimmten Bearbeitungsprozessen durch Zuhörer und Sprecher sowie Merkmalen, welche die Verständigung entscheidend beeinflussen.

Gesprochene Sprache passt sich stärker der Kommunikationssituation und dem Zweck eines Gesprächs an und ist deshalb vielfältiger und varianzreicher als geschriebene Sprache. Da mehrere Personen an der mündlichen Kommunikation beteiligt sind, ist sie ein kooperativer Prozess und entwickelt sich in der Zeit. Bei der in Echtzeit ablaufenden Verständigung steht für die Verarbeitung und Produktion von Gesprächsbeiträgen nur eine begrenzte Zeitspanne zur Verfügung. Die Gesprächsbeteiligten müssen in der Regel beinahe gleichzeitig Gesagtes aufnehmen, verarbeiten und daraufhin ihren Beitrag planen und erstellen. Diese Prozesse erfordern von den Sprechenden, aber auch von der Zuhörerschaft bestimmte Gedächtnisleistungen. Auch rein beobachtende Personen, die nicht am Gespräch selbst partizipieren, müssen eine gewisse Gedächtniskapazität für die Rezeption auditiven Materials bereithalten, da die eintreffenden Informationen mental im Gedächtnis repräsentiert werden müssen und nicht schriftlich fixiert vorliegt. Bei der Bearbeitung von Testaufgaben zum Zuhören und zum Lesen dürfte hier der gravierendste Unterschied liegen. Im Gegensatz zum Leser muss der Zuhörer den Sprachfluss unmittelbar in Elemente gliedern, die vom Kurzzeitgedächtnis aufgenommen werden können. Diese phonische Gliederung ist für das Lesen nicht notwendig, da der Leser jederzeit Textstellen wiederholt aufsuchen kann, um bestimmte Informationen zu erhalten.

Ein Gespräch entsteht immer im Zusammenspiel unterschiedlichster Teilhandlungen, wobei die Beteiligten gemeinsam die Bearbeitung der Gesprächsaufgaben übernehmen. Die Verständigung ist abhängig von den jeweiligen Gesprächspartnern, die in Interaktion miteinander stehen und abwechselnd verschiedene Gesprächsrollen einnehmen. Die typische Rollenverteilung in einem Gespräch ist die Zwei-Parteien-Kommunikation. Bei den IQB-Aufgaben ist sie aufgrund der Testsituation allerdings nicht gegeben. Die Schüler stellen eine reine Zuhörerschaft dar, ohne die Möglichkeit, sich an der Kommunikation zu beteiligen. Aus diesem Grund ist auch die wechselseitige Wahrnehmung der Kommunikationssituation, bei der sich alle Gesprächsbeteiligten gegenseitig registrieren und im Idealfall aufeinander eingehen und voneinander lernen, während des Tests für die Schüler nicht gegeben. Einerseits spielt dies eine zu vernachlässigende Rolle, da sich die Jugendlichen nicht aktiv an der Kommunikation beteiligen und deshalb auf Signale des Gesprächspartners auch nicht zur Aufrechterhaltung des Gesprächs reagieren müssen. Andererseits werden aber ihre eigenen Signale, die z. B.

Unverständnis ausdrücken, ebenfalls nicht wahrgenommen und das Gespräch, das die Schüler hören, wird ohne Rücksicht vollzogen.

Bei der Untersuchung von mündlicher Kommunikation müssen auch ihre Teilbereiche nonverbale, wahrnehmungs- und inferenzgestützte sowie verbale Kommunikation Beachtung finden. Nonverbale, wahrnehmungs- und inferenzgestützte Kommunikation haben keine direkte Entsprechung in der textbasierten Verständigung. Bei der Analyse der Audio- Materialien im Rahmen dieser Arbeit können die Wahrnehmungen, die für die wahrnehmungs- und inferenzgestützte Kommunikation eine Rolle spielen, sowie paraverbale Elemente und Geräusche nicht erfasst werden. Registriert werden können in einem gewissen Rahmen jedoch die umgesetzten Ergebnisse, wie beispielsweise Ellipsen oder Pausen. Die Analysen dieser Arbeit werden sich im Wesentlichen auf die Ergebnisse der verbalen Kommunikation, d. h. die Tondokumente und weniger auf ihre Entstehung beschränken.

Die Produktion einer Äußerung wird von den Gesprächspartnern wechselseitig in ihrer zeitlichen Abfolge wahrgenommen und Diskurse sind als gemeinsames Ergebnis aller Gesprächsbeteiligten zu sehen. Die besonderen Merkmale dieser Ergebnisse sind überwiegend auf lexikalischer und syntaktischer Ebene zu finden, weshalb bei den Analysen für diese Arbeit auf diesen Gebieten ein Schwerpunkt liegt. Besonderes Augenmerk wird dabei auf typische Merkmale gesprochener Sprache gelegt, wie Ellipsen, Referenz-Aussage-Strukturen und Konstruktionsabbrüche, sogenannte Anakoluthe. Auch deiktische Ausdrücke werden genauer im vorliegenden Stimulusmaterial ermittelt. Durch die gemeinsame Umgebung und die Kopräsenz der Gesprächsparteien während eines Gesprächs werden deiktische Ausdrücke möglich und die Redekonstellation und das Verhältnis der Sprechenden zueinander spielt eine direktere Rolle als bei der schriftlichen Kommunikation. Da die Schüler an den Gesprächen jedoch nur als Zuhörer teilhaben, und die Bezugspunkte für die deiktischen Ausdrücke inferieren müssen, könnten gerade diese Ausdrücke das Verständnis der IQB-Aufgaben erschweren.

Die Wahrnehmung akustischer Informationen erfolgt zeitlich gegliedert und alle Verarbeitungsprozesse treten erst nach der Reizaufnahme ein. Die Sprachwahrnehmung wird durch die unterschiedlichen Sprachvarietäten und individuellen Ausprägungen gesprochener Sprache erschwert. Im Gegensatz zur geschriebenen Sprache spielen bei der Verarbeitung gesprochener Sprache auch paralinguistische Merkmale eine wichtige Rolle. Es wird deutlich, dass während des Zuhörens nicht nur der sprachliche Input wahrgenommen und verarbeitet werden muss, sondern dass zum erfolgreichen Zuhören auch die Bildung und Aufrechterhaltung einer Intention zur Selektion, die Wahrnehmung und Verarbeitung paraverbaler Informationen, der Sprechermerkmale sowie der Situationsmerkmale gehören. Die Verarbeitung auditiver Informationen erweist sich dadurch als äußerst komplex und stellt hohe Anforderungen an den Zuhörer. Besondere Bedeutung fällt dabei der Anzahl der Sprecher und den Hintergrundgeräuschen zu. Es ist zu vermuten, dass Stimuli, bei denen mehrere Sprecher auftreten, schwieriger zu verstehen sind. Da ein deutlicher Sprecherwechsel durch sehr unterschiedliche Stimmen jedoch gliedernd wirkt, die Aufmerksamkeit auf sich zieht und dadurch eher konzentrationsfördernd wirkt, ist eine Schwierigkeitssteigerung nur bei Aufgaben mit mehreren sehr ähnlichen Stimmen wahrscheinlich. Ein höherer Anteil an Hintergrundgeräuschen

führt wahrscheinlich zu einer Zunahme der Itemschwierigkeit, da die Aufmerksamkeit der Zuhörer stärker beansprucht und deshalb weniger lang aufrechterhalten werden kann. Um diesen Effekt zu vermeiden, werden die Stimuli noch vor ihrem Einsatz im Test am IQB auf ihre Klangqualität hin geprüft und so weit wie möglich mithilfe eines professionellen Tonstudios optimiert. Um dennoch mögliche Effekte zu erkennen, geben die Schüler zusätzlich während des Tests für alle Stimuli auf einer fünfstufigen Skala an, wie gut sie die Texte/Diskurse hören konnten.

Entscheidend für die Aufnahme und Verarbeitung der eintreffenden Informationen ist zunächst die Zuwendung von Aufmerksamkeit auf den Sprachfluss. Sie wird durch Signalfaktoren und motivationale Faktoren beeinflusst. Die Informationen, die aufgrund von Aufmerksamkeitszuwendung wahrgenommen werden, werden zunächst in verschiedenen kognitiven Systemen im Arbeitsgedächtnis bearbeitet. Bezüglich des Arbeitsgedächtnisses unterscheidet sich Zuhören vom Lesen zum einen dadurch, dass für die Verarbeitung auditiver Informationen nur ein einziges Arbeitssystem zuständig ist, zum anderen aber auch durch eine erhöhte Störanfälligkeit der Arbeitsprozesse. In Baddeleys Modell spielt besonders die Gedächtnisspanne eine Rolle für die Kapazität des Arbeitsgedächtnisses. Eine Überlastung der Systeme durch zu viele Reize führt zu einem sogenannten Flaschenhals, einem Engpass in der Informationsverarbeitung. Aufgrund der hohen Präsentationsgeschwindigkeit können auditiv häufig nicht alle Informationen erfasst werden. Flaschenhälse können durch Übung und Automatisierung vermieden werden. Es ist zu vermuten, dass Stimuli mit sehr vielen, dicht aufeinanderfolgenden Informationen eher zu einem Flaschenhals führen. Die Dichte der aufeinanderfolgenden Informationen wird anhand der Anzahl von Propositionen ermittelt. Es wird erwartet, dass Stimuli, die viele Propositionen aufweisen, schwieriger sind als Stimuli mit wenigen Propositionen.

Für die Arbeit mit den IQB-Aufgaben sind Studien zu Doppelaufgaben, bei denen zwei unterschiedliche Tätigkeiten gleichzeitig ausgeführt werden müssen, von größter Bedeutung. Da die Schüler bei den IQB-Aufgaben die Geschwindigkeit der eintreffenden Informationen während des Zuhörens nicht beeinflussen können und bei einigen Aufgaben die Items noch während des Hörens lösen müssen, handelt es sich hierbei auch in gewisser Weise um Doppelaufgaben. Der Fokus selektiver Aufmerksamkeit liegt i. d. R. innerhalb einer Modalität. Auch Doppelaufgaben können jedoch gut bewältigt werden, wenn die Modalitäten getrennt voneinander sind. Es gilt zu prüfen, ob das Bearbeiten der Items während des Hörens einen Einfluss auf die Itemschwierigkeit hat, oder ob die Prozesse des parallelen Lesens und Schreibens so weit automatisiert sind, dass kein Effekt zu beobachten ist. Um Ermüden und Leistungsabfall zu vermeiden und die Aufmerksamkeit der Schüler auf einem hohen Niveau zu halten, wird bei einer reinen Testzeit von insgesamt 120 Minuten in den Testheften alle 20 Minuten der Kompetenzbereich gewechselt sowie nach 60 Minuten eine 15-minütige Pause eingelegt.

Nach der Verarbeitung des auditiven Inputs in der artikulatorischen Schleife werden die Informationen im Idealfall ins Langzeitgedächtnis übertragen. Da für die Verarbeitung eines Lautstroms nur die artikulatorische Schleife und nicht das gesamte Arbeitsgedächtnis zur Verfügung stehen, dauert die Aufnahme von auditiven Informationen i. d. R. länger als von ge-

schriebenen. Im Langzeitgedächtnis werden die Informationen als mentale Repräsentationen gespeichert, wobei verbale Informationen tendenziell eher linear gespeichert werden. Andere Theorien gehen davon aus, dass Wissen in Begriffen gespeichert wird, die in semantischen bzw. propositionalen Netzwerken organisiert sind und assoziativ abgerufen werden können. Sowohl bei den semantischen als auch bei den propositionalen Netzwerken spielt es eine Rolle, wie ausgeprägt die Netzwerke beim jeweiligen Zuhörer sind, d. h. wie viel Erfahrung er mit dem entsprechenden Themengebiet bereits gemacht hat oder wie gut er sich damit auskennt. Bei den IQB-Tests haben die Schüler die Möglichkeit zu den Aufgaben auf einer fünfstufigen Skala anzugeben, ob sie bereits vor dem Test Kontakt mit den Stimuli hatten (wenn sie einen Stimulus beispielsweise schon im Radio gehört haben) und wie gut sie sich mit der darin entfalteten Thematik auskennen. Es wird vermutet, dass bei Vorkenntnissen auch die entsprechenden Netzwerke stärker ausgebildet sind und die Informationsverarbeitung schneller erfolgen müsste. Konkret würde dies bedeuten, dass diese Schüler dem Stimulus mehr Informationen entnehmen könnten und dementsprechend auch die dazugehörigen Items besser beantworten müssten.

Bei der Speicherung von Informationen sind jedoch nicht nur die begriffliche Organisation der Informationen von Bedeutung, sondern auch die den Informationen zugrunde liegenden Skripte und Schemata. Diese sind Voraussetzung für angemessenes Handeln und in Fällen, in denen sie nicht mit den Skripten und Schemata des Sprechers übereinstimmen, kommt es im Verstehensprozess zu Missverständnissen. Schlussfolgerungen werden auf der Basis von gespeichertem Wissen gezogen. Je mehr Vorwissen der Zuhörer zu einem Thema hat, desto passendere Schemata und Skripte stehen ihm zur Verfügung, die ihm korrekte Schlussfolgerungen erlauben. Die Abfrage, ob die Stimuli bereits bekannt sind und wie gut sich die Schüler mit den Themen auskennen, ist also auch aus der Perspektive der Skript- und Schema-Theorie sinnvoll.

Beim Abruf von Informationen wurde beobachtet, dass Informationen, die wiedererkannt werden müssen im Gegensatz zu Informationen, die frei reproduziert werden sollen, leichter erinnert werden. Diese Ergebnisse sprechen dafür, dass geschlossene Aufgabenformate, insbesondere das Format Multiple-Choice, einfacher sind als halboffene oder offene Formate. Wird eine Information in einem Item genauso abgefragt, wie sie im Stimulus präsentiert wird, so ist in diesen Fällen ein (halb-)offenes Format einfacher. Bei den IQB-Aufgaben werden deshalb die Itemformate untersucht und mit den Itemschwierigkeiten korreliert. Zusätzlich dazu werden auch die Anforderungsbereiche betrachtet, da diese Aufschluss über die kognitiven Aktivitäten geben, die zur Beantwortung der Items notwendig sind.

Durch die Darstellung der beim Sprachverstehen relevanten Prozesse wird deutlich, in welchen Bereichen vor allem Schwierigkeiten bei der Verarbeitung auditiven Materials auftreten können. Verständnisprobleme können beispielsweise durch mangelndes Wissen über Wortbedeutungen auftreten. Dabei ist dies i. d. R. eher ein Problem von Fremd- und Zweitsprachenlernern als von Muttersprachlern, denn es ist anzunehmen, dass sie aufgrund von fehlenden lexikalischen und grammatikalischen Kenntnissen häufiger Schwierigkeiten bei der Identifikation von Wortgrenzen und der Zuordnung der korrekten Wortbedeutung haben. Außerdem schei-

nen Mehrdeutigkeiten, Negationen, zur Beantwortung eines Items notwendige Inferenzen oder Hintergrundwissen sowie Referenzen, die Kohärenz im Stimulus erzeugen, die Verarbeitung von Informationen zu erschweren. Alle genannten Aspekte werden an den IQB-Stimuli untersucht.

3.2. Modell des Zuhörens als Grundlage für diese Arbeit

Modelle zur Informationsverarbeitung wurden in der Vergangenheit von unterschiedlichen Forschergruppen entwickelt. Die einflussreichsten Modelle zum Leseverstehen stammen von Kintsch und van Dijk (1983) sowie von Just und Carpenter (1992). Zum Verständnis von Multiple-Choice-Items, bestehen u. a. Verarbeitungstheorien von Embretson und Wetzel (1987) sowie Sheehan und Ginther (2001).

Die Modelle von Kintsch und van Dijk (1983) („Construction-Integration-Modell") und von Just und Carpenter (1992) („Capacity Theory of Comprehension") legen den Fokus auf kognitive Prozesse, Vorwissen und Ressourcen bzw. Arbeitsgedächtniskapazität des Zuhörers und nicht primär auf das zu verarbeitende Material. Dennoch lassen die beiden Theorien vermuten, dass das Verständnis von Stimuli schwieriger wird, je mehr Propositionen aufzunehmen sind, denn lange bzw. dichte Stimuli beanspruchen auch das Arbeitsgedächtnis des Zuhörers in stärkerem Maße als kurze. Überschreitet die Anzahl der Propositionen die Arbeitsgedächtniskapazität, so kommt es zu einem Flaschenhals bei der Informationsaufnahme und das Verstehen ist gestört. Es ist demnach anzunehmen, dass die Länge der Stimuli mit der Itemschwierigkeit korreliert, denn mit zunehmender Länge erhöht sich i. d. R. auch die Anzahl der Propositionen. Auch die Entwicklung einer kohärenten Repräsentation dürfte sich erschweren, je mehr Informationen dafür berücksichtigt werden müssen. Hier spielt natürlich eine Rolle, wie kohärent der Stimulus z. B. aufgrund von Junktionen oder kohärenzstiftenden Ausdrücken überhaupt ist. Je mehr derartige Anhaltspunkte ein Stimulus aufweist, umso leichter dürfte es für den Rezipienten sein, Kohärenz zu erkennen. Nicht zuletzt spielt auch eine Rolle, wie leicht der Attraktor bzw. die Distraktoren auf Basis der Stimulusinformation als solche zu erkennen sind.

Die beiden kognitive Prozessmodelle für das Lösen von Multiple-Choice Leseverstehensitems von Embretson und Wetzel (1987) und Sheehan und Ginther (2001) gehen von der Annahme aus, dass der Anteil der zum Finden der vom Item verlangten Information im Stimulus benötigten kognitiven Arbeit maßgeblich die Schwierigkeit von Items beeinflusst. Die Autoren legen daher ihren Schwerpunkt stärker auf den Umgang mit den Items und weniger auf das Verständnis des Stimulus. Beispielsweise unterscheidet das Modell von Embretson und Wetzel (1987) zwischen dem eigentlichen Verstehen des Lesetextes und dem Lösen der Items. Das Modell von Sheehan und Ginther (2001) fokussiert hingegen auf dem Zusammenspiel von Stimulus und Items. Für die Schwierigkeit ist nach diesem Modell die Lokalisation der zur Itemlösung relevanten Information mit ausschlaggebend. Die Unterscheidung zwischen den beiden Arten von Lokalisationseffekten scheint eher theoretisch. Praktisch wäre die Schwierigkeit von Items in beiden Fällen dadurch zu variieren, wie räumlich nah die Formulierungen bzw. Inhalte in Fragestamm und Antwortoptionen im Stimulus beieinander stehen. Nach Sheehan und Ginther (2001) wirkt es sich ferner auf die Itemschwierigkeit aus, wie sehr sich aufgrund der Lektüre mit dem im Item erfragten Aspekt beschäftigt wurde, also wie stark dieser elaboriert worden ist.

Die beschriebenen Modelle beziehen sich in erster Linie auf Studien zum Leseverstehen. Eignen sie sich auch dazu, die Prozesse des Hörverstehens zu verdeutlichen? Um diese Frage zu beantworten soll noch einmal kurz auf die wichtigsten Bedingungen gesprochener und geschriebener Sprache und die daraus resultierenden Prozesse eingegangen werden. Mündliche Kommunikation zeichnet sich im Gegensatz zu schriftlicher Kommunikation dadurch aus, dass die Sprachbeiträge in der laufenden Interaktion verankert sind und aufgrund z. T. sehr schneller Sprecherwechsel i. d. R. nur wenig Zeit zur Verarbeitung des Gesagten zur Verfügung steht. Aufgrund dessen ist gesprochene Sprache viel stärker situationsgebunden und integriert häufig implizit Bezugspunkte der Sprechsituation, beispielsweise durch deiktische Elemente oder nonverbale Hinweise. Die Kommunikationssituation findet natürlich auch in der geschriebenen Sprache Beachtung, wird dort aber gewöhnlich explizit verbalisiert. Aus der Übertragung der Inhalte gesprochener Sprache durch Schallwellen resultiert zum einen Flüchtigkeit des Gesagten und zum anderen das Merkmal der Prosodie. Schriftliche Sprache ist dagegen dauerhafter immer wieder nachlesbar und bedient sich nicht prosodischer Elemente, wie Tonhöhe, Lautstärke oder Sprechgeschwindigkeit.

Gedächtniskapazität spielt also bei der Verarbeitung gesprochener Sprache eine viel wichtigere Rolle, da der Lautfluss in kurzer Zeit verarbeitet werden muss und nach der Darbietung nicht nochmals konsultiert werden kann. Zusätzliche Reize, wie die Differenzierung mehrerer ähnlich klingender Gesprächsteilnehmer oder das Filtern von störenden Hintergrundgeräuschen benötigen deshalb häufig mehr Verarbeitungsressourcen als ähnliche Faktoren bei der Verarbeitung geschriebener Sprache erfordern würden. Nach Baddeleys Modell des Arbeitsgedächtnisses (2002) wird der eingetroffene Lautflusses nach einer ersten sehr kurzen Speicherung im echoischen Gedächtnis in der artikulatorischen Schleife verarbeitet. Visuelle Informationen hingegen werden zunächst im ikonischen Gedächtnis zwischen gespeichert um dann im räumlich-visuellen Notizblock verarbeitet zu werden. Abgesehen davon unterscheiden sich die Verarbeitungsprozesse jedoch nicht voneinander. Für Informationen beider Modalitäten gilt, dass für sie mentale Repräsentationen gebildet werden müssen, die in die bestehenden kognitiven Strukturen integriert werden müssen. Auch Annahmen über die inhaltliche und/oder grammatische Natur weiterer eintreffender Informationen werden bei der Verarbeitung sowohl gesprochener als auch geschriebener Sprache getätigt.

Abbildung II-3.2. gibt einen zusammenfassenden Überblick über die beim Zuhören ablaufenden Prozesse: Auf den Zuhörer treffen flüchtige Schallwellen, die Äußerungsinformationen des Sprechers enthalten. Aufgrund von Faktoren wie Motivation oder der Absicht, Informationen auszuwählen und aufzunehmen, wird dem Sprachfluss Aufmerksamkeit zugewendet. Erst dann wird er zeitlich gegliedert wahrgenommen und nach kurzer Speicherung im echoischen Gedächtnis ins Kurzzeitgedächtnis übertragen. Der Lautfluss wird einerseits vom Sprecher phonisch gegliedert dargeboten, muss aber auch vom Zuhörer phonisch gegliedert wahrgenommen werden, um weiter in der artikulatorischen Schleife verarbeitet werden zu können. Bereits hier spielen Faktoren wie Sprechermerkmale, Situationsmerkmale oder Hintergrundwissen des Zuhörers für die weitere Verarbeitung eine Rolle, denn der Lautstrom enthält i. d. R. nonverbale und verbale Informationen sowie Informationen, die auf Wahrnehmungen und Inferenzen des Zuhörers basieren. Während der Verarbeitung der eintreffenden Informationen

werden im Wesentlichen mentale Repräsentationen davon erstellt, und diese werden soweit möglich in die bestehenden kognitiven Strukturen integriert. Gleichzeitig werden ständig inhaltliche und grammatische Annahmen über die noch erwarteten Informationen gebildet. Gesicherte Inhalte, d. h. Informationen, die als mentale Repräsentationen in die bestehenden kognitiven Strukturen integriert werden konnten, werden ins Langzeitgedächtnis übertragen.

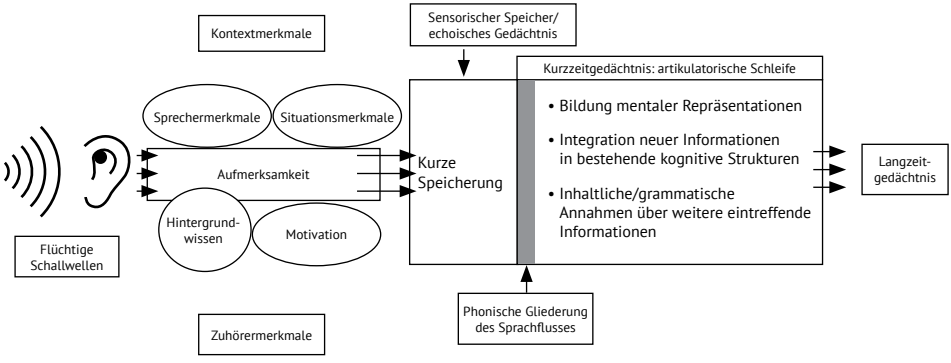


Abbildung II-3.2.: Zuhörprozesse

Ob die Verarbeitung gesprochener und geschriebener Sprache, bis auf Unterschiede in der Wahrnehmung, tatsächlich unterschiedlich verläuft, bleibt weiterhin ungeklärt. Derzeit gibt es noch keinen Konsens darüber, ob unterschiedliche Dimensionen der Sprachkompetenz überhaupt existieren, welche es sein könnten und wie sie klassifiziert werden könnten. (Alderson, 2000: 10) Ebenso ungeklärt ist, ob die beiden rezeptiven Fertigkeiten Lesen und Zuhören aus den gleichen oder unterschiedlichen Teilkompetenzen bestehen. (Levine & Revers, 1988; Lund, 1991) Diese Fragen haben in letzter Zeit zahlreiche Forschergruppen dazu animiert, Dimensionsanalysen durchzuführen, mit deren Hilfe untersucht wird, ob sich sprachliche Kompetenz auch empirisch in Teilkompetenzen wie Zuhören und Lesen differenzieren lässt.

3.3. Operationalisierung des Konstrukts „Hörverstehen" in Testaufgaben

Modelle kommunikativer Kompetenz verdeutlichen, welche Anforderungen – abgesehen von den Prozessen der Informationsverarbeitung – an die Schüler im Bereich des Hörverstehens gestellt werden und in Sprachtests Beachtung finden müssen und helfen dabei, das zu testende Konstrukt in Aufgaben zu übersetzen. Das Modell von Canale und Swain (1981) berücksichtigt neben dem Kontext der Sprachverwendung auch Gesprächsstrategien und den Diskurs. Cummins (1979) führt zum einen durch seine Unterscheidung in alltagssprachliche Fähigkeiten und schulisch-akademische Sprachfähigkeiten das Konzept der Kompetenzstufe ein, zum anderen betont er, dass zur Bewältigung einer sprachlichen Aufgabe sowohl die Kontexteinbettung als auch die durch die Aufgabe benötigten kognitiven Anforderungen eine Rolle spielen. Damit lenkt er die Aufmerksamkeit zum ersten Mal auch eindeutig auf die Testpersonen selbst. Das gebräuchlichste Modell stammt von Bachman und Palmer (1996), die das Modell von Canale und Swain weiterentwickelten. Sie unterscheiden im Rahmen kommunikativer Sprachkompetenz die Bereiche Sprachwissen, strategische Kompetenz und psychophysiologi-

sche Mechanismen und binden damit das formale System von Sprache an die Besonderheiten des Sprachgebrauchs an. Daneben finden auch die Gesprächsbeteiligten und der Kontext der Sprachverwendung Beachtung. Ähnlich wie Cummins berücksichtigen Bachman und Palmer auch die Aufgaben selbst, die unterschiedliche Anforderungen an die Sprachanwender stellen.

Das Modell von Bachman und Palmer (1996) diente als Grundlage für die Entwicklung von Testaufgaben für den Kompetenzbereich Zuhören. In vereinfachter Form wurden bei der Aufgabengenerierung die in Abbildung II-3.3. dargestellten Aspekte berücksichtigt.

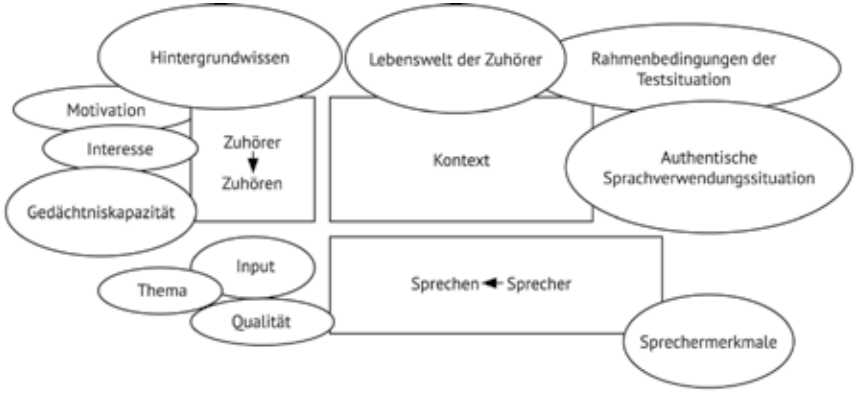


Abbildung II-3.3.: Am Zuhörprozess beteiligte Aspekte

Dementsprechend wurden die Aufgaben so entwickelt, dass die Stimuli thematisch den Interessen und der Lebenswelt der Jugendlichen entsprechen und dadurch zum Zuhören motivieren. Sie setzen im Idealfall kein Hintergrundwissen voraus und sind thematisch weitgehend neutral, ohne starke Emotionen bei den Schülern hervorzurufen. Starke emotionale Reaktionen könnten auch einen Einfluss auf das Testergebnis haben. Die Stimuli sollen authentische Sprachverwendungssituationen enthalten und keine für die Testsituation didaktisierten Materialien sein. Gleichzeitig wurde darauf geachtet, dass sie von ihrer Aufnahmequalität so hochwertig sind, dass unbeabsichtigte Störgeräusche – und damit zusätzliche Beanspruchungen von Gedächtniskapazität – weitgehend ausgeschlossen werden können. Im Idealfall werfen die Items Fragen auf, die auch beim Umgang mit dem Stimulusmaterial außerhalb der Testsituation naheliegend gewesen wären. Die Schüler sollen im Umgang mit den Stimuli so weit wie möglich die Sprachkompetenz einsetzen müssen, die auch außerhalb der Testsituation von ihnen erwartet werden würde.

Den Aufgabenentwicklern wurden thematisch kaum Vorgaben gemacht, in der Hoffnung, auf diese Weise eine große Vielfalt an Aufgaben zu erhalten. Ausgeschlossen wurden lediglich Stimuli, die in der Testsituation verstörend und damit leistungsbeeinflussend auf die Schüler wirken könnten, wie beispielsweise Themen zu Krieg, Verkehrsunfällen, Scheidung der Eltern,

etc. Der entwickelte Aufgabenpool weist eine große Bandbreite an unterschiedlichen Stimuli und Items auf. Zur Erstellung von Kompetenzprofilen und zur Erfassung der in den Bildungsstandards beschriebenen Fertigkeiten, eignen sich diese Aufgaben sehr gut. Generalisierbare Aussagen über den Einfluss bestimmter Merkmale auf die Item- bzw. Aufgabenschwierigkeit sind jedoch schlechter möglich, da die einzelnen Merkmale bei jeder Aufgabe anders miteinander kombiniert sind und konfundierende Einflüsse nicht experimentell auszuschließen sind.

3.4. Ausblick: Ableitung von weiteren Merkmalen

In den folgenden Kapiteln wird ein Überblick über große Schulleistungsstudien gegeben, um zu verdeutlichen, an welcher Stelle die Arbeiten des IQB ansetzen. Bei national und international durchgeführten Studien wie z. B. den Vergleichsarbeiten (VERA) oder PISA gab es in der Vergangenheit kaum Testteile zur Überprüfung des Hörverstehens in der Muttersprache. Merkmale, welche einen Einfluss auf die Aufgabenschwierigkeit haben, wurden bei PISA (OECD, 2002) jedoch für den Kompetenzbereich Lesen identifiziert. In der DESI-Studie (Klieme et al., 2003) wurden die Stimulus- und Aufgabenschwierigkeit beeinflussenden Merkmale für die Hörverstehensaufgaben in Englisch und für die Leseverstehensaufgaben in Deutsch untersucht. Dabei dienen die bei PISA und DESI verwendeten Kategorien unter anderem als Basis für die IQB-Untersuchungen.

Ausführliche Kriterienraster zu schwierigkeitsbeeinflussenden Merkmalen finden sich im fremdsprachlichen Bereich im Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER) (Europarat, 2001) und im Dutch Grid (Alderson et al., 2006). Der GER wurde als Rahmen für Sprachkompetenzmodelle konzipiert und geht neben zur Kommunikation benötigten Kenntnissen und Fertigkeiten auch auf den kulturellen Kontext der Sprachverwendungssituation ein. Er beschreibt für Fremdsprachenlerner für alle Kompetenzbereiche drei verschiedene Referenzniveaus ausführlich mit Deskriptoren. Da sich vor allem im Bereich der Kompetenten Sprachverwendung (Niveau C) die Beschreibungen auch für die Sprachverwendung in der Erstsprache eignen, wird in dieser Arbeit auf die Kriterienraster des GER zurückgegriffen. Daneben gibt es im GER Angaben über Arten gesprochener Stimuli, Höraktivitäten sowie Hörabsichten, die in der Lern- und Testsituation Beachtung finden könnten. Der GER unterscheidet Merkmale, welche Kommunikation erschweren, und zwar nach „Materiellen Bedingungen", „Soziale Bedingungen" und „Zeitdruck". Dabei spielen soziale Bedingungen wie „Relativer Status der Teilnehmenden", „An-/Abwesenheit von Zuhörenden" oder „Soziale Beziehungen zwischen den Teilnehmenden" für die IQB-Aufgaben kaum eine Rolle, da es sich hier um eine Testsituation handelt, in der die Schüler als Zuhörer aufgezeichneten Stimuli folgen. Allein der Aspekt „Anzahl der Gesprächspartner" könnte bei den Analysen relevant werden.

Die vorgestellten Modelle zum Sprachverstehen, die Ergebnisse aus nationalen und internationalen Schulleistungsstudien sowie der GER und das Dutch Grid sind hilfreiche Ergänzungen, da in den Bildungsstandards selbst kaum konkrete Aussagen zum Bereich „Zuhören" gemacht werden. Angaben zu schwierigkeitsbeeinflussenden Merkmalen finden sich in der KMK-Handreichung überhaupt nicht (vgl. KMK, 2004). In dem auf der Grundlage der IQB-Aufgaben entstandenen Kompetenzstufenmodell werden die einzelnen Kompetenzstufen aber durch Niveau typische Merkmale beschrieben, die zum einen Schwierigkeitsunterschiede zwischen der

Darbietung der Stimuli bzw. der Stimuli selbst transparent machen, zum anderen aber auch durch kognitive Operationen verdeutlichen, welche Kompetenzen von den Schülern auf der entsprechenden Stufe beherrscht werden sollten. Diese Merkmale werden im Rahmen dieser Arbeit untersucht. Auf diese Weise kann auch eine Validierung der Beschreibungen der einzelnen Kompetenzstufen erfolgen.

Ein Kompetenzstufenmodell wurde auch in Luxemburg in Anlehnung an die Modelle des GER für die deutsche Sprache entwickelt. Deutsch ist in Luxemburg Alphabetisierungs- und Unterrichtssprache, nicht aber Verkehrssprache. Die Besonderheit des Modells ist, dass es für alle Schulformen und Ausbildungszyklen konzipiert ist und für die speziellen Bedürfnisse derselben angepasst werden kann. Dafür werden Merkmale genannt, die es erlauben, das Modell in seiner Komplexität und Schwierigkeit zu erhöhen oder zu verringern. Diese Merkmale beziehen sich auf die Art und Komplexität des Inhalts, die Komplexität des Stimulus und die Art und den Schwierigkeitsgrad der Aufgaben und sind auch für diese Arbeit relevant.

In der Deutschdidaktik wurde der Fokus erst in neuerer Zeit dem Kompetenzbereich Zuhören zugewendet, der hier meist als Teilbereich von Unterrichtsgesprächen, Diskussionen und Debatten auftaucht (vgl. Vogt, 2002), vereinzelt aber auch im Sinne einer Hörerziehung Beachtung findet (z. B. Kinder-/Jugendliteratur und Medien, in Forschung, Schule und Bibliothek/Kjl&m, 08.3., 60(3), 2008). Um die Bedeutung des Hörverstehens für den Unterricht zu erfassen, werden zunächst bundesländerspezifische Lehrplan- und Curriculavorgaben mit den länderübergreifenden Bildungsstandards verglichen. Dabei zeigt sich, dass die Lehrpläne die sehr allgemein formulierten Bildungsstandards konkretisieren, aber kaum Angaben machen, die für die Analyse von Stimuli und Aufgaben hilfreich sein könnten. Obwohl Zuhören und Hörverstehen einen Großteil des Unterrichts einnehmen (vgl. die Videostudien des IPN und zu TIMSS), werden sie zumindest gemäß der Vorgaben in den Lehrplänen im Unterricht kaum thematisiert und vermittelt. Zuhörkompetenzen werden bei den Schülern also vorausgesetzt und die Schüler müssen bereits über Zuhörstrategien verfügen, um beispielsweise Lehrerinstruktionen zum Wissenserwerb nutzen zu können.

3.5. Ausblick: Studien zu schwierigkeitsbeeinflussenden Merkmalen

Die Untersuchung von schwierigkeitsbeeinflussenden Merkmalen zur Vorhersage von Itemschwierigkeiten mit Regressionsmodellen macht empirische Aussagen, welche Item- oder Stimulusmerkmale die Itemschwierigkeit beeinflussen. Itemschwierigkeiten werden durch Merkmale erklärt und die Menge der Itemschwierigkeiten wird auf Regressionsparameter der Merkmale reduziert. Die kausale Interpretation eines einzelnen Prädiktors ist streng genommen nur bei experimenteller Variation des Item- oder Stimulusmaterials im Hinblick auf das zu untersuchende Merkmal manipulierbar. Durch die Analysen wird eine homogene mittlere Wirkung eines Item- bzw. Stimulusmerkmals im Hinblick auf die gesamte Schülerstichprobe untersucht. Die Merkmale werden damit psychometrisch auf die Itemseite geschoben. Mehrdimensionale Modelle (Buck & Tatsuoka, 1998) hingegen fassen Itemmerkmale mit korrespondierenden Attributen als Personenfähigkeiten auf, sodass Attribute auf der Personenseite heterogen (oder differenziell) ausgeprägt sind. Demzufolge gibt es nicht die eindimensionale Zuhörkompetenz, sondern es wird von einem mehrdimensionalen Konstrukt ausgegangen.

Bei der Beschreibung von Merkmalen, die einen Einfluss auf die Itemschwierigkeit haben, gibt es im Wesentlichen zwei unterschiedliche Vorgehensweisen: text- und adressatenzentrierte Ansätze. Bei textzentrierten Ansätzen werden textspezifische Eigenschaften, wie z. B. linguistische Merkmale oder der thematische Aufbau verwendet, um die Aufgabenschwierigkeit vorherzusagen. Adressatenzentrierte Ansätze ermitteln Merkmale, wie z. B. die vom Item verlangte kognitive Operation zur Beschreibung der mithilfe von Normierungsstudien gewonnenen Itemschwierigkeiten, beispielsweise mit einer Regressionsanalyse. Die statistischen Analysen in dieser Arbeit belaufen sich sowohl auf text- als auch auf adressatenzentrierte Ansätze. Einige der beschriebenen Merkmale sind in den Hörtexten selbst verortet, manche beziehen sich überwiegend auf die Items, andere beruhen auf einer Interaktion zwischen Stimulus und Items. Auch einige personenbezogene Merkmale werden im Rahmen dieser Arbeit berücksichtigt. Beispielhaft für einen adressatenzentrierten Ansatz zur Bestimmung der Aufgabenschwierigkeit wird die sogenannte „Rule-Space-Analysis" von Buck und Tatsuoka (1998) dargestellt. Dabei werden die Determinanten der Aufgabenschwierigkeit sowohl theoretisch als auch empirisch begründet und zu allen Determinanten gibt es exakte Angaben über ihre relative Bedeutung unter Einschluss von Wechselwirkungen.

Merkmale, welche einen Einfluss auf die Stimulusschwierigkeit haben können, sind z. B. thematische Merkmale der Stimuli, wie die thematische Entfaltung oder die propositionale Dichte. Die vorgestellten Befunde zu schwierigkeitserklärenden Stimulusmerkmalen weisen darauf hin, dass Aufgaben zu propositional dichten Stimuli schwieriger sind als zu propositional weniger dichten Stimuli. Merkmale, die sich auf die Präsentation der Stimuli beziehen sind beispielsweise Sprechereigenschaften, wie die Sprechgeschwindigkeit, akustische Eigenschaften der Stimuli, wie die Tonqualität der Aufnahme und ihre Lautstärke, sowie das Merkmal, ob der Stimulus einmal oder mehrmals vorgespielt wird. Aus Gründen der Testfairness gilt generell, dass ein Stimulus die bestmögliche Qualität besitzen soll und mehrmals vorgespielt werden sollte. Dabei ist noch nicht abschließend geklärt, ob die ablaufenden Zuhörprozesse bei mehrmaligem Hören immer dieselben sind. Aus Gründen der Authentizität wird hingegen i. d. R. das einmalige Vorspielen präferiert. Zu den linguistischen Merkmalen der Stimuli gehören beispielsweise die Länge und Anzahl von Stimulusfragmenten, die Komplexität des Wortschatzes, Negationen, Struktur und Kohärenz. Auch die Worthäufigkeit spielt hier eine Rolle. Dabei müssen häufig gebrauchte Wörter aber nicht immer bedeuten, dass der Stimulus auch einfacher ist, da sie z. T. viele Bedeutungen haben können, die erst im Rahmen des Kontextes richtig interpretiert werden können. Aber auch die Komplexität des Wortschatzes kann einen Einfluss auf die Schwierigkeit haben. Je höher der Anteil an Fachvokabular, an abstrakten Begriffen oder an Inhaltswörtern ist, desto schwieriger wird häufig der Stimulus. Bei einem hohen prozentualen Anteil an Inhaltswörtern wird davon ausgegangen, dass diese Stimuli inhaltlich sehr dicht sind und hohe Ansprüche an die Verarbeitungskapazitäten der Rezipienten stellen. Das Vorkommen von Negationen in den Itemformulierungen sowie dem zur Lösung relevanten Stimulusabschnitt erhöht meist die Schwierigkeit, da bei Negationen i. d. R. erst eine positive mentale Repräsentation erstellt wird, die dann in einem zweiten Schritt negiert wird. Für das Merkmal „Kohärenz" gibt es überwiegend Untersuchungen zum Leseverständnis. In mehreren Studien konnte dabei gezeigt werden, dass die Testpersonen bereits über gewisse Lesefertigkeiten bzw. Hintergrundwissen zur Textart bzw. zum Thema verfügen

müssen, um von kohärenteren Stimuli zu profitieren. (vgl. Abrahamsen & Shelton, 1989; Best et al., 2006) Die Abhängigkeit des Verständnisses von der Stimulusstruktur und der Kohärenz untersuchte Yin-Kum (1995). In der Studie wurde zwischen den Struktur-Typen „Kausalität", „Vergleich" und „Aufzählung" differenziert. Die Probanden erinnerten bei Stimuli des Typs „Vergleich" deutlich mehr als bei den anderen Struktur- Typen. Daraus wurde gefolgert, dass vergleichende Stimuli den Aufbau kohärenter mentaler Repräsentationen erleichtern und kohärente Stimuli einfacher zu verstehen sind als weniger kohärente Stimuli.

Zu den Merkmalen, die auf Itemeigenschaften beruhen, gehört z. B. das Itemformat. Bei den IQB-Aufgaben wird zwischen geschlossenen, halb-offenen und offenen Itemformaten unterschieden. Besonderes Forschungsinteresse fiel in der Vergangenheit auf das Multiple-Choice Format, da es zwar das bevorzugte Format im Large-Scale-Assessment ist, aber auch den größten Formateffekt zu haben scheint. Bei der Untersuchung von Multiple-Choice-Items werden die Position des Attraktors im Item und die Plausibilität der Distraktoren berücksichtigt. Zur dritten Kategorie der Merkmale, die auf einer Interaktion des Stimulus mit den dazugehörigen Items beruhen, gehören Merkmale, wie der Zeitpunkt der Itempräsentation, der Grad der Überlappung der Itemformulierungen mit dem Stimulus und die zur Beantwortung eines Items notwendige Information (NI). Von einer Präsentation der Items vor dem Zuhören wird generell abgeraten, da dies eine zusätzliche Belastung des Arbeitsgedächtnisses impliziert, auch wenn die Probanden diese Variante präferierten. Auch die Bearbeitung der Items während des Zuhörens erhöht die kognitive Belastung für die Testpersonen, da simultan mehrere Tätigkeiten ausgeführt werden müssen. Die Präsentation der Items nach dem Zuhören führte zu den besten Ergebnissen, insbesondere dann, wenn der Stimulus danach ein weiteres Mal angehört werden konnte. Auch eine Überlappung der Itemformulierungen mit dem Stimulus führte dazu, dass die Itemschwierigkeit sank. Zur Einschätzung des Abstraktheitsgrades der NI nutzen Evetts und Gauthier (2005) ein fünfstufiges Schema. Sie gehen davon aus, dass die Schwierigkeit steigt, je abstrakter die NI ist. Die Schwierigkeit der Items hängt auch von der kognitiven Operation ab, die notwendig ist, um die NI zu erhalten. Es kann beispielsweise schwieriger sein, die NI durch eine Schlussfolgerung zu erhalten, als sie direkt (oder paraphrasiert) aus dem Stimulus zu entnehmen. Aber auch die Position der NI im Stimulus (am Anfang, in der Mitte, am Ende) und die Auftretenshäufigkeit der NI spielen eine Rolle. Angenommen wird z. B. dass ein Item einfacher wird, je häufiger die benötigte NI im Stimulus dargeboten wird. Schließlich wurden Hinweise darauf gefunden, dass sich Zusammenhänge zur Itemschwierigkeit ergeben, wenn die NI explizit oder implizit im Stimulus enthalten ist. Variablen, wie der kognitive Anspruch bzw. der Inferenztyp, stufen diesen Aspekt feiner ab und ergänzen ihn: Items, zu deren Lösung explizit im Stimulus gegebene Information gefunden werden muss, sind i. d. R. leichter, weil Inferenzen nicht erforderlich sind und der kognitive Anspruch somit gering ist. Das Entnehmen von expliziten aber paraphrasierten Informationen ist kognitiv häufig etwas anspruchsvoller. Die Itemschwierigkeit erhöht sich weiter, wenn mehr oder weniger komplexes Schlussfolgern, das Entnehmen nur implizit enthaltener Information oder sogar das Einbeziehen von Vorwissen verlangt werden.

Neben den Merkmalen der Stimuli und der Items bzw. ihrer Interaktion spielen auch Merkmale der Testpersonen eine entscheidende Rolle. Für diese Arbeit finden die Motivation bzw. das

Interesse der Schüler an den Themen der Aufgaben, ihr Hintergrundwissen dazu, ihre Arbeitsgedächtniskapazität sowie ihre Sprachkenntnisse Beachtung. Zusätzlich wird berücksichtigt, wie gut die Schüler die Stimuli akustisch verstehen konnten.

4. Hörverstehen - Eine kompetenzdiagnostische Perspektive

Im Kapitel „Hörverstehen – Eine kompetenzdiagnostische Perspektive" wird dargestellt, welche nationalen und internationalen Vorarbeiten aus dem Bereich der Didaktik und der empirischen Bildungsforschung für die Identifikation von schwierigkeitsbeeinflussenden Merkmalen berücksichtigt wurden. Um Sprachkompetenz für die Operationalisierung mithilfe von Aufgaben zunächst vereinfacht darstellen zu können, erwies sich insbesondere das Modell der kommunikativen Sprachkompetenz von Bachman und Palmer (1996) als hilfreich (Kapitel 4.1.). Es berücksichtigt neben Sprachwissen auch psycho-physiologische Mechanismen der Testpersonen, strategische Kompetenzen sowie den Kontext der Sprachverwendung. Schwierigkeitsbeeinflussende Merkmale wurden ferner im Rahmen großer Schulleistungsstudien wie DESI (Deutsch Englisch Schülerleistungen International) (Kapitel 4.3.1.) oder PISA (Programme for International Student Assessment) (Kapitel 4.3.2.) identifiziert. Diese Merkmale geben einen Anhaltspunkt für die Beschreibung und Kategorisierung der in dieser Arbeit überprüften Merkmale. Aber auch der Gemeinsame Europäische Referenzrahmen für Sprachen (GER) fand mit seinen Deskriptoren zur Beschreibung von Sprachständen Beachtung (Kapitel 4.2.). Die Einbeziehung zahlreicher zusätzlicher Dokumente in die Untersuchung neben den Bildungsstandards (Kapitel 4.4.1.), den dazu gehörenden Kompetenzstufenmodellen (Kapitel 4.4.2.) und den Rahmenplänen der Länder (Kapitel 4.4.3.) erwies sich als notwendig, da sowohl die Bildungsstandards als auch die Lehrpläne und Curricula Hörverstehen nur am Rande behandeln und kaum bzw. sehr ungenaue Angaben dazu machen. Obwohl der Kompetenzbereich „Sprechen und Zuhören" vor allem in Lehrplänen neueren Datums verstärkt Beachtung findet, fokussieren die Lehrpläne vor allem auf Angaben zum Kompetenzbereich „Gesprächskompetenz", insbesondere zum „Sprechen". In den Bildungsstandards sind die Formulierungen im Bereich „Zuhören" so knapp und vage, dass für die Umsetzung methodische und inhaltliche Anhaltspunkte häufig fehlten. Das auf der Grundlage der IQB-Aufgaben gewonnenen Daten entwickelte Kompetenzstufenmodell „Zuhören" beschreibt neben Mindest- und Regelanforderungen auch motivierende Leistungserwartungen zur schulischen Weiterentwicklung. Die Stufenbeschreibungen enthalten sowohl verstehens- als auch modalitätsspezifische Aspekte des Hörverstehens. Ergänzend dazu wird ein Blick auf die Kompetenzstufenmodelle in Luxemburg geworfen. Da die Modelle curricular übergreifend für alle Schulformen und Ausbildungszyklen einen Orientierungsrahmen abgeben sollen, werden in ihnen auch Angaben dazu gemacht, wie für die unterschiedlichen Leistungsniveaus die Art und Komplexität des Inhalts, die Komplexität des Textes sowie die Art und Schwierigkeit der Aufgaben angepasst werden können.

4.1. Messung von Sprachkompetenz

Sprachkompetenz nimmt im Rahmen schulischer Lehr- und Lernprozesse eine Schlüsselrolle

ein, insofern als durch sie einerseits Sprachkönnen ausgedrückt wird und sie andererseits die Voraussetzung und das Mittel zum Wissenserwerb darstellt. Insbesondere mündliche Kommunikation besitzt im Unterricht die Funktion von Lernmedium, Lerngegenstand und Lernziel, denn die sprachlichen Äußerungen der Lehrkräfte, vor allem bei der Instruktion, und der Lernenden selbst sollen Lernprozesse auslösen. (Rost, 2002) Daneben werden aber auch emotionale Aspekte wie Lob, Kritik oder Mahnungen durch verbale oder nonverbale Kommunikation transportiert (vgl. Becker-Mrotzek & Vogt, 2001). Zuhörkompetenzen spielen in diesem Zusammenhang also eine wichtige Rolle, da ohne sie eine Beteiligung an der Unterrichtskommunikation (und sei es auch nur als Zuhörer) nicht möglich ist.

In der DESI-Studie wird Sprachkompetenz definiert als „Komplex von Teilfähigkeiten, die durch den schulischen Unterricht vermittelt werden sollen." (Jude & Klieme, 2007: 14) Der Begriff wird funktional als kognitive Disposition verstanden, mit deren Hilfe situative Anforderungen erfolgreich bewältigt werden können. Auch Weinert (1999) sieht Sprachkompetenz dementsprechend als Kompetenz mit unterschiedlichen Subdomänen und differenziert Sprachkompetenz in allgemeine Fähigkeiten, funktional bezogene Leistungsdispositionen, motivationale Voraussetzungen, anforderungsbezogene Handlungsfähigkeiten sowie Strategiewissen und übergreifende Kompetenzen. Gerade die motivationalen, volitionalen und sozialen Aspekte spielen beim Erwerb und der Überprüfung von Sprachkompetenz eine wichtige Rolle.

Differentialdiagnostische Ansätze gehen davon aus, dass sich Sprachkompetenz in spezifische Subdomänen aufschlüsseln lässt, wie z. B. in linguistische Elemente wie Morphologie, Syntax oder Lexik oder aufgrund von übergeordneten Strukturen, wie produktive oder rezeptive Teilkompetenzen. Durch die empirische Erfassung dieser Subdomänen werden individualdiagnostische Aussagen möglich. (vgl. Bachman et al., 1990; Bolton, 2000; Butler & Stevens, 2001) Eine Möglichkeit, die Teilbereiche sprachlicher Kompetenz darzustellen, wird in Tabelle II- 4.1. gegeben. Sie unterscheidet Sprachkompetenz einerseits nach ihrem Interaktionsgrad und andererseits nach dem Kommunikationsmodus. Grammatikalisches, soziolinguistisches und pragmatisches Sprachwissen wird dabei in allen Teilbereichen benötigt.

Tabelle II-4.1.: Klassifikation von Teilbereichen sprachlicher Kompetenz

		Interaktionsgrad	
		produktiv	rezeptiv
Kommunikationsmodus	auditiv	Sprechen	Zuhören
	visuell	Schreiben	Lesen

Ziel bei der Messung von Sprachkompetenz ist ein Auswertungsdesign, das neben allgemeinen Aussagen über die erreichte sprachliche Kompetenz auch differenzierte Aussagen über die erbrachten Leistungen in den Teilbereichen der Sprachkompetenz ermöglicht. Aus diesem Grund wird in neuerer Zeit kaum mehr rein deklaratives Faktenwissen erfasst, sondern sprachdiagnostische Verfahren versuchen unterschiedliche Testinstrumente zu integrieren.

So können die unterschiedlichen rezeptiven, produktiven sowie wissensbasierten Teilbereiche sprachlicher Kompetenz am besten abgebildet werden. (vgl. Jude & Klieme, 2007)

Um Sprachkompetenz für Tests zu operationalisieren, behalf man sich mit Modellen, in denen kommunikative Kompetenz vereinfacht dargestellt werden sollte. Ein einflussreiches Modell wurde in den späten 80er Jahren von Canale und Swain (1981) entwickelt. Darin umfasst die kommunikative Kompetenz vier Komponenten: die linguistische, die soziolinguistische, die diskursive und die strategische Kompetenz. Die strategische Kompetenz interagiert mit den anderen drei Kompetenzen und umfasst im Wesentlichen die Beherrschung verbaler und nonverbaler Kommunikationsstrategien zur Aufrechterhaltung des Gesprächs. Die linguistische Kompetenz besteht hauptsächlich aus Sprachwissen, aus der Beherrschung des formalen Systems einer Sprache. Die Kombination dieses Sprachwissens mit den entsprechenden Bedeutungen zu einem sinntragenden Text fällt in den Bereich der diskursiven Kompetenz. Die soziolinguistische Kompetenz hat zusätzlich die Situation der Sprachverwendung im Blick. Hier geht es darum, in verschiedenen Kontexten unter Berücksichtigung beeinflussender Faktoren, wie z. B. dem Status der Gesprächsbeteiligten, angemessene Äußerungen zu produzieren. Dieses Modell berücksichtigte im Gegensatz zu den Ansätzen der 60er und 70er Jahre nicht nur den Kontext der Sprachverwendung, sondern auch Gesprächsstrategien und den Diskurs und erweitert damit das Blickfeld der Testentwicklung.

Das in den späten 80er Jahren veröffentlichte *Modell der kommunikativen Sprachkompetenz/ Communicative Language Ability* (CLA) von Bachman (1991) ist von den Studien von Canale und Swain beeinflusst. Es geht davon aus, dass kommunikative Sprachkompetenz aus Sprachwissen, einer strategischen Kompetenz und psycho-physiologischen Mechanismen besteht. Damit sind Faktoren wie Motivation, die Aufmerksamkeit der Sprachbeteiligten oder ihre anatomischen Voraussetzungen gemeint. In diesem Modell erhalten also Merkmale der Testpersonen mehr Aufmerksamkeit. Sprachwissen umfasst einerseits strukturelles Wissen, wie grammatisches und textuelles Wissen, und andererseits pragmatisches Wissen, wie illokutives und soziolinguistisches Wissen. Die Fähigkeit, dieses Wissen angemessen umzusetzen und in der Interaktion die Sprachverwendung zu bewerten, zu planen und auszuführen wird als strategische Kompetenz bezeichnet. Alle drei Bereiche interagieren mit dem Kontext und den Wissensstrukturen der Sprechenden. Da sowohl das Verständnis von Kontext, als auch die relevanten kognitiven Anforderungen bei den Testpersonen variieren, gilt es für jeden Test gründlich die Testpopulation zu prüfen und auf die Lernenden als Individuen einzugehen.

Bachman differenzierte den Bereich des Sprachwissens noch genauer und stellte 1996 zusammen mit Palmer eine revidierte Version des Modells vor (Bachman & Palmer, 1996) (vgl. Abbildung II-4.1.). In diesem Modell ist das Wissen über das formale System von Sprache angebunden an die Besonderheiten des Sprachgebrauchs. Die erfolgreiche Anwendung des Sprachwissens wird beeinflusst von der Interaktion mit Wissensschemata, dem Wissen von und der Erfahrung mit der Welt sowie affektiven Schemata, wie dem emotionalen Gedächtnis. Dabei ist formalsprachliches Wissen (language knowledge) die Grundlage jeder sprachlichen Interaktion und von rein kommunikativem Sprachhandeln (communicative performance) abzugrenzen. In einem dynamischen, metakognitiven Prozess der Bewertung, Planung und

Zielsetzung interagieren diese Bereiche miteinander. Das Modell berücksichtigt also metakognitive Prozesse, Sprachwissen, affektive Schemata und Wissensschemata des Sprachverwendenden, aber auch die Gesprächsbeteiligten sowie den Kontext der Sprachverwendung. Aufgrund seiner Komplexität eignet sich das Modell sehr gut für die Testentwicklung. Daher bildet es auch in der vorliegenden Arbeit eine Grundlage für die Definition der zu prüfenden Kompetenzbereiche.

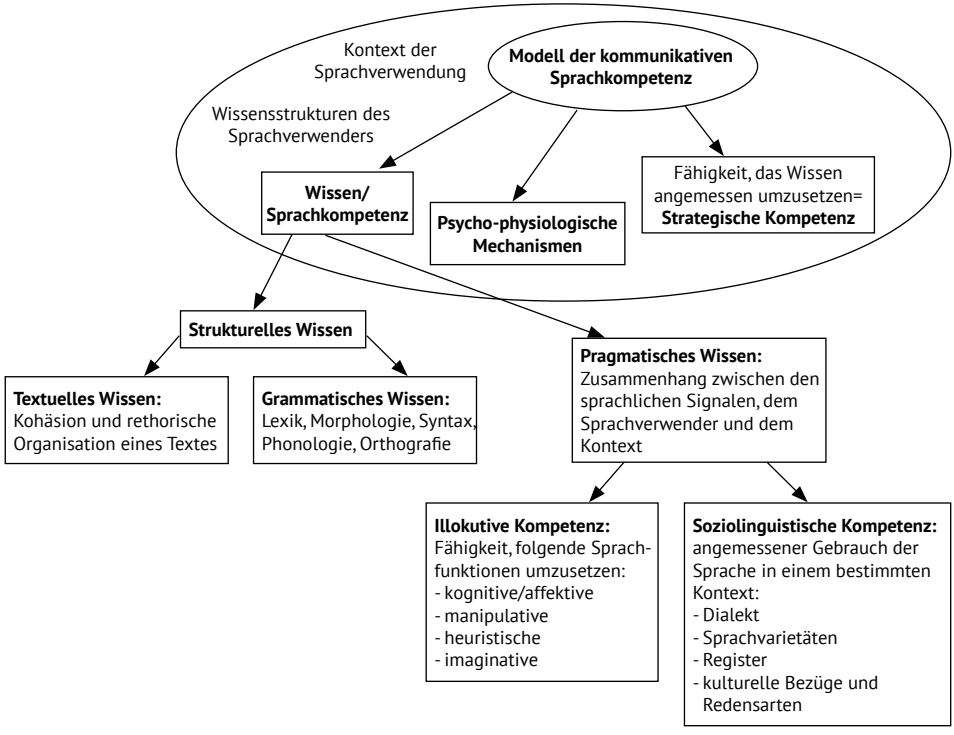


Abbildung II-4.1.: Modell der kommunikativen Sprachkompetenz (Bachman & Palmer, 1996)

Neben der Bestimmung von sprachlichen Kompetenzen legen Bachman und Palmer (1996) in ihren Arbeiten den Fokus auch auf die Beschreibung von Aufgabenmerkmalen. Es wird zwischen grammatikalischen und pragmatisch soziolinguistischen Faktoren der Sprachkompetenz unterschieden, wobei gerade die pragmatisch soziolinguistischen Faktoren stark von den Testaufgaben und der Testmodalität abhängen. Durch die Beachtung von Aufgabenmerkmalen sollen durch eine Modellierung der Aufgaben zusätzlich Hinweise auf die zur Lösung benötigten Teilkompetenzen erhalten werden. Für die Aufgaben können daraus Anforderungsprofile mit unterschiedlichen Schwierigkeitsniveaus erstellt werden, die helfen, die kommunikativen Kompetenzen der Lernenden abzubilden.

„we suggest that the construct definition includes only the relevant components of language ability, and that the >skill< elements be specified as characteristics of the tasks in which language ability is demonstrated" (Bachman & Palmer, 1996: 128)

„[R]elevant components of language ability" (Bachman & Palmer, 1996: 128) im Kontext der Sprachverwendung zu testen, wird oftmals als Forderung verstanden, „authentische" Aufgaben einzusetzen. Dabei kann der Begriff „authentisch" Unterschiedliches bedeuten. Eine Lesart von „authentisch" impliziert beispielsweise, dass ein authentischer Stimulus nicht spezifisch für den Einsatz in einem Test erstellt wurde. Unter einer authentischen Aufgabe versteht man oft eine Aufgabe, deren Merkmale sich stark mit den Merkmalen einer Anforderung oder eines zu lösenden Problems in der Zielsprache decken. Diese Art von Stimuli bieten kulturelle und linguistische Informationen, die in pädagogischen Stimuli häufig verloren gegangen sind. Authentizität in diesem Sinne ist nach Bachman und Palmer (1996) bedeutend, um die erbrachten Testleistungen hinsichtlich ihrer Bedeutung für die Lösung analoger Probleme außerhalb der Testsituation generalisieren zu können. Außerdem erhöht Authentizität die Augenscheinvalidität des Tests, indem die Testpersonen die Aufgaben als relevant betrachten.

Dies kann nicht zuletzt Auswirkungen auf ihre Testleistungen haben. Buck und Tatsuoka (1998) betonen, dass ein Test zur Erhebung von sprachlichen Leistungen Prozesse replizieren sollte, die auch außerhalb der Testsituation zur Kommunikation notwendig sind. Die verwendeten Stimuli sollten dem Material ähneln, mit dem die Testpersonen auch im Alltag konfrontiert werden und die Aufgaben sollten verhältnismäßig realistisch sein, d. h. einen klar definierten Kommunikationszweck in einem bekannten Kontext erfüllen.

Obwohl die Forderung nach authentischen Aufgaben die Qualität sprachlicher Tests deutlich steigerte, gibt es auch Warnungen davor, Authentizität zum alleinigen Gütemaßstab eines Tests zu machen. So fand Lewkowicz (1997) bei ihren empirischen Untersuchungen zur Authentizität von Hörverstehensaufgaben heraus, dass nur für wenige Testpersonen die Authentizität einer Testaufgabe eine wichtige Rolle spielte. Sie folgerte daraus, dass auch der Einsatz authentisch wirkender pädagogischer Stimuli im Rahmen von Testes vertretbar sei. Auch Grotjahn (2000) plädiert dafür, dem Kriterium der Authentizität nicht zu viel Gewicht beizumessen. Er argumentiert im Sinne von Bachman und Palmer, dass hohe Authentizität die Akzeptanz des TESTDAF Tests bei Testadministratoren und Lehrenden erhöhen dürfte, eine Auswirkung auf die Validität der Testergebnisse aber nicht gesichert sei. Er weist darauf hin, dass Sprachtests immer einer gewissen Nicht-Authentizität unterliegen, da sie in einer Testsituation Reaktionen verlangten, bei denen es nicht darum geht, Informationen zu geben, sondern Wissen zu zeigen.

Der GER empfiehlt eine Vorgehensweise, die der einzelnen Situation angepasst ist. Der Referenzrahmen schließt auch für die Lernsituation geschriebene und veränderte Materialien nicht kategorisch aus, sondern weist darauf hin, dass von Fall zu Fall über deren Einsatz entschieden werden muss. (vgl. Europarat, 2001: 145ff)

4.2. Hörverstehen im Gemeinsamen Europäischen Referenzrahmen für Sprachen (GER)

Der *Gemeinsame Europäische Referenzrahmen für Sprachen* (GER), als internationale Leitlinie für das Lernen, Unterrichten und Testen von Sprachen in Europa, stellt in ganz Europa eine gemeinsame Basis für die Entwicklung von zielsprachlichen Lehrplänen, curricularen Richtlinien, Prüfungen, Lehrwerken usw. dar. Er beschreibt umfassend, welche Kenntnisse und Fertig-

keiten entwickelt werden müssen, damit erfolgreiche Kommunikation möglich wird. Auch der kulturelle Kontext, in den Sprache eingebettet ist, findet dabei Berücksichtigung. (Europarat, 2001: 14) Der GER wurde nicht für die Erstellung von Sprachtests konzipiert, sondern liefert einen Rahmen für Sprachkompetenzmodelle. Die Deskriptoren und Niveaubeschreibungen wurden theoretisch und empirisch entwickelt und beschreiben einerseits quantitativ Merkmale der Sprachverwendungssituation und andererseits qualitativ das Bewältigen derselben. So wird es möglich, differenzierte Sprachniveaus zu beschreiben. Der GER ist primär ein Dokument für den Fremdspracherwerb. Da die Deskriptoren aber Niveaustufen der Sprachverwendung beschreiben, welche auch für die Sprachverwendung der Erstsprache relevant sind, findet er hier Beachtung. Berücksichtigung findet dabei auch die Tatsache, dass für eine nicht unbedeutende Zahl an Lernenden Deutsch nicht die Erstsprache darstellt.[6]

Der GER beschreibt für alle Kompetenzbereiche drei verschiedene Referenzniveaus: das Niveau der Elementaren Sprachverwendung (A), das der Selbstständigen Sprachverwendung (B) und das Niveau der Kompetenten Sprachverwendung (C). Diese Referenzniveaus sind jeweils in eine höhere (2) und eine niedrigere (1) Sprachverwendungsstufe unterteilt. Dabei gilt, dass die auf jeder Stufe erworbenen Kompetenzen stets auch für die nachfolgenden, höheren Stufen benötigt werden. Für alle Referenzniveaus gibt es verschiedene Skalen mit Deskriptoren, welche das typische oder wahrscheinliche Verhalten der Lernenden auf diesen Niveaus beschreiben. Für diese Beschreibungen werden zumeist positive Formulierungen verwendet, es wird jedoch auch dargestellt, was der Lernende ggf. noch mit Einschränkungen tun kann, wenn dies für ein bestimmtes Niveau kennzeichnend ist. (Europarat, 2001: 46) Der GER unterscheidet nicht zwischen Sprachverwendenden und Sprachlernenden, sondern betrachtet beide Gruppen als „sozial Handelnde". (ebd., 21) Die mangelnde Unterscheidung zwischen dem Lernen, dem Erwerb und der Verwendung einer Sprache führt dazu, dass einerseits die Aussagen und Deskriptoren für die Kommunikation in der Fremdsprache teilweise recht global ausfallen, andererseits die Beschreibungen der Stufen z. T. auch auf die Kommunikation in der Erstsprache zutreffen.

Eine allgemeine Skala gibt zunächst Hinweise globaler Art zur Zuhörkompetenz. Es folgen dann Skalen für die oben beschriebenen Höraktivitäten. Für jede Skala wurde versucht, die wesentliche Veränderung zu erfassen, welche die einzelnen Niveaustufen voneinander unterscheidet. Dabei fällt auf, dass sich die Niveaus nur schwierig explizit mithilfe eindeutiger Kriterien voneinander trennen lassen. Die Deskriptoren für die GER-Niveaus sind äußerst vage, da beispielsweise modifizierende Adjektive oder Adverbien wie „längere" Redebeiträge, „einigermaßen" vertraute Themen etc. statt genauer quantifizierbaren Angaben verwendet werden. In diesem Einteilungssystem wird das Niveau C2 als kompetente Sprachverwendung bezeichnet. Obwohl das Niveau C2 nicht (fast) muttersprachliche Sprachkompetenz beschreibt, ist anzunehmen, dass sich im schulischen Kontext des Deutschunterrichts über die Schularten verteilt Lerner befinden, deren Leistungsfähigkeiten in Deutsch im Bereich B2-C2 des GER anzusiedeln wären. Vereinzelt dürften sich sogar Schüler, vor allem im Bereich der Hauptschule, mit B1 Fähigkeiten finden. (Europarat, 2001: 54ff)

6 29.4% der 10–15-jährigen weisen in Deutschland Migrationshintergrund auf. (Statistisches Bundesamt, 2010)

Im GER werden außerdem Arbeitstechniken beschrieben, welche die Lerner für einen erfolgreichen Umgang mit Sprache beherrschen müssen. Die für das Hörverstehen relevanten Techniken sind laut GER „Notizen machen" und „Hinweise identifizieren/erschließen". Außerdem macht der GER vielseitige Vorschläge zu den Arten gesprochener Texte, die sich für den Einsatz im Unterricht eignen sowie zu den damit verbundenen Höraktivitäten und Hörabsichten. Dabei fällt auf, dass die Arten der vorgeschlagenen Texte stark lebensweltlichen Bezug haben und dem Lernenden sowohl authentisches Material als auch authentische Situationen offerieren. Zuhören wird im GER als Prozess zwischen Parteien betrachtet und die Höraktivitäten werden bezüglich der Involviertheit des Zuhörers und der Anwendungssituation unterschieden. Zu jedem Bereich gibt es differenzierte Skalen, mit denen die Fertigkeiten der Lernenden beschrieben werden können. Die mit dem Zuhören verbundenen Absichten können entsprechend ihrer Globalität gegliedert werden, sind jedoch nicht einem Bereich spezifisch zuzuordnen. Es handelt sich dabei um folgende Hörabsichten (Europarat, 2001: 71):

- Selektiv verstehen (eine ganz bestimmte Information erhalten)
- Detailliert verstehen (das Gesprochene in allen Einzelheiten verstehen)
- Global verstehen (erfahren, was insgesamt gemeint ist)
- Schlussfolgerungen ziehen können.

Der Gemeinsame Europäische Referenzrahmen beschreibt ferner Merkmale und Bedingungen, welche die Kommunikation erschweren und einschränken (Europarat, 2001: 55ff) (vgl. Abbildung II-4.2.). Es handelt sich dabei um materielle Bedingungen, soziale Bedingungen und Zeitdruck.

Materielle Bedingungen	des Mündlichen	- Klarheit der Aussprache - Störgeräusche - Hintergrundgeräusche - Verzerrungen - Wetterbedingungen
	des Schriftlichen	- Schlechte Druckqualität - Schlecht lesbare Handschrift - Schlechte Beleuchtung
Soziale Bedingungen		- Anzahl der Gesprächspartner und der Grad ihrer Vertrautheit - Relativer Status der Teilnehmenden (Macht und Solidarität, etc.) - An-/Abwesenheit von Zuhörenden - Soziale Beziehungen zwischen den Teilnehmenden
Zeitdruck		- Unterschiedlicher Zeitdruck für Sprechende/Hörende (Echtzeit) und Schreibende/ Lesende (flexibler) - Vorbereitungszeit für Reden, Berichte, u. s. w. (z. B. improvisiert vs. routiniert vs. vorher vorbereitet) - Zeitbeschränkungen für Redebeiträge und Interaktionen (z. B. durch Regeln, Kosten, gleichzeitige Ereignisse und Verpflichtungen etc.) - Andere Arten von Druck: finanziell, Spannung oder Angst erzeugende Situationen (z. B. Prüfungen)

Abbildung II-4.2.: GER Schwierigkeitsbeeinflussende Merkmale

Ergänzend zum GER wurde von einer Forschergruppe um Charles Alderson im Rahmen des Projekts The Dutch CEFR Construct Project (Alderson et al., 2006) für die Kompetenzbereiche „Zuhören" und „Lesen" ein webbasiertes Raster zur Beschreibung von Textmerkmalen sowie zur Einschätzung von Verarbeitungsprozessen beim Zuhörer/Leser entwickelt. Die Beschreibungen sind an die Deskriptoren des GER angebunden und systematisieren diese. Mithilfe des Dutch Grids wird die Kompetenzniveaubestimmung einer Aufgabe möglich, basierend einerseits auf Merkmalen des Hör-/Lesestimulus und andererseits auf der entsprechenden Aufgabenstellung. Das Dutch Grid erweitert die GER Skalen vor allem durch die Beschreibung von Testaufgaben und Aufgabencharakteristika, die sich auf die Schwierigkeit von Aufgaben auswirken. Es enthält folgende Merkmale hinsichtlich des Stimulus: Ursprung/Textart, Authentizitätsgrad, Diskurs-Typ, Domain/Kontext, Themenbereich, Art des Inhalts, Länge, Vokabular, Grammatik sowie die Verständlichkeit für Lerner auf GER-Level. In Bezug auf die Items werden genannt: Itemtyp und Denkoperationen beim Itemlösen.

4.3. Zuhören im Large-Scale-Assessment

Mitte der 90er Jahre vollzog sich die sogenannte „empirische Wende" (Kohler, 2005) in den Erziehungswissenschaften, in deren Rahmen sich Deutschland verstärkt an vergleichenden nationalen und internationalen Schulleistungsstudien beteiligte. Vor allem die PISA-Studien rückten den Bereich der schulischen Bildung ins Zentrum öffentlicher Aufmerksamkeit. Der Bereich Hörverstehen im Fach Deutsch wurde von diesen Studien jedoch bislang kaum in den Mittelpunkt gestellt. Im Folgenden werden die wichtigsten Erhebungen, bei denen auch im Fach Deutsch getestet wurde, kurz vorgestellt. Dabei werden vor allem die DESI Studie und PISA etwas genauer beschrieben. Obwohl in ihnen im Fach Deutsch nur Leseverstehen und 63 nicht Hörverstehen erhoben wurde, geben sie hilfreiche Impulse bei der Beschreibung und Kategorisierung der identifizierten Merkmale.

4.3.1. Nationale Studien

Bei der *Hamburger Studie Aspekte der Lernausgangslage und Lernentwicklung* (LAU) wurde in einer Längsschnittsuntersuchung seit 1996 in zweijährigen Abständen bei einer Vollerhebung in Hamburg der Lernfortschritt in den Fächern Deutsch, Mathematik und der Ersten Fremdsprache geprüft. Zusätzlich ermittelt wurden fächerübergreifende Kompetenzen. LAU dokumentiert die jeweils erreichten Lernstände, Lernentwicklungen und schulbezogenen Einstellungen. Im Fach Deutsch wurden die Teilbereiche „Leseverständnis" und „Rechtschreibung" erfasst, im Durchgang LAU 9 auch die Teilbereiche „Sprache", „Leseverständnis", „Rechtschreiben" und „Textproduktion". Ein Teil der Schüler befand sich im Durchgang LAU 11 bereits in der gymnasialen Oberstufe, ein anderer Teil in einer beruflichen Ausbildung. Aus diesem Grund wurde LAU durch die *Untersuchung der Leistung, Motivation und Einstellungen zu Beginn der beruflichen Ausbildung* (ULME) ergänzt. (Lehmann et al., 2001)

Auch die Studie *Kompetenzen und Einstellungen von Schülerinnen und Schülern* (KESS) wurde als flächendeckende Langzeituntersuchung in Hamburg erstmals im Jahr 2003 mit Schülern am Ende der 4. Jahrgangsstufe in den Fächern Deutsch, Mathematik, Englisch und den Naturwissenschaften durchgeführt. (Bos & Pietsch, 2005) Im Rahmen des Englisch-Tests bearbeiteten die Schüler auch zwei Hörverstehensteile. Zusätzlich wurde der Einfluss von Rah-

menbedingungen (z. B. Schichtzugehörigkeit, Migrationshintergrund und Motivation) auf die Lernentwicklung untersucht. Die Studie wurde in ungefähr zweijährigem Abstand im September 2005 mit denselben Testpersonen als KESS 7 wiederholt, eine dritte Erhebung fand als KESS 8 im Juni 2007 statt. (Bos et al., 2010)

Die Vergleichsarbeiten (VERA) werden in der 3., 6. und der 8. Jahrgangsstufe in den Fächern Deutsch und Mathematik (VERA 3, 6 und 8) sowie der Ersten Fremdsprache Englisch (VERA 6 und 8) bzw. Französisch (VERA 8) durchgeführt. Den Vergleichsarbeiten liegen die Inhalte der Bildungsstandards zugrunde. Ziel ist es, den Lehrkräften ein Instrument an die Hand zu geben, mit dem sie die Leistungen ihrer Klassen schulübergreifend einordnen können. Um den Schulen mehr Zeit zu geben, die Schüler im Anschluss an den Test gezielt zu fördern, werden die Vergleichsarbeiten seit dem Schuljahr 2006/07 in der 3. und 8. Jahrgangsstufe durchgeführt und nicht mehr in der 4. und 9. Jahrgangsstufe. Das VERA-Projekt verfolgt als Ziele Unterrichtsentwicklung, Standardsicherung und –entwicklung und die Erfassung und Verbesserung der Diagnosegenauigkeit. Zusätzlich dazu sollen die Lehrkräfte Informationen zur Beratung der Eltern erhalten und eine größere Vertrautheit im Umgang mit dem PC und dem Internet erwerben. Die Aufgaben für VERA 8 werden seit dem Schuljahr 2008/09 vom IQB erstellt, für den Primarbereich seit dem Schuljahr 2009/10. (http://www.iqb.hu-berlin.de/vera2) Die Erstellung der Aufgaben für VERA 6 ist ab dem Schuljahr 2011/12 durch das IQB vorgesehen. Die Länder legen fest, in welchen Fächern die Arbeiten verpflichtend geschrieben werden. Im Jahr 2010 wurden in den Vergleichsarbeiten für das Fach Deutsch zum ersten Mal auch Aufgaben zum Hörverstehen eingesetzt. Diese Aufgaben orientieren sich an den Bildungsstandards für den Hauptschulabschluss oder für den Mittleren Abschluss (http://www.iqb.hu-berlin.de/vera2) und beziehen sich auf das von der KMK bestätigte Kompetenzstufenmodell zum Zuhören.

Die Vergleichsarbeiten waren in Nordrhein-Westfalen seit 2004/05 als landesweite *Lernstandserhebungen* bekannt, die im ersten Schulhalbjahr in der Jahrgangsstufe 9 durchgeführt wurden. Ebenso wie die Vergleichsarbeiten dienten die zentralen Lernstandserhebungen dazu, die Leistungen von Schulklassen schulübergreifend einzuordnen und an ausgewiesenen Anforderungen und Standards zu messen. Auch die Lernstandserhebungen orientierten sich an den Bildungsstandards und bezogen sich deshalb nicht nur auf den unmittelbar vorausgegangenen Unterrichtsstoff sondern auf längerfristig aufgebaute Kompetenzen. Der Deutsch-Test erfasste jährlich zwei bis drei der Kompetenzbereiche des Kernlehrplans für das Fach Deutsch („Sprechen und Zuhören", „Schreiben", „Lesen - Umgang mit Texten und Medien" und „Reflexion über Sprache"). Eine thematische Klammer verband die Aufgaben der einzelnen Bereiche. Die Teilnahme war für alle Schüler der 9. Jahrgangsstufe aus Haupt-, Real-, Gesamtschulen und Gymnasien verpflichtend. Im Durchlauf 2005/06 wurden bei den Lernstandserhebungen erstmals auch Aufgaben zum „Zuhören und Verarbeiten" gestellt. Für die Rückmeldung der Ergebnisse wurde auf ein vierstufiges Kompetenzstufenmodell Bezug genommen, das sich auf die Inhalte der Kernlehrpläne NRWs bezieht und im Groben die Stufen „Einfache Einzelinformationen aus Hörbeiträgen erfassen und wiedergeben", „Übersichtlich präsentierte Informationen aus Hörbeiträgen erfassen, identifizieren und zuordnen", „Informationen differenziert erfassen und komplexe Verarbeitungs-

leistungen erbringen“ und „Informationen selbstständig und differenziert erfassen und eigenständig verarbeiten“ umfasst. Jede Stufe ist durch weitere Deskriptoren aufgeschlüsselt. (http://www.standardsicherung.schulministerium.nrw.de/lernstand8/)

Im Mai/Juni 2009 fand zum ersten Mal im Fach Deutsch ein Ländervergleich statt. Die Erhebung mit einer repräsentativen Stichprobe umfasste für die Schüler einen Leistungstest und eine Befragung. Die Befragung soll Aufschluss darüber geben, welche Rolle schulische und außerschulische Lerngelegenheiten für die Schüler spielen und inwiefern Rahmenbedingungen für die Optimierung von Lernprozessen genutzt werden können. Zu diesem Zweck wurden zusätzlich zu den Schülern auch die in den teilnehmenden Klassen unterrichtenden Fachlehrkräfte befragt. Der Ländervergleich ist eine von mehreren Maßnahmen, die die Kultusministerkonferenz in ihrer Gesamtstrategie zum Bildungsmonitoring beschlossen hat. Er dient dazu, festzustellen, in wie weit die Schüler die Zielvorgaben der Bildungsstandards erreichen bzw. wo konkreter Steuerungsbedarf besteht. Mit dem Ländervergleich, der auf den für alle Bundesländer verbindlichen Bildungsstandards und entsprechenden Kompetenzstufenmodellen basiert, ist erstmals eine vergleichende Bestandsaufnahme möglich. Der Ländervergleich soll im Fach Deutsch in der Sekundarstufe I alle sechs Jahre durchgeführt werden und den Ländern Rückmeldungen darüber geben, in welchen Bereichen Handlungsbedarf besteht, sodass gezielt bildungspolitische Maßnahmen ergriffen werden können. Diese Vergleiche bilden Teil eines umfassenden Systemmonitorings. Es liefert Informationen, die zur gezielten Förderung bestimmter Leistungsgruppen bzw. zur Weiterentwicklung des Unterrichts dringend benötigt werden. Bundesweit nahmen am Ländervergleich 1500 Schulen teil. Im Fach Deutsch wurden Aufgaben zu den Kompetenzbereichen „Lesen“, „Zuhören“ und „Orthografie“ eingesetzt. (http://www.iqb.hu-berlin.de/aktuell?pg=a_7)

Die Studie *Deutsch Englisch Schülerleistungen International* (DESI) (DESI-Konsortium, 2007) untersuchte ergänzend zum nationalen Bildungsmonitoring durch TIMSS und PISA rezeptive und produktive sprachliche Leistungen in den Fächern Deutsch und Englisch bei Schülern der 9. Jahrgangsstufen aller Schulformen in allen 16 Bundesländern. Im Gegensatz zur DESI-Studie dient PISA dem internationalen Vergleich und beachtet deshalb nicht detailliert die Besonderheiten des deutschen Bildungssystems. Außerdem wurde bei PISA 2000 im sprachlichen Bereich nur Lesekompetenz erfasst wurde. DESI analysierte für die Testentwicklung hingegen die Lehrpläne der einzelnen Länder und testete die Teilbereiche „Diktat“, „Kommunikation“, „Lesen“, „Stil“, „Grammatik“, „Wortschatz“ und „Schreiben“. In der DESI-Studie werden die sprachlichen Fähigkeiten der Schüler auf an den GER angelehnten Kompetenzniveaus berichtet, wobei für jede Teildimension eine eigene Kompetenzskala existiert. Die Teilkompetenzen interagieren miteinander und ergeben zusammen ein Kompetenzprofil der Lernenden, wobei sie eine übergeordnete Gesamtsprachkompetenz mit unterschiedlichen Ausprägungen der Fähigkeiten abbilden. Da bei DESI vor allem die im Unterricht ähnlich vermittelten und überprüften Kompetenzen im Vordergrund stehen, liegt der Schwerpunkt auf der Analyse von Teilfähigkeiten, denen ähnliche kognitive Prozesse zugrunde liegen und/oder bei denen die Aneignung weitgehend parallel abläuft. In diesem Sinne werden die Bereiche der produktiven (Sprechen und Schreiben) sowie der rezeptiven (Zuhören und Lesen) Sprachkompetenz fokussiert. (Jude & Klieme, 2007)

Die Kategorien der Textschwierigkeiten lassen sich Abbildung II-4.3.1a. entnehmen (vgl. Klieme et al., 2003: 38ff):

	Leicht	Mittel	Schwer
Syntax	Linearer Satzbau	Nicht durchgängig linear, mit z.B. Fragen/ Verneinungen	Enthält Kombinationen von Abweichungen (Spitzenstellung, Ausgliederung, verneinte Fragen etc.)
	Ohne Nebensätze	Integriert einfache Nebensätze (Relativsätze, einfache Kausalsätze)	Enthält komplexe Nebensätze (eingeschoben, adversative, konzessive, etc.)
	Unterhalb der 3-Sekunden-Grenze	Füllt das 3-Sekunden-Fenster aus	Überschreitet das 3-Sekunden-Fenster
Lexik	Konkret, häufig gebraucht, im Kontext redundant	Seltene Konkretion, häufigere Abstrakta oder Fremdwörter, im Kontext nur einfach redundant wiederholt	Abstrakta, seltenere Fach-/ Fremdwörter, geringe oder keine Unterstützung durch den Kontext
Vertextung	Die Abfolge im Text entspricht dem Genre	Einschübe oder Umstellungen im Aufbau des Textes	Die Abfolge des Genre ist deutlich umgestellt
	Klare und direkte Thema-Rhema Anbindung		
	Ein Schlüsselwort dirigiert einen Absatz	Zwei Schlüsselwörter dirigieren den Absatz	Absätze werden nicht durch Schlüsselwörter dirigiert
	Die Abfolge ist durch mehrere Gliederungssignale klar gemacht	Wenige Gliederungssignale	Keine Gliederungssignale und kein Roter Faden
	Wenige Absätze	Der Text geht über mehrere Seiten und Absätze: ein roter Faden verbindet erkennbar die Absätze	
	Ein Beispiel gliedert sich in die Argumentation ein	Mehrere Beispiele gliedern sich ein	Keine Beispiele
		Einfache kognitive Dissonanzen sind vorhanden - wirken aber anregend	Kognitive Dissonanzen oder rhetorische Mittel sind ausgiebig benutzt
		Rhetorische Mittel (Metaphern, Ironie, Topoi)	

Abbildung II-4.3.1a.: Merkmale der DESI-Leseverstehensaufgaben Deutsch (vgl. Klieme et al., 2003: 38ff)

Hörverstehen wurde bei DESI nur im fremdsprachlichen Bereich erhoben. Das Konzept des Hörverstehens im Fach Englisch basiert in Anlehnung an Bachmann und Palmer (1996) auf den Theorien von Buck (2001) sowie dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (Klieme et al., 2003: 76). Hörverstehenskompetenz besteht demnach überwiegend aus sprachlicher und strategischer Kompetenz, zu welcher auch das Hintergrundwissen des Zuhörers gehört. Mithilfe der sprachlichen Kompetenzen, der Aktivierung sprachlicher Strukturen und dem Wissen über Laute, Wort- und Satzbedeutungen, können die Anforderungen der Aufgabe vor allem im sprachlichen Bereich bewältigt werden. Auch pragmatisches Wissen kommt hier zum Einsatz, um Funktionen von Äußerungen, kommunikative Kontexte, Sprechsituationen und –intentionen einschätzen zu können. Die sprachlichen und strategischen Kompetenzen sind die Grundlage für die erforderlichen Verstehens- und Behaltensleistungen sowie die notwendigen Fähigkeiten zur Informationsverarbeitung. Die involvierten Teilkompetenzen werden in Abbildung II-4.3.1b. veranschaulicht. Gleichwohl die Grafik den Eindruck vermittelt, dass die Kompetenzen des Zuhörers für jeweils einen Anforderungsbereich der Aufgabe eingesetzt werden, findet sich bei DESI der explizite Hinweis darauf, dass die drei fundamentalen Kompetenzgruppen „parallel und interdependent eingesetzt" (Klieme et al., 2003: 76) werden.

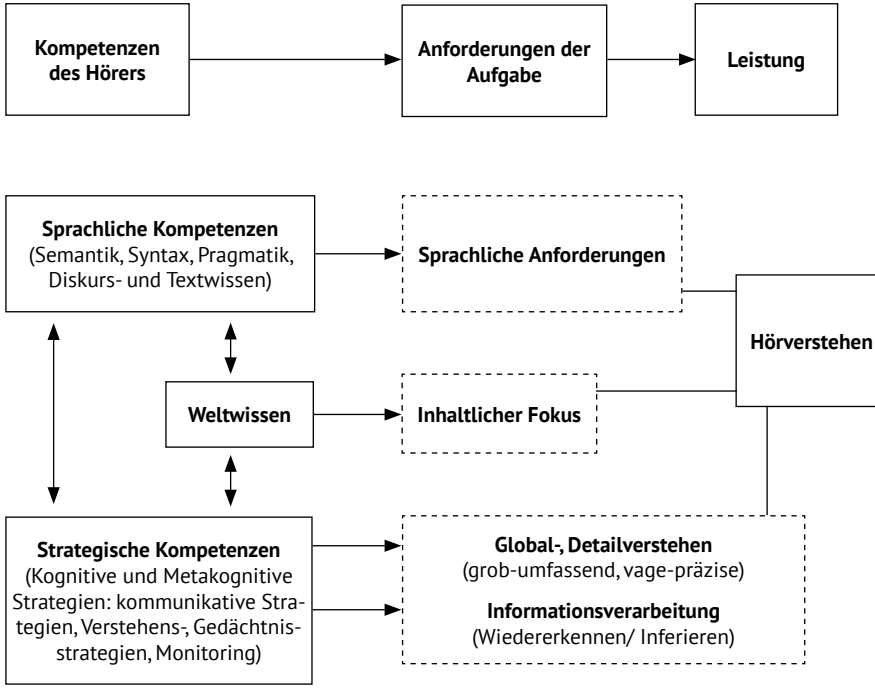


Abbildung II-4.3.1b.: DESI Testkonzept Hörverstehen (Klieme et al., 2003: 76)

Auch für die Hörverstehensaufgaben im Fach Englisch wurden bei DESI Merkmale identifiziert, die einen Einfluss auf die Itemschwierigkeit haben. Dazu gehören die sprachlichen Anforderungen des Textmaterials, der inhaltliche Fokus der Aufgabe, die zum Lösen der Items erforderliche Verstehens- bzw. Behaltensleistung sowie die dazugehörigen Informationsverarbeitungsprozesse. Diese Merkmale werden in Abbildung II-4.3.1c. zusammenfassend dargestellt:

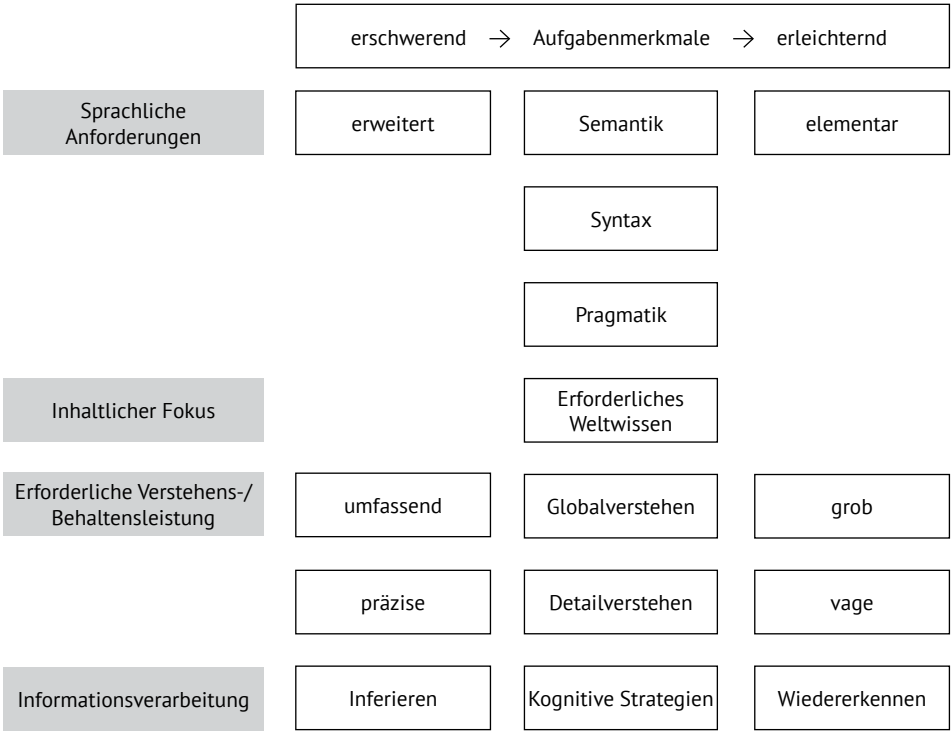


Abbildung II-4.3.1c: Merkmale der DESI-Hörverstehensaufgaben Englisch (Klieme et al., 2003: 79)

4.3.2. Internationale Studien

Internationale Studien werden i. d. R. entweder von der *International Association for the Evaluation of Educational Achievement* (IEA) oder der *Organisation for Economic Co-Operation and Development* (OECD) durchgeführt. Die Tradition internationaler Vergleichsstudien reicht bis in die 70er Jahre, als die IEA die sogenannte *Six-Subject-Study* (in den Fächern Naturwissenschaften, Leseverstehen, Literatur, Englisch als Fremdsprache, Französisch als Fremdsprache und Staatsbürgerkunde) realisierte (http://www.iea.nl/readingcomprehension.html). Die erste internationale Studie, an der sich auch Deutschland beteiligte, war 1990/91 die *Reading Literacy Study*. Sie wurde mit 9- und 14-jährigen in rund 30 Staaten

durchgeführt, um in den unterschiedlichen Bildungssystemen der teilnehmenden Staaten Aufschluss über die Leseleistungen und die außerschulischen Leseaktivitäten zu erhalten. (http://www.iea.nl/reading_literacy.html) Die OECD führte 1994 zusammen mit Statistics Canada eine vergleichende Studie zur Lesekompetenz an 15- bis 65-jährigen Testpersonen erstmals in acht Staaten durch, die *International Adult Literacy Survey* (IALS). Das Ziel war es, vergleichbare Literalitätsprofile über nationale, linguistische und 68 kulturelle Grenzen hinweg zu erhalten. Mittlerweile beteiligen sich rund 30 Staaten an der Untersuchung. (http://www.hrsdc.gc.ca/eng/hip/lld/nls/Surveys/ialsintro.shtml)

Die *Internationale Grundschul-Lese-Untersuchung* (IGLU) (vgl. Bos et al., 2003), welche die deutsche Teilstudie der internationalen *Progress in International Reading Literacy Study* (PIRLS) ist, erhebt die Bereiche „Lesekompetenz" sowie „Mathematik", „Naturwissenschaften" und "Orthografie" bei Schülern in der 4. Jahrgangsstufe in allen 16 Bundesländern. Dabei werden auch die schulischen und häuslichen Bedingungsfaktoren der Schüler berücksichtigt. Die Testaufgaben zum Leseverstehen haben unterschiedliche Schwierigkeitsgrade. Die Texte umfassen zwei Textsorten, und zwar literarische Texte (z. B. Kurzgeschichten) und informierende Texte (z. B. Faltblätter). Ferner wird eine Schülerbefragung durchgeführt, bei der die Einstellung der Schüler zum Lesen und ihre Lesegewohnheiten erhoben werden. Um Tendenzen in der Entwicklung der Lesekompetenz über die Zeit hinweg feststellen zu können, werden PIRLS und IGLU zyklisch alle fünf Jahre durchgeführt. Deutschland beteiligte sich an den Erhebungen in den Jahren 2001 und 2006. PIRLS findet in 36 Nationen statt. (http://www.bmbf.de/de/6626.php)

Die Studie *Programme for International Student Assessment (PISA)* (Deutsches PISA-Konsortium 2001, 2002, 2004, 2005) wird im Auftrag der OECD durchgeführt. Sie wird von einem internationalen Konsortium unter der Leitung des *Australian Council for Educational Research (ACER)* koordiniert. Der Test erfolgt in einem Drei-Jahres-Rhythmus und fokussiert die Domänen „Lesen", „Mathematik" und „Naturwissenschaften". Die Ergebnisse der PISA-Studie verdeutlichten, dass im deutschen Schulsystem ein sehr enger Zusammenhang zwischen sozialer bzw. ethnischer Herkunft und Schülerleistungen besteht und dass gute und schlechte Schulen gravierende Leistungsunterschiede aufweisen (Deutsches PISA-Konsortium 2001, 2002, 2003, 2004, 2005). Ursache für die sozialen Disparitäten seien das differenzierte Schulsystem, das den Schülern ungleiche Lerngelegenheiten biete sowie mangelnde Fördermaßnahmen der Schulen. Eine Reaktion darauf sind beispielsweise Ganztagesschulen oder Sprachlernklassen.

Der Kompetenzbereich „Zuhören" wurde bei PISA bislang nicht erfasst, „Lesen" spielte hingegen in allen Durchgängen eine Rolle. Der Leseverstehensteil ist dabei entsprechend der Literacy-Konzeption konsequent auf Funktionalität ausgerichtet und die Items zeichnen sich durch hohe Authentizität aus. Lesen wird als interaktiver, konstruktiver Prozess verstanden, wobei gerade die Wichtigkeit des reflektierenden und verstehenden Lesens betont und die Bedeutung verschiedener Leseanlässe und Zwecke herausgestellt wird. (vgl. Artelt et al., 2008)

In der PISA-Studie wurden keine A-priori-Einschätzungen der Leseaufgaben nach einzelnen Anforderungsmerkmalen auf der Basis eines Kompetenzstufenmodells durchgeführt. Die

Aufgaben wurden überwiegend nach textimmanenten (z. B. ein allgemeines Verständnis des Textes entwickeln, Informationen ermitteln, eine textbezogene Interpretation entwickeln) und wissensbasierten (z. B. über den Inhalt des Textes reflektieren, über die Form des Textes reflektieren) Verstehensleistungen eingeteilt. Diese wurden weiterhin nach ihrer Komplexität und ihrer formalen Anforderung unterschieden. Diese Merkmale sollen jedoch ausdrücklich nicht als schwierigkeitsbeeinflussende Merkmale verstanden werden, auch wenn damit 76% der Variabilität der Aufgabenschwierigkeit vorhergesagt werden konnte. (vgl. Artelt et al., 2004)

Die PISA-Aufgaben wurden hinsichtlich vier Kategorien eingeschätzt. Im Anschluss daran wurde mithilfe einer Regressionsanalyse überprüft, welche dieser Kategorien in welchem Ausmaß zur Erklärung der Schwierigkeit der Aufgaben beitragen. Die erste Kategorie ist ein dreistufiges „Globalurteil" hinsichtlich der Aufgabenschwierigkeit. Lehrplanexperten der Länder schätzten ein, ab welcher Klassenstufe die Bewältigung der Aufgaben von den 15-jährigen erwartet wird. Für die anderen drei Kategorien wurden Einschätzungen auf fünfstufigen Skalen vorgenommen. Es handelt sich dabei um die Kategorien „Entscheidungsspielraum", „Integrationsgrad" und „Präzision". Die Kategorie „Entscheidungsspielraum" beschreibt die Freiheit, die die Schüler bei der Beantwortung der Aufgaben haben. Je nachdem ob verschiedene Antwortmöglichkeiten selbst generiert werden und deren Wahl begründet werden, oder eine eindeutig definierte Information aus dem Stimulus entnommen werden muss, wird der Entscheidungsspielraum als hoch oder gering eingestuft. Der „Integrationsgrad" beschreibt die Form der Kohärenzbildung, die für die korrekte Aufgabenbearbeitung gebraucht wird. Ein geringer Integrationsgrad ist dann gegeben, wenn der Stimulus nur auf lokale Kohärenz hin geprüft werden muss. Ein hoher Integrationsgrad liegt vor, wenn der Text auf globale Kohärenz geprüft werden muss, wenn also Verbindungen zwischen Informationen gebildet werden müssen, die im Stimulus weit voneinander entfernt liegen. Die Kategorie „Präzision" gibt Aufschluss über die für die Aufgabenlösung notwendige Genauigkeit, mit der alle relevanten Informationen aus Text und Item mit einzubeziehen sind. Hohe Präzision ist besonders dort gefordert, wo Informationen an wenig prominenten Stellen beachtet werden müssen. (vgl. Artelt et al., 2004: 156)

4.4. Hörverstehen im Bildungswesen Deutschlands

Welche Vorgaben hinsichtlich des Kompetenzbereichs Zuhören werden von den Rahmendokumenten des deutschen Bildungswesens, wie den Lehrplänen der Länder und den länderübergreifenden Bildungsstandards gemacht?

4.4.1. Hörverstehen in den Bildungsstandards

Bei den Bildungsstandards der Kultusministerkonferenz (KMK) (KMK, 2004) handelt es sich um Leistungsstandards, die als Regelstandards ein mittleres Anforderungsniveau beschreiben. Anders als Lehrpläne sind die Bildungsstandards keine Sammlungen zu vermittelnder Inhalte, sondern sie benennen fachbezogen und fachübergreifend mit thematisch-inhaltlichen Bezügen den Ausprägungsgrad von Kompetenzen, über den Kinder und Jugendliche am Ende eines Bildungsganges empirisch überprüfbar verfügen sollen. Die Kompetenzen werden einerseits durch Aufgaben konkretisiert und andererseits durch diese Aufgaben überprüft. Es wird angenommen, dass sich der Erwerb einer Kompetenz und ihr jeweiliger Ausprägungs-

grad in der Fähigkeit zeigen, Aufgaben zu lösen, deren Anforderungsprofil transferorientiert Problemlösungen fordert. Es wird also nicht einfach erlerntes Wissen abgefragt, sondern ein Verstehen des jeweiligen Zusammenhangs vorausgesetzt. Kompetenzen werden in diesem Zusammenhang verstanden als Verbindung von Wissen und Können.

Die Deskriptoren der Bildungsstandards sind durchgängig als Aussagesätze formuliert, z. B.: „Die Schülerinnen und Schüler bewältigen kommunikative Situationen in persönlichen, beruflichen und öffentlichen Zusammenhängen situationsangemessen und adressatengerecht." (KMK, 2004: 8) Dabei werden trotz dieser konstativen Art der Formulierung wünschenswerte Zustände definiert, also Normen gesetzt. Mithilfe von Indikatoren, den Testaufgaben, soll geprüft werden, ob und in welcher Ausprägung die Norm erfüllt ist.

Die Bildungsstandards für das Fach Deutsch in der Sekundarstufe I lassen sich in vier zentrale Kompetenzbereiche gliedern (vgl. Abbildung II-4.4.1.). In allen Kompetenzbereichen werden dabei auch für den Kompetenzerwerb grundlegende Methoden und Arbeitstechniken berücksichtigt. Der Kompetenzbereich „Sprache und Sprachgebrauch untersuchen" beschreibt Fertigkeiten und Fähigkeiten, die als Voraussetzung für gelingende Kommunikation in den anderen Kompetenzbereichen gelten. Dieser Kompetenzbereich wird als eigener Bereich gesehen, der jedoch in die anderen Kompetenzbereiche integriert werden soll. Demgegenüber wird „Orthografie" unter den Kompetenzbereich „Schreiben" subsumiert. Im Rahmen der IQB-Arbeiten zeigte sich, dass es sowohl für den Primarbereich als auch für die Sekundarstufe I sinnvoll ist, die Aufgabenentwicklung für den Bereich „Orthografie" vom Bereich „Schreiben" zu separieren und von zwei getrennten Kompetenzbereichen auszugehen. Im Unterschied zur Ersten Fremdsprache werden die Bildungsstandards für das Fach Deutsch „Sprechen und Zuhören" zu einem Kompetenzbereich zusammengefasst. Obwohl dies gerade aus der Perspektive des Deutschunterrichts und der Psycholinguistik sinnvoll ist, in denen „Sprechen" und „Zuhören" Bestandteile von „Gesprächskompetenz" sind, erweist es sich aus testtheoretischer Perspektive als günstiger, „Zuhören" (wie „Orthografie") als eigenen Kompetenzbereich zu behandeln.

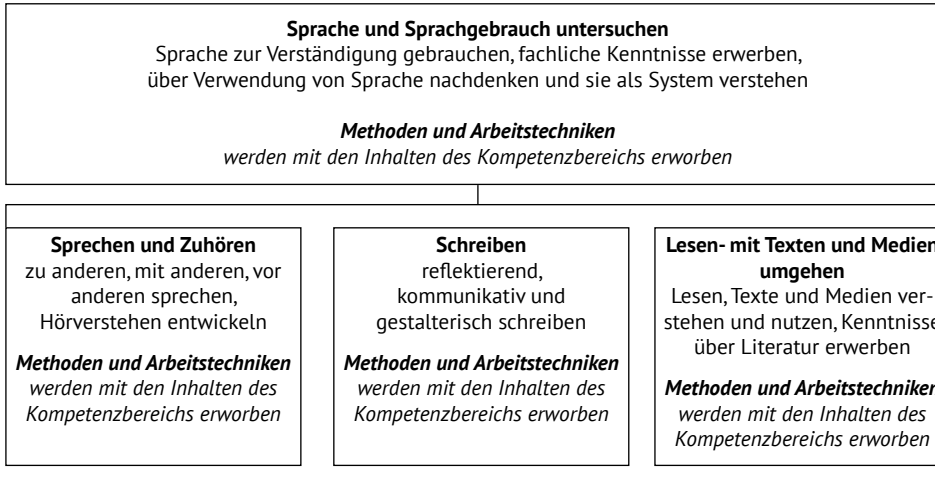


Abbildung II-4.4.1.: Kompetenzmodell der Bildungsstandards (KMK, 2004: 8)

Die wesentlichen Inhalte des Kompetenzbereichs „Zuhören" werden in den Bildungsstandards wie folgt beschrieben (KMK, 2004: 8):

> ***Sprechen und Zuhören***
>
> *Die Schülerinnen und Schüler bewältigen kommunikative Situationen in persönlichen, beruflichen und öffentlichen Zusammenhängen situationsangemessen und adressatengerecht.*
>
> Sie benutzen die Standardsprache. Sie achten auf gelingende Kommunikation und damit auch auf die Wirkung ihres sprachlichen Handelns. Sie verfügen über eine Gesprächskultur, die von aufmerksamem Zuhören und respektvollem Gesprächsverhalten geprägt ist.

Diese Beschreibung zeigt, dass das Konstrukt „Zuhören" in den Bildungsstandards nur sehr vage definiert wird. Auch die Aufschlüsselung des Kompetenzbereichs in Subkompetenzen („Zu anderen sprechen", „Vor anderen sprechen", „Mit anderen sprechen", „Verstehend zuhören" und „Szenisch spielen") verdeutlicht, dass sich ein Großteil der relevanten Kompetenzen auf den Teilbereich „Sprechen" bezieht. Nur der Gliederungspunkt „Verstehend zuhören" bezieht sich auf den Bereich „Zuhören". Explizit gefordert wird darin (ebd. 10):

- Gesprächsbeiträge anderer verfolgen und aufnehmen
- Wesentliche Aussagen aus umfangreichen gesprochenen Texten verstehen, diese Informationen sichern und wiedergeben
- Aufmerksamkeit für verbale und nonverbale Äußerungen (z. B. Stimmführung, Körpersprache) entwickeln.

Da bislang Sprechleistungen noch nicht ökonomisch in großen Schulleistungsstudien gemessen werden können, wurde bei der Aufgabenentwicklung zunächst überwiegend der Bereich „Verstehend zuhören" berücksichtigt. Obwohl die übergreifenden Angaben in den Bildungsstandards zum Kompetenzbereich „Zuhören" recht knapp ausfallen, sind in den Beschreibungen der einzelnen Substandards noch ergänzende Hinweise zu finden. Sie wurden in einer Synopse (vgl. Anhang F) zusammengefasst, in der die relevanten Teilkompetenzen für den Bereich „Zuhören" für den Hauptschulabschluss (HSA) und den Mittleren Schulabschluss (MSA) dargestellt werden.

Generell lässt sich festhalten, dass die Beschreibungen für Hör- und Leseverstehen hinsichtlich der Kategorien und der zugrundeliegenden Kompetenzen sehr ähnlich ausfallen. Auch beim Zuhören bietet sich die Unterscheidung in literarische und informierende Stimuli an, diskontinuierliche Texte mit Tabellen und Graphen spielen hier jedoch keine Rolle. Wie beim Lesen sind auch beim Zuhören Prozesse wie Informationen ermitteln, Interpretieren sowie Reflektieren und Bewerten relevant. In vielen Hörsituationen wird Alltagssprache verwendet und die entstehenden Diskurse weisen deshalb ein höheres Maß an Redundanz, paraverbalen Merkmalen (z. B. Stimmführung, Lautstärke, etc.), Interjektionen (aha, soso, etc.) und Verzögerungslauten (ähm, äh, etc.) auf.

Ein Vergleich der Standards aus dem Bereich „Sprechen und Zuhören“ für den Hauptschulabschluss und den Mittleren Schulabschluss zeigt im Wesentlichen eine Erweiterung der Stimuli um Stimulusarten, Länge und Komplexität.

Standard 1.1. „Zu anderen sprechen“:
Die Schüler, die den Mittleren Schulabschluss anstreben, sollen vor allem in der Lage sein, sach- und situationsangemessen zu agieren. Sie sollen aktiver und selbstständiger in Erscheinung treten und beispielsweise auch in der Lage sein, Gespräche zu leiten und Sprechsituationen zu gestalten. Gefordert wird die Verfügbarkeit eines umfangreichen und differenzierten Wortschatzes, die Begriffe „umfangreich“ und „differenziert“ werden jedoch nicht weiter erklärt. Explizit wird in den Standards für den Mittleren Schulabschluss die Kenntnis der mündlichen Darstellungsformen „Schildern“ und „Erörtern“, sowie bei der Beachtung der Redeweise der Aspekt „Klangfarbe“ erwartet.

Standard 1.2. „Vor anderen sprechen“.
Dieser Standard ist für die Testaufgabenentwicklung nicht relevant, da die unter diesen Standard subsumierten Standards derzeit nicht im Large-Scale-Assessment überprüfbar sind.

Standard 1.3. „Mit anderen sprechen“:
Die Erwartungen für beide Schulabschlüsse sind weitgehend identisch. Im Bereich des Mittleren Schulabschlusses wird der Schwerpunkt stärker auf eine plausible Begründung der eigenen Meinung und eine Reflektion der Standpunkte gelegt.

Standard 1.4. „Verstehend zuhören“:
Für den Hauptschulabschluss wird erwartet, dass Redebeiträge kritisch hinterfragt werden, für den Mittleren Schulabschluss sollen aus umfangreichen Beiträgen Informationen gesichert werden können.

Im Rahmen der Methoden und Arbeitstechniken sollen die Schüler für den Mittleren Schulabschluss nicht nur in der Lage sein, Notizen anzufertigen. Erwartet wird vielmehr die Fähigkeit anschließend mit dem Notierten weiterzuarbeiten. Der Bereich der Gesprächsformen wird durch Dialoge und Debatten erweitert. Dabei wird einerseits die Praktizierung dieser Formen erwartet und andererseits auch die kritische Beobachtung und Reflektion derselben.

Tabelle II-4.4.1.: In den IQB-Items realisierte Bildungsstandards

Codebeschreibung: Geprüfter Standard (BS)
BS113: verschiedene Formen mündlicher Darstellung unterscheiden und anwenden
BS141: Gesprächsbeiträge anderer verfolgen und aufnehmen
BS142: wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben
BS143: Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln

Im Wesentlichen wurden die Standards in Tabelle II-4.4.1. in Testitems umgesetzt. Allgemein lässt sich festhalten, dass die große Offenheit bzw. Allgemeinheit bei den Standardformulierungen Freiheiten und Risiken für die Umsetzung birgt. Standard BS141 „Gesprächsbeiträge anderer verfolgen und aufnehmen" lässt beispielsweise unterschiedliche Möglichkeiten zur Umsetzung zu. Wie sollen die Gesprächsbeiträge verfolgt werden? Was genau wird unter dem Begriff „Gesprächsbeitrag" verstanden? Welche Fragen können im Sinn von Items gestellt werden, um die „Verfolgung" und „Aufnahme" der Gesprächsbeiträge zu überprüfen? Mit der Umsetzung der Standards in Items werden bei derart vage formulierten Vorgaben immer auch Entscheidungen über relevante Facetten getroffen und Formulierungen werden ausgelegt und interpretiert. Diese Entscheidungen und Auslegungen hätten jedoch häufig auch anders getroffen werden können. Bremerich-Vos et al. (2009: 207) sprechen hier von einem „Breitband- Genauigkeitsdilemma" (Bandwidth Fidelity Dilemma). Dieser Umstand ist bei der Interpretation der Ergebnisse der Analysen dieser Arbeit zu berücksichtigen, da sich die Analysen ausschließlich auf die Aufgaben beziehen, die im Rahmen der IQB-Arbeiten zur Illustration der Standards in der Sekundarstufe I erstellt wurden.

Im Rahmen der KMK-Bildungsstandards (KMK, 2004) wird zwischen drei Anforderungsbereichen für die verschiedenen Kompetenzen zur Beschreibung der kognitiven Operationen unterschieden, die zur Itembeantwortung notwendig sind. In der Einteilung der Anforderungsbereiche orientieren sich die Bildungsstandards an den in der Schulpraxis gebräuchlichen Kategorien „Wiedergeben" (AB I) – „Anwenden" (AB II) – „Reflektieren und Beurteilen" (AB III). Im Allgemeinen geht man davon aus, dass die Aufgaben von AB I nach AB III schwieriger werden. Ihr Schwierigkeitsgrad ist u. a. abhängig von der Komplexität der Aufgabenstellung, den Anforderungen an die sprachliche Darstellung sowie der Komplexität und dem Umfang der gewünschten Reflexion oder Bewertung. Die Anforderungsbereiche sollen über alle Kompetenzbereiche hinweg die Leistung klassifizieren, die zur Lösung einer Aufgabe erbracht werden muss. Allerdings ist es fraglich, ob die Differenzierung nach Bereichen im Sinne einer trennscharfen Stufung bzw. Graduierung von Schwierigkeiten verstanden werden kann. Die Einteilung in Anforderungsbereiche ist ein erster Versuch, die zur Itembeantwortung notwendige kognitive Operation zu kategorisieren. Sie wurde jedoch vom Kompetenzstufenmodell „Zuhören" abgelöst.

4.4.2. Kompetenzstufenmodelle Zuhören

Kompetenzstufenmodelle beschreiben Abstufungen und Entwicklungsverläufe von Kompetenzen. So wird es möglich, Lehr- und Lernergebnissen festzustellen. Ein Kompetenzstufenmodell ist die Voraussetzung für die Definition von Mindest-, Regel- und Maximalstandards. Mindeststandards beziehen sich auf Leistungserwartungen, die alle Schüler mindestens bis zu einem bestimmten Zeitpunkt in ihrer Ausbildung erworben haben sollen. Regelstandards beschreiben dagegen die Kompetenzen, die im Durchschnitt von den Schülern erworben sein sollten. Dabei kann es für Schulen hilfreich sein, mit einem Leistungsbereich zu arbeiten, der über die Regelstandards hinausgeht und als Regelstandard plus bezeichnet wird. Dieser zusätzliche Leistungsbereich beschreibt Ziele, die im Rahmen der Weiterentwicklung von Unterricht angestrebt werden können. Zusätzlich dazu beschreiben Maximalstandards Kompetenzen, die die Erwartungen der Regelstandards weit übertreffen. Unter exzellenten indi-

viduellen Lernvoraussetzungen und gelingenden schulischen und außerschulischen Lerngelegenheiten können diese jedoch u. U. erworben werden. (Kompetenzstufenmodell zu den Bildungsstandards im Kompetenzbereich Sprechen und Zuhören – hier Zuhören – für den Mittleren Schulabschluss, 2009)

Die in den Bildungsstandards beschriebenen Kompetenzen (außer dem Kompetenzbereich „Sprechen") wurden mittels der dazu entwickelten Testaufgaben im Jahr 2008 in einer repräsentativen Erhebung überprüft. Auf der Grundlage dieser Daten wurden im Jahr 2009 Kompetenzstufenmodelle für die Kompetenzbereiche „Lesen", „Zuhören" und „Orthografie" entwickelt.

4.4.2.1. Das Kompetenzstufenmodell der Bildungsstandards

Bei der Entwicklung von Kompetenzstufenmodellen spielen testtheoretische, curriculare, fachdidaktische und bildungspolitische Kriterien eine wichtige Rolle. Im Fall des Kompetenzstufenmodells für den Kompetenzbereich „Zuhören" wurden aus diesem Grund sowohl die KMK Bildungsstandards berücksichtigt, als auch, soweit möglich, Kompetenzstufenmodelle aus nationalen und internationalen Vorarbeiten (z. B. PISA). So wird im IQB-Modell von fünf gleich breiten, die Deskriptoren der KMK Bildungsstandards aufgreifenden Kompetenzstufen ausgegangen. Die Stufenbeschreibungen geben angemessene Leistungserwartungen wieder, die auch die Leistungsstreuung innerhalb und zwischen den Ländern berücksichtigen. Neben den Mindest- und Regelanforderungen wird einerseits ein Leistungsminimum beschrieben, das von allen Schülern erreicht werden kann, andererseits werden aber auch Leistungsressourcen der Schüler verdeutlicht, die u. U. gefördert werden müssen. Die Stufenbeschreibungen (insbesondere die Regelstandards Plus) sind als motivierende Leistungserwartungen formuliert, die zur schulischen Weiterentwicklung genutzt werden können, um so eine breite bildungspolitische Akzeptanz zu erhalten. (Kompetenzstufenmodell zu den Bildungsstandards im Kompetenzbereich Sprechen und Zuhören – hier Zuhören – für den Mittleren Schulabschluss, 2009)

Aufgrund der hohen konzeptionellen Ähnlichkeit zum Leseverstehen liegt es nahe, bei der Beschreibung der Kompetenzstufen auch Modelle zum Leseverstehen zu berücksichtigen. Um die spezifischen Aspekte des Zuhörens in die Stufenbeschreibungen mit aufzunehmen, finden sich dort verstehensspezifische Aspekte, aber auch modalitätsspezifische Ergänzungen. Bereits in den Vorbemerkungen zum Kompetenzstufenmodell „Zuhören" wird darauf hingewiesen, dass beim Zuhören, im Unterschied zum Lesen, Gedächtnisleistung eine größere Rolle spielt und Informationen häufiger erinnert und wiedererkannt werden müssen, um ein Item zu lösen. (ebd.: 7) Im Fall von Lesen handelt es sich bei Items zur Informationsentnahme tendenziell eher um die kognitive Aktivität des Lokalisierens. Das Erinnern von Informationen hängt zum Teil von personenspezifischen Variablen, aber auch von der Aufgabenstellung ab. Die Aufgabenstellung steuert einerseits die Intention und den Aufmerksamkeitsfokus des Zuhörers und gibt andererseits die Häufigkeit der Stimuluspräsentation und den Zeitpunkt derselben an. (Kompetenzstufenmodell zu den Bildungsstandards im Kompetenzbereich Sprechen und Zuhören – hier Zuhören – für den Mittleren Schulabschluss, 2009)

Jede Kompetenzstufe ist mit einer Überschrift gekennzeichnet. So geht es auf Kompetenzstufe I überwiegend um das „Wiedererkennen und Erinnern prominenter Einzelinformationen". Schüler auf Stufe II können „Benachbarte Informationen miteinander verknüpfen und den Text genrespezifisch zuordnen". Stufe III zeichnet sich dadurch aus, dass die Schüler hier „Verstreute Informationen miteinander verknüpfen, der Vorlage paraverbale Informationen abgewinnen und den Text ansatzweise im Ganzen erfassen" können. Auf Stufe IV können die Schüler „Auf der Ebene des Textes wesentliche Zusammenhänge erkennen, die Gestaltung reflektieren und versteckte Einzelinformationen erinnern". Schüler, die sich auf Stufe V befinden, können „Interpretieren, Begründen, Bewerten und anspruchsvolle Erinnerungsleistungen" erbringen. In Tabelle II-4.4.2.1. werden die für jede Stufe kennzeichnenden Merkmale zusammenfassend kurz aufgeführt. (ebd., 8ff)

Tabelle II-4.4.2.1.: Kompetenzstufenmodell Bildungsstandards

Kompetenzstufe	Benennung	Merkmale
I	Wiedererkennen und Erinnern prominenter Einzelinformationen	• strukturell eher einfache Hörstimuli • Beantwortung der Items beim Hören • Zweimaliges Hören des Stimulus • Items zu nur gelegentlich paraphrasierten Einzelinformationen • Starke Lenkung der Aufmerksamkeit • Selten interpretative Anforderungen • Überwiegend geschlossene Itemformate • Wenige halboffene Itemformate • Kaum offene Itemformate
II	Benachbarte Informationen miteinander verknüpfen und den Text genrespezifisch zuordnen	• Verknüpfen benachbarter Informationen • Kaum Verknüpfen verstreuter Informationen • Items zum Stimulusgenre • Teilweise Items zur Funktion der Stimuli bzw. einzelnen Aspekten • Schlüsse aufgrund der akustischen Eigenschaften der Stimuli • Items zu Einzelinformationen • Kaum noch entlastende Vorinformationen • Beantwortung der Items nach dem Hören • Weniger geschlossene Itemformate • Mehr halboffene und offene Itemformate
III	Verstreute Informationen miteinander verknüpfen, der Vorlage paraverbale Informationen abgewinnen und den Text ansatzweise im Ganzen erfassen	• Verknüpfen verstreuter Informationen • Erfassen des Stimulus als Ganzes • Erkennen des Themas • Wahrnehmen und Explizieren paraverbaler Informationen (z. B. Tonfall) • Dazu einfache Schlüsse bzw. unplausible Antwortalternativen • Items zu weniger zentralen Einzelinformationen

IV	Auf der Ebene des Textes wesentliche Zusammenhänge erkennen, die Gestaltung reflektieren und versteckte Einzelinformationen erinnern	• Erkennen und Reflektieren wesentlicher Zusammenhänge auf Stimulusebene • Items zur Funktion von Stimulusteilen • Items zu Sprachlichen Mitteln • Items zur Wirkung des Stimulus • Items zu versteckten Einzelinformationen • Längere Hörstimuli • Einmaliges Hören • Keine Lenkung der Aufmerksamkeit
V	Interpretieren, Begründen, Bewerten und anspruchsvolle Erinnerungsleistungen	• globales Verstehen des gesamten Stimulus/ längerer Abschnitte • ungewöhnliche Itemformate (Organigramm) • offene Itemformate • begründete Entscheidungen bzgl. Merkmale des Gehörten • Items zu Geräuschen (Erinnern und Funktion) • Items zu Unterschiede und Gemeinsamkeiten von Stimulusaspekten • Interpretationsaufgaben • Items zur Struktur der Stimuli • Items zu nicht zentralen Einzelinformationen

Die Stufen unterscheiden sich voneinander dadurch, dass aufsteigend komplexere und längere Stimuli verstanden werden müssen. Die Items zeichnen sich durch abnehmende Lenkung (Multiple-Choice-Item vs. Offenes Item) und zunehmende Tiefe der kognitiven Operation (Informationen finden vs. Interpretieren/Bewerten) aus. Die zu extrahierende Information ist mit zunehmender Kompetenzstufe zunehmend schwieriger zu entnehmen und ggf. über mehrere Stellen des Stimulus hinweg zu finden und zu verbinden. Auch paraverbale Informationen spielen auf den höheren Kompetenzstufen eine Rolle.

4.4.2.2. Exkurs: Luxemburg

Luxemburg ist insofern für die Arbeit mit Sprachtests interessant, als es historisch und geographisch bedingt über drei offizielle Landessprachen verfügt: Luxemburgisch, Französisch und Deutsch. Deshalb und aufgrund des hohen Anteils an ausländischen Mitbürgern wird in Luxemburg in allen sprach- und kulturpolitischen Konzepten auf die Wertschätzung und den Erhalt der Mehrsprachigkeit Wert gelegt. So wird die Mehrsprachigkeit auch im Schulsystem konsequent eingehalten. Die Alphabetisierung der Kinder erfolgt zunächst auf Deutsch. Deutsch ist im Enseignement primaire auch Unterrichtssprache in den Fächern Mathematik und Sachkunde. Jedoch kommt bereits im zweiten Halbjahr der zweiten Klasse Französischunterricht hinzu. Im zweiten Teil des Enseignements secondaire löst Französisch dann Deutsch als Unterrichtssprache ab. (vgl. Ministère de l'Education nationale et de la Formation professionnelle, 2008)

Deutsch ist aufgrund der Sprachverwandtschaft zum Luxemburgischen keine Fremdsprache für die Schüler. Da es in Luxemburg jedoch nicht als kommunikative Verkehrssprache gebraucht wird, kann es auch nicht als Zweitsprache bezeichnet werden. Deutsch ist in Luxemburg

Alphabetisierungssprache sowie Kommunikationssprache in der Schule. Man begegnet der deutschen Sprache in der Literatur und im Fernsehen. Deshalb ist der Deutschunterricht in Luxemburg weder mit dem muttersprachlichen Deutschunterricht in Deutschland noch mit dem fremdsprachlichen Deutschunterricht eines anderen Landes vergleichbar. (vgl. ebd., 16ff)

Die Kompetenzstufenmodelle in Luxemburg beschreiben die Sprachhandlungskompetenz der Schüler im Schriftlichen und im Mündlichen, und zwar für die Kompetenzbereiche „Leseverstehen", „Hörverstehen", „Texte schreiben" sowie „Sprechen, Reden und Zuhören". Die Bereiche „Sprache gebrauchen und Sprachgebrauch untersuchen" und „Methoden und Arbeitstechniken" sind in diese basalen Kompetenzbereiche integriert. Für diese Arbeit sind aus Gründen der mangelnden Vergleichbarkeit mit den deutschen Kompetenzstufenmodellen jedoch nicht die Modelle als solche interessant, sondern die Tatsache, dass die Beschreibungen curricular übergreifend für alle Schulformen und Ausbildungszyklen einen Orientierungsrahmen abgeben sollen. Daher werden in den Luxemburgischen *Bildungsstandards Sprachen* (vgl. Ministère de l'Education nationale et de la Formation professionnelle, 2008) Angaben darüber gemacht, wie die Bildungsstandards für die einzelnen Schulformen und Ausbildungszyklen angepasst werden können. Um das Anforderungsniveau zu erhöhen bzw. zu verringern, sollen „Art und Komplexität des Inhalts", die „Komplexität des Textes" sowie „Art und Schwierigkeitsgrad der Aufgabenstellung" variiert werden. Dabei werden zu jedem Bereich mehrere Unterpunkte (z. B. Relevanz für die Lernenden, Thema und Inhalt des Textes) mit unterschiedlichen Ausprägungspolen (vertraut vs. neu) oder zu berücksichtigenden Aspekten (z. B. Motivation, notwendiger Verstehenshorizont) angegeben. Den Lehrkräften soll das Schema als Anleitung dienen, Stimuli und Aufgaben hinsichtlich verschiedener Kriterien zu prüfen und ggf. für ihre Lerngruppe zu verändern. (vgl. ebd., 61ff; Abbildung II-4.4.2.2.)

Art und Komplexität des Inhalts

- Die Relevanz für die Lernenden (Motivation, notwendiger Verstehenshorizont, Vor- und Kontextwissen, Lebensweltbezug)
- Thema und Inhalt des Textes (vertraut vs. neu, komplex vs. einfach, beschreibend vs. problemorientiert)

Komplexität des Textes

- Textsorte (z. B. vertraut, bekannt, neu; Gebrauchs-, Sach- und Medientexte vs. literarische Texte)
- Die sprachliche Komplexität (z. B. einfacher vs. abstrakter oder bildlicher Ausdruck, rhetorische Besonderheiten, sprachliche Register)
- Die besondere (Themen)Struktur und Kohärenz des Textes (z.B. konkret vs. abstrakt, zeitnah vs. historisch, linear, chronologisch, episodisch, verschränkt, Vor- und Rückblenden, Montagen, Intertextualität, explizite vs. implizite Textbezüge; erzählend, beschreibend, argumentativ, u. s. w.)
- Die mediale Realisierung (gesprochen vs. geschrieben, Sprechtempo, gleichzeitiges Sprechen, Text-Bild-Verhältnis)
- Die Textlänge

Art und Schwierigkeitsgrad der Aufgaben

- Entlastung und Orientierungshilfen vor der Bearbeitung von Aufgaben
- Vertrautheit mit den Arbeitsanweisungen und Aufgabenstellungen
- Zeit für die Beantwortung der Aufgaben
- Anzahl der Aufgaben
- Verschränkung von Aufgaben und Teilaufgaben, Aufgabenkomplexe
- Aufgabenformate (offene, halboffene oder geschlossene Aufgabentypen)
- Formative oder summative Aufgabenstellungen
- Reproduktive, produktive oder interaktive Aufgabentypen
- Sozialformen: Einzel-, Partner- oder Gruppenarbeit

Abbildung II-4.4.2.2.: Schwierigkeitsbeeinflussende Merkmale Luxemburg (vgl. Ministère de l'Education nationale et de la Formation professionnelle, 2008: 61ff)

4.4.3. Hörverstehen in den Rahmenplänen der Länder

Neben den Bildungsstandards und den dazugehörigen Kompetenzstufenmodellen sind für die Lehrkräfte in Deutschland noch die Lehrpläne und Curricula ihres jeweiligen Bundeslandes verbindlich. Von den 16 Ländern der Bundesrepublik Deutschland wurden für die vorliegende Arbeit insgesamt 41 Lehrpläne der jeweils vertretenen Schularten untersucht. Bei der Analyse wurden einerseits die fachspezifischen Zielbeschreibungen berücksichtigt, aber auch die fächerübergreifenden Fertigkeiten und überfachlichen Kompetenzen. In vielen Lehrplänen werden bei den Zielbeschreibungen verbindliche und fakultative Ziele unterschieden. Dieser Verbindlichkeitsgrad variiert jedoch über die untersuchten Dokumente hinweg. Für die Analyse waren deshalb sowohl verbindliche als auch unverbindliche Zielbeschreibungen relevant. Die Unterschiedlichkeit der Zielbeschreibungen führte dazu, dass keine explizite Aussage darüber möglich ist, wie gut die Schüler ein Lernziel beherrschen müssen. Auch Aussagen über Kompetenzstufen bzw. Kompetenzunterschiede über die Jahrgangsstufen hinweg können deshalb nicht gemacht werden.

Im Allgemeinen fällt auf, dass der Kompetenzbereich „Sprechen und Zuhören" zwar vor allem in neueren Lehrplanversionen verstärkt Beachtung findet, insgesamt jedoch nach wie vor recht vage Angaben dazu gemacht werden. Zuhören (analog zum Lesen) wird meist noch nicht als eigenständige Teilkompetenz betrachtet, sondern eher als Teilfähigkeit, die dienenden Charakter hat. Sie ist zwar für den Erwerb „höherer" Kompetenzen, wie der Gesprächskompetenz, notwendig, wird aber in der 9. Jahrgangsstufe im Wesentlichen vorausgesetzt.

So werden beispielsweise im Lehrplan für den Mittleren Schulabschluss im Fach Deutsch in der 9. Jahrgangsstufe in Sachsen aus dem Jahr 2004 die allgemeinen fachlichen Ziele „Entwickeln des Leseverstehens", „Entwickeln der mündlichen Sprachfähigkeit", „Entwickeln der schriftlichen Sprachfähigkeit" und „Entwickeln der Reflexionsfähigkeit über Sprache" unterschieden. Dazu werden fünf Lernbereiche und drei Wahlpflichtbereiche definiert. Angaben zum Hörverstehen finden sich über die Lernbereiche hinweg verstreut, und zwar (vgl. Sächsisches Staatsministerium für Kultus, 2001: 31 – 34):

> *Lernbereich 1 „Gewusst wie"*
> Anwendung von Methoden der Informationsbeschaffung (Befragung → Ergebnisse zusammenfassen; Mitschrift → Unterrichtssequenzen)
>
> *Lernbereich 4 „Botschaften verstehen"*
> Kennen von Funktion und Wirkung der sprachlichen Kommunikation → Strategien der Texterschließung, Diskussion (Inhalts- und Beziehungsaspekte der Kommunikation →Sprecher und Zuhörer; konkrete Kommunikationssituationen untersuchen → Vorstellungsgespräch, Prüfungssituation, privater und öffentlicher Bereich, nonverbale Kommunikation; Verwendung von Sprache in Bezug auf Absicht und Wirkung untersuchen → Information und Manipulation, Werbung, Talkshow, Nachrichtensendung, Chat)

Wahlpflichtbereich 1 „Soundcheck"
Sich positionieren zu Texten aktueller Musikproduktionen → Liedtexte als literarische Texte (Lesen und hören → Diskussion)

Wahlpflichtbereich 3 „Abenteuer Sprache"
Beurteilung von Sprache in verschiedenen Erscheinungsformen (Sprachvarietäten in mündlichen und schriftlichen Äußerungen erkennen → Wandelbarkeit von Sprache; Angemessenheit von Umgangssprache → Situationsbezug, Personenbezug, Zweckmäßigkeit)

Der bayrische Lehrplan aus dem Jahr 2009 für das Fach Deutsch in der 9. Jahrgangsstufe in der Realschule unterscheidet für den Kompetenzbereich „Sprechen und Zuhören" die Teilbereiche „Verständlich und sinntragend sprechen", „Aktiv zuhören", „Anderen etwas mitteilen" und „Miteinander sprechen". Bereits der einleitende Passus macht deutlich, dass Kompetenzen im Bereich „Zuhören" höchstens als Teil einer allgemeineren Gesprächskompetenz gesehen werden. „Die Schüler vertiefen die schon in früheren Jahren angebahnte Gesprächs- und Diskussionsfähigkeit. Vor allem im Rahmen der beruflichen Orientierung erproben sie Gesprächsstrategien und Verhaltensregeln für besondere Situationen. Der Gebrauch der Standardsprache und die mündliche Ausdrucksfähigkeit werden weiter gefördert." (Bayerisches Staatsministerium für Unterricht und Kultus, Realschule: Lehrpläne/Standards Realschule R6: 404) Differenziertere Angaben zum Zuhören werden überwiegend in den Bereichen „Aktiv zuhören" und „Anderen etwas mitteilen" gemacht, und zwar:

Aktiv zuhören - anderen auch über einen längeren Zeitraum konzentriert zuhören - zum Gehörten Fragen stellen und Stellung nehmen - Informationen aufnehmen und differenziert verarbeiten, z. B. für die Zusammenfassung von Diskussionsergebnissen, für ein Protokoll, für Argumentationen

Anderen etwas mitteilen - Informationen einholen, zusammenfassen und wiedergeben, z. B. Möglichkeiten oder Verlauf des Betriebspraktikums, aktuelle Ereignisse aus Medien, Diskussionsergebnisse, Textinhalte

Die Vorgaben aller Lehrpläne wurden in einer Synopse für den Hauptschulabschluss und einer für den Mittleren Schulabschluss zusammengefasst und den KMK Bildungsstandards gegenübergestellt. Dabei ergaben sich aufgrund der inhaltlichen Vorgaben für beide Synopsen die Themenbereiche „Diskutieren", „Informationen entnehmen und informieren" sowie „Sprachbetrachtung". Es fällt auf, dass die meisten Lehrpläne Hörverstehen bei der Formulierung der von den Schülern zu erlangenden Kompetenzen implizieren und diesen Kompetenzbereich i. d. R. anwendungsbezogen im Zusammenhang mit der Kompetenz „Sprechen" im Rahmen von Gesprächskompetenz behandeln. Dies wird in Forderungen wie „Störungen in Gesprächsabläufen erkennen und Verbesserungsvorschläge erarbeiten" oder „Anteil nehmen durch Bestätigen, Rückfragen, Widerspruch" deutlich. Explizit in Teilkompetenzen aufgeschlüsselt und beschrieben werden zu erwartende Fähigkeiten, die allein in den Kompetenzbereich Hörverstehen fallen, jedoch kaum[7].

7 vgl. dagegen beispielsweise die Gliederung des GER (Europarat, 2001: 71) „Selektiv verstehen", „Detailliert verstehen", „Global verstehen" und „Schlussfolgerungen ziehen können"

Da die Dimensionalität von Hörverstehen jedoch noch nicht abschließend untersucht worden ist und Hörverstehen in der authentischen Sprachverwendungssituation tatsächlich meist im Rahmen von Gesprächen relevant wird, erscheint der eingeschlagene Weg durchaus plausibel. Die Forderungen, die in den Bildungsstandards unter dem Standard 1.4. subsumiert sind, entsprechen im Wesentlichen den in Tabelle II- 4.4.3a. zusammengefassten Lehrplanformulierungen:

Tabelle II-4.4.3a.: Vergleich BS 1.4. mit Lehrplanformulierungen

Standard	Bildungsstandards	Ergänzung aus den Lehrplänen
BS141	Gesprächsbeiträge anderer verfolgen und aufnehmen	• Genau zuhören • Fremde Meinungen aufgreifen und berücksichtigen • Informationen aufnehmen
BS142	Wesentliche Aussagen aus umfangreichen gesprochenen Texten verstehen, diese Informationen sichern und wiedergeben	• Wesentliche Aussagen zusammenfassen (Sachtexte, thematisch eingegrenzte überschaubare Gespräche/ Diskussionen) • Techniken des Mitschreibens • Verlaufs- und Ergebnisprotokoll
BS143	Aufmerksamkeit für verbale und nonverbale Äußerungen (z. B. Stimmführung, Körpersprache) entwickeln.	• Verbale und nonverbale Gestaltungsmittel der Vortragenden wahrnehmen und benennen

Die Angaben der Lehrpläne ergänzen die Deskriptoren der Bildungsstandards und machen häufig konkrete Angaben dort, wo die Bildungsstandards relativ knapp bleiben. Dies wird am Beispiel in Tabelle II-4.4.3b. deutlich. Der ausführliche Vergleich der Bildungsstandards mit den Lehrplanangaben findet sich in Anhang A.

Tabelle II-4.4.3b.: Vergleich Bildungsstandard – Lehrplan

Formulierung der Bildungsstandards	Ergänzung aus den Lehrplänen der Länder
Gesprächsregeln einhalten/ Redestrategien einsetzen (Fünfsatz, Anknüpfungen formulieren)	• Gesprächsregeln/-strategien entwickeln: - aktiv zuhören - ausreden lassen - Fragestellung erfassen - Standpunkte klären - Nachfragen - Strittiges klären - Sich auf Vorredner beziehen - Nach Lösungen suchen - Körpersprache - Sprachebene - Angemessene Selbstdarstellung - Steuerung des Gesprächsverlaufs - zusammenfassen • Höflich miteinander umgehen: - Höflichkeitsformen - Beschwerden - Kritik üben • Fair, zielgerichtet, kriterienorientiert und wirkungsvoll diskutieren • Unfaires Diskussionsverhalten abwehren

Die Lehrpläne geben aus diesem Grund wertvolle Impulse für die Aufgabenentwicklung, indem sie inhaltlich explizieren, was genau mit einem recht global formulierten Bildungsstandard gemeint sein kann. Da es sich dabei jedoch insgesamt überwiegend um den Bereich Gesprächskompetenz handelt, werden kaum Methoden und Inhalte genannt, mit denen speziell „Zuhören" und „Hörverstehen" geübt werden können. Daher sind die Lehrpläne sowohl für die Aufgabenentwicklung im Bereich „Zuhören" als auch für die Suche nach schwierigkeitsbeeinflussenden Merkmalen nur bedingt hilfreich.

5. Ausgewählte Merkmale und ihr Einfluss auf die Itemschwierigkeit

Das Kapitel „Ausgewählte Aufgabenmerkmale und ihr Einfluss auf die Itemschwierigkeit" enthält einen Überblick über ausgewählte Studien zu schwierigkeitsbeeinflussenden Merkmalen. Vorgestellt werden Faktoren, die nachweislich einen Einfluss auf die Schwierigkeit von Hör- und Leseverstehensaufgaben hatten. Da angenommen wird, dass sich, bis auf einige modalitätsspezifische Subdimensionen, die Prozesse des Lesens und des Zuhörens vor allem im Bereich des „Verstehens" sehr ähnlich sind, wird auch auf Studien Bezug genommen, die sich mit dem Leseverstehen beschäftigen. Viele Studien fanden unter experimentellen Bedingungen statt, sodass die isolierten Merkmale bei den sehr heterogenen IQB-Aufgaben möglicherweise keinen Effekt zeigen.

Die Schwierigkeit eines Textes wird nach Klein-Braley (1994: 184) von mindestens drei Faktoren beeinflusst: Den Fähigkeiten des Lesers, dem Textinhalt und der Textstruktur. Dementsprechend werden neben den Stimulusmerkmalen („Textinhalt" und „Textstruktur") auch Personenmerkmale, wie das Hintergrundwissen der Probanden zum Thema oder ihre Arbeitsgedächtniskapazität, beschrieben. Zu den Merkmalen, die an den Stimuli festzumachen sind, gehören u. a. die Komplexität des Wortschatzes und andere sprachliche Merkmale. Ferner spielt auch die Stimulusstruktur, der Aufbau der Stimuli nach wiederkehrenden Mustern, eine Rolle. Eine Beschreibung der Stimuli nach Relationstypen oder nach ihrem thematischen Aufbau führt zu Kriterien, wie der Wiederaufnahmestruktur der Stimuli oder ihrem Grad an Kohärenz. Im Rahmen einer Befragung von Lehrkräften zu den Stimuli wurden zusätzlich subjektive Einschätzungen zu bestimmten Merkmalen wie Gesamteindruck oder Wirkung eingeholt. Unter Itemmerkmalen werden Merkmale verstanden, die sich auf die Items beziehen. Dazu gehört zum Beispiel das Merkmal „Itemformat", wobei insbesondere auf die Merkmale von Multiple-Choice-Items eingegangen wird. Zu diesem Itemformat gibt es aus der angloamerikanischen Forschung relativ viel Literatur, da es für dieses Item eine lange Testtradition gibt. Viele Itemmerkmale sind jedoch „interaktiv" und können nicht losgelöst vom Stimulus betrachtet werden. Dazu zählen einerseits Merkmale wie der Zeitpunkt der Itembearbeitung, aber auch der Grad der Überlappung der Item-Formulierungen mit dem Stimulus. Andere Merkmale verlangen eine kognitive Operation von der Testperson, die nur mithilfe des Stimulus erbracht werden kann. Relativ zentral ist dabei die Information, die zur Beantwortung des Items notwendig ist (NI). Diese Information wird in ihrer Art, ihrer Position und ihrer Auftretenshäufigkeit im Stimulus beschrieben. Nicht unberücksichtigt sollen auch die Eigen-

schaften der Testpersonen bleiben, da angenommen wird, dass unterschiedliche Probanden unterschiedlich auf den gleichen Test reagieren. Untersucht wird deshalb, ob Motivation hinsichtlich der Testbearbeitung und Interesse an den Stimuli einen Einfluss auf die Testleistung haben. Auch das Hintergrundwissen der Schüler zu den ausgewählten Themen, ihre Sprachkenntnisse sowie ihre Arbeitsgedächtniskapazität werden in die Analysen mit einbezogen.

Aufgrund der hohen Korrelationen zwischen Lese- und Hörverstehen wird heute davon ausgegangen, dass beide Kompetenzbereiche über eine gemeinsame Subdimension „Verstehenskompetenz" verfügen, sich jedoch auch hinsichtlich einiger Subdimensionen voneinander unterscheiden, insbesondere in Bereichen, die Dekodierprozesse betreffen. (vgl. Hale et al., 1989; Buck, 1992; Bae & Bachman, 1998) Obwohl beide rezeptiven Fertigkeiten viele Gemeinsamkeiten aufweisen, unterscheiden sie sich beispielsweise darin, dass Zuhören stärker auf ein holistisches Verständnis der Botschaft abzielt als Lesen. Kürschner und Schnotz zeigen in einem integrierten Modell des Hör- und Leseverstehens (Kürschner & Schnotz, 2008: 144), dass sich beide Kompetenzbereiche auf bestimmten Verarbeitungsebenen unterscheiden und unter bestimmten Verarbeitungsbedingungen auch zu unterschiedlichen Lernleistungen führen können. Diese Ergebnisse rechtfertigen die Berücksichtigung der wissenschaftlichen Literatur zu schwierigkeitsbeeinflussenden Merkmalen auch im Bereich des Leseverstehens. Selbst wenn sich die Prozesse des Zuhörens und des Lesens voneinander unterscheiden, weisen beide Domänen so viele Gemeinsamkeiten auf, dass die folgenden Untersuchungen für die Analysen in dieser Arbeit berücksichtigt werden.

Aus der Literatur z. B. zum fremdsprachlichen Leseverstehen (z. B. Freedle & Kostin, 1993a; Evetts & Gauthier, 2005), zum fremdsprachlichen Hörverstehen (z. B. Freedle & Kostin, 1996) und zum Leseverstehen in der Erstsprache (z. B. Chalifour & Powers, 1989; Nold & Rossa, 2007) ergeben sich bestimmte Merkmale, von welchen vermutet wird, dass sie sich als Prädiktoren für die Itemschwierigkeit erweisen. Dabei wird zwischen Merkmalen unterschieden, die sich auf die Stimuli oder die Items beziehen oder die auf einer Interaktion vom Stimulus mit den Items beruhen. (Buck & Tatsuoka, 1998)

In der Regel wird unter Itemschwierigkeit das gemittelte Ergebnis einer Gruppe von Testpersonen bei einem Item verstanden. Itemschwierigkeit kann aber auch die Interaktion von latentem Faktor und der Performanz bei einem bestimmten Item sein. In jedem Fall hängt die Schwierigkeit aber in den üblichen Messmodellen von der Testleistung ab und ist kein „reines" Merkmal der Items. Mithilfe von Itemmerkmalen empirische Itemschwierigkeiten vorherzusagen ist problematisch, weil diese Itemstatistiken selbst eine Funktion der Interaktionen zwischen den Testpersonen und den Items sind. Korrelationen zwischen den Itemmerkmalen und der empirischen Itemschwierigkeit beruhen deshalb in gewisser Weise immer auf Autokorrelationen. Auch Bachman (2002) weist darauf hin, dass die Itemschwierigkeit keine Eigenschaft des Items ist. Vielmehr sei sie immer relativ zur getesteten Person zu sehen ist, da Personenfähigkeiten mit den Anforderungen der Items interagieren. Deshalb ist nach Bachman Itemschwierigkeit auch kein isolierter Faktor, sondern beruht auf dem Zusammenspiel unterschiedlicher Komponenten im Test (z. B. Personenfähigkeiten, Itemanforderungen, etc.). Die Vorhersagekraft bestimmter Itemmerkmale ist demnach immer zweideutig: Sie kann auf

der Prädiktorfunktion des Merkmals beruhen, aber auch darauf, dass das Merkmal mit der abhängigen Variable interagiert. Bachman rät deshalb gründlich zwischen unterschiedlichen Kategorien von Merkmalen zu trennen und beispielsweise Merkmale, die sich nur auf den Stimulus beziehen von solchen zu unterscheiden, bei denen die Testpersonen involviert sind. Auch Buck und Tatsuoka (1998) weisen darauf hin, dass Merkmale stets aus zwei unterschiedlichen Perspektiven betrachtet werden können: aus der Perspektive des Items und aus der Perspektive der Testperson, von der bestimmte Fähigkeiten zur Lösung eines Items erwartet werden. Da jedoch die Eigenschaften eines Stimulus auch immer mit den Merkmalen der Items und der Testpersonen interagieren, sind „reine" Vorhersagen der Stimulus- oder der Itemschwierigkeit nicht möglich und die Ergebnisse müssen immer ganzheitlich (d. h. im Zusammenspiel Stimulus – Item – Testperson) interpretiert werden. (vgl. Grotjahn, 2000)

5.1. Stimulusmerkmale

Jensen et al. (1997) kamen in ihren Arbeiten zum Ergebnis, dass Stimulus-Merkmale kaum Einfluss auf die Itemschwierigkeit haben, solange es sich nicht um sehr technische Texte handelt. Stattdessen erwiesen sich Merkmale, die auf einer Interaktion der Items mit den Stimuli und den Testpersonen beruhen als stärkere Prädiktoren. Dennoch gibt es zahlreiche Studien, die sich vor allem mit Stimulus-Merkmalen im Bereich des fremdsprachlichen Leseverstehens (z. B. Keshavarz (2007) und Nold & Rossa (2007) mit ESL Lernern) und des Leseverstehens in der Erstsprache auseinandersetzen (z. B. Evetts & Gauthier (2005) und Best et al. (2006) mit englischsprachigen Lernern sowie Yin-Kum (1995) mit chinesischen Probanden). Im Bereich des Hörverstehens überwiegen Studien mit L2-Lernern, z. B. Bacon (1992) mit Spanisch-Lernern; Cervantes & Gainer (1992) und Field (2003) mit ESL Lernern.

Derartige Studien sind seltener für das Deutsche und i. d. R. weniger stark auf einzelne Merkmale fokussiert. Grotjahn (2000) untersucht schwierigkeitsbeeinflussende Merkmale im DaF Bereich, Bremerich-Vos et al. (2009) untersuchen sprachliche Kompetenzen im Fach Deutsch im Primarbereich und Böhme & Robitzsch (2009) beschreiben methodische Aspekte der Erfassung der Lesekompetenz. Gerade für den Kompetenzbereich „Zuhören" gibt es derzeit noch verhältnismäßig wenig Literatur.

5.1.1. Sprachliche Merkmale des Stimulus

Zu den sprachlichen Merkmalen der Stimuli zählen neben der Anzahl und der Länge von Stimulusfragmenten, wie Wörter oder Silben, auch die Worthäufigkeit und die Komplexität des Wortschatzes. Die Komplexität des Wortschatzes wird häufig am Grad der Konkretheit und der Menge an Inhalts- und Funktionswörtern, aber auch an der Anzahl und Länge von Wörtern oder Silben bestimmt. Die Anzahl der Silben wird dabei repräsentativ für die Wortlänge angesehen. Im Gegensatz zur Gliederung von Sprache in Morpheme, die kleinsten bedeutungstragenden Einheiten, ist eine intuitive Gliederung einer Wortform in Silben meist sehr einfach, denn alle Silben weisen einen zentralen Vokal auf, der von unterschiedlichen Konsonantenkonstellationen umgeben ist. (Imhof, 2003: 83) Die Anzahl und Länge der Wörter oder Silben wurden als Ausdruck der Komplexität des Wortschatzes im Rahmen der Lesbarkeitsforschung in den 70er Jahren recht häufig untersucht (vgl. Klare, 1963; Abram & Dowling, 1979).

Klein-Braley (1994) beruft sich bei ihrer Verwendung des Merkmals „Wortlänge" als Indikator bezüglich der Häufigkeit des Wortschatzes auf das Zipf-Mandelbrot-Gesetz: Demnach sind häufig auftretende Wörter kürzer. Seltene und kaum gebrauchte Wörter sind den Lesern/Zuhörern oft nicht bekannt und diese Wörter können das Verständnis des Stimulus erschweren. Dementsprechend zeigten Embretson und Wetzel (1987), dass schwierigere Items eines Tests zum Textverstehen auch einen höheren Anteil an seltener gebrauchten Wörtern aufweisen. Einen positiven Zusammenhang zwischen dem Wortschatz von Testpersonen und deren Leseverstehenskompetenz wies Carroll (1993) nach. Dieser Zusammenhang kann damit erklärt werden, dass ein reicher Wortschatz bzw. die Geschwindigkeit, mit der beim Lesen darauf zugegriffen werden kann, einen Einfluss auf das Leseverstehen hat. (Daneman, 1988; Perfetti, 1994)

Freedle und Kostin (1993a) konnten für den Bereich des Leseverstehens nachweisen, dass die Itemschwierigkeit durch kürzere Wörter mit ein bis zwei Silben gesenkt wurde. Bei Hörverstehensaufgaben senkten jedoch längere Wörter mit drei oder mehr Silben die Itemschwierigkeit. Rowe und McNamara (2008) weisen darauf hin, dass die Auftretenshäufigkeit allein noch kein Hinweis auf die Schwierigkeit von Wörtern sein muss. In der Regel haben häufig gebrauchte Wörter mehrere Bedeutungen und der Kontext muss stärker berücksichtigt werden.

Studien zur Passagenlänge von Newsome und Gaite (1971) und zur Satzlänge von Klare (1974-75) zeigen, dass die Anzahl der Textfragmente (Silben, Wörter, Sätze, Absätze) die Textschwierigkeit beeinflussen kann. Beispielsweise berichtet Goh (2000), dass ihre Testpersonen Schwierigkeiten hatten, lange Sätze zu verstehen und Just und Carpenter (1992) zeigten, dass die Verarbeitung längerer Sätze größere Ansprüche an das Arbeitsgedächtnis stellt. Embretson und Wetzel (1987) verwendeten eine entsprechende Variable in ihren Untersuchungen und konnten signifikante Korrelationen mit der Itemschwierigkeit nachweisen. Freedle und Kostin (1993a) bestätigten zunächst die Ergebnisse von Embretson und Wetzel, konnten in einer weiteren Studie (Freedle & Kostin, 1996) jedoch keinen Einfluss der durchschnittlichen Satz- und Passagenlänge mehr auf die Itemschwierigkeit feststellen. Die Itemschwierigkeit stieg jedoch mit ansteigender Wortzahl im Aufgabenstamm, in den Distraktoren und in der NI. Perkins und Brutten (1993), Gorin und Embretson (2006) sowie Nissan et al. (1996) fanden keinen signifikanten Zusammenhang zwischen der Länge des Textabschnitts und der Itemschwierigkeit.

Trotz dieser gegensätzlichen Befunde wird angenommen, dass Satzlänge und Satzkomplexität korrelieren, da zumeist komplexe Sätze auch länger sind. Ferner beinhalten längere Sätze mehr linguistische Variablen, wie z. B. Einschübe oder adjektivische Ergänzungen, welche als schwieriger zu verarbeiten gelten. Es wird angenommen, dass sich der kognitive Aufwand beim Lösen eines Items mit der Länge des dafür relevanten Stimulusabschnitts erhöht, gerade wenn der Stimulus nur einmal gehört wird. Der Grund dafür liegt darin, dass mit zunehmender Länge mehr Informationen simultan verarbeitet werden müssen. Außerdem besteht die Gefahr, dass die Aufmerksamkeit der Testpersonen bei längeren Sätzen oder Passagen nachlässt. (vgl. Kapitel 2.1.1.2. *Baddeleys Modell des Arbeitsgedächtnisses*)

Die Komplexität des Wortschatzes ist aber nicht nur abhängig von der Länge und der Anzahl von Stimulusfragmenten, sondern auch von Faktoren, wie dem Anteil an Fachvokabular, abstrakten oder konkreten Begriffen, Inhaltswörtern, Synonymen, der Relation Unter- und Oberbegriff u. ä. Perfetti (1985) fand heraus, dass ein Textabschnitt mit vielen unbekannten, selten vorkommenden Wörtern schwieriger zu verstehen ist. Auch Nissan et al. (1996) fanden für die Variable „Infrequent Vocabulary" eine signifikante Korrelation mit der Itemschwierigkeit.

Ein weiterer Prädiktor der Schwierigkeit ist der Abstraktionsgrad der Stimuli. Von abstrakten Themen handelnde Stimuli sind schwieriger zu verstehen als Stimuli, die sich thematisch nahe an der Lebenswelt der Lernenden befinden. Freedle und Kostin (1996) konnten zeigen, dass erwartungsgemäß konkrete Stimuli, die nicht in erster Linie von abstrakten Konzepten handelten, einfacher waren als abstrakte Stimuli. Dieses Ergebnis erhielten sie auch für Lesetexte. Konkrete Informationen, wie Angaben über Personen und Aktionen, müssen unter Umständen nur im Sinn eines Abgleichs von Item- und Stimulusformulierung lokalisiert werden. Um bestimmte abstrakte Informationen, wie Gründe oder Motivationen, zu erschließen, kann es notwendig sein, zu generalisieren oder zu inferieren. Die Entnahme von abstrakteren Informationen ist also in der Regel schwieriger. (Mosenthal, 1996) Eine Einschätzung des Abstraktheitsgrads von Informationen wurde von Evetts und Gauthier (2005) mit einem fünfstufigen Schema vorgenommen. Das Schema erfasst den „Type of Requested Information" (TORI) und wurde auch für die Untersuchung der IQB-Aufgaben verwendet.

Ein weiteres aussagekräftiges Merkmal ist die Unterscheidung von Inhalts- und Funktionswörtern. Inhaltswörter sind bedeutungstragende Wörter wie Nomen, Verben, Adjektive und Adverbien. Es wird angenommen, dass die Verarbeitung von Texten mit vielen Inhaltswörtern schwieriger ist als mit wenigen Inhaltswörtern. Aus Studien zum Leseverstehen ist bekannt (vgl. Lüer, 1988; Willenberg, 1995), dass Leser Inhaltswörter länger fixieren als Funktionswörter. Dies spricht dafür, dass die Verarbeitung von Inhaltswörtern mehr kognitive Energie in Anspruch nimmt und deshalb schwieriger ist als die Verarbeitung von Funktionswörtern. (vgl. Imhof, 2003: 85) Davey (1988) konnte nachweisen, dass der Anteil an Inhaltswörtern im Attraktor und den Distraktoren signifikant mit der Itemschwierigkeit korrelierte. In Bezug auf den Anteil an Inhaltswörtern im Lesetext und der Itemschwierigkeit fand sich kein signifikanter Zusammenhang. Dieser Zusammenhang wurde jedoch von Embretson und Wetzel (1987) nachgewiesen. Durch eine höhere Anzahl an Inhaltswörtern steigt auch die Anzahl der zu verarbeitenden Propositionen und die Ansprüche an das Arbeitsgedächtnis nehmen zu. Die Befunde von Embretson und Wetzel (1987) wurden von Perkins und Brutten (1993) bestätigt.

Einige weitere Merkmale erschweren die Verarbeitung des sprachlichen Inputs. Beispielsweise wird beim Verstehen eines Negativsatzes zuerst die inferierte Annahme extrahiert und dann die enthaltene Negation verarbeitet. (Anderson, 2001: 400ff) So konnten Freedle und Kostin für das Hör- und das Leseverstehen zeigen (1996: 23; 1993a), dass Negationen im Stimulus, im Attraktor und in den Distraktoren die Item- und die Aufgabenschwierigkeit erhöhen. Es wurden signifikante Korrelationen für die Variablen „Verneinungen in den korrekten Antwortoptionen" und „Verneinungen in den inkorrekten Antwortoptionen" gefunden. Für die Anzahl der Verneinungen im Stimulus konnte jedoch kein signifikanter Zusammenhang mit

der Itemschwierigkeit nachgewiesen werden. Nissan et al. (1996) erhielten für die Variable „Negative in Stimulus" nur dann signifikante Ergebnisse, wenn mehr als eine Negation im Stimulus auftrat. Allerdings wurden in dieser Studie TOEFL Dialog-Items zum Hörverstehen untersucht, bei denen die Stimuli zwischen 2.79 und 15.42 Sekunden dauern, also sehr kurz sind. Negationen in den Multiple-Choice-Optionen hatten keinen Einfluss auf die Itemschwierigkeit.

5.1.2. Struktur der Stimuli

Bei der Beschreibung der Stimuli spielen auch ihre grammatische und thematische Struktur eine wichtige Rolle. Die Ausführungen zur Struktur stammen aus der Textlinguistik und beziehen sich dementsprechend auf Texte. Für die Analysen der IQB-Stimuli wurden die relevanten Merkmale jedoch so adaptiert, dass sie sich auch für die Untersuchung der Diskurse eignen. Unter dem Begriff „Textstruktur" versteht man „the logical connections among ideas and subordination of some ideas to others." (Meyer & Freedle, 1984: 127) Der Begriff „Textstruktur" stammt aus der Textlinguistik, soll aber im Folgenden unter dem Begriff „Stimulusstruktur" sowohl für Texte als auch für Diskurse angewendet werden. Stimuli sind i. d. R. nach bestimmten Mustern aufgebaut und können nach immer wiederkehrenden, strukturellen Relationstypen in größere Abschnitte organisiert werden. Die Relationstypen geben an, wie ein Satz bzw. eine Äußerung auf den gesamten Stimulus bezogen werden soll. Relationstypen können auf allen Ebenen des Stimulus existieren. Für Abschnitte können also genauso wie für Unterabschnitte Relationen gefunden werden und die Inhaltselemente eines längeren Stimulus können hierarchisch organisiert werden. Anderson (2001: 414) stellt eine Auswahl der wichtigsten Relationstypen vor (vgl. Tabelle II-5.1.2.):

Tabelle II-5.1.2.: Relationstypen nach Anderson 2001

Relationstyp	Beschreibung
1. Antwort	Auf eine gestellte Frage folgt eine Antwort, oder ein Problem wird vorgestellt und seine Lösung folgt.
2. Spezifizierung	Nach einer allgemeineren Darstellung werden spezifische Informationen gegeben.
3. Erklärung	Für einen Sachverhalt wird eine Erklärung gegeben.
4. Beweis	Zur Unterstützung einer Aussage werden Beweise vorgelegt.
5. Reihenfolge	Argumente werden in ihrer zeitlichen Abfolge zusammenhängend dargeboten.
6. Ursache	Ein Ereignis wird als Ursache eines anderen Ereignisses dargestellt.
7. Ziel	Ein Ereignis wird als Ziel eines anderen Vorgangs dargestellt.
8. Aufzählung	Sachverhalte werden in loser Struktur aufgeführt.

Die grammatische Verknüpfungsstruktur eines Stimulus verweist auf seine thematische Struktur. Der Begriff „Thema" bezieht sich einerseits auf den kommunikativen Hauptgegenstand und andererseits auf den Grund- oder Leitgedanken eines Stimulus. Das Thema ist dann der Kern des Stimulus und kann dabei als die größtmögliche Kurzfassung des Inhalts entweder in einem bestimmten Segment realisiert sein oder es muss aus dem Inhalt abstrahiert werden. In der Regel enthält ein Stimulus mehrere Themen mit unterschiedlicher Relevanz, wobei die Bestimmung des Themas vom Verständnis des jeweiligen Lesers abhängt. (Brinker,

2005: 55ff) Die thematischen Zusammenhänge werden durch die Relationen der Propositionen hergestellt und durch explizite oder implizite Wiederaufnahmen erzeugt. Bei expliziten Wiederaufnahmen ist in aufeinanderfolgenden Sätzen bzw. Äußerungen eines Textes bzw. Diskurses die Referenzidentität bestimmter sprachlicher Ausdrücke gewährleistet, die sich auf das gleiche außersprachliche Objekt beziehen. Dies bedeutet, dass ein bestimmter Ausdruck durch die Wiederholung dieses Substantivs, durch andere substantivische Wörter oder durch bestimmte Pronomen nachfolgend wieder aufgenommen wird. Wiederaufnahmen können auch durch Ausdrücke mit größerem Bedeutungsumfang, sogenannten Oberbegriffen, erfolgen. Referenzen können das Verständnis des Stimulus erschweren, da vom Ausdruck auf das Objekt geschlossen werden muss. Bei impliziten Wiederaufnahmen besteht zwischen dem wiederaufnehmenden Ausdruck und dem wiederaufgenommenen Ausdruck keine Referenzidentität. Die beiden Ausdrücke stehen jedoch in Beziehung zueinander, indem beispielsweise ein Ausdruck im anderen enthalten oder ein Teil davon sein kann. (Brinker, 2005)

In der Wiederaufnahmestruktur wird die thematische Progression des Stimulus sichtbar, wie beispielsweise eine Anordnung nach dem Prinzip des Nebeneinanders oder eine Verschiebung in der thematischen Perspektive. (Brinker, 2005: 45ff) Ein Ansatz, die thematische Struktur zu erfassen, ist das Makro- und Superstrukturkonzept von van Dijk (1980), das sich an der Generativen Transformationsgrammatik und ihrer Unterscheidung von Oberflächen- und Tiefenstruktur orientiert. Nach van Dijk repräsentiert die semantische Tiefenstruktur die globale Bedeutung eines Stimulus. Durch Verfahren der paraphrasierenden Reduktion nach bestimmten Makroregeln können aus den Propositionen des konkreten Oberflächentextes die Makropropositionen der Tiefenstruktur gewonnen werden. Van Dijk geht davon aus, dass die Makrostruktur und ihr Aufbau in einem psychologischen Prozess-Modell des Textverstehens eine entscheidende Rolle spielen. Van Dijks Ergebnisse flossen in ein Modell des Textverstehens ein, das er zusammen mit Walter Kintsch erstellte (vgl. Kapitel 2.2.3. *Das Construction-Integration-Modell*).

Texte mit stärker zusammenhängender Textstruktur werden häufig als qualitativ besser und einfacher zu lesen empfunden. Im Rahmen dieser zusammenhängenden Textstruktur, dem Textzusammenhang, wird zwischen „Kohäsion" und „Kohärenz" unterschieden: (vgl. Beaugrande & Dressler, 1981) Von „Kohäsion" spricht man dann, wenn grammatisches Wissen verwendet wird, um einen Zusammenhang auf Satz- und auf Textebene herzustellen, wie beispielsweise Kongruenz im Tempus oder grammatische Verknüpfungsmöglichkeiten, wie verknüpfende Konjunktionen oder Wiederaufnahmen. Bei „Kohärenz" wird ein Textzusammenhang durch kulturelles Wissen hergestellt, z. B. durch ein einheitliches Thema oder kausale Verknüpfungen.

Die meisten Studien zur Kohärenz wurden im Bereich Leseverstehen durchgeführt (z. B. Cain, 2003; Best et al., 2005; 2006). Unter einem wenig kohärenten Stimulus werden dabei Stimuli verstanden, bei denen viele Inferenzen für das Bilden von kohärenten Repräsentationen notwendig sind. Hoch kohärente Stimuli beinhalten Elemente, die mehr explizite Hinweise auf Zusammenhänge innerhalb von und zwischen Sätzen bzw. Äußerungen geben. (vgl. Best et al., 2005; 2006) Die Kohärenz des Stimulus scheint gerade bei weniger geübten Lesern mit

einem geringeren Wortschatz eine wichtige Rolle für das Textverständnis zu spielen. Ein wenig kohärenter Stimulus mit einer geringen Anzahl erklärenden Verknüpfungen stellt höhere Anforderungen an den Leser beim Bilden einer globalen, zusammenhängenden Repräsentation, da die Informationen aus dem Stimulus mit Hintergrundwissen integriert werden müssen.

Abrahamsen und Shelton (1989) fanden in ihren Studien mit lernbehinderten Erwachsenen heraus, dass Textverständnis leichter erzielt werden konnte, wenn in den Texten Referenzbezüge durch Nominalphrasen ersetzt wurden. Best et al. (2006) arbeiteten mit Schülern der vierten Jahrgangsstufe, die unterschiedlich kohärente narrative und expositorische Texte lasen und sowohl Detailverstehensfragen als auch globale Fragen (z. B. zum Thema) beantworten sollten. Die weniger kohärenten Textversionen waren die Texte in ihrer originalen Fassung, in der stärker kohärenten Version wurden diese Texte durch kausale Konnektoren und erklärende Verbindungen zwischen den einzelnen Konzepten angereichert. Die Schüler wurden entsprechend ihres Vorwissens in zwei Gruppen eingeteilt. Narrative Texte wurden von den Kindern besser verstanden als expositorische Texte. Dies mag daran liegen, dass ihnen die Inhalte der narrativen Texte vertrauter waren. Bei expositorischen Texten fiel den Schülern vor allem die Beantwortung von globalen Fragen schwierig, bei denen Textinformationen stärker als bei Detailfragen integriert werden müssen. Möglicherweise fehlt den Kindern noch Wissen zur Verarbeitung von Inhalten aus expositorischen Texten. Diese Annahme wird auch durch die Beobachtung gestützt, dass Schüler mit mehr Vorwissen beim Verständnis von expositorischen Texten besser abschnitten. Kohärentere Texte förderten das Textverständnis insbesondere bei narrativen Texten, zu denen globale Fragen gestellt wurden. Bei expositorischen Texten wurde kein Effekt der kohärenteren Texte auf das Textverständnis festgestellt. Dies legt die Vermutung nahe, dass erhöhte Kohärenz nur dann förderlich wirkt, wenn bereits ein gewisses Textverständnis vorliegt.

Davey (1988) operationalisierte einen Kohärenz-Faktor in Lesetexten und nahm an, dass das Ausmaß der Verbundenheit von Informationen im Text durch Pronomen, Konnektive und Konjunktionen auch die Textverarbeitungsprozesse beeinflusst. Davey schätzte die Stimuli hinsichtlich der dichotomen Variable „Der Bezug von Konnektiven, Pronomen o. ä. ist (nicht) klar und eindeutig" ein. Es ergaben sich allerdings keine signifikanten Korrelationen mit der Itemschwierigkeit. Freedle und Kostin (1996: 23) konnten zeigen, dass Referenzausdrücke im Item-Stamm die Itemschwierigkeit signifikant senkten. Referenzausdrücke im Attraktor erhöhten die Itemschwierigkeit. Referenzausdrücke in nicht-akademischen Texten erleichterten das Textverständnis, was nach Einschätzung der Autoren allerdings auch damit zusammenhängen kann, dass diese Texte insgesamt einfacher waren.

Im Rahmen des *International Adult Literacy Surveys* (IALS) (Evetts & Gauthier, 2005) wird von der Annahme ausgegangen, dass Informationen in Texten in wiederkehrenden Organisationsmustern erscheinen. Dabei bestimmt die Organisation der Information im Wesentlichen die Komplexität des Textes. Unterschieden wird zwischen narrativen und darstellenden Modellen, die in weitere Sub-Modelle differenziert werden. Eine Untersuchung zum Einfluss der Stimulusstruktur auf die Behaltensleistung von Testpersonen wurde beispielsweise von Meyer und Freedle (1984) durchgeführt. Sie ließen College-Studenten Passagen lesen, die alle fast glei-

che Inhalte hatten, sich jedoch in ihrer Struktur unterschieden. Die Studenten konnten mehr Ideeneinheiten wiedergeben, wenn die Informationen vergleichend/kontrastierend oder kausal strukturiert waren als wenn sie aneinandergereiht wurden. Leser scheinen während des Lesens also bestimmte Struktur-Schemata auszuwählen und nutzen das entsprechende Wissen für die Textverarbeitung. Aktivieren Leser entsprechende Schemata, fallen ihre Testergebnisse i. d. R. besser aus als bei Lesern, die keine entsprechenden Schemata aktivieren. Nach Kletzien (1992) beeinflusst das Vorwissen der Leser zum Thema des Textes auch ihren Einsatz von Struktur-Schemata. Je mehr die Testpersonen über ein Thema wissen, desto weniger aktivieren sie Struktur-Schemata.

Auch Freedle und Kostin (1996) untersuchten bei den TOEFL Minitalk Aufgaben verschiedene rhetorische Organisationsmuster und unterschieden die Muster „Argumentieren", „Auflisten/ Beschreiben", „Kausalisieren", „Vergleichen" und „Problematisieren/Lösen". Von diesen Möglichkeiten zur rhetorischen Organisation erleichtert das Muster „Auflisten/ Beschreiben" die Items, wohingegen „Vergleichen" und „Problematisieren/Lösen" die Itemschwierigkeit erhöhte. Diese Befunde wurden für die Organisationsmuster „Auflisten/ Beschreiben" und „Problematisieren/ Lösen" in ganz ähnlicher Weise für Leseverstehensaufgaben erzielt (Freedle & Kostin, 1993b).

Yin-Kum (1995) untersuchte den Einfluss der drei Textstruktur-Typen „Causation", „Comparison" und „Collection" an 224 Schülern einer Secondary School in Hong Kong. Die Schüler, die eine Passage des Typs „Comparison" lasen, schnitten bei der Textwiedergabe insgesamt deutlich besser ab und konnten insbesondere mehr Leitgedanken wiedergeben als Schüler, die Passagen des Typs „Causation" oder „Collection" lasen. Schüler mit Texten des Typs „Causation" schnitten am schlechtesten ab. Dies hängt möglicherweise damit zusammen, dass Texte des Typs „Comparison" mehr die Aufmerksamkeit lenkende Begriffe enthalten, wie z. B. „ähnlich", „unterschiedlich", „im Gegensatz dazu", etc. Die Propositionen dieses Texttyps scheinen dadurch stärker miteinander verbunden zu sein, was die Erstellung einer mentalen Repräsentation erleichtert. Texte des Typs „Collection" sind zwar ebenso deutlich durch Begriffe wie „erstens", „zweitens", „außerdem", etc. gegliedert, hier sind die Propositionen jedoch einfach aneinander gereiht und nicht wie im ersten Fall dicht miteinander verknüpft. Dies bedeutet eine größere Beanspruchung des Arbeitsgedächtnisses, da mehr einzelne Elemente aktiv gehalten werden müssen. Im Falle der Texte des Typs „Causation" werden kompliziertere Verhältnisse dargestellt, die zwar häufig auch durch Begriffe wie „deshalb" oder „also" eingeleitet werden, insgesamt jedoch weniger deutlich markiert sind. Zum Teil sind Inferenzen notwendig, um die Beziehungen zwischen den Fakten zu erkennen. Yin-Kum (1995) konnte zeigen, dass die besser abschneidenden Testpersonen i. d. R. der originalen Textstruktur bei der Wiedergabe des Textinhalts folgen. Unabhängig vom bearbeiteten Texttyp konnten sich diese Probanden an mehr Inhalte erinnern. Auch Thorndyke (1977) fand heraus, dass ein Text besser behalten wird, wenn der Textaufbau der natürlichen Struktur des Textes entspricht. Texte, deren ursprüngliche inhaltliche Reihenfolge vertauscht wurde, konnten schlechter erinnert werden.

Auch das Bewusstsein für die Textstruktur hat bei den Testpersonen einen Einfluss auf die Behaltensleistung (z. B. Meyer & Freedle, 1984; Richgels et al., 1987; Carrell, 1992). Testper-

sonen, denen die entsprechenden Textschemata bewusst sind, konnten nicht nur insgesamt mehr Informationen wiedergeben, sondern erinnerten sich auch an mehr Leitideen als Testpersonen, die kein Bewusstsein für die Textstruktur hatten. Dies mag damit zusammenhängen, dass die Testpersonen beim Lesen leichter der Textstruktur folgen und wichtige Gedächtniskapazitäten auf die Informationsentnahme lenken können, da sie mit der Textstruktur vertraut sind. Dabei erweist sich jede Art von Vorwissen für die Leser als hilfreich (Meyer & Freedle, 1984). Richgels et al. (1987) führten eine ähnliche Studie mit Schülern der 6. Jahrgangsstufe durch. Die Schüler wurden daraufhin untersucht, wie bewusst ihnen die Textstrukturen „Aneinanderreihung", „Vergleich/Kontrast", „Kausalität" und „Problem/Lösung" sind. Zusätzlich dazu wurde erhoben, wie viele Informationen aus Texten dieser Strukturen wiedergegeben werden konnten. Auch Richgels et al. (1987) fanden über alle Aufgaben hinweg starke Bewusstheit für die Struktur „Vergleich/Kontrast". Im Gegensatz zu Meyer und Freed88 les Ergebnissen fand sich bei Richgels et al. jedoch keine Bewusstheit für die kausale Textstruktur. Ein Grund dafür könnte sein, dass die College-Studenten bei Meyer und Freedle mehr Erfahrung mit kausalen Textstrukturen haben als Schüler der 6. Jahrgangsstufe und sie demnach auch leichter erkennen können.

5.1.3. Thematische Merkmale

Der Grad der inhaltlichen Komplexität eines Stimulus wird häufig an seiner propositionalen Dichte festgemacht. Unter propositionaler Dichte wird die Anzahl an Propositionen dividiert durch die Wortzahl eines Stimulus verstanden. Die propositionale Dichte beeinflusst die Ansprüche, die an die Informationsverarbeitung gestellt werden und damit auch die Itemschwierigkeit. Kintsch und Keenan (1973) konnten zeigen, dass die Anzahl der Propositionen in einem Satz die Lesezeit stärker als die Anzahl der Wörter beeinflusst. Wenn die Anzahl der Wörter konstant gehalten wird, sind Texte mit mehr Propositionen schwieriger. Übergeordnete Propositionen können besser erinnert werden als Propositionen, die strukturell untergeordnet sind.

In einer Studie, bei der Testpersonen Propositionen aus Lese- oder Hörstimuli wiedergeben mussten, stellten Kintsch et al. (1975) keine signifikanten Unterschiede zwischen den Bereichen Hören und Lesen fest. Zwar beeinflusste die Veränderung der Stimuluslänge und die Anzahl der Propositionen die Wiedergebensleistung der Personen, diese war aber unabhängig von der auditiven oder visuellen Präsentation des Stimulus. Die Forschergruppe folgerte daraus, dass beim Lesen und Hören wahrscheinlich sehr ähnliche Prozesse beteiligt sind.

Studien zum Einfluss unterschiedlicher Themenbereich auf die Itemschwierigkeit sind relativ selten. Ein Ergebnis liegt dazu jedoch beispielsweise von Freedle und Kostin (1996) vor. Die Autoren zeigten, dass nicht-akademische Themen aufgrund ihrer größeren allgemeinen Vertrautheit einfacher sind als akademische Themen.

5.1.4. Präsentationsmerkmale

Zu den akustischen Eigenschaften der Hörstimuli gehören beispielsweise die Tonqualität der Aufnahme, die Lautstärke, mit der die Stimuli vorgespielt werden, aber auch die Rahmenbedingungen der Testsituation, wie die Lautstärke im Testraum. Mattys et al. (2009: 203) unter-

scheiden bezüglich beeinträchtigender Rahmenbedingungen zwischen „energetic masking" und „informational masking". Beim Energetic Masking führen die Rahmenbedingungen (z. B. Hintergrundgeräusche) dazu, dass der Lautfluss nur gestört wahrgenommen wird und dass lexikalisch-semantisches Wissen in den Hintergrund tritt, denn die Testpersonen achten verstärkt auf herausstechende akustische Details. Beim Informational Masking werden aufgrund der Rahmenbedingungen (z. B. aufgrund von zwei simultan ablaufenden Tätigkeiten) Verarbeitungsressourcen unabhängig von der Qualität des Lautflusses blockiert. Die Testpersonen achten in diesen Fällen bei der Dekodierung verstärkt auf lexikalisch-semantische Informationen und weniger auf akustische Details.

Zu den akustischen Merkmalen der Stimuli zählt auch die Sprechgeschwindigkeit. Sie wird definiert als die Anzahl der Wörter, die pro Sekunde geäußert wird. Flowerdew und Miller (1992) fanden für L2 Studenten, dass eine Steigerung der Sprechgeschwindigkeit die Itemschwierigkeit für die Studenten erhöhte. Auch Griffiths (1990; 1991) erhielt in seinen Studien mit semi-naturwissenschaftlichen Texten und Geschichten analoge Befunde. Die Testpersonen scheinen also Schwierigkeiten im Umgang mit schnell gesprochener Sprache zu haben. Sehr schnell nacheinander eintreffende Informationen erfordern eine hohe Aufmerksamkeitsleistung der Testpersonen. Wenn die Aufmerksamkeit an einer Stelle nachlässt, können auch nachfolgende Informationen manchmal nicht mehr richtig eingeordnet werden. Gerade schlechtere Schüler, die sich auf bottom-up-Strategien verlassen, können sehr schnell gesprochene Stimuli nur schlecht verarbeiten.

Auch die Anzahl der Stimuluspräsentationen kann einen Einfluss auf die Schwierigkeit einer Aufgabe haben. Untersuchungen, ob die Anzahl der Stimuluspräsentationen das Testergebnis verbessert, wurden vor allem im fremdsprachlichen Bereich durchgeführt. Mit der kommunikativen Wende erfolgte auch der Umschwung von mehrmaliger Stimuluspräsentation hin zum einmaligen Vorspielen des Stimulus. Von den Befürwortern des einmaligen Vorspielens wird meist mit der größeren Authentizität dieses Verfahrens argumentiert, da die identische Wiederholung eines Beitrags außerhalb der Unterrichts- und Testsituation selten vorkommt. In der Realität bietet sich meist nur die Gelegenheit zum einmaligen Anhören. Aus diesem Grund sollten derartige Stimuli auch im Hörverstehenstest nur einmal vorgespielt werden. So soll nicht nur der authentischen Sprachverwendungssituation Rechnung getragen werden, sondern auch die Hörverstehenskompetenz unter Alltagsbedingungen beobachtet werden, zu denen eben auch unvorhersehbare Störungen (wie Husten, ein vorbeifahrendes Auto etc.) gehören. Buck (2001) stellt zur Überlegung, ob nicht gerade die Fähigkeit, aufgrund der Redundanz des auditiven Inputs Inferenzen zu ziehen, einen wesentlichen Teil der Hörverstehenskompetenz ausmacht. Er weist darauf hin, dass Hörverstehen ein automatischer Prozess ist und bislang unklar ist, welche zusätzlichen Faktoren beim mehrmaligen Hören aktiviert werden. Aus diesem Grund präferiert Buck eine einmalige Stimuluspräsentation (Buck, 2001).

Im Sinne der Testvalidität ist es wünschenswert, eine große Anzahl an Aufgaben im Test zur Messung eines Konstrukts zu haben. Wird ein Stimulus nur einmal vorgespielt, wird wertvolle Testzeit gewonnen, die mit weiteren Aufgaben gefüllt werden kann. Mehrmaliges Vorspielen dient jedoch zur Erhöhung der Testfairness, da in der Testsituation Störungen kompensiert

werden können. Wird der Input in der Testsituation durch Geräusche beeinträchtigt oder unterbrochen, haben die Testpersonen beim zweiten Hören die Chance, die versäumten Stellen doch noch aufzunehmen. Gerade bei einem Test, der als Large-Scale-Assessment durchgeführt wird, sollte im Sinn der Testreliabilität gewährleistet sein, dass alle Teilnehmer die gleichen Bedingungen bei der Durchführung haben. In diesem Sinne ist auch dafür zu sorgen, dass die Testsituation den Teilnehmern erlaubt, ihre Fähigkeiten optimal zu zeigen. Zu wissen, dass sie den Stimulus nur einmal hören werden, kann ein großer Stressfaktor sein und die Testleistung beeinträchtigen. Allerdings gibt es für diese Vermutung bislang keine empirischen Befunde. Selbst bei größtmöglichem Bemühen um authentische Aufgaben und optimale Testbedingungen wird die Testsituation leider niemals einer authentischen Anwendungssituation gleichen. Die Fragen zu einem Input zielen i. d. R. auf einen viel intensiveren und weiter reichenden Umgang damit ab, als sich die Testpersonen außerhalb des Tests mit dem Stimulus beschäftigt hätten. Es läge nahe, ihnen deshalb die Möglichkeit zu bieten, sich auch intensiver mit dem Stimulus auseinanderzusetzen, als sie dies außerhalb der Testsituation würden. Aber geht eine mehrmalige Stimuluspräsentation auch mit höheren Testleistungen einher?

Studien von Henning (1991) sowie Brindley und Slatyer (2002) brachten keine Hinweise darauf, dass es für die Qualität des Tests oder der Testergebnisse sinnvoll wäre, den Stimulus mehrmals zu präsentieren. Henning (1991) untersuchte im Rahmen des TOEFL fünf Variablen danach, ob sie die Diskriminanz zwischen guten und schlechteren Lernenden beeinflussen. Dabei zeigte sich, dass eine mehrfache Stimuluspräsentation für die Diskriminanz der Aufgaben keine Rolle spielte. Brindley und Slatyer (2002) analysierten mehrere Variablen bei einer Studie mit erwachsenen ESL Lernern. Sie konnten im Gegensatz zu Sprechgeschwindigkeit und Testformat bei der Häufigkeitsvariation der Stimulusdarbietung keinen Effekt auf die Aufgabenschwierigkeit finden.

Dagegen berichtet Lund (1991), dass erwachsene Lerner der deutschen Sprache als Zweitsprache ihre Leistungen bei der Wiedergabe von Propositionen und Wortschatz durch ein wiederholtes Anhören des Stimulusmaterials verbessern konnten. Je besser die Lerner waren, desto offensichtlicher wurde der Ergebniszuwachs. Bei Dupuy (1999) profitierten vor allem Lerner mit sehr geringen Sprachfähigkeiten von einer wiederholten Stimuluspräsentation und steigerten dadurch ihre Leistungen um ca. 30%. Für Lerner auf einem mittleren Fähigkeitsniveau lag die Leistungssteigerung bei nur noch ca. 10%. Cervantes und Gainer (1992) arbeiteten mit erwachsenen Testpersonen, die Englisch als Fremdsprache lernen. Die Testpersonen schnitten im Lückendiktat besser ab, wenn sie den Stimulus ein zweites Mal anhören konnten. Eine Vergleichsgruppe bearbeitete eine syntaktisch modifizierte leichtere Version des Diktats und schnitt dabei nicht signifikant besser oder schlechter ab. Cervantes und Gainer argumentieren auf dieser Grundlage für den Einsatz authentischer Stimuli, welche zur Schwierigkeitsreduktion wiederholt vorgespielt statt, sprachlich vereinfacht werden sollten.

Die Ergebnisse von Lund (1991) sowie Cervantes und Gainer (1992) sind jedoch vorsichtig zu interpretieren, da von der Wiedergabe von Formulierungen und Ausdrücken nur sehr bedingt auf allgemeine Hörverstehenskompetenz geschlossen werden kann. Die Verarbeitung von auditivem Input erfolgt ja auf einer semantischen Ebene, gespeichert wird also in erster Linie

der Inhalt einer Botschaft und nicht die genaue Formulierung derselben. Es scheint nahe liegend, dass in der Studie von Lund isolierte Begriffe oder Formulierungen besser erinnert werden, wenn die Gelegenheit besteht, sie mehrfach zu hören. In diesem Fall handelt es sich jedoch weniger um Hörverstehen als um eine Gedächtnisleistung. Auch in der Studie von Cervantes und Gainer wird weniger Hörverstehen als Worterkennung geprüft. Tatsächlich müssen die Testpersonen nicht die Bedeutung der Lückenwörter oder den Inhalt des Stimulus kennen, um in dieser Aufgabe erfolgreich abzuschneiden. Relevant ist für ein gutes Testergebnis lediglich die korrekte Schreibweise des Wortes. Es verwundert nicht, dass das mehrmalige Hören der Begriffe für die Überprüfung, Korrektur und richtige Schreibung hilfreich ist.

Bislang ist nicht bekannt, ob die Hörverstehensprozesse beim zweiten Hören die gleichen sind wie beim erstmaligen Hören. Die Vermutung liegt nahe, dass die Prozesse nicht identisch sind und unterschiedliche Verstehensstrategien bemüht werden. Hörverstehen ist durch das automatische Verarbeiten der sprachlichen Informationen gekennzeichnet. Es ist dann notwendig, den Stimulus mehrmals vorzuspielen, wenn er z. B. mangels sprachlicher Fähigkeiten eben nicht automatisch verarbeitet wurde. Das zweite Hören ermöglicht den Einsatz von Strategien, um diesen Mangel zu kompensieren. Wird der Stimulus ein zweites Mal dargeboten, so ist davon auszugehen, dass bei der Bearbeitung der Items in stärkerem Maß als beim ersten Hören Strategien eine Rolle spielen.

5.1.5. Subjektive Einschätzungen der Stimuli

Die Lesbarkeitsforschung geht davon aus, dass Texte im Idealfall so geschrieben sind, dass die anvisierte Zielgruppe ohne Schwierigkeiten ihre beabsichtigte Bedeutung erfassen kann. Methoden zur Bestimmung der Schwierigkeit sollen eine optimale Passung von Text und Zielgruppe gewährleisten, wobei Verstehen mit der Lesbarkeit gleichgesetzt wird. Die Lesbarkeitsforschung basiert auf quantitativ auswertbaren und objektiv feststellbaren Textmerkmalen. Diese werden statistisch zueinander in Beziehung gesetzt und mit bestimmten Lesbarkeitsformeln wird die stilistische Textschwierigkeit errechnet. Die Vorhersagen über die Lesbarkeit müssen über eine Reihe von Textarten und Schwierigkeiten möglich sein, ohne die Texte dafür zu lesen oder Versuchspersonen befragen zu müssen. Dabei spielen persönliche Faktoren der Rezipienten, wie Zeit, Ort, Intelligenz oder Motivation keine Rolle. Auch inhaltliche Aspekte wie Anschaulichkeit, Einfachheit oder strukturelle Faktoren, wie Gliederung und Übersichtlichkeit werden nicht berücksichtigt. (vgl. Lehrndorfer, 1996)

Den Lesbarkeitsindizes wird vorgeworfen, höchstens Texte mit komplexer Syntax oder schwierigem Wortschatz zu identifizieren, jedoch i. d. R. keine Hinweise darauf zu geben, wie schwierig ein Text wirklich ist. Kurze Wörter und einfache Sätze sind letztendlich keine Garantie für Textverständlichkeit. Dementsprechend wurden die herkömmlichen Lesbarkeitsformeln von einigen Forschergruppen um weitere, subjektivere Merkmale ergänzt. Beispielsweise ließ die Hamburger Forschergruppe um Langer, Schulz von Thun und Tausch (vgl. Langer et al., 1974) Texte zusätzlich zu den objektiven Merkmalen der Lesbarkeitsforschung auf einer siebenstufigen Likert-Skala nach unterschiedlichen Kriterien einschätzen. Anschließend reduzierten sie die Dimensionen mittels Faktorenanalysen auf vier, und zwar „Einfachheit – Kompliziertheit", „Gliederung/Ordnung – Ungegliedertheit/Zusammenhang-losigkeit", „Kürze/Prägnanz – Weit-

schweifigkeit“ und „zusätzliche Stimulanz – keine zusätzliche Stimulanz“. Mit der letzten Dimension werden im weitesten Sinn auch motivationale Aspekte erfasst, da diese Dimension für Merkmale wie „anregend“, „interessant“ und „abwechslungsreich“ steht.

Auch Carroll (1964) verwendete objektive Maße, wie Auszählungen oder Verhältnisse sowie subjektive Maße. Insgesamt arbeitete er mit 29 adjektivischen Skalen und acht Ratern. Jede Textpassage wurde mehrfach eingeschätzt. Die Reliabilität der subjektiven Skalen war dann hoch, wenn das zu bewertende Kriterium relativ objektiv (z. B. humoristisch) war. Wurden jedoch persönliche Bewertungen der Rater verlangt (z. B. angenehm – unangenehm), sank die Reliabilität der subjektiven Maße. Carroll isolierte sieben Faktoren: General Stylistic Evaluation, Personal Affect, Ornamentation, Abstractness, Seriousness, Characterization und General Stylistic Evaluation. Für den Faktor General Stylistic Evaluation konnte er keine signifikante Korrelation mit irgendeinem objektiven Faktor finden, was er als Indiz dafür betrachtete, dass ein Globalurteil nicht maschinell vergeben werden kann.

5.2. Itemmerkmale

5.2.1. Itemformat

In Bezug auf Testfairness und -validität liegt großes Interesse beim Einfluss unterschiedlicher Itemformate auf die Itemschwierigkeit. In der Regel geht man davon aus, dass viele unterschiedliche Itemformate den Einfluss minimieren, den ein bestimmtes Format auf die Itemschwierigkeit haben kann. Eine Vielzahl unterschiedlicher Itemformate hat außerdem den Vorteil, dass damit unterschiedliche Aspekte der Schülerleistung erhoben werden und so die Konstrukt-Validität des Tests erhöht wird. (In’nami, 2006)

Jedes Item besteht in der Regel aus zwei Teilen. Im ersten Teil werden bestimmte Informationen gegeben, der zweite Teil verlangt bestimmte Informationen von der Testperson. Bei der Bearbeitung eines Items wird die gegebene Information dazu verwendet, die gesuchte Information zu finden. Häufig sind dafür Inferenzen notwendig. Bei den IQB-Items wird zwischen geschlossenen, halb-offenen und offenen Itemformaten unterschieden.

Alle geschlossenen Itemformate werden dichotom kodiert. Bei den Richtig-Falsch-Items müssen die Schüler zwischen einer richtigen und einer falschen Antwortalternative wählen. Wesentliche Vorteile liegen in der relativ einfachen Entwicklung und Auswertbarkeit sowie der kurzen Zeit zur Bearbeitung. Problematisch hingegen ist die Notwendigkeit, das zu testende Thema auf eine Entscheidungsfrage zu reduzieren, die eindeutig beantwortet werden kann. Daher wird dieses Itemformat vorrangig dann verwendet, wenn es gilt, Faktenwissen wie Namen, Daten oder Definitionen abzuprüfen. Ein zweiter Nachteil dieses Itemformats ist die hohe Wahrscheinlichkeit einer zufällig richtigen Lösung. Die Chance, ein Item ohne Wissen allein durch Raten richtig zu lösen, liegt hier bei 50%. Aus diesem Grund werden bei Richtig-Falsch-Items meist mehrere Einzelfragen oder -aussagen zu einem Item zusammengefasst. Das Item gilt nur dann als vollständig richtig gelöst, wenn eine Mindestanzahl der Einzelfragen korrekt beantwortet wurde. In diesem Zusammenhang kann auch erwogen werden, eine Aufgabe als teilweise richtig gelöst zu bewerten, sobald ein bestimmter Prozentsatz der

Einzelfragen korrekt bearbeitet worden ist. Auch Multiple-Choice-Items zählen zu den geschlossenen Itemformaten. Auf das Format Multiple-Choice wird im nächsten Kapitel genauer eingegangen.

Weitere geschlossene Itemformate sind „Zuordnung" und „Umordnung". Das Zuordnungs- Format besteht aus zwei Listen von Begriffen oder Aussagen, die einander zugeordnet werden müssen (z. B. sind einer Liste mit Autoren die entsprechenden Bücher zuzuordnen). Umordnungsitems verlangen, dass einzelne, ungeordnet aufgelistete Elemente nach bestimmten Kriterien in eine Reihenfolge gebracht werden. Bei den Elementen kann es sich um Buchstaben, Wörter, Wortgruppen oder ganze Abschnitte handeln.

Halboffene Items werden komplexer ausgewertet und unterschiedliche Schülerlösungen können als richtig akzeptiert werden. Bei Kurzantwort-Items werden keine Antwortalternativen zur Auswahl vorgegeben. Sie verlangen eine eigenständig formulierte kurze Antwort, die auch nur aus einem Wort oder wenigen Wörtern bzw. Sätzen bestehen kann. Häufig werden halboffene Items auch als Lückentext angeboten. Der Begriff Lückentext-Aufgabe ist ein Sammelbegriff für eine Vielzahl von Aufgabentypen. Die gemeinsame Grundlage ist die Verwendung eines Textes, aus dem Wörter gelöscht werden, die vom Probanden zu rekonstruieren sind. Die lückenhaften Sätze werden mit der Aufforderung präsentiert, die fehlenden Wörter zu ergänzen. Dabei kann es notwendig sein, ein Kriterium für richtige Lösungen anzugeben. Dieses Vorgehen eignet sich dazu, das Verständnis des Textzusammenhangs zu messen. Im Idealfall gibt es für jede Lücke nur eine richtige Lösung. Um die Anzahl der richtigen Antworten zu beschränken, können potentielle Lösungen vorgegeben werden, aus denen die jeweils richtige ausgewählt werden muss. Damit wechselt man vom halboffenen zum geschlossenen Aufgabenformat. Distraktoren für eine Lücke dürfen nicht mit dem Attraktor einer anderen Lücke übereinstimmen.

Items mit offenem Antwortformat erlauben den Schülern, auf ein bestimmtes Item frei zu antworten. Für die Kodierung sind klare Lösungshinweise notwendig, um richtige von falschen Antworten zu unterscheiden. Offene Schreibaufgaben lassen sich nach „Freiheitsgraden" unterscheiden, die oft mit den erwarteten Textumfängen korrespondieren. Dabei sind offene Itemformate i. d. R. sehr aufwändig auszuwerten und die Beurteilerreliabilität ist häufig nur gering ausgeprägt. Aus didaktischer Perspektive wird offenen Itemformaten häufig ein höherer Wert zugesprochen. So hält beispielsweise Nitko (2004: 181) offene Items für besonders wertvoll, da die Schüler in ihnen dazu angeleitet werden können, ihre Antworten auch zu begründen, was Rückschlüsse auf höhere Denkleistungen ermöglicht.

Zur Rolle des Itemformats in Hinblick auf die Itemschwierigkeit gibt es mehrere Untersuchungen: Die Studien von Tuinman (1973-74) und Katz et al. (1990) legen nahe, dass der Stimulus keine Rolle für die Bearbeitung der Items spielt. Die Aufgabenschwierigkeit wird stattdessen allein von der Gestalt der Items bestimmt. Katz et al. (1990) konnten zeigen, dass Textstruktur und Textinhalt mit der Lösungshäufigkeit nicht korrelieren, wohingegen Itemstruktur und Iteminhalt eine wichtige Funktion in Bezug auf die Itemschwierigkeit haben. Zu diesem Ergebnis kommen auch Drum et al. (1981). Sie sind der Ansicht, dass Item- Variablen die wichtigsten Prädiktoren bezüglich der Schwierigkeit von Leseverstehensaufgaben sind.

5.2.2. Multiple-Choice-Items

Die vielleicht bekannteste Variante des geschlossenen Formats stellt das Multiple-Choice (MC) Item dar. Ein Multiple-Choice-Item besteht aus einem Itemstamm und i. d. R. aus genau einem Attraktor und drei Distraktoren. Die Beschränkung auf vier Antwortoptionen liegt darin begründet, dass damit eine maximale Ratewahrscheinlichkeit von 25% gegeben ist. Ein Schüler, der die richtige Antwort nicht erkennt, kann diese also höchstens mit einer 25%-igen Wahrscheinlichkeit erraten. Diese Genauigkeit genügt, um wissende von ratenden Schülern unterscheiden zu können. Die Distraktoren erfüllen einige, jedoch nicht alle durch die Frage vorgegebenen Suchkriterien oder sie scheinen die Frage zu beantworten, sind jedoch falsch.

Das MC-Format wird im Folgenden genauer dargestellt, da es die Grundlage für zahlreiche internationale Begleitforschung darstellt. Vor allem im Bereich des Leseverstehens wurden MC-Aufgaben gründlich untersucht. Das Multiple-Choice-Format nimmt gerade im Large-Scale-Assessment einen wichtigen Platz ein, da es kosten- und zeitsparend eingesetzt werden kann und durch die Möglichkeit der elektronischen Auswertung leichter reliable Ergebnisse zu erzielen sind als bei halboffenen oder offenen Itemformaten. Auch bei den Testpersonen scheinen Multiple-Choice-Items beliebt zu sein. Die Missing-Anteile bei dem Format Multiple- Choice liegen in jedem Fall deutlich unter denen offener Formate. In der Studie von Hollingworth et al. (2007) lagen die Missing-Anteile im Leseverstehenstest bei den offenen Items bei über 30%, wohingegen sie bei den Multiple-Choice-Items lediglich 1% betrugen. Ferner wird häufig als Vorteil von Multiple-Choice-Items betrachtet, dass sie konfundierende Effekte, beispielsweise durch verlangte Schreibleistungen, weitgehend ausschließen. (Martinez, 1991)

Embretson und Wetzel (1987) zeigten mit ihrem Informationsverarbeitungsmodell für MC-Items zum Leseverstehen, dass die Entscheidungsprozesse, die für die Wahl der richtigen Antwortoption notwendig sind, die Itemschwierigkeit in stärkerem Maße beeinflussen als die Prozesse zum Aufbau einer mentalen Repräsentation des Textes.

Aber auch spezielle Einzelaspekte des Formats Multiple-Choice wurden untersucht. Beispielsweise wurde geprüft, ob die Position des Attraktors im Item oder die Plausibilität der Distraktoren einen Einfluss auf die Itemschwierigkeit haben. Bei Golub-Smith (1987) beeinflusste die Veränderung der Attraktor-Position bei Multiple-Choice-Items die IRT-Parameter der Hörverstehensitems, wenn alle anderen Variablen konstant gehalten wurden. Die Ursache dafür könnte darin liegen, dass die Testpersonen alle anderen Optionen nicht mehr lesen müssen, wenn der Attraktor an erster Stelle kommt und direkt von den Testpersonen als korrekt identifiziert wird. Dies spart Zeit, da die Testpersonen direkt zum nächsten Item gehen können, und entlastet das Arbeitsgedächtnis, da die verschiedenen Optionen nicht mehr miteinander abgeglichen werden müssen. Nissan et al. (1996) fand für die Variable „Position of Correct Answer" keine signifikanten Korrelationen mit der Itemschwierigkeit, wobei hier die anderen Variablen nicht konstant gehalten wurden.

Je mehr ein Distraktor mit den gegebenen oder den gesuchten Informationen gemeinsam hat, desto plausibler ist er. (Evetts & Gauthier, 2005) Items sollten umso einfacher sein, je deutlicher die Distraktoren als falsch zu identifizieren sind. Entsprechend konnten Embretson und

Wetzel (1987) eine signifikante Korrelation zwischen Itemschwierigkeit und „Falsifikation" nachweisen. Das Maß für die Falsifikation wurde von Perkins und Brutten (1993) übernommen, sie fanden aber keine signifikanten Korrelationen. Auch in den Analysen von Gorin und Embretson (2006) zeigte diese Variable keine signifikanten Effekte. Davey (1988) arbeitete mit einer ähnlichen Variable, der „Plausibilität der Distraktoren". Je plausibler die Distraktoren erscheinen, desto schwieriger ist ein MC-Item zu lösen. Aber auch Davey konnte keine signifikanten Zusammenhänge mit der Itemschwierigkeit nachweisen.

Die Kritik am MC-Format besteht darin, dass die Items wenig valide seien und sich für die Überprüfung von Textverstehen kaum eigneten, da sie keine natürlichen Lese- oder Zuhörsituationen widerspiegeln. (Nissan et al., 1996; Buck, 2001) Nach Hambleton und Murphy (1992) förderten MC-Items die Erwartung einer einzigen richtigen Lösung, sie wirken einengend bei der Umsetzung von Lehrplaninhalten in Testitems, fokussieren nur auf diskreten Fähigkeiten und werden insbesondere schlechteren Schülern nicht gerecht. Auch Valencia und Pearson (1988) halten das MC-Format nur bedingt dazu geeignet, Verstehenskompetenzen zu überprüfen. Das Format zwinge dazu, eine Antwort aus vorgegebenen Alternativen zu wählen, wenn es vielleicht andere Lösungsoptionen gegeben hätte. Aus diesem Grund wird sogar angenommen, dass durch das Format Multiple-Choice andere Konstrukte als durch (halb-)offene Formate erfasst würden. Diese Überlegungen bestätigten sich in empirischen Untersuchungen zum Itemformat jedoch nicht (z. B. Bridgeman, 1992), allerdings weisen beispielsweise Birenbaum und Tatsuoka (1987) darauf hin, dass offene Itemformate gegenüber dem Format Multiple-Choice viele zusätzliche diagnostische Informationen beinhalten.

An Multiple-Choice-Aufgaben zum Testen von Textverständnis wird auch kritisiert, dass die Testpersonen den Stimulus nicht lesen bzw. anhören müssen, um die Items zu beantworten. Katz et al. (1990) wiesen nach, dass Schüler Multiple-Choice-Items im Scholastic Assessment Test (SAT), einem Test zur Studienplatzvergabe an amerikanischen Universitäten, mit über der Ratewahrscheinlichkeit liegender Häufigkeit lösen konnten, ohne die dazugehörigen Texte gelesen zu haben. Die Lösungshäufigkeit steigt jedoch signifikant, wenn die Testpersonen den Stimulus zur Bearbeitung der Aufgabe hinzuziehen dürfen, auch wenn ein Teil der Items mit über der Ratewahrscheinlichkeit liegender Lösungshäufigkeit ohne den Stimulus gelöst werden konnte (Katz et al., 1990).

Auch nach Royer (1990) eignen sich Multiple-Choice-Aufgaben nicht dazu, Textverständnis zu überprüfen, da sie andere Fähigkeiten der Testpersonen testen, wie Weltwissen und die Fähigkeit, Schlussfolgerungen zu ziehen. Eine starke Korrelation einiger Textvariablen mit der Itemschwierigkeit weist jedoch darauf hin, dass die Testpersonen durchaus Textinformationen bei ihrer Auswahl der entsprechenden MC-Option berücksichtigen. Würden Multiple- Choice-Items, wie von Royer (1990) angenommen, tatsächlich überwiegend Weltwissen und die Fähigkeit, zu Inferieren testen, so dürfte die Präsentation des Stimulus die Lösungshäufigkeit der Items nicht signifikant beeinflussen.

Ferner wird kritisiert, dass mit dem MC-Format überwiegend reproduktive Leistungen und kaum produktive Fähigkeiten wie schlussfolgerndes Denken, komplexe Operationen und Pro-

blemlösen erfasst werden. Da die Wiedererkennensleistung normalerweise höher als die Reproduktionsleistung ist, wird angenommen, dass Multiple-Choice-Aufgaben tendenziell einfacher sind als offene Itemformate. So berichtet Shohamy (1984), dass Leseverstehensitems im MC-Format einfacher sind als halboffene oder offene Items. Yi'an (1998) fand in 95 seiner Studie mit chinesischen EFL Lernern durch Think-aloud-protocols heraus, dass das MC-Format die Informationsverarbeitung für fortgeschrittene Studenten erleichterte. Die Antwortoptionen wirkten fokussierend, da viele Studenten die Stimuli betreffende Erwartungen aufbauen konnten. Schwächere Studenten profitierten jedoch nicht vom MC-Format. Sie wurden durch die Distraktoren vom Attraktor abgelenkt oder wählten den Attraktor aus falschen Gründen.

5.2.3. Zeitpunkt der Itembearbeitung

Bei der Frage, zu welchem Zeitpunkt des Tests die Probanden die Items sehen dürfen, werden meist drei verschiedene Modelle untersucht: Die Items werden vor oder nach dem Hören des Stimulus oder zwischen zwei Hördurchgängen präsentiert. Diskutiert wird, ob die Kenntnis der Items die Aufmerksamkeit für relevante Informationen im Stimulus steuert bzw. ob sie die Verarbeitungsprozesse vielleicht dadurch behindert, dass zusätzliche Informationen im Arbeitsgedächtnis behalten werden müssen. Ferner könnte es sein, dass Aufgaben, die Informationen auf Satz- oder Wortebene verlangen, die Aufnahme des Stimulus als Ganzen beeinträchtigen. (Sherman, 1997) Dieser Effekt könnte gerade für das Hörverstehen relevant sein, da top-down Verarbeitungsstrategien hier durch die Flüchtigkeit der gesprochenen Sprache besonders wichtig sind (Lund, 1991). Shohamy und Inbar (1991) vermuten, dass die Kenntnis der Items Zuhörstrategien fördert, die mit dem Wissen des Zuhörers verbunden sind und weniger vom Stimulus gesteuert werden. Bacon (1992) wies dagegen nach, dass derartige top-down Strategien besonders bei schwierigerem Audio-Input kaum eine Rolle spielten, sondern sich die Testpersonen dann eher auf bottom-up Strategien verließen.

Gernsbacher (1994) konnte zeigen, dass Versuchspersonen Stimuli nur unvollständig verarbeiten, wenn ihnen keine spezifische Aufgabenstellung für die Verarbeitung genannt wird. Dies spricht dafür, den Testpersonen die Items vor dem Hören des Stimulus zu präsentieren. Sherman (1997) untersuchte, ob der Zeitpunkt der Itempräsentation bei zweimaligem Vorspielen des Stimulus die Ergebnisse der Testpersonen beeinflusst. Signifikant bessere Ergebnisse ergaben sich nur, wenn die Fragen zwischen der ersten und zweiten Stimuluspräsentation vorgelegt wurden. Allerdings bevorzugten die Testkandidaten nach Angaben in einem Zusatzfragebogen deutlich eine Präsentation der Items vor dem Hören, auch wenn dies keinen Einfluss auf die Testergebnisse hatte. Dabei sprachen sich leistungsschwächere Schüler stärker als leistungsstärkere Schüler für die Präsentation der Items vor dem Hören aus. Sherman räumt die Möglichkeit ein, dass die Kenntnis der Items vor dem Hören hilfreiche Informationen liefern kann. Gleichzeitig befürchtet sie, dass die Lernenden ihre Konzentration nicht auf die relevanten Informationen fokussieren, sondern sich vielmehr unter Druck setzen, die richtigen Stellen nicht zu verpassen und letztendlich den Stimulus weniger tief aufnehmen.

Field (2003) zeigte die Items vor der Stimuluspräsentation und ließ sie während des Hörens bearbeiten. Er kommt zu dem Schluss, dass eine Bearbeitung der Items während des Hörens für die Lernenden schwierig sein kann. I. d. R. werden die Items in der Reihenfolge darge-

boten, in der die Informationen auch im Stimulus erscheinen. Dies kann dazu führen, dass die Lernenden eine verpasste Information zu einem Item noch immer erwarten, obwohl sie bereits die nachfolgenden Items bearbeiten sollten. Eine besondere Herausforderung stellt die Bearbeitung der Items während des Hörens auch aufgrund der notwendigen Koordination mehrerer Tätigkeiten dar. So müssen die Testpersonen gleichzeitig zuhören, die Informationen des Stimulus aufnehmen und verstehen, die Items lesen und verstehen, die zur Bearbeitung der Items relevanten Informationen aus dem Stimulus finden, die Antworten formulieren und nicht zuletzt schreiben (Sherman, 1997).

In einer Studie mit japanischen Englisch-Lernenden untersuchten Yanagawa und Green (2008), ob es zu Unterschieden in der Itemschwierigkeit führt, wenn bei Multiple-Choice-Items vor dem Hören das ganze Item (Frage und Antwortoptionen), nur die Frage oder ob nur die Antwortoptionen gezeigt werden. Dabei ergaben sich keine signifikanten Unterschiede in den Testergebnissen, wenn entweder das ganze Item oder nur die Frage gezeigt werden. Allerdings sind Items, bei denen im Vorfeld die Antwortoptionen gesehen werden dürfen, signifikant schwieriger. Dies mag damit zusammenhängen, dass die Zuhörer versuchen, das Gehörte lexikalisch auf die Itemoptionen zu beziehen.

Im Rahmen der Bemühungen, die Testsituation der authentischen Sprachverwendungssituation anzupassen, ist für eine Präsentation der Items vor dem Hören zu plädieren. In einer authentischen Hörsituation ist die Hörabsicht definiert und es ist davon auszugehen, dass die Testpersonen auch im Test aufmerksamer zuhören, wenn sie den Grund für das Zuhören kennen. Fragwürdig ist an dieser Stelle, ob die Fragestellungen der Items ein adäquates Äquivalent für authentische Hörabsichten sind, oder ob nicht das Bestreben, gut im Test abzuschneiden, die tragende Hörabsicht darstellt.

5.2.4. Überlappung der Item-Formulierungen mit dem Stimulus

Freedle und Fellbaum (1987) fanden heraus, dass Multiple-Choice-Items im TOEFL Hörverstehenstest wesentlich von der Überlappung von Formulierungen aus dem Stimulus in Items/Auswahloptionen in ihrer Schwierigkeit beeinflusst werden. Die Testpersonen entscheiden sich i. d. R. für die Auswahlmöglichkeit, welche der Formulierung im Stimulus am Ähnlichsten ist. Meist ist dies auch die korrekte Antwort. Weitere Variablen zur Überlappung der Formulierungen, die auch in den Studien zum Leseverstehen eingesetzt wurden (Freedle & Kostin, 1993a), zeigten signifikant stärkere Ausprägungen für das Lesen. Dies könnte daran liegen, dass es einfacher ist, gleiche Wortgruppen in einem Lesetext zu identifizieren, als sie im Falle der Zuhöraufgaben im Gedächtnis zu behalten und zu erinnern. Die Überlappung der Formulierung bewirkt in jedem Fall eine Veränderung in der Aufgabenschwierigkeit, die zugrunde liegenden Prozesse unterscheiden sich jedoch für das Hör- und das Leseverstehen.

Freedle und Kostin (1996) führen eine spezielle Variable für das Hörverstehen ein, welche die Attraktivität der korrekten Option erfasst. Kodiert wird die Attraktivität des Attraktors im Vergleich zum am häufigsten gewählten Distraktor. Dabei wird einerseits die Anzahl der überlappenden Inhaltswörter bei Attraktor und Stimulus im Vergleich zu ihrer Anzahl bei dem am häufigsten gewählten Distraktor und dem Stimulus erhoben. Außerdem wurde die Anzahl des

am häufigsten auftretenden Inhaltsworts in Attraktor und Stimulus sowie im beliebtesten Distraktor und Stimulus ermittelt. Je attraktiver ein Attraktor im Vergleich zu einem Distraktor erscheint, desto einfacher wird das Item. Je höher also die Anzahl der überlappenden Inhaltswörter bei Attraktor und Stimulus und je häufiger die Inhaltswörter des Attraktors im Stimulus erscheinen, desto stärker sinkt die Itemschwierigkeit.

5.2.5. Die zur Beantwortung eines Items notwendige Information (NI)

Buck und Tatsuoka verstehen unter dem Begriff „necessary information" (NI) „the information in the text which the reader must understand to be certain of the correct answer." (1998: 134) Der dahinterliegende Gedanke ist, dass die Verständlichkeit der Information, die speziell zur Beantwortung eines Items notwendig ist, einen entscheidenden Einfluss auf die Schwierigkeit des Items hat. Buck und Tatsuoka (1998) konnten zeigen, dass tatsächlich Merkmale der NI den größten Einfluss auf die Itemschwierigkeit haben, gefolgt von Merkmalen des die NI umgebenden Textes. Merkmale des gesamten Textes hatten einen vergleichsweise geringen Einfluss auf die Itemschwierigkeit.

5.2.5.1. Art der NI

Die Art der vom Item geforderten Informationsentnahme stellt unterschiedliche kognitive Anforderungen an die Testpersonen. Bei deskriptiven Informationsfragen wird eher das Verständnis auf der Ebene der Textbasis gemessen, wohingegen kausale Inferenzfragen das Vorhandensein eines kohärenten mentalen Modells überprüfen. (vgl. Grotjahn, 2000) Bei der Beantwortung eines Items muss die im Item gegebene Information von den noch benötigten Informationen getrennt werden und mithilfe der gegebenen Informationen müssen die zur Beantwortung des Items benötigten Informationen im Stimulus gefunden werden. Diese Informationen müssen z. B. im Fall von Multiple-Choice-Items im Attraktor wiedererkannt werden, im Fall von offenen Items reproduziert oder inferiert werden. Beim Wiedererkennen wird eine Information als vertraut identifiziert, wenn sie nach dem Lernen erneut gesehen oder gehört wird. (vgl. Hayes, 1995: 31) Dieser Vorgang wird als leichter eingeschätzt als Reproduktion. Wirken jedoch kontextuelle Faktoren unterstützend, kann die Reproduktionsrate auch höher als die Wiedererkennungsrate liegen (vgl. Anderson, 2001: 231).

Um aus unvollständigen Informationen ein mentales Modell zu bilden, müssen häufig auch Schlussfolgerungen bzw. Inferenzen gezogen werden, denn „(Listening) Comprehension is the process of relating language to concepts in one's memory and to references in the real world." (Rost, 2002: 59) Crothers (1979) unterscheidet zwei grundlegende Arten der Schlussfolgerung: Zum einen werden fehlende Elemente in der Oberflächenstruktur des Stimulus ergänzt, um Kohärenz zwischen neuen Informationen und Bekanntem herzustellen. Dazu gehört beispielsweise auch die Verbindung einzelner Sätze/Äußerungen zu einem globalen Ganzen. Dabei spielen häufig semantische Hinweise auf der Basis von Wortfeldern eine Rolle. Zum anderen werden elaborierende Schlussfolgerungen gezogen, die aus dem Stimulus herausreichen und ihn mit dem Wissen des Lesers/Zuhörers verbinden um eine erweiterte mentale Repräsentation des Stimulus zu erstellen. (vgl. Kapitel 2.2.3. *Das Construction-Integration- Modell*) Auf diese Weise wird Kohärenz mit den Wissensstrukturen der Leser/Zuhörer hergestellt. Zu Inferieren erfordert immer die Speicherung von Informationen aus vorhergehenden Sti-

mulusteilen zusammen mit der Verarbeitung von aktuellen Sätzen. Das Arbeitsgedächtnis ist dabei der entscheidende Faktor für die Fähigkeit, verschiedene Informationen über einen Text hinweg zu integrieren. Die Kapazität des Arbeitsgedächtnisses wird häufig sogar entscheidend dafür gesehen, wie erfolgreich Inferenzen getätigt werden und mit welchem Erfolg das mentale Modell erstellt werden kann. (vgl. Daneman, 1988; Biemiller & Slonim, 2001; Just & Carpenter, 1992) Einen Zusammenhang zwischen der Fähigkeit korrekt zu Inferieren und dem Leseverstehen zeigten Cain et al. (2002, 2004).

Die Annahme, dass Items leichter sind, wenn der Stimulus die richtige Option direkt bestätigt, wurde in mehreren Studien belegt. Embretson und Wetzel (1987) fanden eine signifikante Korrelation für die Variable „Konfirmation" (die korrekte Option wird (nicht) vom Text bestätigt). Dieses Ergebnis konnte jedoch von Perkins und Brutten (1993) nicht repliziert werden. Davey (1988) operationalisierte eine Variable, die Aufschluss über die Explizitheit/Implizitheit der zur Lösung benötigten Information gibt. Diese Variable korrelierte moderat mit der Itemschwierigkeit und signifikant mit der Variable „Inferenztyp". Bei der Variable „Inferenztyp" wurde eingeschätzt, wie kognitiv anspruchsvoll die zur Lösung notwendigen Inferenzen sind. Eine ähnliche Variable wurde von Perkins und Brutten (1993) eingesetzt. Sie ließen einschätzen, ob die benötigte Information wörtlich im Stimulus zu finden ist oder ob kognitive Operationen (z. B. schlussfolgern, abwägen, negieren, etc.) durchgeführt werden müssen, um die Information zu erhalten. Diese Variable „Kognitiver Anspruch" korrelierte signifikant mit der Itemschwierigkeit. Auch Nissan et al. (1996) fanden für die Variable „Implicit Versus Explicit Information Tested" einen Zusammenhang mit der Itemschwierigkeit. Nissan et al. vermuteten, dass besonders das Multiple-Choice-Format zur Überprüfung von Inferenzen ungeeignet ist, da häufig mehrere verschiedene logische Inferenzen möglich sind und die Testpersonen im Multiple-Choice-Format auf eine bestimmte Antwort festgelegt werden. Plausible Inferenzen der Testpersonen müssen jedoch nicht unbedingt in den Antwortoptionen auftauchen und die Testpersonen müssen dann ihre eigenen Inferenzen mit den vorgegebenen Antwortoptionen abgleichen.

Evetts und Gauthier (2005) kategorisieren Informationsentnahmestrategien in vier Stufen: „locating", „cycling", „integrating" und „generating". Der Prozess, die gegebene Information in der Frage mit einer Stimulusstelle abzugleichen, um die gesuchte Information zu finden, wird als „locating" bezeichnet. Je mehr sich gegebene Informationen und Stimulusstelle gleichen, desto einfacher ist es, die gesuchte Information zu ermitteln. Entsprechen sich beispielsweise die grammatischen Strukturen von gegebener und gesuchter Information im Stimulus, so ist die Beantwortung dieser Items einfacher. Je mehr Informationen gegeben werden, desto schwieriger wird das Item, da mehr Informationen mit dem Stimulus abgeglichen werden müssen. Beim „cycling" müssen mehrere Stimulusstellen gefunden werden, um Informationen daraus zu erhalten. Items, die zyklisches Suchen erfordern, sind leichter, wenn genau angegeben wird, wie viele Informationen gefunden werden müssen. Die Strategie des „integrating" baut auf dem zyklischen Suchen auf. Die daraus gewonnenen Informationen werden beim Integrieren miteinander verglichen, um Gemeinsamkeiten und Unterschiede zu ermitteln. Dabei fällt es meist leichter, Gemeinsamkeiten festzustellen. Beim „generating" müssen die Leser/Zuhörer Hintergrundwissen (z. B. über Kategorien) aktivieren, um gegebenes Wissen mit dem

Stimulus in Verbindung bringen zu können bzw. die richtige Antwortalternative auswählen zu können.

5.2.5.2. Position der NI

Ein wichtiger Teil der Sprachverarbeitung ist die Fähigkeit, aus unterschiedlichen Quellen stammende Informationen miteinander zu verbinden. Wichtige Informationen müssen im Arbeitsgedächtnis (vgl. Kapitel 2.1.1.2. *Baddeleys Modell des Arbeitsgedächtnisses*) aktiviert gehalten werden, während eintreffende Informationen verarbeitet werden. Je größer der Abstand zwischen zusammengehörigen Informationen ist, desto größer wird der Anspruch an das Arbeitsgedächtnis, da die Informationen lange aktiviert gehalten werden müssen und stets weitere zu verarbeitende Informationen eintreffen. (Chang, 1980; Jarvella, 1971; Daneman, 1988)

Kieras (1985) und Hare et al. (1989) wiesen nach, dass es einfacher ist, Informationen zu lokalisieren, die zu Beginn eines Satzes/einer Äußerung, eines Absatzes oder des Stimulus stehen, als Informationen, welche am Ende desselben zu finden sind. Am schwersten fällt es Lernern Informationen aufzunehmen, die mittig stehen. Diese Annahmen werden auch in den Untersuchungen von Conrad (1989) bestätigt. Er führte eine Studie zum Hörverstehen in der Zweitsprache durch, bei der eine Liste mit Sätzen erinnert werden sollte. Conrad konnte zeigen, dass es den Testpersonen leichter fiel, sich an Sätze am Anfang und zum Ende der Liste zu erinnern.

Diese Ergebnisse gehen zurück auf den von Ebbinghaus untersuchten Primacy- und Recency- Effekt. Bereits im 19. Jahrhundert experimentierte Ebbinghaus mit sprachlichem Material in Form von Listen. Die Versuchspersonen mussten diese Listen auswendig lernen und das gelernte Material dann wiedergeben. Dabei zeigte sich, dass die Wiedergabe des Materials vom Anfang und vom Ende der Liste deutlich besser ist, als die Wiedergabe des mittig stehenden Materials. Die bessere Wiedergabe des Anfangsmaterials wird als Primacy-Effekt, die bessere Wiedergabe des Endmaterials wird als Recency-Effekt bezeichnet. Der Primacy-Effekt liegt darin begründet, dass die anfänglichen Elemente der Liste häufiger wiederholt werden können und dadurch zum Teil sogar im Langzeitgedächtnis gespeichert werden. Der Recency-Effekt lässt sich dadurch erklären, dass sich die zuletzt memorierten Elemente noch im Arbeitsspeicher befinden und deshalb leichter wiedergegeben werden können. (vgl. Oberauer et al., 2006: 125f)

Analog zu den Ergebnissen zum Leseverstehen von TOEFL-Items (Freedle & Kostin, 1993a) fanden Freedle und Kostin (1996) diese Ergebnisse für das Hörverstehen bestätigt: Items, welche sich auf Informationen am Anfang oder am Ende des Stimulus berufen, sind einfacher als Items, die sich auf mittige Informationen beziehen. Obwohl die Ergebnisse für beide Kompetenzbereiche sehr ähnlich sind, ist anzunehmen, dass die Gründe dafür divergieren. Bei den Leseverstehensaufgaben sind sowohl der Stimulus als auch die ganze Aufgabe zu jedem Zeitpunkt sichtbar und immer wieder für die Testpersonen abrufbar. Mittige Informationen sind mit visuellen Suchstrategien im Leseprozess schwieriger aufzufinden. Im Fall der Hörverstehensaufgaben sehen die Probanden nur die Itemoptionen im Testheft abgedruckt.

Der Stimulus und auch der Itemstamm werden den Testpersonen vorgespielt. Die Konzentration auf Informationen am Anfang und am Ende beruht in diesem Fall auf der Schwierigkeit, den gesamten auditiven Input im Gedächtnis zu behalten. Auch Sheehan und Ginther (2001) untersuchten die TOEFL Leseverstehensaufgaben und konnten zeigen, dass die Itemschwierigkeit signifikant von der Häufigkeit des Auftretens und dem Ort der gesuchten Information im Text beeinflusst wurde.

5.2.5.3. Auftretenshäufigkeit der NI

Eine wichtige Besonderheit gesprochener Sprache ist ihre Redundanz. Forschungsarbeiten dazu stammen beispielsweise von Chiang und Dunkel (1992). Sie fanden heraus, dass leistungsstarke chinesische Englischlerner von Redundanzen im Stimulus profitierten, wohingegen bei leistungsschwachen Lernern kein Unterschied zu beobachten war. Obwohl der Einfluss von Redundanzen anscheinend vom Leistungsniveau der Testpersonen abhängt, spielt er auch insgesamt eine Rolle beim Textverständnis. Chiang und Dunkel unterschieden bei ihrer Studie zwischen einer weiteren Ausführung einer Information und einer mehrfachen wörtlichen oder sinngemäßen Wiederholung von Informationen. Auch Parker und Chaudron (1987) erhielten bei ihren Arbeiten mit L2-Lernern ähnliche Ergebnisse. Ausführende Erweiterungen, welche eine Wiederholung von Informationen beinhalten und die thematische Struktur eines Stimulus klar gliedern, verbesserten das Textverständnis der Testpersonen. Pica et al. (1987) untersuchten mit ESL-Schülern den Einfluss von Instruktionen auf die Verstehensleistung. Eine Teilgruppe erhielt Instruktionen, die in grammatischer Komplexität reduziert wurden und deren Redundanz erhöht wurde. Die zweite Teilgruppe erhielt unveränderte Instruktionen, bei denen jedoch die Möglichkeit bestand, nachzufragen. Es zeigte sich, dass Redundanz die Verständnisleistung erhöhte, wohingegen die grammatische Komplexität der Stimuli keinen Einfluss auf das Verständnis zu haben schien. Die Länge der Stimuli erwies sich als Möglichkeit für Redundanz auch als wichtiges Merkmal. Den größten Einfluss auf die Verstehensleistung hatte jedoch die Interaktion mit dem Instruktor. Bestand die Möglichkeit nachzufragen, erhöhte sich die Verstehensleistung am deutlichsten.

Freedle und Kostin (1996) untersuchen die Variable „Wiederholung der NI". Von Bedeutung ist hier die Überlappung von Inhalten aus dem Stimulus mit dem Attraktor. Untersucht wird bei dieser Variable einerseits, ob die Information ausgeschmückt, paraphrasiert oder wörtlich wiedergegeben wird, und andererseits ob sie an einer oder mehreren Stimulusstellen auftaucht. Dabei spielt auch eine Rolle, ob die Informationen in einem Verhältnis 1:1 paraphrasiert werden, oder ob sich die Paraphrase über den ganzen Stimulus verteilt und die Testpersonen die Teilinformationen erst wieder zusammenfügen müssen. Bezog sich die notwendige Information zur Lösung des Items auf mehr als einen Satz und/oder enthielt einer der Sätze zusätzliche, ausschmückende Informationen, so erhöhte dies die Aufgabenschwierigkeit. Eine reine Wiederholung oder Paraphrasierung schien keinen Einfluss auf die Aufgabenschwierigkeit zu haben (Freedle & Kostin, 1996: 26).

5.3. Merkmale der Testpersonen

Die Schwierigkeit von Stimuli und Items kann stets nur in Abhängigkeit der Leistungsfähigkeit der Testpersonen betrachtet werden. Merkmale, die einen Einfluss auf die Testleistung ha-

ben können, sind u. a. das Arbeitsgedächtnis der Testpersonen, linguistische Fähigkeiten (z. B. Sprachkenntnisse) und kognitive Fähigkeiten, wie verbale und nonverbale Strategien (z. B. die Fähigkeit, sich unbekannte Wortbedeutungen aus dem Kontext zu erschließen), aber auch Motiviertheit und Vorwissen zu und Interesse an einem Thema. (Yousif, 2006)

5.3.1. Motivation und Interesse

Motivationale Variablen, wie z. B. extrinsische und intrinsische Motivation oder Leistungsmotivation, steuern den Prozess der Informationsverarbeitung. Bei intrinsischer Motivation liegt die Motivation, eine Aktivität durchzuführen, in der Aktivität selbst. Sie wird als befriedigend bzw. belohnend empfunden. Extrinsische Motivation ist durch Ereignisse begründet, die als Folgen einer Handlung erwartet werden. (vgl. Schmalt & Skolowski, 2006) Imhof (2003) betont, dass Zuhören eine intentionale Handlung ist und demnach der Umfang und die Qualität der Informationsverarbeitung durch den aktiven Beschluss zuzuhören, beeinflusst werden. Störungen werden erkannt und ausgeblendet und konkurrierende Handlungsimpulse werden unterdrückt.

Schiefele (1996; Schiefele & Krapp, 1996) untersuchte die Bedeutung des thematischen Interesses für das Textlernen. Hohes Interesse an einem Thema geht i. d. R. mit positiven Gefühlen und hoher persönlicher Bedeutsamkeit einher. Individuelle Interessen drücken Orientierungen hinsichtlich bestimmter Themen aus und können die Motivation, sich mit bestimmten Themen zu beschäftigen, stark beeinflussen. Schiefele konnte zeigen, dass thematisches Interesse ein bedeutender motivationaler Prädiktor des Textlernens ist.

5.3.2. Hintergrundwissen

Auch Hintergrundwissen spielt eine Rolle bei der Informationsverarbeitung, denn bei der Erstellung eines mentalen Modells muss häufig auch auf Hintergrundwissen im Sinn von allgemeinem Wissen, Kenntnissen und Erfahrungen über Umwelt und Gesellschaft zurückgegriffen werden. Durch Hintergrundwissen wird es möglich, neue Tatsachen einzuordnen und entsprechend zu handeln, auch wenn detaillierte Informationen fehlen. (Evetts & Gauthier, 2005) Dass Hintergrundwissen, zusammen mit dem spezifischen Wortschatz, einen Einfluss auf das Textverständnis hat, wurde bereits in den 80er Jahren gezeigt. (vgl. Spilich et al., 1979; Wittrock et al., 1975)

In der Regel wird zwischen themenbezogenem (inhaltlichem) und metakognitiven Wissen unterschieden. Zu metakognitivem Wissen zählt u. a. Strategiewissen, das unabhängig vom konkreten Inhalt für den Umgang mit Texten eingesetzt werden kann. Themenbezogenes Wissen („domain-specific knowledge“ bei Best et al., 2005) hängt damit zusammen, wie schnell Stimulusmaterial verstanden wird, indem relevantes Wissen aktiviert wird. Für die Bildung eines Situationsmodells ist die vorhandene Wissensstruktur der Personen von Bedeutung, denn Hintergrundwissen beeinflusst vor allem die Verarbeitung auf den höheren semantischen Ebenen. Je mehr Hintergrundwissen eine Person zu einem Thema hat, desto differenzierter kann das Situationsmodell aufgebaut werden, da Inferenzen und Assoziationen leichter möglich werden, ohne sich weiterhin stark auf die explizite Textbasis beziehen zu müssen. (vgl. Imhof, 2003: 146; Best et al., 2005: 68) Demnach müsste Hintergrundwissen einen positiven

Einfluss auf die Verarbeitung von Texten und Diskursen haben.

Keshavarz et al. (2007) untersuchten den Einfluss von Hintergrundwissen auf das Leseverständnis von L2 Texten an 240 Studenten aus dem Iran. Jede Testperson bearbeitete Aufgaben zu zwei biographischen Texten, und zwar über einen bekannten Islamischen Geistlichen und über einen nicht Islamischen Geistlichen. Die Autoren fanden heraus, dass Vertrautheit mit dem Inhalt signifikant mit den Testergebnissen des Leseverständnistests korrelierte.

McNamara und Kintsch (1996) evaluierten die Interaktion von Kohärenz und Hintergrundwissen beim Textverstehen, indem sie die Kohärenz von Texten manipulierten und das Ausmaß des Verstehens bei Lesern mit viel und wenig Hintergrundwissen erhoben. Sie konnten zeigen, dass Leser mit wenig Hintergrundwissen eher von einem kohärenten Text profitieren als Leser mit viel Hintergrundwissen. Ein Grund dafür könnte darin liegen, dass die Redundanz der hoch-kohärenten Texte zusammen mit dem Hintergrundwissen Personen mit viel Hintergrundwissen eher zu einer oberflächlicheren Verarbeitung verleitete. Ein weiterer Grund könnte auch sein, dass Personen mit viel Hintergrundwissen bei inkohärenten Stimuli eher kompensatorische Strategien einsetzen und dadurch oft zu einem gründlicher aufgebauten mentalen Modell gelangen (und damit zu tieferem Textverständnis) als bei kohärenten Stimuli. Rubin (1994) weist darauf hin, dass sich Hintergrundwissen in manchen Fällen sogar negativ auswirken kann, wenn beispielsweise Informationen falsch interpretiert werden, weil sie dem Hintergrundwissen der Testperson widersprechen. Im Gegensatz dazu fanden jedoch O'Malley et al. (1989) und Lund (1991) mithilfe von Laut-Denk-Protokollen heraus, dass die Testpersonen gerade bei Verständnisschwierigkeiten allgemeines Wissen aktivierten.

5.3.3. Arbeitsgedächtnis

Unterschiede in der Textverarbeitungsfähigkeit von Individuen können u. a. auf die Kapazität des Arbeitsgedächtnisses zurückgeführt werden. (vgl. Kapitel 2.1.1.2. *Baddeleys Modell des Arbeitsgedächtnisses*) Das Arbeitsgedächtnis wird zur gleichzeitigen Verarbeitung und Speicherung von Informationen benötigt und es wird davon ausgegangen, dass die für diese Operationen zur Verfügung stehende Kapazität begrenzt ist. (z. B. Tuholski et al., 2001; Baddeley, 1986; Kyllonen & Christal, 1990) Der reibungslose Ablauf der Verarbeitungsprozesse hängt davon ab, wie viele Informationen wie schnell vom Arbeitsgedächtnis verarbeitet werden müssen und wie viele Verarbeitungsressourcen dafür zur Verfügung stehen. Angenommene individuelle Unterschiede in der Arbeitsgedächtniskapazität von Personen müssten im Umkehrschluss auch Unterschiede in den höheren kognitiven Prozessen erklären können.

Erste Untersuchungen dazu wurden von Daneman und Carpenter (1980) zum Leseverstehen bei Erwachsenen durchgeführt. Da für die Texterschließung Informationen während des Lesens zwischengespeichert werden müssen (z. B. Just & Carpenter, 1980; Kintsch & van Dijk, 1978), gilt die Arbeitsgedächtniskapazität für die Lesekompetenz als relevant und es wird angenommen, dass Menschen mit einer größeren Arbeitsgedächtniskapazität bei Aufgaben zum Leseverstehen bessere Ergebnisse erzielen müssten. Bei den Aufgaben der ersten Studien in diesem Bereich wurde i. d. R. reine Reproduktion von Informationen verlangt. Die Ergebnisse korrelierten nicht oder nur sehr schwach mit dem Leseverstehen (z. B. Perfetti & Goldman, 1976).

Daneman und Carpenter (1980) vermuteten daraufhin, dass zur Vorhersage von Leseverstehenskompetenz Aufgaben benötigt werden, die gemäß des Modells von Baddeley und Hitch (1974) neben der Speicherung von Informationen auch deren Verarbeitung erfordern. Sie entwickelten deshalb den Aufgabentyp „Reading Span", bei dem die Testpersonen voneinander unabhängige Wörter memorieren und wiedergeben sowie die verbalen Informationen aktiv verarbeiten müssen. Mit diesen Aufgaben konnten Daneman und Carpenter erstmals Zusammenhänge zwischen der Lesekompetenz und dem Arbeitsgedächtnis nachweisen. Daneman und Merikle (1996) bestätigten diese Ergebnisse an einer größeren Probandenstichprobe. Die gefundenen Korrelationen beruhen jedoch nicht nur auf der verbalen Ähnlichkeit der „Reading Span" und der Leseverstehensaufgaben, sondern konnten auch mit numerischem Material gezeigt werden. Turner und Engle (1989) arbeiteten z. B. mit strukturell äquivalenten Aufgaben mit numerischem Material. Auch diese Aufgaben eigneten sich gut als Prädiktoren für das Leseverstehen. Daneman und Tardif (1987) konnten erneut zeigen, dass sowohl verbale als auch numerische Arbeitsgedächtnisaufgaben signifikant mit verbalen Fähigkeitsmaßen zusammenhängen.

Für die Erfassung des Kurzzeitgedächtnisspeichers, in dem lediglich passiv Informationen zwischengespeichert werden, werden häufig einfache Spannenaufgaben verwendet. Bei diesen Aufgaben ist keine Verarbeitung der Informationen notwendig. (z. B. Baddeley, 1995; de Jonge & de Jong, 1996) Zur Erfassung der Arbeitsgedächtniskapazität werden hingegen häufig Aufgaben gewählt, für die eine gleichzeitige Speicherung und Verarbeitung erforderlich ist, sogenannte komplexe Spannenaufgaben. (z. B. Oberauer, 2005; Oberauer et al., 2000) Um die Konfundierung sprachlicher Fähigkeiten mit den Ergebnissen des Tests zur der Arbeitsgedächtniskapazität zu vermeiden, wurden im Rahmen dieser Arbeit Aufgaben zur Erfassung einer Zahlenspanne eingesetzt. Dabei wurden den Testpersonen auditiv einzelne Ziffern vorgegeben, die sie anschließend in aufsteigender Reihenfolge wiedergeben sollten. Neben der kurzzeitigen Speicherung war also auch eine Verarbeitung der eingetroffenen Informationen notwendig.

5.3.4. Sprachkenntnisse

Auch der Grad der Beherrschung der getesteten Sprache hat einen Einfluss darauf, welche Merkmale von den Testpersonen als schwierigkeitsbeeinflussend empfunden werden. Nur 85% der Bevölkerung in Deutschland sind Sprecher, deren Erstsprache Deutsch ist (Graefen & Liedke, 2008: 23) und 42% Kinder, das sind 1,7 Mio., wachsen in Familien mit Migrationshintergrund auf (Autorengruppe Bildungsberichterstattung, 2010: 6). Obwohl Jugendlichen mit Migrationshintergrund, die im Rahmen ihrer familiären Situation zweisprachig aufwachsen, durch die Zweisprachigkeit eine Reihe von Vorteilen für ihre sprachliche und kognitive Entwicklung erwachsen können, bedarf es eine explizite Berücksichtigung und Förderung dieser sprachlichen Voraussetzungen, um sie weiterzuentwickeln. (vgl. Bund-Länder- Kommission für Bildungsplanung und Forschungsförderung. Heft 107. Förderung von Kindern und Jugendlichen mit Migrationshintergrund. 2003: 44) Verschiedene große internationale Schulleistungsstudien wie z. B. PISA (Baumert & Schümer, 2001; Ramm et al., 2004; Walter & Taskinen, 2007), DESI (Beck & Klieme, 2007; DESI-Konsortium, 2007), IGLU (Schwippert et al., 2007) oder TIMSS (Bonsen et al., 2008) zeigten, dass Jugendliche mit Migrationshintergrund

im deutschen Bildungssystem in den Domänen Lesen, Mathematik und Naturwissenschaften häufig benachteiligt sind. Für die Erklärung dieser migrationsbedingten Disparitäten hat sich die Beherrschung der Verkehrssprache des Aufenthaltslandes als bedeutsam erwiesen. Die Beherrschung der deutschen Sprache ist für das Verständnis des Unterrichts in Deutschland und damit für den Erwerb schulischer Kompetenzen essentiell, da der Unterricht i. d. R. auf Deutsch durchgeführt wird. (vgl. Esser, 2006; Ramm et al., 2004). Deshalb wird in Schulleistungsstudien (z. B. Ländervergleich Sprachen; Köller et al., 2010) die Erstsprache und die regelmäßig im Elternhaus gesprochene Sprache ermittelt.

Fehlende phonetische Grundlagen bei der Sprachbeherrschung können zu Problemen im Bereich des Hörverstehens (vgl. Pallier et al., 1997: 129; Hirschfeld, 1992: 17) aber auch des Leseverstehens (vgl. Cain et al., 2004; Yuill et al. 1989) führen. Allerdings gibt es auch Hinweise darauf, dass dysfunktionale Hörmuster durch wissensgeleitete Inferenzen zum Teil ausgeglichen werden können. Im Rahmen der L2-Forschung wurde gezeigt, dass begrenzte Sprachkenntnisse häufig durch top-down Inferenzen kompensiert werden, die auf außersprachlichem Wissen basieren. (vgl. Goh, 2000; Freebody & Anderson, 1983)

Für L2-Lerner spielt auch die Kapazität des Arbeitsgedächtnisses eine stärkere Rolle als für muttersprachliche Schüler. Die sprachlichen Informationen werden nach dem Aufnehmen im Arbeitsgedächtnis zu mentalen Repräsentationen verarbeitet, wobei die neu eintreffenden Informationen mit bereits bekannten Informationen und Schemata in Beziehung gesetzt werden. (vgl. Kapitel 2.1. *Verarbeitung eintreffender Informationen*) Die Ergebnisse einer Lerner-Befragung von Goh (2000) weisen darauf hin, dass tatsächlich eine Überlastung des Arbeitsgedächtnisses ein häufiges Problem für L2-Lerner im Zusammenhang mit Hörverstehensaufgaben darstellt.

III

Darstellung der Untersuchung

III Darstellung der Untersuchung

Bevor die Analysen ausführlich besprochen werden, werden einige Vorinformationen zu den Testinstrumenten, der Datengrundlage und den verwendeten Forschungsmethoden gegeben. In Kapitel 1 werden die dieser Arbeit zugrunde liegenden wissenschaftlichen Fragestellungen behandelt. In Kapitel 2 werden die identifizierten Variablen in Bezug auf die Stimuli, Items und Personen vorgestellt und beschrieben. In einem dritten Kapitel wird dann eine allgemeine Abhandlung über die verwendeten Forschungsmethoden gegeben. Diese allgemeinen Informationen werden in den Abschnitten zu den einzelnen Analysen nur noch ergänzt und präzisiert, um umfangreiche und unübersichtliche Beschreibungen zu vermeiden und dafür inhaltliche Fragen stärker zu fokussieren.

1. Wissenschaftliche Fragestellungen

Ziel dieser Arbeit ist es, die Bedeutung von Stimulus- und Itemmerkmalen besser zu verstehen und ihren Einfluss auf die Itemschwierigkeit zu bestimmen. Für die theoriegeleitete Aufgabenkonstruktion ist die Kenntnis von Stimulus- und Itemmerkmalen wichtig, um geeignete, d. h. in ihrer Schwierigkeit für die Zielgruppe passende Stimuli für die Entwicklung von Items zu finden. Auch die Items selbst können durch die Kenntnis von Merkmalen, die einen Einfluss auf die Itemschwierigkeit haben, gezielt für bestimmte Zielgruppen angepasst werden. Auf diese Weise wird nicht nur der Prozess der Aufgabenentwicklung effizienter und stärker professionalisiert, das Wissen über derartige Prädiktoren ist auch für die Beschreibung der einzelnen Stufen in einem Kompetenzstufenmodell von (mindestens deskriptiver) Bedeutung.

1.1. Forschungsfragen

1. Welche Einzelmerkmale der Items und der Stimuli eignen sich dazu, die Item- bzw. die Aufgabenschwierigkeit im Hörverstehensprozess vorherzusagen?

2. Welche Merkmalsgruppen der Items und der Stimuli eignen sich dazu, die Item- bzw. Aufgabenschwierigkeit im Hörverstehensprozess vorhersagen?

3. Ist eine Einschätzung der Stimulusschwierigkeit auch mittels eines Globalurteils bzw. mithilfe eines Fragebogens reliabel?

4. In welcher Weise beeinflussen personenbezogene Merkmale die Testleistung?

1.2. Hypothesen

1.2.1. Hypothese 1 - Einzelmerkmale

Es wird angenommen, dass Einzelmerkmale der Items und der Stimuli, die Item- und die Aufgabenschwierigkeit im Hörverstehensprozess vorhersagen können.

Stimulusmerkmale

Es wird folgender Einfluss von Stimulusmerkmalen auf die Item- bzw. Aufgabenschwierigkeit angenommen (vgl. auch Tabelle III-1.2.1a.):

Stimulusmerkmale: Gruppe I: Komplexität des Wortschatzes und sprachliche Merkmale
Je höher die Ausprägungen der Variablen „Wortlänge (WLS)", „Lange Wörter (PLW)", „Mehrsilbige Wörter (PMW)", „Wiederaufnahmen (WIE)", „Inhaltswörter (IWH)" und „Substantive/Eigennamen/Appellative (SUB)" sind, desto schwieriger werden die Stimuli und die Items, da längere Wörter bzw. Inhaltswörter sowie Substantive häufig komplexer sind und deren Verarbeitung mehr Gedächtniskapazität in Anspruch nimmt. Ein Stimulus mit einer dichten Wiederaufnahmestruktur weist auf hohe Kohärenz hin und ist häufig schwieriger zu verstehen, da die Wiederaufnahmen nicht explizit erfolgen und auch Informationen vom Stimulusanfang im Arbeitsgedächtnis aktiv gehalten werden müssen, damit später eintreffende Informationen (Wiederaufnahmen) verstanden werden müssen.

Für die Merkmale „Einsilbige Wörter (PEW)" und „Worthäufigkeit (GWS)" wird ein gegenteiliger Effekt erwartet. Je stärker der Wortschatz eines Stimulus mit dem Grundwortschatz überlappt und je kürzer die Wörter in diesem Stimulus sind, desto einfacher sollte der verwendete Wortschatz sein und umso einfacher sind diese Stimuli zu verstehen. Eine höhere Ausprägung der Variablen „Deixis (DEI)" und „Referenzen (REF)" führt zu einem Anstieg der Item- bzw. Aufgabenschwierigkeit, da sie Inferenzleistungen vom Zuhörer erfordern. Auch von den Ausprägungen der Variable „Rhetorische Mittel (RHE)" (RHE1: „Bildliche Darstellungsformen", RHE2: „Uneigentliches Sprechen" und RHE3: „Neudeutsch/Anglizismen") und der Variable „Negationen (NEG)" wird ein Anstieg der Schwierigkeit erwartet, da die genannten sprachlichen Merkmale die Verarbeitung der Informationen erschweren können. Obwohl neudeutsche Ausdrücke und Anglizismen bevorzugt von Jugendlichen verwendet werden, wird davon ausgegangen, dass u. U. nicht alle in den Stimuli vorkommenden Begriffe gleichermaßen bei allen Jugendlichen bekannt sind und dadurch ebenfalls schwierigkeitssteigernd wirken.

Größere Ausprägungen der Variablen „Strukturbestimmung (STR)", „Jugendsprache/Umgangssprache (RHE4)" und „Verben (VER)" sollten negativ mit der Item- und der Aufgabenschwierigkeit korrelieren. Die Ausprägungen der Variable STR sind Merkmale gesprochener Sprache und es wird vermutet, dass Diskurse im Gegensatz zu gesprochenen Texten weniger dicht

und einfacher zu verstehen sind. Auch die Ausprägung der Variable RHE4 sollte deshalb eher schwierigkeitssenkend wirken. Davon ausgenommen könnten jedoch die Merkmale „Ellipse (STR2)", „Anakoluthe (STR4)" und „Adjazenzstrukturen (STR3)" sein. Durch die Unvollständigkeit der Strukturen „Ellipse" und „Anakoluth" ist der Zuhörer gezwungen, Informationen selbst zu ergänzen. Der Rückschluss vom Gesagten auf das Gemeinte könnte in diesen Fällen schwieriger zu leisten sein. Adjazenzstrukturen tragen zwar einerseits zur Kohärenz der Stimuli bei, andererseits erfordern sie vom Zuhörer erhöhte Aufmerksamkeit, da das vorher Gesagte im Arbeitsgedächtnis aktiv gehalten werden muss um die Struktur zu verstehen. Deshalb wird angenommen, dass diese beiden Variablen, die Aufgabenschwierigkeit eher erhöhen. Stimuli mit einem höheren Anteil an Verben (Verbalstil) sind i. d. R. einfacher zu verstehen als Stimuli mit einem höheren Anteil an Substantiven (Nominalstil).

Stimulusmerkmale: Gruppe II: Präsentationsmerkmale
Für die Variablen „Länge in Minuten (LST)", „Wortzahl (AWS)", „Anzahl der Sprecher (ASP)", „Sprechgeschwindigkeit (SGS)" und „Akzent/Dialekt/Aussprache (AST)" gilt, dass höhere Ausprägungen bzw. Anteile die Item- und Stimulusschwierigkeit erhöhen, da längere Stimuli mit mehreren Sprechern, einer hohen Sprechgeschwindigkeit und Akzent-/Dialektanteilen die Verarbeitung im Arbeitsgedächtnis erschweren. Eine höhere Ausprägung der Variable „Anzahl der Stimuluspräsentationen (AHO)" senkt die Schwierigkeit jedoch, da durch wiederholtes Anhören des Stimulus die entstandene mentale Repräsentation überprüft und ggf. korrigiert bzw. ergänzt werden kann.

Stimulusmerkmale: Gruppe III: Inhaltlich-thematische Merkmale
Höhere Ausprägungen der Variable „Hintergrundwissen (WEL)" erhöhen die Schwierigkeit, da das geforderte Hintergrundwissen ggf. nicht bei allen Schülern in gleichem Ausmaß verfügbar ist. Die Schwierigkeit ist auch dann höher, wenn literarische Stimuli (SLT) vorliegen. Es wird angenommen, dass literarische Stimuli dichter sind und außerdem höhere Ausprägungen der Variable RHE („Bildliche Darstellungsformen", „Uneigentliches Sprechen" und „Neudeutsch/Anglizismen") besitzt. Die Variablen „Hörkontext (HKO)", „Stimulusfunktion (TFU)" und „Thema (THE)" erfassen thematische Eigenschaften der Stimuli. Dabei wird angenommen, dass Stimuli und die entsprechenden Items dann schwieriger werden, wenn eher abstrakte, der Lebenswirklichkeit der Schüler wenig entsprechende Stimuli vorliegen.

Stimulusmerkmale: Gruppe IV: Struktur der Stimuli und propositionale Dichte
Für die Variablen „Länge der Propositionen (MLP)", „Anteil der Propositionen (PRW)" und „Anzahl der Propositionen (PRO)" gilt, dass höhere Ausprägungen bzw. Anteile die Schwierigkeit erhöhen müssten, da sie ein Maß für die benötigte Gedächtniskapazität sind. Hinsichtlich der Variable „Relationstypen (REL)" wird erwartet, dass die Relationstypen „Erklärung/Beweis/Ursache (REL4) ", „Ziel/Bedingung (REL6) " und „Spezifizierung (REL3)" die Schwierigkeit eher erhöhen, da sie komplexere Argumentationsstrukturen beinhalten, wohingegen die Typen, „Antwort (REL2)", „Frage/Impuls/Themensetzung (REL1)" und „Reihenfolge/Aufzählung (REL5)" die Schwierigkeit eher senken. Diese Relationstypen beinhalten stark gliedernde Elemente, wodurch das Verständnis entsprechender Stimuli erleichtert wird. Ein höherer Anteil der Variable „Schlussfolgerungen/Inferenzen (SFI)" steigert die Schwierigkeit, da es kognitiv anspruchsvoll ist, korrekt zu inferieren.

Stimulusmerkmale: Gruppe V: Globalurteil
Je höher die Stimulusschwierigkeit durch die Aufgabenentwickler (Variable „Stimulusschwierigkeit (TSA)" eingeschätzt wird, desto höher liegt auch die empirischen Stimulusschwierigkeit.

Tabelle III-1.2.1a.: Übersicht über den vermuteten Einfluss der Stimulusmerkmale

Variable	Stimulusmerkmale	Einfluss auf die Schwierigkeit	Variable	Stimulusmerkmale	Einfluss auf die Schwierigkeit
Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale					
WLS	Wortlänge	+	STR5	Verberststellung	-
PEW	Einsilbige Wörter	-	RHE1	Bildliche Darstellungsformen	+
PLW	Lange Wörter	+	RHE2	Uneigentliches Sprechen	+
PMW	Mehrsilbige Wörter	+	RHE3	Neudeutsch/Anglizismen	+
IWH	Inhaltswörter	+	RHE4	Jugendsprache/Umgangssprache	-
GWS	Worthäufigkeit	-			
STR1	Referenz-Aussage-Strukturen	-	DEI	Deixis	+
			WIE	Wiederaufnahmen	+
STR2	Ellipsen	+	REF	Referenzen	+
STR3	Adjazenzstrukturen	-	NEG	Negationen	+
STR4	Anakoluthe	+	SUB	Substantive/Eigennamen/ Appellative	+
STR5	Nähezeichen	-			
			VER	Verben	-
Merkmalsgruppe II: Präsentationsmerkmale					
LST	Länge in Minuten	+	SPG	Sprechgeschwindigkeit	+
AWS	Wortzahl	+	AST	Akzent/Dialekt/Aussprache	+
ASP	Anzahl der Sprecher	+	AHO	Anzahl der Stimulus-Präsentationen	-
Merkmalsgruppe III: Inhaltlich-thematische Merkmale					
SLT	Literarischer Stimulus	+	WEL	Hintergrundwissen	+
HKO	Hörkontext – abstrakt	+	THE	Thema – abstrakt	+
TFU	Funktion – fern der Lebenswelt der Schüler	+			
Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte					
REL1	Relationstyp: Frage/ Impuls/Themensetzung	-	REL5	Relationstyp: Ziel/ Bedingung	+
REL2	Relationstyp: Antwort	-	SFI	Schlussfolgerungen/ Inferenzen	+
REL3	Relationstyp:Spezifizierung	-			
REL4	Relationstyp: Erklärung/ Beweis/Ursache	+	PRO/ PRW	Anzahl/Anteil der Propositionen	+
REL5	Relationstyp: Reihenfolge/ Aufzählung	-	MLP	Länge der Propositionen	+
Merkmalsgruppe V: Globalurteil					
TSA	Stimulusschwierigkeit				+

Anmerkungen: +: größere Anteile des Merkmals erhöhen die Schwierigkeit;
- : größere Anteile des Merkmals senken die Schwierigkeit

Itemmerkmale

Es wird angenommen, dass die einzelnen Itemmerkmale folgenden Einfluss auf die Schwierigkeit haben (vgl. auch Tabelle III-1.2.1b.):

Itemmerkmale: Gruppe I: Itemformat
Je offener das Itemformat (IFA) bzw. (IFK) ist, desto schwieriger werden die Items, da die Schüler neben Kompetenzen im Bereich Zuhören auch Fähigkeiten im Bereich Schreiben aufweisen müssen. Ferner ist es i. d. R. schwieriger, eine Antwort selbst zu formulieren als eine Antwortalternative von mehreren auszuwählen.

Itemmerkmale: Gruppe II: Merkmale der Itempräsentation
Je später die Items beantwortet werden (Variable „Position des Items innerhalb der Aufgabe (PIA)"), desto schwieriger sind sie, da die Informationen länger im Arbeitsgedächtnis aktiv gehalten werden müssen. Liegt der „Zeitpunkt der Itembearbeitung (ZIB)" nach dem Anhören des Stimulus, wird die Aufgabe einfacher. In diesem Fall müssen zwar die Informationen im Arbeitsgedächtnis aktiv gehalten werden, was jedoch weniger Arbeitsgedächtniskapazität in Anspruch nimmt, als die gleichzeitige Durchführung der für die Itembeantwortung während des Anhörens benötigten Tätigkeiten Lesen des Items und Schreiben der Antwort.

Itemmerkmale: Gruppe III: Merkmale von MC-Items
Je plausibler die Distraktoren sind, desto schwieriger sollte es sein, den Attraktor zu finden und dementsprechend sollte auch die Itemschwierigkeit zunehmen (Variablen „Mittlere Plausibilität der Distraktoren (MPD)" und „Größte vorkommende Plausibilität der Distraktoren (PDI)"). Je später der Attraktor innerhalb des MC-Items erscheint, desto schwieriger ist er zu identifizieren, da auch mentale Repräsentationen der Distraktoren im Arbeitsgedächtnis aktiv gehalten werden müssen (Variable „Position des Attraktors im MC-Item (PMC)").

Itemmerkmale: Gruppe IV: Kognitive Anforderungen der Items
Je stärker kognitive Operationen wie Schlussfolgern oder Transferieren gefordert sind, desto schwieriger werden die Items, da diese Operationen das Arbeitsgedächtnis in verstärktem Maße beanspruchen. Dieser Aspekt wird durch die Variablen „Anforderungsbereich (AFB)",„Geprüfter Standard (BS)" und „Typ der NI (TNI)" erfasst. Die Itemschwierigkeit steigt auch, wenn mehrere Informationen zur Lösung benötigt werden (Variable „Anzahl der benötigten NI pro Item (ANI)"), die an unterschiedlichen Stellen im Stimulus auftreten (Variable „Position der NI auf Stimulusebene (PST)") und nur einmal genannt werden (Variable „Auftretenshäufigkeit der NI (ARN)"). Ferner tragen Abstraktheits- bzw. Konkretheitsgrad (Variable „Konkretheit der NI (TCO/TOR)") und Umfang (Variable „Wortzahl der NI (WNI)") der gesuchten Information und die Notwendigkeit Hintergrundwissen zum Verständnis der Items zu aktivieren (Variable „Hintergrundwissen (HGW)") zur erhöhten Schwierigkeit der Items bei.

Itemmerkmale: Gruppe V: Globalurteil
Je höher die Itemschwierigkeit durch die Aufgabenentwickler (Variable „Itemschwierigkeit (SEA)") eingeschätzt wird, desto höher liegt auch die empirische Itemschwierigkeit.

Tabelle III-1.2.1b.: Übersicht über den vermuteten Einfluss der Itemmerkmale

Variable	Itemmerkmale	Einfluss auf die Schwierigkeit
Merkmalsgruppe I: Itemformat		
IFK/IFA	Itemformat	je offener, desto schwieriger
Merkmalsgruppe II: Merkmale der Itempräsentation		
ZIB	Zeitpunkt der Itembearbeitung	einfacher nach dem Anhören
PIA	Position des Items innerhalb der Aufgabe	je später, desto schwieriger
Merkmalsgruppe III: Merkmale von MC-Items		
PMC	Position des Attraktors im MC-Item	je später, desto schwieriger
PDI/MPD	Plausibilität der Distraktoren	je plausibler, desto schwieriger
Merkmalsgruppe IV: Kognitive Anforderungen der Items		
AFB	Anforderungsbereich	je höher, desto schwieriger
STA	Geprüfter Standard	je kognitiv anspruchsvoller, desto schwieriger
ANI	Anzahl der benötigten NI pro Item	je höher, desto schwieriger
ARN	Auftretenshäufigkeit der NI	je höher, desto einfacher
PST	Position der NI auf Stimulusebene	am Anfang und am Schluss einfacher
WNI	Wortzahl der NI	je höher, desto schwieriger
TCO/TOR	Konkretheit der NI	je abstrakter, desto schwieriger
HGW	Hintergrundwissen	je mehr erfordert, desto schwieriger
TNI	Typ der NI	je kognitiv anspruchsvoller, desto schwieriger
Merkmalsgruppe V: Globalurteil		
SEA	Itemschwierigkeit	je höher, desto schwieriger

1.2.2. Hypothese 2: Merkmalsgruppen

Es wird angenommen, dass Einzelmerkmale der Items und der Stimuli, die Item- und die Aufgabenschwierigkeit im Hörverstehensprozess vorhersagen können.

Die Item- und Stimulusmerkmale wurden nach unterschiedlichen Gesichtspunkten gruppiert:

- Die am höchsten mit der Item- bzw. Aufgabenschwierigkeit korrelierenden Merkmale
- Faktorenanalytisch bestimmte Merkmalsgruppen
- Thematisch gebildete Merkmalsgruppen
- Ausgewählte Stimulusmerkmale aus den Zusammenhangsanalysen

Es wird davon ausgegangen, dass für Gruppen zusammengefasster Merkmale ein höherer Einfluss auf die Item- bzw. Aufgabenschwierigkeit zu beobachten ist, als für einzelne Merkmale.

1.2.3. Hypothese 3: Globalurteil und Fragebogen

> Es wird angenommen, dass eine Einschätzung der Schwierigkeit auch mittels eines Globalurteils bzw. aufgrund einer Einschätzung der Stimuli mittels eines Fragebogens reliabel ist.

Es wird untersucht, ob sich auch eine sehr globale Einschätzung der Stimuli (Merkmal „Stimulusschwierigkeit (TSA)“) und der Items (Merkmal „Itemschwierigkeit (SEA)“) bzw. eine stärker qualitativ orientierte Einschätzung mittels eines Fragebogens dazu eignet, die Schwierigkeit der Stimuli relativ präzise vorherzusagen. Es wird davon ausgegangen, dass die globale Schwierigkeitseinschätzung mit den empirischen Schwierigkeiten korrespondiert und die mithilfe des Fragebogens vorgenommenen Einschätzungen dazu dienen, einen verlässlichen, zutreffenden Eindruck über die Schwierigkeit der Stimuli zu erhalten.

1.2.4. Hypothese 4: Personenmerkmale

> Es wird angenommen, dass personenbezogene Merkmale die Testleistung beeinflussen.

Es wird angenommen, dass die untersuchten Personenmerkmale einen starken Einfluss auf die Testleistung haben und damit auch implizit die Item- bzw. Aufgabenschwierigkeit beeinflussen. Bei allen vier Merkmalen („Motivation/Interesse (MOT)“, „Bekanntheitsgrad des Stimulus (BEK)“, „Verständlichkeit (VST)“ und „Vertrautheit mit dem Thema (VTS)“) wird davon ausgegangen, dass höhere Werte zu besseren Testleistungen bzw. einer niedrigeren Schwierigkeit führen. Ferner wird angenommen, dass eine höhere Arbeitsgedächtniskapazität zu besseren Testergebnissen führt sowie dass Schüler, die Deutsch als Muttersprache gelernt haben, besser im Test abschneiden.

2. Instrument und Datengrundlage

Die IQB-Aufgaben prüfen die von den KMK-Bildungsstandards vorgegebenen Inhalte und Kompetenzen und sollen dazu beitragen, Stärken und Schwächen im Leistungsprofil der Schüler aufzudecken, damit schwächere Bereiche im Unterricht verstärkt Beachtung finden können. Durch die Integration interkultureller und kommunikativer Aspekte bei der Aufgabenentwicklung sollen auch diese Bereiche verstärkt Eingang in den Unterricht finden.

Weder der individuelle Zuhörprozess noch das daraus resultierende Hörverstehen können unmittelbar beobachtet werden. Es ist nur möglich, Hörverstehen auf der Basis beobachteten Verhaltens zu erschließen. Bei der Erstellung von objektiven, reliablen und validen Testaufgaben muss sichergestellt sein, dass die Aufgaben allgemeingültige Kriterien erfüllen. Die Aufgaben sollen nur dann gelöst werden können, wenn die Schüler über relevante Kompetenzen,

Fertigkeiten und Fähigkeiten verfügen. Die Lösung soll im Idealfall nicht mithilfe von Weltwissen oder durch Raten antizipiert werden können. Deshalb beweist der Schüler unabhängig vom gewählten Lösungsweg, dass er die geforderten Kompetenzen besitzt, wenn er die Aufgabe korrekt löst. Der Lösungsweg fließt nicht in die Bewertung ein. Bei vielen Aufgaben werden jedoch konkret Vorschläge gemacht, welche Strategien und Methoden sich anbieten würden (z. B. Stichwörter mitnotieren) oder es werden in der Arbeitsanweisung Hinweise gegeben, welche Kompetenzen von der Aufgabe gefordert werden (z. B. „Höre genau zu…" zielt auf Detailverstehen). Dabei geht es um die kommunikative Bewältigung der Aufgabe und Rechtschreibfehler und grammatikalische Schwächen werden nur berücksichtigt, wenn das Verständnis der Antwort durch diese Fehler gravierend beeinträchtigt wird.

Die Aufgaben wurden in enger Kooperation mit Vertretern der Fachdidaktik sowie der empirischen Bildungsforschung durch Lehrkräfte entwickelt. Im Auftrag des IQB wurden unter fachdidaktischer Federführung von Prof. Dr. Albert Bremerich-Vos (Universität Duisburg- Essen) von Lehrkräften aus ganz Deutschland aus dem Sekundarbereich I in vier regionalen Arbeitsgruppen seit März 2007 Aufgaben zu den in den KMK-Bildungsstandards ausgewiesenen Kompetenzbereichen „Sprechen und Zuhören", „Schreiben", „Lesen – mit Texten und Medien umgehen" und „Sprache und Sprachgebrauch untersuchen" generiert (vgl. KMK, 2004: 8). Im Rahmen regelmäßig stattfindender Treffen zur Optimierung der Aufgaben wurden die Items durch ein externes wissenschaftliches Berater-Team von Experten aus dem Bereich der Deutschdidaktik und der empirischen Bildungsforschung begutachtet und ggf. überarbeitet. Noch im selben Jahr wurden die erstellten Aufgaben pilotiert und im Folgejahr wurde eine Auswahl geeigneter Aufgaben in einer weiteren Untersuchung normiert.

Die Aufgabenentwickler wurden bei der Wahl der Stimuli kaum eingeschränkt. Sie wurden lediglich dazu angehalten, keine Stimuli zu kontroversen und emotional stark bewegenden Themen zu verwenden, da in der Testsituation im Unterschied zu einer Behandlung im Unterricht nicht auf die Reaktionen der Schüler auf diese Stimuli eingegangen werden kann. Es ist nicht auszuschließen, dass derartige Stimuli auch einen Einfluss auf die Testleistungen der Schüler besitzen. Die Offenheit bei der Wahl der Stimuli führte zu sehr unterschiedlichen Aufgaben. Dies wurde explizit befürwortet, da auf diese Weise unterschiedliche Facetten der Bildungsstandards erfasst werden können. Bei den Stimuli handelt es sich um sprachliche Beiträge, die entweder zunächst nur phonisch vorlagen und dann verschriftlicht wurden oder in anderen Fällen auch auf einer schriftlichen Grundlage basieren. Sie sind zum Teil konzeptionell schriftlich, d. h. sie weisen mehr Merkmale geschriebener Sprache als gesprochener Sprache auf. Im Verhältnis gibt es bei den Aufgaben weniger Diskurse als mündliche Texte (6 Diskurse, 13 Texte, 11 Beiträge mit Diskurs- und Textanteilen). Obwohl es sich also um Zuhöraufgaben handelt, dominiert die geschriebene und nicht die gesprochene Sprache.

Ziel einer Pilotierung ist es einerseits, psychometrische Informationen über die eingesetzten Aufgaben zu erhalten, um schließlich hochwertige Items für die finale Testung auswählen zu können. Die ausgewählten Items müssen psychometrischen Kriterien wie auch fachdidaktischen Anforderungen gerecht werden, um größtmögliche Validität zu erreichen. Andererseits sollen im Rahmen einer Pilotierung ausführliche Kodieranleitungen für die halboffenen

und offenen Aufgaben entwickelt und erprobt werden. Dazu wurde für diese Aufgaben eine Mehrfachkodierung durchgeführt, um Maße der Beurteilerübereinstimmung als ein Kriterium für die Güte der Kodieranweisungen ermitteln zu können. Aufbauend auf diesen Informationen wurden die Kodieranweisungen zielgerichtet optimiert. Eine Feststellung der Kompetenzstände der Schüler wird zu diesem Zeitpunkt noch nicht angestrebt. Dennoch werden Hintergrundinformationen erfragt und Angaben in Hinblick auf die Wahrnehmung der Testinstrumente durch die Schüler eingeholt. Die Auswertung der Testdaten aus den Pilotierungsstudien erfolgte mit Methoden der klassischen und probabilistischen Testtheorie. Auf Basis der Auswertung der Pilotierungsdaten wurden einzelne Testaufgaben für die finale Normierungsstudie übernommen, revidiert oder vom weiteren Einsatz in Leistungstestungen ausgeschlossen. Insgesamt wurden 150 Deutschaufgaben pilotiert, darunter 34 Aufgaben aus dem Bereich Zuhören. Insgesamt nahmen ca. 5300 Schüler der 8. bis 10. Klassenstufen an Hauptschulen, Realschulen, Gymnasien, Gesamtschulen sowie Berufsschulen aus 14 Bundesländern an den Pilotierungsstudien teil.

Im April und Mai des Jahres 2008 erfolgte die Normierung der überarbeiteten Aufgaben an einer national repräsentativen Stichprobe. Ziel der Normierungsstudie war unter anderem, die Schülerkompetenzen auf nationalen Skalen abbilden zu können. Es sollte festgestellt werden, wie viele Schüler die in den Bildungsstandards formulierten Leistungserwartungen erreicht haben. Die Bedeutung der Normierungserhebung für die Entwicklung des IQB-Itempools besteht jedoch primär darin, dass für die fraglichen Aufgaben zuverlässige Kennwerte ermittelt wurden. So wurde eine fundierte Auswahl von Aufgaben für zukünftige Studien ermöglicht.

An der Normierung nahmen bundesweit ca. 7900 Schüler aus 105 Hauptschulen, 54 Realschulen, 64 Gymnasien, 49 Integrativen Gesamtschulen und 48 Schulen mit mehreren Bildungsgängen teil. Die Studie umfasste für die Schüler einen Leistungstest und eine Befragung sowie eine Befragung der zuständigen Fachlehrkräfte. Sie fand im Frühjahr 2008 statt und wurde von geschulten externen Testleitern an 169 Schulen durchgeführt. Im Einzelnen umfasste die Gesamtstichprobe 71 Klassen der 8. Jahrgangsstufe (ca. 1560 Schüler), 167 Klassen der Jahrgangsstufe 9 (ca. 3960 Schüler) und 96 Klassen der 10. Jahrgangsstufe (ca. 2400 Schüler). Innerhalb der gezogenen Schulen wurde je eine Klasse bzw. ein Kurs der vorgesehenen Jahrgangsstufen für die Untersuchung gezogen. Die Testhefte wurden insgesamt von 6701 Schülern bearbeitet. Auf Grundlage eines so genannten komplexen Multi-Matrix- Designs wurden Testaufgaben in Blöcken zusammengefasst und im nächsten Schritt verschiedenen Testheften zugewiesen, wobei einzelne Schüler immer nur Teilmengen des gesamten Itempools bearbeiteten. Aufgrund dieses Multi-Matrix-Designs schwankt die Fallzahl für die Zuhöraufgaben zwischen 289 und 898. Die Testhefte waren durch bestimmte, in mehreren Testheften auftretende Ankeraufgaben miteinander verbunden. Auf die verschiedenen Schulen und Schulklassen wurden die Testhefte so verteilt, dass die unterschiedlichen inhaltlichen Kompetenzen empirisch zueinander in Beziehung gesetzt werden konnten. Es kamen insgesamt 72 Testhefte zum Einsatz. Jedes Testheft beinhaltete sechs Aufgabenblöcke, wobei jeder Block für 20 Minuten konzipiert war und je nach Aufgabenumfang eine oder mehrere Aufgaben zu einer bestimmten Kompetenz enthielt. Insgesamt kamen 97 Deutschaufgaben aus allen fünf Kompetenzbereichen zum Einsatz, darunter 30 Aufgaben zum Zuhören mit

insgesamt 384 Items (vgl. Tabelle III-2a.). Jede Aufgabe besteht aus einem ggf. auch in mehreren Teilen präsentierten Stimulus, zu dem Items zu bearbeiten sind.

Auf Beschluss der Kultusministerkonferenz vom 15.12.05 wurde schließlich für das Jahr 2009 ein Ländervergleich zur Überprüfung des Erreichens der Bildungsstandards in den Fächern Deutsch und Englisch in der Jahrgangsstufe 9 vorgesehen. An der Studie nahmen insgesamt 1.500 Schulen mit etwa 42.500 Schülern teil. Die Auswahl der Schulen wurde über die Ziehung einer Zufallsstichprobe vorgenommen. An 201 der Schulen erfolgte die Testung der Bildungsstandardaufgaben im Rahmen der PISA-Studie an einem zweiten Testtag. An allen anderen Schulen wurde die Erhebung an einem Testtag durchgeführt. Das gesamte Design, die Erstellung der Testhefte sowie die Auswertung der Daten aus der Normierungsstudie und dem Ländervergleich oblag dem IQB, wobei die logistische Durchführung der Testung dem IEA Data Processing and Research Center (DPC) in Hamburg übertragen wurde. Für die Auswertung der Items wurden je zehn Kodierer an fünf Schulungsterminen geschult.

Die gewonnenen Daten stammen aus zwei unterschiedlichen Aufgabenstichproben:

1) Die detaillierten linguistischen Ratings der Item- und Stimulusmerkmale sowie die Angabe zweier Globalurteile (für die Items und die Stimuli) durch die Aufgabenentwickler erfolgten in Bezug auf die 30 bei der Normierung eingesetzten Aufgaben (384 Items) (vgl. Tabelle III.2a.).
2) Die Schülerangaben zu den Stimuli, der Aufmerksamkeitstest und die Sprachkenntnisse sowie die Einschätzungen der Stimuli durch die Deutschlehrkraft der getesteten Klasse basieren auf den Aufgaben des Ländervergleichs (vgl. Tabelle III.2b.). Die 16 Aufgaben mit insgesamt 178 Items im Ländervergleich wurden unverändert, z. T. sogar in identischen Blöcken aus der Normierungsstudie eingesetzt. Deshalb können diese Angaben auf die Itemkennwerte der Normierung übertragen werden.

Die Aufgaben unterscheiden sich stark in der Länge ihrer Stimuli (30 Sek. – 11 Min. 20 Sek.), der Anzahl ihrer Items (3 – 35) sowie den Aufgabenschwierigkeiten (-2.36 - 0.77). Auch die Bandbreite der leichtesten und schwierigsten Items innerhalb einer Aufgabe variiert stark. Die meisten Aufgaben sind insofern authentisch, als die Stimuli nicht eigens für die Testsituation aufgenommen wurden, sondern mitgeschnittene Radiobeiträge sind. Die Stimuli für die Aufgaben „G106_Schulvorfall“ und „G119_Installateur“ wurden von den Aufgabenentwicklern als Skript eingereicht und erst nachträglich im Tonstudio aufgenommen. 14 der Aufgaben wurden primär für den Hauptschulabschluss entwickelt, 16 der Aufgaben sind Aufgaben für den Mittleren Schulabschluss konzipiert. 19 Aufgaben wurden in der Normierung für beide Abschlüsse eingesetzt.

Die linguistischen Analysen wurden von der Arbeitsgruppe Deutsch anhand der verschriftlichten Texte/Diskurse am IQB durchgeführt. Ein Teil der Ratings wurde von zwei oder mehr Ratern unabhängig voneinander vorgenommen. Bei Nicht-Übereinstimmung wurden die Zweifelsfälle diskutiert und so verbindliche Rating-Kriterien entwickelt bzw. die Rating- Skala angepasst.

Tabelle III-2a.: Übersicht über die Hörverstehens-Aufgaben (Normierung)

Aufgabe	N	P	P_{min}	P_{max}	Aufgabe	N	P	P_{min}	P_{max}
G100_Promiquiz	26	-0.90	-2.73	1.29	G116_Mund	31	-1.23	-2.97	-1.01
G101_Radioolympiade	23	-0.42	-2.71	0.91	G117_Maike	15	-1.51	-2.92	-1.15
G102_Raumprojekt	12	0.56	-1.72	2.08	G118_Leila	4	-1.84	-2.41	-1.35
G103_Raupe	9	0.48	-1.35	3.24	G119_Installateur	4	-1.04	-2.26	-1.24
G104_Reklame	23	-0.58	-2.98	3.19	G120_Ida Ehre spricht	4	0.35	-1.04	1.72
G105_Schulbesuch	11	-0.88	-3.93	2.21	G121_Friseur	9	-1.47	-3.03	-1.13
G106_Schulvorfall	8	-2.36	-3.82	-1.20	G122_Frauenfußball	13	-1.84	-3.76	-1.24
G107_Spießer	12	-0.17	-2.36	2.65	G123_Fönproblem	35	-0.96	-3.03	3.96
G108_Sprache	19	-0.67	-3.11	1.88	G124_Dreck	15	-1.64	-3.96	1.00
G109_Torhüter	7	-1.12	-2.16	1.33	G126_Bild im Ohr	18	0.55	-1.89	3.95
G110_Tour de France	23	-0.80	-2.24	2.63	G127_Bauchredner	15	-0.79	-4.74	2.59
G111_Vorstellungsgespräch	8	0.77	-1.63	3.11	G130_Arme Ritter	3	-0.07	-0.87	0.48
G112_Wetterbericht	7	-0.17	-1.40	1.12	G131_Altweibersommer	5	0.73	0.11	1.62
G113_Wetterman	6	-0.38	-2.43	1.02	G132_allerbeste Frage	7	-1.69	-3.13	-1.50
G114_Pinguin	5	-1.02	-1.94	-1.15	G133_Aber sonst gesund	7	-1.87	-2.87	-1.15

Anmerkungen: N = Anzahl der Items; P = Aufgabenschwierigkeit (= aggregierte Itemschwierigkeit);
P_{min} = leichtestes Item; P_{max} = schwierigstes Item

Tabelle III-2b.: Übersicht über die Hörverstehens-Aufgaben (2. Testtag, Ländervergleich)

Aufgabe			
G100_Promiquiz	G107_Spießer	G114_Pinguin	G121_Friseur
G103_Raupe	G109_Torhüter	G118_Leila	G122_Frauenfußball
G104_Reklame	G110_Tour de France	G119_Installateur	G124_Dreck
G105_Schulbesuch	G111_Vorstellungsgespräch	G120_Ida Ehre spricht	G131_Altweibersommer

2.1. Übersicht über die identifizierten Variablen

Aus den vorhergehenden Kapiteln ergeben sich die Merkmale in den Tabellen III-2.1a., III-2.1.b. und III.2.1c., von denen angenommen wird, dass sie einen Einfluss auf die Item- und Aufgabenschwierigkeit haben könnten.

Tabelle III-2.1a.: Übersicht über alle untersuchten Stimulusmerkmale

Variable	Stimulusmerkmal	Variable	Stimulusmerkmal
IQB - Ratings			
LST	Länge in Minuten	REL	Relationstypen
AWS	Wortzahl	REL1	Frage/Impuls/Themensetzung
WLS	Wortlänge	REL2	Antwort
PEW	Einsilbige Wörter	REL3	Spezifizierung
PLW	Lange Wörter	REL4	Erklärung/Beweis/Ursache
PMW	Mehrsilbige Wörter	REL5	Reihenfolge/Aufzählung
IWH	Inhaltswörter	REL6	Ziel/Bedingung
GWS	Worthäufigkeit	RHE	Rhetorische Mittel
ASP	Anzahl der Sprecher	RHE1	Bildliche Darstellungsformen
SGS	Sprechgeschwindigkeit	RHE2	Uneigentliches Sprechen
AST	Akzent/Dialekt/Aussprache	RHE3	Neudeutsch/Anglizismen
AHO	Anzahl der Stimuluspräsentationen	RHE4	Jugendsprache/Umgangssprache
SLT	Literarischer Stimulus	DEI	Deixis
HKO	Hörkontext	WIE	Wiederaufnahmen
TFU	Funktion	REF	Referenzen
THE	Thema	NEG	Negationen
STR	Strukturbestimmung	SUB	Substantive/Eigennamen/Appellative
STR1	Referenz-Aussage-Strukturen	VER	Verben
STR2	Ellipsen	SFI	Schlussfolgerungen/Inferenzen
STR3	Adjazenzstrukturen	WEL	Hintergrundwissen
STR4	Anakoluthe	PRO/PRW	Anzahl/Anteil der Propositionen
STR5	Nähezeichen	MLP	Länge der Propositionen
STR6	Verberststellung		
Lehrer - Einschätzungen			
VTH	Hintergrundwissen	TON	Ton
WAK/WFA	Wortschatz	AGU/AKE	Ausdrucksweise
GRA	Grammatik	INF	Informationsebene
KOH	Kohärenz	WNL/WEH	Wirkung
GIU/GEU/GAE	Gesamteindruck		
Einschätzung der Aufgabenentwickler			
TSA	Stimulusschwierigkeit		

Tabelle III-2.1b.: Übersicht über alle untersuchten Itemmerkmale

Variable	Stimulusmerkmal	Variable	Stimulusmerkmal
IQB - Ratings			
IFK	Itemformat: Kodierleistung	AFB	Anforderungsbereich
IFK.GA	Geschlossen - Ankreuzen	BS	Geprüfter Standard
IFK.GK	Geschlossen - Kodieren	BS141	Gesprächsbeiträge anderer verfolgen und aufnehmen
IFK.01	Einfache Antwort – 0/1 Kodierung	BS142	Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, sichern und wiedergeben
IFK.1P	Antwort mit mehreren Codes	BS143	Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln
IFA	Itemformat: Ausfüllleistung	BS113	Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden
IFA.MC	Multiple Choice	PDI	Größte vorkommende Plausibilität der Distraktoren
IFA.RF	Richtig-Falsch	MPD	Mittlere Plausibilität der Distraktoren
IFA.ZO	Zuordnung	PMC	Position des Attraktors im MC-Item
IFA.RI	Reihenfolge	ANI	Anzahl der benötigten NI pro Item
IFA.HO	Halboffen	ARN	Auftretenshäufigkeit der NI
IFA.OI	Offen	TNI	Typ der NI
ZIB	Zeitpunkt der Itembearbeitung	WNI	Wortzahl der NI
PIA	Position des Items innerhalb der Aufgabe	TCO/TOR	Konkretheit der NI (5- und 33-stufig)
HGW	Hintergrundwissen		
Einschätzung der Aufgabenentwickler			
SEA	Stimulusschwierigkeit		

Tabelle III-2.1c.: Übersicht über alle untersuchten Personenmerkmale

Variable	Stimulusmerkmal	Variable	Stimulusmerkmal
IQB - Ratings			
MOT	Motivation/Interesse	VTS	Vertrautheit mit dem Thema
BEK	Bekanntheitsgrad des Stimulus	ARB	Arbeitsgedächtnis
VST	Verständlichkeit	SPR	Sprachkenntnisse

2.2. Beschreibung der identifizierten Variablen

2.2.1. Stimulusmerkmale aus den IQB-Ratings

Die identifizierten Stimulusmerkmale, mit denen die IQB-Ratings durchgeführt wurden, lassen sich unter didaktischen und sprachwissenschaftlichen Gesichtspunkten in folgende thematische Untergruppen kategorisieren:

Gruppe I: Komplexität des Wortschatzes und sprachliche Merkmale:
„Wortlänge (WLS)", „Einsilbige Wörter (PEW)", „Lange Wörter (PLW)", „Mehrsilbige Wörter (PMW)" „Inhaltswörter (IWH)", „Worthäufigkeit (GWS)", „Strukturbestimmung (STR)", „Rhetorische Mittel (RHE)", „Deixis (DEI)", „Wiederaufnahmen (WIE)", „Referenzen (REF)", „Negationen (NEG)", „Substantive/Eigennamen/Appellative (SUB)", „Verben (VER)",

Gruppe II: Präsentationsmerkmale:
„Länge in Minuten (LST)", „Wortzahl (AWS)" „Anzahl der Sprecher (ASP)", „Sprechgeschwindigkeit (SGS)", „Akzent/Dialekt/Aussprache (AST)", „Anzahl der Stimuluspräsentationen (AHO)"

Gruppe III: Inhaltlich-thematische Merkmale:
„Literarischer Stimulus (SLT)", „Hörkontext (HKO)", „Funktion (TFU)", „Thema (THE)", „Hintergrundwissen (WEL)",

Gruppe IV: Struktur der Stimuli und propositionale Dichte:
„Relationstypen (REL)", „Schlussfolgerungen/Inferenzen (SFI)", „Anzahl der Propositionen (PRO)", „Anteil der Propositionen (PRW)", „Länge der Propositionen (MLP)"

Gruppe V: Globalurteil
„Stimulusschwierigkeit (TSA)"

Im Folgenden werden die theoretisch angenommenen Ausprägungen bzw. Abstufungen der einzelnen Merkmale genauer beschrieben und in numerische, den Ratings zugrunde liegende, Einordnungen übertragen. Diese Übertragungen erfolgten im Idealfall mit dem Ziel, einen positiven Zusammenhang zur Schwierigkeit zu erhalten. Die Einschätzungen waren zum Teil kategorial im Sinn von „vorliegend" – „nicht vorliegend" und zum Teil quantitativ im Sinn von Klassifikationen auf einer mehrstufigen Skala.

2.2.1.1. Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale

Wortlänge (WLS):
Das Merkmal der mittleren Wortlänge wurde dichotom kategorisiert, wobei fünf Buchstaben die Grenze zwischen den Codes darstellen.

Einsilbige Wörter (PEW):
Der Prozentanteil einsilbiger Wörter wurde dichotom in den Stufen ⩽ 55% und > 55% kategorisiert. Lange Wörter (PLW): Der Prozentanteil langer Wörter, d. h. von Wörtern mit mehr als sechs Buchstaben, wurde in drei Stufen erfasst: < 20%, 20% - 25% und > 25%

Mehrsilbige Wörter (PMW):
Der Prozentanteil mehrsilbiger Wörter mit mehr als drei Silben wurde dichotom in zwei Stufen erfasst, wobei 5% die Grenze zwischen den Codes darstellt.

Inhaltswörter (IWH):
Die Kodierung des Merkmals Prozentanteil der Inhaltswörter erfolgte dreistufig mit den Grenzen 45% und 50%.

Worthäufigkeit (GWS):
Die Worthäufigkeit (GWS) als prozentuale Überlappung mit dem Grundwortschatz wurde auf der Grundlage des Grundwortschatzes Leipzig (http://wortschatz.uni-leipzig.de/) ermittelt. Für die Analysen wurden zwei Codestufen mit dem Grenzwert 60% gebildet.

Strukturbestimmung (STR):
Da es sich bei den IQB Stimuli zum Teil um Diskurse mit höheren Anteilen gesprochener Sprache handelt, ist eine valenzgrammatische Untersuchung der Stimuli nicht sinnvoll. Stattdessen wurden die Stimuli entsprechend den Kriterien von Hennig (2006) analysiert. Die Struktur wurde für jede Proposition bestimmt. Insgesamt wurden 15 Strukturtypen bei den Stimuli eingeschätzt, allerdings ergaben sich nur bei den folgenden sechs dichotomen Strukturtypen mit den angegebenen Codegrenzen ausreichend hohe Fallzahlen (vgl. Tabelle III- 2.1.1.1a.), sodass weitere Analysen sinnvoll erschienen:

Tabelle III-2.1.1.1a.: Übersicht über die Einzelcodes der Variable STR

Code	Code-grenze	Kurzbeschreibung	Langbeschreibung
STR1	1%	Referenz-Aussage-Strukturen	Referenz-Aussage-Strukturen weisen ein Referenzobjekt auf und eine Einheit, mit der eine Aussage über das Referenzobjekt gemacht wird. Diese Einheit enthält oft ein Element, das auf das Referenzobjekt zurückweist, z. B. „Das Wetter, das ist heute ganz ausgezeichnet."
STR2	15%	Ellipsen	„grammatisch unvollständig, aber kommunikativ vollständig" (Hennig, 2006: 198), d. h. obwohl aus grammatischer Perspektive Teile fehlen, sind Ellipsen dennoch durch den sprachlichen Kontext verstehbar.
STR3	1%	Adjazenzstrukturen	Für adjazente Gesprächsstrukturen gibt es i. W. zwei Varianten: 1. Ein Gesprächspartner beginnt eine Gesprächseinheit durch eine adjazente Strukturierung, die dann vom anderen weitergeführt wird (turnübergreifend). 2. Eine Einheit wird von einem Gesprächspartner beendet (turnintern) und der andere Gesprächspartner schließt eine bzw. mehrere weitere Einheit(en) daran an.
STR4	1%	Anakoluthe	Im Laufe der Versprachlichung werden Projektionen nicht erfüllt und begonnene syntaktische Konstruktionen werden nicht oder anders zu Ende geführt (Fiehler, 2005: 1238ff), z. B. „Ich, ja, weil es, och, das sind ja Fragen, mein lieber Schwan."
STR5	1%	Nähezeichen	Alle sprachlichen Ausdrücke, die nicht als Satz, als Anakoluth oder Ellipse qualifizieren. Dazu gehören z. B. Responsive, Engführungssignale, Rederechtssignale, Zögerungssignale sowie die Operatoren in Operator-Skopus-Strukturen. Sie bauen keine syntaktischen Projektionen auf. (vgl. Hennig, 2006: 101)
STR6	1%	Verberststellung	z. B. „Sind interessante Themen drin"

Rhetorische Mittel (RHE):
Bei der Variable RHE wurden vier Unterkategorien kodiert, die alle dichotom mit einer Grenze bei 1% in die Analysen einflossen. Tabelle III-2.1.1.1b. gibt einen Überblick über die einzelnen Kategorien der Variable RHE.

Tabelle III-2.1.1.1b.: Übersicht über die Einzelcodes der Variable RHE

Code	Codegrenze	Codebeschreibung	Beispiel
RHE1	1%	Bildliche Darstellungsformen	Metaphern, Bilder, Redewendungen
RHE2	1%	Uneigentliches Sprechen	Ironie Übertreibungen, Wortspiele
RHE3	1%	Neudeutsch/Anglizismen	
RHE4	1%	Jugendsprache/Umgangssprache	

Deixis (DEI):
Im weitesten Sinne sprachliche Einheiten, die ihre Bedeutung erst im Kontext einer bestimmten Sprechsituation erlangen, wie z. B. „ich", „du", „jetzt", „dann", „hier", „da". Die Variable DEI wurde dichotom mit der Grenze 10% zwischen den Stufen kategorisiert.

Wiederaufnahmen (WIE):
Bei den Wiederaufnahmen wurde zunächst zwischen expliziten und impliziten Wiederaufnahmen unterschieden. Vorab-Analysen zeigten zwischen den Kategorien jedoch keinen Effekt, so dass sie zu einer Kategorie zusammengefasst wurden. Die Variable WIE wurde in den Analysen dreistufig mit den Grenzwerten 10% und 20% zwischen den Codes verwendet.

Referenzen (REF):
Die Variable REF erfasst die kodierten Referenzen im Verhältnis zur Wortzahl und wurde dreistufig mit Abstufungen bei 20% und 50% für die Analysen verwendet. Referenzen wurden ursprünglich innerhalb eines Satzes, Satzübergreifend und über die Grenzen des Stimulus hinweg erfasst. Diese Kategorisierung wurde jedoch aufgrund mangelnden Einflusses auf die Itemschwierigkeit in den Vorab-Analysen aufgegeben.

Negation (NEG):
Negationen heben eine satzförmige Äußerung auf, indem die Aussage als nicht gültig bezeichnet wird. Diese Aufhebung einer Aussage kann durch Negation als Erscheinung der Wortbildung (z. B. „unglücklich") und als Negation auf Ebene der Syntax (z. B. „nicht", „kein", etc.) erfolgen. Bezüglich der Wortnegation wurden die Präfixe „wider" und „miss" generell, d. h. auch bei lexikalisierten Wörtern, als negierende Vorsilbe angesehen. Auch lexikalisch ist eine derartige Negationswirkung möglich (z. B. durch das Verb „leugnen"). Lexikalisch implizite Negation, wie beim Verb „leugnen", wurde nicht erfasst. (vgl. Graefen & Liedke 2008: 109) Obwohl die Variable NEG zunächst mit den Subkategorien „Wortnegation", „Satznegation" und „Leere Negation" erfasst wurde, wurden diese Subkategorien aufgrund zu vernachlässigender Effekte in den Vorab-Analysen zusammengefasst und die Variable wurde dichotom („Negation tritt seltener als 2% auf" und "Negation tritt häufiger als 2% auf") kategorisiert.

Substantive/Eigennamen/Appellative (SUB):
Ermittelt wurde der prozentuale Anteil an Substantiven, Eigennamen und Appellativen in den Stimuli. Diese Variable ist dichotom mit einer Grenze bei 20%.

Verben (VER):
Die Variable VER beschreibt den prozentualen Anteil der Verben in den Stimuli, wobei diese Variable dichotom mit einer Grenze bei 15% Verwendung findet.

2.2.1.2. Merkmalsgruppe II: Präsentationsmerkmale

Länge in Minuten (LST):
Dieses Merkmal wurde dreistufig kategorisiert. Die Grenzen liegen bei 2 und 4 Minuten.

Wortzahl (AWS):
Die Wortzahl wurde in vier Stufen erfasst, wobei die Grenzen 200, 400 und 600 Wörter sind. Die syntaktischen Wörter wurden nach den Regeln von Gloy (1973) ermittelt.[8]

Anzahl der Sprecher (ASP):
Die Anzahl der Sprecher wurde in die vier Stufen „Ein Sprecher", „Zwei Sprecher", „Drei Sprecher" und „Mehr als drei Sprecher" kategorisiert. Da bei mehr als drei Sprechern kein gemeinsames Gespräch mehr stattfindet, sondern die Sprecher i. d. R. nacheinander befragt werden, werden mehr als drei Sprecher in eine Kategorie zusammengefasst.

Sprechgeschwindigkeit (SGS):
Die mittlere Sprechgeschwindigkeit wurde dichotom als die Anzahl der gesprochenen Wörter pro Sekunde kodiert. Als Grenze zwischen den Kategorien werden zwei Wörter pro Sekunde veranschlagt. Akzent/Dialekt/Aussprache (AST): Bei dieser Variable wurden „Standardsprache", „leichter regionaler Dialekt", „starker regionaler Dialekt" und „ausländischer Akzent" erfasst. Kein Stimulus ist durchgängig dialektal gefärbt oder mit ausländischem Akzent gesprochen ist. Ausprägungen dieser Variable wurden daher für einzelne Gesprächsbeiträge kodiert.

Anzahl der Stimuluspräsentation (AHO):
Bei der Variable AHO wird unterschieden, ob der Stimulus einmal oder zweimal vorgespielt wird.

---

8 1. Jede Buchstabengruppe, die durch ein Leerzeichen voneinander getrennt wurde, ist ein Wort.
2. Jede Zahl ist ein Wort.
3. Wörter mit Bindestrich zählen als ein Wort, wenn sie so in einem bestimmten Wörterbuch verzeichnet sind. Sonst zählt dieses Wort als zwei Worte.
4. Zwei Wörter sind unterschiedliche Wörter, wenn sie unterschiedlich geschrieben werden.
5. Zusammenziehungen werden als ein Wort gezählt.
6. Abkürzungen werden als ein Wort gezählt, und zwar ein anderes als die voll ausgeschriebene Version dieses Wortes.
7. Rechtschreibfehler werden korrigiert, außer die Fehlschreibung ist ein beabsichtigter Neologismus.

2.2.1.3. Merkmalsgruppe III: Inhaltlich-thematische Merkmale

Literarischer Stimulus (SLT):
Die Variable SLT erfasst dichotom, ob ein Stimulus literarisch oder nicht-literarisch ist.

Hörkontext (HKO):
Die Variable HKO erfasst die (Mach-)Art des Stimulus. Unterschieden wurden folgende Kategorien: „Interview" (auch wenn nur ein Sprecher zu hören ist, jedoch davon auszugehen ist, dass er auf eine Frage antwortet), „Hörspiel", „Reportage" (auch Kurzbeiträge und Podcasts), „Servicesendung" (wie z. B. Verkehrsfunk oder Wetterbericht), „vorgetragene Lyrik/Prosa", „nachgestellte Situation" (diese Beiträge wurden auf der Grundlage eines Skripts im Tonstudio aufgenommen, vermitteln jedoch den Eindruck von Unmittelbarkeit und Authentizität) und „Vortrag".

Funktion (TFU):
Die Variable TFU unterscheidet vier verschiedene Funktionen (vgl. Tabelle III-2.2.1.3.), die der Stimulus überwiegend haben kann, und zwar eine argumentative Funktion, eine informierende/beschreibende Funktion, eine narrative, erzählende Funktion und eine kommunikative Funktion.

Tabelle III-2.2.1.3.: Übersicht über die Einzelcodes der Variable TFU

Code	Kurzbeschreibung	Langbeschreibung
1	Argumentativ	argumentierende Äußerungen von Personen in verschiedenen Situationen, Pro und Contra zu einem Thema, Meinungen, formale Argumentationen, z. B. eine formelle Debatte
2	Informierend/beschreibend	Anordnungen oder Anweisungen, Ankündigungen, Kurzdefinitionen, Programmansagen im Radio, Zusammenfassungen, technische oder impressionistische Beschreibungen, Gestaltung von Räumen, äußeres Erscheinungsbild, etc.
3	Narrativ/erzählend	Geschichten, Witze, Anekdoten, Dokumentationen, etc.
4	Kommunikativ	Kommunikation herstellen, Plaudern, Smalltalk, etc.

Thema (THE):
Die IQB-Aufgaben wurden entsprechend den Themenvorschlägen des Gemeinsamen Europäischen Referenzrahmens eingestuft. (Europarat 2001: 51) Diese Themenvorschläge wurden jedoch z. T. erweitert und modifiziert, sodass weitere Kategorien zur Verfügung standen. Die IQB-Stimuli wurden nach folgenden 13 Themen eingeschätzt: „Information zur Person", „Orte", „Wohnen und Umwelt", „Pflanzen und Tiere", „Dienstleistungen", „Gesundheit und Hygiene", „Freizeit und Unterhaltung", „Sprache", „Ausbildung/Beruf", „Wetter", „Menschliche Beziehungen", „Tägliches Leben", „Essen und Trinken".

Hintergrundwissen (WEL):
Erfasst wurden in den Stimuli alle Begriffe und Formulierungen, für deren Verständnis Hintergrundwissen notwendig ist.

Tabelle III-2.2.1.4.: Übersicht über die Einzelcodes der Variable REL

Code	Codegrenze	Kurzbeschreibung	Langbeschreibung
REL1	10%	Frage/Impuls/ Themensetzung	Bei Radiobeiträgen z. B. auch die Überleitungen zu neuen Abschnitten der Sendungen, wie Verabschiedungen oder Zusammenfassungen, die eindeutig einen Übergang innerhalb der Sendung darstellen.
REL2	20%	Antwort	Auf eine gestellte Frage folgt eine Antwort, oder ein Problem wird vorgestellt und seine Lösung folgt.
REL3	10%	Spezifizierung	Nach einer allgemeineren Darstellung werden spezifische Informationen gegeben.
REL4	10%	Erklärung/Beweis/ Ursache	Für einen Sachverhalt wird eine Erklärung gegeben. Zur Unterstützung einer Aussage werden Beweise vorgelegt. Ein Ereignis wird als Ursache eines anderen Ereignisses dargestellt.
REL5	30%	Reihenfolge/ Aufzählung	Argumente werden in ihrer zeitlichen Abfolge zusammenhängend dargeboten. Sachverhalte werden in loser Struktur aufgeführt.
REL6	20%	Ziel/Bedingung	Ein Ereignis wird als Ziel eines anderen Vorgangs dargestellt.

2.2.1.4. Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte

Relationstypen (REL): Die Variable REL umfasst die in Tabelle III-2.2.1.4. dargestellten Teilbereiche, die für jede Sprecheinheit untersucht wurden. Die einzelnen Codes flossen dichotom mit den angegebenen Grenzen in die Analysen ein.

Schlussfolgerungen/Inferenzen (SFI):
Diese Variable beschreibt, wie hoch der prozentuale Anteil an Begriffen oder Formulierungen in den Stimuli im Verhältnis zur Wortzahl ist, die eine Schlussfolgerung oder eine Inferenz von Seiten des Zuhörers erfordern. Die Grenze bei dieser dichotomen Variable liegt bei 2% Kodierungen Inferenzen im Verhältnis zur Anzahl der Sätze.

Anzahl der Propositionen (PRO):
Die Variable PRO beschreibt die Anzahl der Propositionen in einem Stimulus in drei Stufen, wobei die Grenze zwischen den Stufen bei 50 und 100 Propositionen liegt.

Anteil der Propositionen (PRW):
Die Variable PRW macht Angaben über den Anteil an Propositionen in einem Stimulus und wird berechnet aus dem Quotient aus der Wortzahl pro Stimulus und der Anzahl der Propositionen. (PRO/AWS)

Länge der Propositionen (MLP):
Die Variable MLP erfasst die mittlere Länge der Propositionen eines Stimulus in vier Stufen: weniger als 5 Wörter, 5 bis 6 Wörter, 6 bis 7 Wörter und mehr als 7 Wörter.

2.2.1.5. Merkmalsgruppe V: Globalurteil

Stimulusschwierigkeit (TSA):
Die Stimuli wurden von Aufgabenentwicklern in die Kategorien „leicht“,„mittel“ und „schwierig“ eingestuft.

2.2.2. Stimulusmerkmale aus dem Lehrerfragebogen

Für die folgenden Variablen wurde ein Fragebogen entwickelt, mit dem die Deutschlehrkräfte der getesteten Schüler eine Einschätzung der Stimuli auf einer fünfstufigen Likert-Skala abgeben konnten. Der Fragebogen fokussiert einerseits auf der Vertrautheit der Schüler mit den Stimuli (Vertrautheit mit dem Thema) und andererseits auf Kategorien, zu denen auch quantitative Analysen durchgeführt werden, wie Wortschatz, Grammatik und Kohärenz. Die Lehrkräfte hörten zusammen mit den Schülern die Stimuli von CD und hatten, ebenso wie die Schüler, nicht die Möglichkeit die verschriftlichte Version der Stimuli einzusehen.

Aufgabe: _____

(An dieser Stelle wird der Titel der jeweiligen Hörverstehensaufgabe eingetragen, die die Schüler jeweils bearbeiten. Zu jeder Hörverstehensaufgabe werden unten stehende Fragen gestellt.)

Bitte hören Sie zunächst zusammen mit den Schülerinnen und Schülern den Hörtext an. Schätzen Sie dann ein, wie der Hörtext wahrscheinlich hinsichtlich der folgenden Kriterien von den Schülern wahrgenommen wird.

		1	2	3	4	5	
Vertrautheit mit dem Thema	vertraut	☐	☐	☐	☐	☐	gar nicht vertraut
Wortschatz	abstrakt	☐	☐	☐	☐	☐	konkret
	fachspezifisch	☐	☐	☐	☐	☐	alltäglich
Grammatik	schwierig	☐	☐	☐	☐	☐	einfach
Kohärenz/ Textzusammenhang*	stark ausgeprägt	☐	☐	☐	☐	☐	wenig ausgeprägt
Gesamteindruck	interessant	☐	☐	☐	☐	☐	uninteressant
	elegant	☐	☐	☐	☐	☐	unbeholfen
	abwechslungsreich	☐	☐	☐	☐	☐	eintönig
Ton	persönlich/gefühlsbetont	☐	☐	☐	☐	☐	unpersönlich/sachlich
Ausdrucksweise	gewählt	☐	☐	☐	☐	☐	umgangssprachlich
	komplex/ausschweifend	☐	☐	☐	☐	☐	einfach/knapp
Informationsebene	spitzfindig/tiefgründig	☐	☐	☐	☐	☐	offensichtlich/oberflächlich
Wirkung	regt zum Nachdenken an	☐	☐	☐	☐	☐	regt zum Lachen an
	ernsthaft	☐	☐	☐	☐	☐	humorvoll

* Kohärenz kann zum Beispiel durch ein einheitliches Thema, Kongruenz im Tempus, kausale Verknüpfungen oder grammatische Verknüpfungsmöglichkeiten, wie textverknüpfende Konjunktionen oder Wiederaufnahmen erzeugt werden.

Abbildung III-2.2.2.: Lehrerfragebogen

2.2.3. Itemmerkmale aus den IQB-Ratings
Auch für die Itemmerkmale lassen sich analog zu den Stimulusmerkmalen Variablengruppen bilden, die auf einer inhaltlichen und thematischen Ähnlichkeit beruhen.

Gruppe I: Itemformat:
„Itemformat/Ausfüllleistung (IFA)", „Itemformat/Kodierleistung (IFK)"

Gruppe II: Merkmale der Itempräsentation:
„Zeitpunkt der Itembearbeitung (ZIB)", „Position des Items innerhalb der Aufgabe (PIA)"

Gruppe III: Merkmale von MC-Items:
„Position des Attraktors im MC-Item (PMC)", „Größte vorkommende Plausibilität der Distraktoren (PDI)", „Mittlere Plausibilität der Distraktoren (MPD)"

Gruppe IV: Kognitive Anforderungen der Items:
„Anforderungsbereich (AFB)", „Geprüfter Standard (STA)", „Anzahl der benötigten NI pro Item (ANI)", „Auftretenshäufigkeit der NI (ARN)", „Position der NI auf Stimulusebene (PST)", „Wortzahl der NI (WNI)", „Konkretheit der NI (TCO/TOR)", „Hintergrundwissen (HGW)", „Typ der NI (TNI")

Gruppe V: Globalurteil:
„Eingeschätzte Itemschwierigkeit durch die Aufgabenentwickler (SEA)"

2.2.3.1. Merkmalsgruppe I: Itemformat

Itemformat (IFK) und (IFA):
Bei der Untersuchung des Itemformats wurden zwei Variablen eingesetzt. Dies ist zum einen die Variable Itemformat IFK, die auf den Eingaben der IQB-Datenbank beruht (vgl. Tabelle III-2.2.3.1a.) und zum anderen die Variable IFA, die auf dem Offenheitsgrad der Items beruht (vgl. Tabelle III-2.2.3.1b.). Bei der Variable IFK werden die Items nach dem Grad ihres Kodieraufwands eingeschätzt. Bei den Codes GA und GK wird die Schülerantwort stark gelenkt und es besteht lediglich die Möglichkeit ein Kreuz bei verschiedenen Antwortoptionen zu setzen (GA) oder die Antwort aus einer vorgegebenen Auswahl auszuwählen (GK). Bei Code 0/1 wird vom Schüler eine Kurzantwort von meist drei bis fünf Wörtern erwartet. Bei Code 1P kann der Schüler etwas ausführlicher formulieren, wobei auch hier Rechtschreibung und Grammatik für die Beurteilung der Antwort keine Rolle spielen.

Eine zweite Einschätzung mittels der Variable IFA beruht auf dem Offenheitsgrad der Items und differenziert noch genauer zwischen den einzelnen Itemtypen (vgl. Tabelle III-2.2.3.1b.). Diese Variable drückt stärker bereits die kognitive Anforderung aus, die das Item an die Schüler bei der Beantwortung stellt.

Tabelle III-2.2.3.1a.: Übersicht über die Einzelcodes der Variable IFK

Itemformat (IFK)		
Code	**Kurzbeschreibung**	**Langbeschreibung**
IFK.GA	Geschlossen – Ankreuzen	Geschlossenes Antwortformat, das über Computer (Schwärzungsgrad) ausgewertet werden kann
IFK.GK	Geschlossen – Kodieren	Geschlossenes Antwortformat, das durch eine Person kodiert werden muss; z. B.: Zuordnungsaufgaben, Unterstreichen
IFK.01	Einfache Antwort - 0/1 Kodierung	Kurze Antwort, die sich sehr schnell mit 0 oder 1 kodieren lässt; z. B.: Zahl, Datum, Stadt usw.
IFK.1P	Antwort mit mehreren Codes	Komplexere Antwort, die Interpretationsspielraum bei der Auswertung bietet und eine Kodiererschulung erfordert

Tabelle III-2.2.3.1b.: Übersicht über die Einzelcodes der Variable IFA

Itemformat (IFK)	
Code	**Codebeschreibung**
IFA.MC	Multiple-Choice-Item
IFA.RF	Richtig-Falsch-Item
IFA.ZO	Zuordnungs-Item
IFA.RI	Reihenfolge-Item
IFA.HO	Halboffenes Item
IFA.OI	Offenes Item

2.2.3.2. Merkmalsgruppe II: Merkmale der Itempräsentation

Zeitpunkt der Itembearbeitung (ZIB):
Der Zeitpunkt der Itembearbeitung wird dichotom kategorisiert, und zwar mit den Ausprägungen „Itembearbeitung während des Hörens“ und „Itembearbeitung nach dem Hören“.

Position des Items innerhalb der Aufgabe (PIA):
Bei der Variable PIA wurden die Einzelcodes wie in der Tabelle III-2.2.3.2. zusammengefasst. Grundlage für die Zusammenfassung, war die Überlegung, dass nicht die genaue Position des Items (z. B. an fünfter Stelle) ausschlaggebend sein dürfte, sondern vielmehr seine Positionierung eher zu Beginn einer Aufgabe, in der Mitte oder zum Ende einer Aufgabe. Items an Position 1 besitzen eine Sonderrolle, indem sie den Einstieg in eine Aufgabe darstellen. Items in dieser Position werden demnach getrennt erfasst. Eine fünfte Kategorie gibt Aufschluss über sehr lange Aufgaben mit mehr als 8 Items. Zum Teil tauchen Items bis in Position 17 auf. Kategorie 5 wird insgesamt jedoch sehr schwach besetzt.

2.2.3.3. Merkmalsgruppe III: Merkmale von MC-Items

Position des Attraktors im MC-Item (PMC):
Bei der Variable PMC wird erfasst, ob der Attraktor im Multiple-Choice-Item an erster, zweiter, dritter oder vierter Stelle steht.

Tabelle III-2.2.3.2.: Übersicht über die Einzelcodes der Variable PIA

Position des Items innerhalb der Aufgabe (PIA)	
Code	**Codebeschreibung**
1	An 1. Stelle
2	Zu Beginn der Aufgabe: An 2./3. Stelle
3	In der Mitte der Aufgabe: An 4./5. Stelle
4	Zum Ende der Aufgabe: An 6. – 8. Stelle
5	Ab 9. Stelle

Plausibilität der Distraktoren (PDI) und (MPD):
Die Distraktoren bei MC-Items sollen von der richtigen Antwort ablenken, indem sie zwar eindeutig falsch sind, aber Schülern, die die richtige Antwort nicht kennen, dennoch als plausible Antwortmöglichkeit erscheinen. Bei den Analysen wurde zwischen zwei unterschiedlichen Varianten unterschieden: Zunächst wurde die größte vorkommende Plausibilität der Distraktoren (PDI) untersucht. Es wird davon ausgegangen, dass wahrscheinlich der attraktivste Distraktor vom Attraktor ablenkt. Die entsprechenden Einzelcodes sind in Tabelle III- 2.2.3.3. dargestellt. In einer zweiten Analyse wurde die mittlere Plausibilität der Distraktoren (MPD) ermittelt, um einen Eindruck davon zu erhalten, wie attraktiv die Distraktoren insgesamt im Vergleich zum Attraktor sind. Die Plausibilität der Distraktoren wurde für alle Items im MC4-Format (d. h. Multiple-Choice-Items mit vier Ankreuzmöglichkeiten) von zwei Ratern, Rater 1 und Rater 2, unabhängig voneinander in die in der Tabelle beschriebenen Kategorien eingeschätzt. Distraktoren mit dem Code 1 können am leichtesten von Schülern, die die richtige Antwort kennen, als falsch identifiziert werden. Distraktoren mit den Codes 2 und 3 sind in etwa gleich schwierig als falsch zu identifizieren und könnten ggf. auch zusammengefasst werden. Für die folgenden Analysen liegen die Einschätzungen von Rater 2 (der Autorin dieser Arbeit) zugrunde.

Tabelle III-2.2.3.3.: Übersicht über die Einzelcodes der Variable PDI

Größte vorkommende Plausibilität der Distraktoren (PDI)		
Code	**Kurzbeschreibung**	**Langbeschreibung**
1	Distraktoraussage im Stimulus: wenig plausibel	Die Distraktoraussage im Stimulus ist nur fälschlich durch eine Verbindung mit dem Weltwissen ins mentale Modell des Stimulus integrierbar.
2	Distraktoraussage nicht im Stimulus: plausibel	Die Distraktoraussage ist nicht im Stimulus, jedoch durch eine Verbindung mit dem Weltwissen passend ins mentale Modell des Stimulus integrierbar.
3	Distraktoraussage im Stimulus: abwegig	Die Distraktoraussage im Stimulus ist offensichtlich abwegig.
4	Distraktoraussage im Stimulus: plausibel	Die Distraktoraussage im Stimulus ist nicht auf den ersten Blick und ohne kritische Abwägung der anderen Alternativen im Hinblick auf die Fragestellung als falsch einzuschätzen.

2.2.3.4. Merkmalsgruppe IV: Kognitive Anforderungen der Items

Anforderungsbereich (AFB):
Die Variable AFB erfasst, welchem in den Bildungsstandards beschriebenen Anforderungsbereich die kognitive Operation zuzurechnen ist, die von einem Item verlangt wird. Dabei werden die folgenden drei Anforderungsbereiche unterschieden: Anforderungsbereich I: „Wiedergeben", Anforderungsbereich II: „Anwenden" und Anforderungsbereich III: „Reflektieren und Beurteilen".

Geprüfter Standard (BS):
Durch die Variable BS wird ausgedrückt, welcher Bildungsstandard mit den Items abgedeckt wird. Zur Variable STA existieren Einzelcodes, wie in Tabelle III-2.2.3.4a. dargestellt.

Tabelle III-2.2.3.4a.: Übersicht über die Einzelcodes der Variable BS

Geprüfter Standard (BS)	
Code	**Codebeschreibung**
BS141	Gesprächsbeiträge anderer verfolgen und aufnehmen
BS142	Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, sichern und wiedergeben
BS143	Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln
BS113	Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden

Anzahl der benötigten NI pro Item (ANI):
Bei der Variable ANI wird erfasst, wie viele Informationen aus dem Stimulus für die Beantwortung eines Items notwendig sind (vgl. Tabelle III-2.2.3.4b.).

Tabelle III-2.2.3.4b.: Übersicht über die Einzelcodes der Variable ANI

Anzahl der benötigten NI pro Item (ANI)	
Code	**Codebeschreibung**
0	Es wird eine NI benötigt, die weder im Stimulus noch im Item auftaucht.
1	Es wird eine NI benötigt, die im Stimulus gegeben wird.
2	Es werden zwei NIs benötigt, die im Stimulus gegeben werden.
3	Es werden drei oder mehr NIs benötigt, die im Stimulus gegeben werden.
4	Es wird eine NI benötigt, die sich nicht lokalisieren lässt, z. B. eine globale Einschätzung des Beitrags, aber auch Sprecheridentifikationen, Geräusche oder die Bestimmung von Tonfällen der Sprecher.
5	Es wird eine NI benötigt, die aber bereits durch das Item gegeben wird.

Auftretenshäufigkeit der NI (ARN):
Für die Variable ARN wurden Codes vergeben, wie in der Tabelle III-2.2.3.4c. dargestellt. Bei einigen Items wurden mehrere Informationen zur Beantwortung benötigt. In diesen Fällen wurden auch mehrere Auftretenshäufigkeiten angegeben, die durch Unterstriche voneinander getrennt wurden. Die Codevergabe erfolgte chronologisch: Die erste NI erhielt den ersten Code, etc. Es werden maximal drei NIs zur Beantwortung eines Items benötigt. Diese Variablen mit Mehrfachinformationen wurden in die Codes 1 – 4 integriert.

Tabelle III-2.2.3.4c.: Übersicht über die Einzelcodes der Variable ARN

Auftretenshäufigkeit der NI (ARN)	
Code	**Codebeschreibung**
0	NI kommt nicht im Stimulus vor.
1	NI kommt einmal vor (ist also noch nicht redundant).
2	NI kommt zweimal vor.
3	NI kommt dreimal oder häufiger vor.
4	Die Redundanz der NI lässt sich nicht bestimmen, da es sich bei der NI um eine globale Einschätzung des Beitrags handelt.

Position der NI auf Stimulusebene (PST):
Die Position der NI wird auf Stimulusebene in drei Stufen erfasst: Die NI steht am Anfang des Stimulus, die NI steht in der Mitte des Stimulus und die NI steht am Ende des Stimulus.

Wortzahl der NI (WNI):
Die Zahl der Wörter der einzelnen NIs wird in vier Kategorien mit den folgenden Abstufungen angegeben: weniger als 20 Wörter, 21 bis 40 Wörter, 41 bis 60 Wörter und mehr als 60 Wörter.

Konkretheit der NI – 33-stufig (TCO) und 5-stufig (TOR):
Beide Variablen geben Aufschluss über die Art der NI, der zur Lösung des Items notwendigen Information, wobei die Variable TOR eine Zusammenfassung der TORI-Codestufen (”Type of Requested Information”) (Evetts & Gauthier, 2005) ist. Der im originalen TORI-Schema verwendete Code „Orte/Relationen“ wurde in dieser Kategorisierung nicht übernommen. Tabelle III-2.2.3.4d. gibt einen Überblick über die Einzelcodes der Variable TCO.

Tabelle III-2.2.3.4d.: Übersicht über die Einzelcodes der Variable TCO

Konkretheit der NI (TCO)			
Code	**Codebeschreibung**	**Code**	**Codebeschreibung**
1	Personen	**17**	Verifikationen
2	Gruppen/Tiere/Orte	**18**	Stil- und Formelemente des Beitrags
3	Dinge	**19**	Soziale Interaktionen/Zwischenmenschliche Beziehungen
4	Mengenangaben	**20**	Emotionen und Verhaltensbewertungen
5	Zeiten	**21**	Ursache - Effekt/Problem- und Lösungsaufbau
6	Attribute	**22**	Erklärungen/Begründungen/Hintergründe
7	Typen/Arten	**23**	Behauptungen
8	Aktionen/Versuche	**24**	Beweise
9	Art und Weise	**25**	Meinungen
10	Bedingungen	**26**	Ähnlichkeiten
11	Zwecke/Ziele	**27**	Gesamteinschätzung des Beitrags
12	Funktionen	**28**	Thema
13	Kriterien	**29**	Muster-Vorhersage
14	Probleme	**30**	Prozess-Ablauf/Folgen/Sequenzen
15	Lösungen	**31**	Äquivalente/Paraphrasen
16	Pronominale Referenzen/Grammatische Erscheinungen	**32**	Unterschiede

Die Variable TOR fasst die Einzelcodes der Variable TCO wie folgt zusammen: „Hoch konkrete Information (TORICODE 1 - 3)“, „Weniger konkrete Information (TORICODE 4 - 9)“, „Abstrakte Information (TORICODE 10 - 19)“, „Sehr abstrakte Information (TORICODE 20 - 27)“ und „Hoch abstrakte Information (TORICODE 28 - 32)“.

Hintergrundwissen (HGW):
Erfasst wurden in den Items alle Begriffe und Formulierungen für deren Verständnis Hintergrundwissen notwendig ist.

Typ der NI (TNI):
Bei dieser Variable (TNI) wird erfasst, welche Information für die Beantwortung eines Items notwendig ist. Für die Variable TNI wurden ursprünglich 23 verschiedene Codes vergeben. Für jede NI wurde angegeben, welche Art der Information benötigt wurde. Werden mehrere NIs pro Item verlangt, so wurde beispielsweise ein Code 1_5 vergeben. Die erste NI entspricht dem Typ 1, wohingegen die zweite NI dem Typ 5 entspricht. In allen Fällen waren nur sehr wenige Items von einem derartigen Sammelcode betroffen. Diese Sammelcodes wurden deshalb jeweils in einen Hauptcode integriert, und zwar in den Code, der jeweils zuerst vergeben wurde. Der Code 1_5 wurde also in den Code 1 integriert. Die verbleibenden elf Codes, mit denen im weiteren Verlauf dieser Arbeit gerechnet wurde, sind in Tabelle III-2.2.3.4e. dargestellt.

Tabelle III-2.2.3.4e.: Übersicht über die Einzelcodes der Variable TNI

Typ der NI (TNI)	
Code	**Codebeschreibung**
1	Leitidee/Kernaussage
2	Detail
3	Meinung des Zuhörers
4	Stimmung/Einstellung des Sprechers
5	Schlussfolgerung
6	Kommunikativer Zweck/Funktion/Wirkung des Stimulus
7	Struktur des Stimulus/Zusammenhang zwischen Teilen
8	Genre des Stimulus
9	Geräusche/Paraverbales
10	Sprachliche Mittel
11	Sprecheridentifikation

2.2.3.5. Merkmalsgruppe V: Globalurteil

Itemschwierigkeit (SEA):
Die Items wurden von den Aufgabenentwicklern in die Kategorien „leicht", „mittel" und „schwierig" eingestuft.

2.2.4. Personenmerkmale

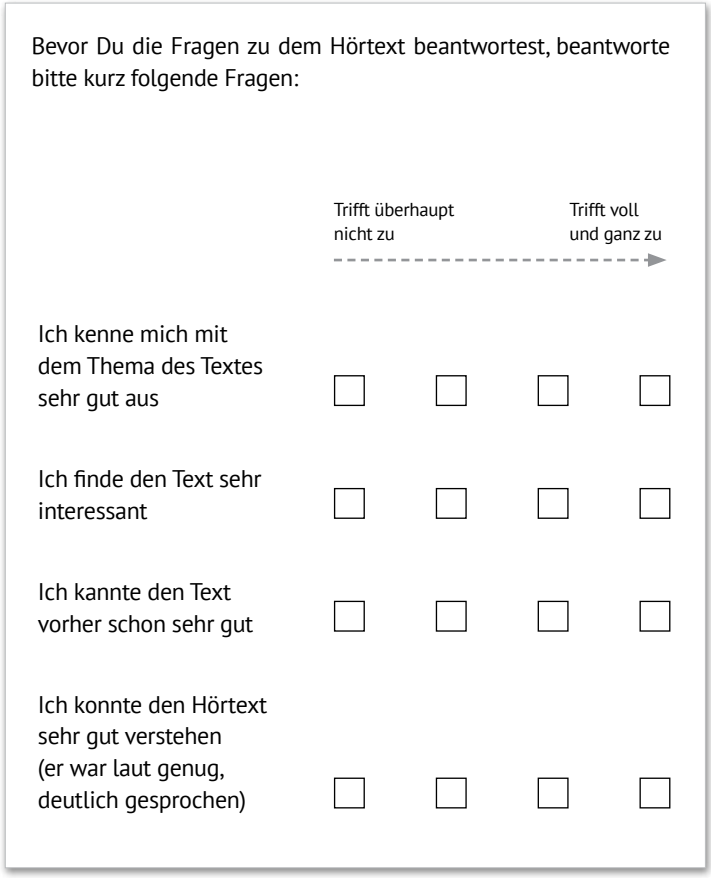
2.2.4.1. Einschätzung der Aufgaben durch die Schüler
Die folgenden Variablen wurden von den Schülern während des Tests im Testheft auf einer fünfstufigen Likert-Skala mit den Polen „Trifft überhaupt nicht zu" – „Trifft voll und ganz zu" bewertet. Die Schüler schätzten die folgenden Aussagen auf ihr Zutreffen hin ein. Dabei wird in der Befragung für alle Stimuli der Begriff „Text" verwendet, da der Begriff „Diskurs" möglicherweise nicht allen Schülern an allen Schularten vertraut ist. Abbildung 2.2.4.1. zeigt das im Testheft abgedruckte Item.

Vertrautheit mit dem Thema (VTS):
Aussage: „Ich kenne mich mit dem Thema des Textes sehr gut aus."

Motivation/Interesse der Schüler (MOT):
Aussage: „Ich finde den Text sehr interessant."

Bekanntheitsgrad des Stimulus (BEK):
Aussage: „Ich kannte den Text vorher schon sehr gut."

Verständlichkeit (VST):
Aussage: „Ich konnte den Hörtext sehr gut verstehen (er war laut genug, deutlich gesprochen)"


Bevor Du die Fragen zu dem Hörtext beantwortest, beantworte bitte kurz folgende Fragen:

	Trifft überhaupt nicht zu			Trifft voll und ganz zu
Ich kenne mich mit dem Thema des Textes sehr gut aus	☐	☐	☐	☐
Ich finde den Text sehr interessant	☐	☐	☐	☐
Ich kannte den Text vorher schon sehr gut	☐	☐	☐	☐
Ich konnte den Hörtext sehr gut verstehen (er war laut genug, deutlich gesprochen)	☐	☐	☐	☐

Abbildung III-2.2.4.1.: Einschätzung der Aufgaben durch die Schüler im Testheft

2.2.4.2. Arbeitsgedächtniskapazität

Um die relevanten Einflussfaktoren voneinander abgrenzen zu können, wurde für den Aufmerksamkeitstest ein Format mit Ziffern gewählt, sodass kein Wortschatzwissen erforderlich ist. Den Schülern werden zunächst drei bis acht Ziffern vorgesprochen, die sie nach dem Zuhören in aufsteigender Reihenfolge notieren sollen. Ein Beispiel des Aufmerksamkeitstests befindet sich in Anhang B.

2.2.4.3. Sprachkenntnisse

Um den Grad der Deutschkenntnisse der Schüler zu erheben, wurden sie gebeten im Schülerfragebogen Auskunft zu folgenden zwei Fragen zu geben:

1. Wie oft sprichst du zu Hause deutsch? („Ich spreche zu Hause immer deutsch" – „Ich spreche zu Hause manchmal deutsch und manchmal eine andere Sprache" – „Ich spreche zu Hause niemals deutsch")

2. Welche Sprache hast du in deiner Familie zuerst gelernt (Muttersprache)? Wenn du zweisprachig aufgewachsen bist, kannst du beide Sprachen ankreuzen. („Deutsch" – „Bosnisch" – „Französisch" – „Englisch" – „Griechisch" – „Italienisch" – „Kroatisch" – „Polnisch" – „Russisch" – „Serbisch" – „Türkisch" – „Kurdisch" – „Eine andere Sprache")

Da für diese Untersuchung nicht von Bedeutung ist, welche Sprache die Schüler als erstes gelernt haben, werden die Angaben zur zweiten Frage folgendermaßen zusammengefasst: „Deutsch" – „nicht Deutsch" – „zwei Sprachen angekreuzt"

3. Verwendete Forschungsmethoden

Für die vorliegende Arbeit stellen korrelations- und regressionsanalytische Methoden in Kombination mit explorativen und konfirmatorischen Faktorenanalysen wichtige Informationsquellen dar. In diesem Kapitel sollen die verwendeten Forschungsmethoden kurz erläutert werden. Zunächst wird ein Einblick in die theoretischen Grundlagen gegeben, wonach dann die einzelnen Methoden beschrieben werden. In Kapitel 3.1.2. werden die deskriptiven Analysen vorgestellt und Kapitel 3.1.3. ist der Item-Response-Theory (IRT) gewidmet. Es werden im Anschluss daran die Analyse von Verteilungen (Kapitel 3.1.4.), die Korrelationsanalysen (Kapitel 3.1.5.), Regressionsanalysen (Kapitel 3.1.6.) sowie Dimensionsanalysen (Kapitel 3.1.7.) beschrieben. Ein Kapitel (3.1.8.) zu den Verfahren zur Erfassung der Beurteilerübereinstimmung beendet den Überblick über die verwendeten Methoden.

3.1. Theoretische Grundlagen

In Leistungstests soll vom beobachteten Verhalten in der Testsituation auf die Fähigkeiten und Kompetenzen der Testperson geschlossen werden. Dabei wird nach Hartig unter Kompetenz das Ergebnis der erfolgreichen Interaktion individueller Fähigkeiten mit spezifischen situationalen Anforderungen verstanden. (vgl. Hartig, 2008: 70) In psychometrischen Modellen kann erfolgreiches Verhalten in Testsituationen unter Berücksichtigung relevanter Fähigkeiten der Testperson sowie relevanter Itemanforderungen modelliert werden. Der Begriff „Itemanforderungen" bezieht sich dabei auf Itemmerkmale, die Anforderungen außerhalb der Testsituation repräsentieren, für welche die getesteten Kompetenzen relevant sind. Ferner wird davon ausgegangen, dass die Itemanforderungen einen Einfluss auf das Lösen der Items haben. Die Itemanforderungen können entweder durch Merkmale des Stimulus (z. B. die Komplexität des Wortschatzes) oder durch die Prozesse beschrieben werden, die zum Lösen des Items notwendig sind. Unter dem Begriff „Fähigkeiten" versteht Hartig einen Sammelbegriff für die individuellen Unterschiede in den kognitiven Ressourcen der Testpersonen, die zum Lösen der Items notwendig sind, wie beispielsweise spezifisches Wissen oder Fähigkeiten zur Durchführung spezifischer kognitiver Operationen. (vgl. Hartig, 2008: 71)

Nach Hartig (2008) sind Konstrukte in der empirischen Bildungsforschung häufig stark an konkrete Situationen gebunden, also situationsspezifisch. Kane (2001) schlägt dazu eine Unterscheidung in „observable attributes" und „theoretical constructs" vor. Unter „observable at-

tributes" versteht er beobachtbare Ausprägungen eines bestimmten Verhaltens. „Theoretical constructs" sind dagegen theoriegeleitet definiert. Gemäß Borsboom et al. (2004) sind diese „theoretical constructs" in reflektiven Messmodellen prüfbar, wobei angenommen wird, dass Änderungen im latenten Konstrukt Änderungen in den Manifestationen (Itemantworten) verursachen.

Unterscheiden sich Items in ihren Anforderungen, so kann dies einerseits Auswirkungen auf die Itemschwierigkeit und andererseits auf die Dimensionalität besitzen. Änderungen in der Itemschwierigkeit beschreiben dabei eine Auswirkung der Itemanforderung, die im Mittel für alle Personen gilt. Andererseits kann sich dadurch auch die Relation zwischen den Antworten verändern, die sich als Korrelation zwischen den erzielten Ergebnissen zu unterschiedlichen Items manifestiert. (vgl. Hartig, 2008: 72) In diesem Fall differenzieren die Itemanforderungen zwischen den Personen und wirken damit nicht nur homogen für alle Personen auf die Itemschwierigkeit, sondern auch heterogen zwischen Personen. So werden mehrdimensionale IRT-Modelle zur Beschreibung zwischen den Itemanforderungen in den entsprechenden Dimensionen erforderlich. Gegenüber den vielfach eingesetzten eindimensionalen IRT-Modellen ist hierbei ein komplexeres Modell im Sinne eines kognitiv-diagnostischen Modells (DiBello et al., 2007) notwendig, während die Auswirkungen auf Änderungen in der Itemschwierigkeit einem item-diagnostischen Ansatz (von Davier, 2009) entsprechen.

Für die meisten durchgeführten Analysen wurde das Programm SPSS 15 (SPSS Inc. 2007) verwendet. Die Auswertung der Daten auf der Basis von Item-Response-Modellen erfolgte mithilfe der Software *ConQuest* (vgl. Wu et al., 2007) Damit ist die simultane Schätzung von Rasch-Modellen für die Teilpopulationen und die Schätzung eines Gesamtmodells für die gesamte Population möglich. Dabei sind vor allem die Itemschwierigkeiten in der Gesamtpopulation und den Teilpopulationen, aber auch die durchschnittliche Leistungsfähigkeiten der einzelnen Populationen von Interesse. (vgl. Knoche & Lind, 2005)

Bei der Skalierung wurden verschiedene Varianten erprobt, wie Aufgaben, bei denen mehrere Einzelfragen zu einem Item zusammengefasst werden, am besten zu bepunkten sind. Für Zuordnungsaufgaben, die aus zwei bis drei Teilfragen bestehen, bei denen „ja" oder „nein" bzw. „stimmt" oder „stimmt nicht" anzukreuzen ist, wurde nur dann ein Punkt vergeben, wenn alle Teilantworten bzw. Variablen richtig beantwortet wurden. Dies ist dadurch begründet, dass die Ratewahrscheinlichkeit für eine Frage mit zwei Antwortalternativen („ja"/"nein") 50% beträgt und eine Zusammenfassung mehrerer Fragen dieses Typs die Ratewahrscheinlichkeit deutlich reduziert. Außerdem wurden Items mit einer relativen Lösungshäufigkeit kleiner 0.05 und größer 0.95 in der Skalierung nicht berücksichtigt. Auch Schüler, die im gesamten Testheft keine Items beantwortet haben sowie Items, die von keinem Schüler bearbeitet wurden, wurden von der Skalierung ausgeschlossen. Zur Berechnung der eigentlichen Item- und Personenparameter wurden die Mittelwerte der Personenfähigkeiten auf den Wert null fixiert. Die Verortung der Skala ist im Raschmodell ein freier Parameter und muss deshalb fixiert werden. Üblich ist das Fixieren der mittleren Personenfähigkeit oder der mittleren Itemschwierigkeit auf den Wert null.

3.2. Deskriptive Itemanalysen

Um Items und Testwertverteilungen deskriptivstatistisch evaluieren zu können, müssen zunächst einige Voranalysen unternommen werden. Dazu gehört beispielsweise die Analyse der empirischen Lösungshäufigkeiten, der Itemschwierigkeiten und der Trennschärfen. Bei Leistungstests ohne Zeitbegrenzung werden die empirischen Lösungshäufigkeiten durch den Schwierigkeitsindex *pi* gekennzeichnet. Er gibt an, wie groß der Anteil der Probanden ist, die das Item richtig lösen. *pi* wird berechnet aus der tatsächlich erreichten Punktsumme aller Probanden und der maximal erreichbaren Punktsumme aller Probanden. *pi* wird für jedes Item einzeln berechnet und kann zwischen 0 und 1 liegen. Je größer der Wert, desto leichter gilt das Item. Um Personenunterschiede sichtbar zu machen, gelten einerseits Items im mittleren Schwierigkeitsbereich als ideal, andererseits wird versucht, Items mit einer möglichst breiten Schwierigkeitsstreuung einzusetzen. (Bortz & Döring, 2002: 218) Die Lösungshäufigkeit eines Items ist der prozentuale Anteil der richtigen Antworten. In der Regel sollte die Lösungshäufigkeit eines Items zwischen 5% bis 95% liegen.

Durch den Trennschärfekoeffizienten *r* wird deutlich, wie gut das gesamte Testergebnis aufgrund der Beantwortung eines einzelnen Items vorhergesagt wird. Die Trennschärfe beschreibt, wie gut ein Item in der Lage ist, zwischen starken und schwachen Merkmalsausprägungen zu unterscheiden. Personen mit einer hohen Punktzahl im Gesamtergebnis weisen demnach auch eine hohe Punktzahl in einzelnen trennscharfen Items auf und umgekehrt. Es wird angenommen, dass Schüler mit hoher Gesamtpunktzahl über stark ausgeprägte Merkmale bzw. über hohe Fähigkeiten auf der zu messenden Skala verfügen und dementsprechend eher in der Lage sind, auch die einzelnen Items richtig zu lösen. Die Trennschärfe drückt also die Korrelation der Beantwortung eines Items mit dem Gesamtwert aus und muss für jedes Item einzeln berechnet werden. Die Trennschärfe kann einen Wert zwischen -1 und +1 annehmen. Ein schwieriges Item mit hoher Trennschärfe, d. h. einem großen positiven Wert, kann von leistungsstarken Schülern mit einer höheren Wahrscheinlichkeit gelöst werden als von leistungsschwachen Schülern. Eine positive hohe Trennschärfe besagt auch, dass das entsprechende Item etwas sehr Ähnliches misst wie der Gesamttest. Liegt die Trennschärfe bei null - als Grenzwert wird dabei meist der Wert 0.2 betrachtet - ist das Item nicht dazu geeignet, zwischen Schülern mit hohem Testwert und mit niedrigem Testwert zu unterscheiden. (vgl. Lienert & Raatz, 1998: 26f)

Theoretisch sind Lösungshäufigkeit/Itemschwierigkeit und Trennschärfe unabhängig voneinander. Empirisch aber zeigt sich zwischen beiden eine umgekehrt U-förmige Beziehung. Daraus lässt sich schließen, dass die höchste Trennschärfe erzielt wird, wenn die Lösungshäufigkeit bei ungefähr 50% liegt, wenn also das Item eine mittlere Schwierigkeit aufweist.

3.3. Item-Response-Theory (IRT)

Die klassische Testtheorie (KTT) dient dazu, Merkmalsunterschiede zwischen Personen exakt und ökonomisch zu erfassen. Sie geht davon aus, dass sich der Messwert einer Person in einem Testitem immer aus der wahren Ausprägung des untersuchten Merkmals und einem zufälligen Messfehler zusammensetzt. Mithilfe der KTT kann auf Basis der Reaktionen in mehreren Items die wahre Ausprägung des Merkmals geschätzt und damit die Messgenauigkeit des

Testergebnisses bestimmt werden. Allerdings sind die Kennwerte der KTT (z. B. Itemschwierigkeit oder Itemtrennschärfe) stichprobenabhängig und die gefundenen Ergebnisse können nur bedingt verallgemeinert werden. Dieser Nachteil wird von der Item-Response-Theorie (IRT) überwunden. Sie beruht auf der Annahme, dass es einen probabilistischen Zusammenhang zwischen den Merkmalsausprägungen und dem beobachtbaren Messwert gibt. Die IRT beschreibt also das Reaktionsverhalten der Probanden in Abhängigkeit von Personen- und Itemparametern. Die IRT erlaubt es, die statistischen Eigenschaften von Items und die Fähigkeiten der Testpersonen so zu schätzen, dass sie von einer bestimmten Gruppe Testpersonen oder einem bestimmten Test unabhängig sind. (vgl. Moosbrugger, 2007) Ein weiterer Vorteil der IRT liegt darin, dass bei der KTT kein expliziter Bezug zwischen der individuellen Leistung einer Person und der Schwierigkeit eines Items hergestellt wird. Bei der IRT werden dagegen individuelle Fähigkeiten und Itemschwierigkeiten auf einer gemeinsamen Skala abgetragen. So können individuelle Testwerte durch ihre Abstände zu den Itemschwierigkeiten interpretiert werden. (vgl. Rauch & Hartig, 2007: 240f)

Durch IRT-Analysen wird es möglich, jedem Schüler nur eine Stichprobe aus einer Gesamtheit homogener Testaufgaben zur Bearbeitung vorzulegen. Obwohl meist nur wenige gemeinsame Items, sog. Anker-Items, zwischen den verschiedenen Testheften eines Multi-Matrix- Designs auftreten, ist es dennoch möglich, alle Schüler und alle Items auf eine Metrik zu bringen. (vgl. Robitzsch, 2009: 54) Bei einem Multi-Matrix-Design erhalten Gruppen von Probanden Testhefte mit unterschiedlichen Aufgaben. Deshalb stehen bei der Messwertberechnung nicht zu allen Aufgaben Daten von allen Testpersonen zur Verfügung. Um dennoch Messwerte berechnen zu können, werden Verfahren der probabilistischen Testtheorie (vor allem mehrdimensionale Rasch-Modelle) eingesetzt. Die Anker-Items verbinden die Testhefte miteinander und positionieren die Items der verschiedenen Testformen auf einer Skala mit einer gemeinsamen Metrik. Aus diesem Grund können für die Probanden nach Bearbeitung unterschiedlicher Testformen geschätzte individuelle Testwerte miteinander verglichen werden.

Die IRT geht explizit der Frage nach, welche Rückschlüsse auf bestimmte Merkmale gezogen werden können, wenn nicht alle Items des Tests beantwortet wurden. Dazu unterscheidet die IRT zwischen manifesten und latenten Variablen. Die manifesten Variablen beschreiben das beobachtbare Antwortverhalten der Testpersonen, bei den latenten Variablen handelt es sich dagegen um nicht beobachtbare Fähigkeiten oder Persönlichkeitsmerkmale. Latente Variablen repräsentieren Unterschiede zwischen den Testpersonen, die Unterschiede im beobachteten Antwortverhalten erklären können. (vgl. Hartig, 2008: 75) Es wird davon ausgegangen, dass das manifeste Verhalten von diesen latenten Merkmalsausprägungen abhängt, was sich in Korrelationen zwischen den manifesten Variablen ausdrückt. Vom Antwortverhalten auf die Items auf ein dahinterliegendes Fähigkeits- oder Persönlichkeitsmerkmal, also von der manifesten auf die potentiell dahinterliegende latente Variable zu schließen, ist dagegen nur bei Itemhomogenität möglich.

Das in den manifesten Variablen beobachtete Verhalten ist also ein Indikator für das in der latenten Variable ausgedrückte dahinterliegende Konstrukt. Dabei gilt es die Ausprägung der latenten Variable zu erschließen. Bei der Anwendung von IRT-Modellen kann für jeden Pro-

banden eine Schätzung seiner individuellen Ausprägung auf der latenten Variable angegeben werden. Dieser geschätzte Personenparameter ist der IRT-basierte Testwert einer Person. (vgl. Moosbrugger, 2007)

3.3.1. Das Rasch-Modell

Das bekannteste IRT-Modell ist das Rasch-Modell. Das dichotome Rasch-Modell versucht eine bestimmte gesetzmäßige Beziehung zwischen der Fähigkeitsausprägung einer Person, der Itemschwierigkeit und der Wahrscheinlichkeit des Probanden zu ermitteln, dieses Item zu lösen. Das dichotome Rasch-Modell ist ein Einparameter-Logistisches Modell (1-PL), da stets nur ein Parameter zur Beschreibung jedes Items dient. (vgl. Robitzsch, 2009: 45) Es geht von drei Grundannahmen aus: lokaler stochastischen Unabhängigkeit der Items, Personenhomogenität und spezifischer Objektivität. Aufgrund der lokalen stochastischen Unabhängigkeit der Items verschwinden die Zusammenhänge zwischen den manifesten Variablen bei Kontrolle der latenten Personenfähigkeit. Personenhomogenität bewirkt, dass die Bestimmung der Itemparameter unabhängig von der gewählten Stichprobe sein soll und aufgrund der spezifischen Objektivität ist auch eine Bestimmung der Personenparameter unabhängig von der Itemstichprobe möglich. Einerseits können also Personenparameter bei beliebigen Testaufgaben und andererseits Itemparameter für die Bearbeitung der Items durch eine beliebige Personenstichprobe ermittelt werden. Das einparametrische Rasch-Modell unterliegt der Annahme der doppelten Monotonizität. Dies bedeutet, dass die Schwierigkeitsreihenfolge der Items für alle Testpersonen gleich angenommen wird. (vgl. Robitzsch, 2009: 46ff) Goldstein, Bonnet und Rocher (2007) stellen mit ihren Analysen zu den PISA 2000 Daten die Annahme gleicher Itemladungen im Rasch-Modell jedoch in Frage. Sie gehen davon aus, dass nicht Eindimensionalität der Daten gezeigt wird, sondern eine Teilmenge von Items gefunden wurde, auf die das Rasch-Modell hinreichend zu passen scheint.

Der Schwierigkeitsparameter ist im Rasch-Modell definiert als die Merkmalsausprägung, bei der die Lösungswahrscheinlichkeit für ein Item genau 50% beträgt. Rauch und Hartig definieren die Schwierigkeit eines Items im Rasch-Modell „als jene Ausprägung auf der Fähigkeitsskala, die erforderlich ist, um das Item mit einer Wahrscheinlichkeit von 50% zu lösen." (Rauch & Hartig, 2007: 242) Personen- und Itemparameter können so auf derselben Skala abgetragen werden. Dabei stellt die Differenz zwischen der individuellen Merkmalsausprägung des Probanden und der Anforderung des jeweiligen Items die entscheidende Größe für die Lösungswahrscheinlichkeit dar. Spezifische Ausprägungen der Personenfähigkeit kann man über die itemcharakteristische Funktion (einer mathematischen Gleichung über die Beziehung zwischen dem manifesten Antwortverhalten auf die Items und der Ausprägung der latenten Merkmale) in Lösungswahrscheinlichkeiten für Items mit bestimmten Schwierigkeiten übertragen. (Rauch & Hartig, 2007) Im Rasch-Modell hängt die Wahrscheinlichkeit der Daten also nicht davon ab, wie viele Personen welche Items gelöst haben, sondern nur davon, wie viele Personen ein Item lösen bzw. wie viele Items von einer Person gelöst wurden. Das Rasch-Modell hat auch den Vorteil, dass die Itemparameter ohne gleichzeitige Berücksichtigung der Personenparameter geschätzt werden können. Durch diese Separierbarkeit der Parameter wird Stichprobenunabhängigkeit erzielt. (vgl. Moosbrugger, 2007)

3.3.2. Zweiparametrische (2-)PL Modelle

Um Items mit geringen Itemladungen zu identifizieren, wird neben dem Schwierigkeits- auch ein Diskriminationsparameter eingeführt. Items mit geringen Itemladungen messen tendenziell in geringerem Ausmaß das Gesamtkonstrukt und differenzieren schlechter zwischen Personen mit hoher und niedriger Kompetenz. Sekundärdimensionen können geringe Diskriminationsparameter verursachen. Es ist deshalb wünschenswert, diese Items zu identifizieren und ggf. aus dem Gesamttest zu entfernen. Dafür wird ein Modell benötigt, das berücksichtigt, dass verschiedene Items unterschiedlich gut zwischen schwächeren und stärkeren Merkmalsausprägungen differenzieren können. Diese Voraussetzung wird vom 2-Parameter Modell (2-PL) erfüllt. (vgl. Birnbaum, 1968) Der Grad der Veränderung der Lösungswahrscheinlichkeiten in Abhängigkeit von der Merkmalsausprägung wird durch die Diskriminationsparameter ausgedrückt. Da sie ein Maß der Sensitivität der Items für Merkmalsunterschiede darstellen, entsprechen sie den Trennschärfekoeffizienten der Itemanalyse. (vgl. Moosbrugger, 2007: 237ff) Mit dem zusätzlichen Einsatz von Diskriminationsparametern entfällt im 2-PL-Modell die Eigenschaft der doppelten Monotonizität. (vgl. Robitzsch, 2009)

IRT-Modelle werden meist so interpretiert, dass die Testpersonen über eine Fähigkeit oder Eigenschaft verfügen, die das im Test durch die Antworten auf die Testitems gezeigte Verhalten verursacht. Das Rasch-Modell kann aber auch dazu eingesetzt werden, eine formative Messung durchzuführen, ohne dass eine latente Variable zugrunde gelegt wurde. In diesem Fall definiert der Summenwert einer Person die Skala und damit das Konstrukt. Die Validität des Konstrukts wird dann durch Experten des jeweiligen Inhaltsbereichs (z. B. bei den IQB-Aufgaben durch die Aufgabenentwickler oder die didaktischen Experten) definiert. (vgl. Rossiter, 2002) Unter der Annahme einer formativen Messung ist der Ausschluss von Items mithilfe von Modellfitkriterien im Rasch- oder 2PL-Modell nicht zwangsläufig (Goldstein, 1982).

3.4. Analyse von Verteilungen

Um ungeeignete Variablen für den Ausschluss von weiteren Analysen zu identifizieren, werden für kategorial eingeschätzte Variablen Häufigkeitsanalysen durchgeführt. In der deskriptiven Statistik werden zentrale Informationen und wesentliche Merkmale einer gegebenen Häufigkeitsverteilung zusammenfassend durch Maßzahlen wie Anteilswerte, Mittelwerte und Streuungsmaße (z. B. Varianz) beschrieben. Neben den absoluten Häufigkeiten und Prozentwerten werden z. B. auch die kumulierten Werte berichtet, d. h. die Anzahl aller Merkmalsträger mit gleichem oder geringerem Wert. Eine weitere dieser Maßzahlen ist der Mittelwert. Der Mittelwert berechnet sich aus der Summe aller Messwerte, die durch die Anzahl aller Messwerte dividiert wird. Die Varianz, eine Kennzahl für die Streubreite oder Dispersion, gibt an, wie weit die Antworten durchschnittlich vom Mittelwert der gegebenen Häufigkeitsverteilung abweichen. Sie ist die Summe der quadrierten Abweichungen der einzelnen Messwerte vom Mittelwert, dividiert durch die Stichprobengröße bzw. den Freiheitsgrad. Die Analyse von Merkmalen als Erklärung für Varianz in den empirischen Itemschwierigkeiten ist nur dann sinnvoll, wenn sich die Schwierigkeiten auch wirklich deutlich voneinander unterscheiden. Die Varianz ist neben der Standardabweichung eines der wichtigsten Maße zur Bewertung der Variabilität eines Merkmals in der Stichprobe. Bei der Berechnung der Varianz entsteht durch das Quadrieren der Werte ein oft schwierig interpretierbares Ergebnis. Aus diesem Grund wird

aus der Varianz die Wurzel gezogen um die Standardabweichung (Streuung des Mittelwerts) zu erhalten. (vgl. Leonhart, 2008: 40ff) Als schwacher Effekt gilt eine Varianzaufklärung von 1%, $Eta^2 \approx 6\%$ gilt als mittlerer Effekt und $Eta^2 \approx 14\%$ als starker Effekt (vgl. Bortz, 2005: 259).

3.5. Korrelationsanalysen

Bei der Berechnung von Korrelationen beschränkt man sich typischerweise auf lineare Zusammenhänge. Korrelationsanalysen sollen mögliche Zusammenhänge zwischen den ausgewählten Variablen und den empirischen Itemschwierigkeiten sowie der Variablen untereinander aufdecken. Diese Ergebnisse dienen dazu, Variablen zu identifizieren, die möglicherweise keinen oder einen nur sehr geringen Einfluss auf die Itemschwierigkeit besitzen. Eine Prüfung der Zusammenhänge der Variablen untereinander könnte gegebenenfalls zu einer Gruppierung von Variablen führen, um als Gruppe besonders gut die Itemschwierigkeit zu prognostizieren. Man unterscheidet zwischen positiven (je mehr, desto mehr) und negativen (je mehr, desto weniger) Zusammenhängen. Der Korrelationskoeffizient gibt an, wie stark eine Korrelation zwischen Variablen auftritt. Ein gebräuchlicher Korrelationskoeffizient für bivariate Korrelationen (d. h. Korrelationen zwischen zwei Variablen) ist Pearsons r. Er liegt zwischen +1 und -1. Dabei gilt nach Cohen (1988) für r 0.1 als kleiner, 0.3 als mittlerer und 0.5 als großer Effekt. Sind die Werte von r negativ, so besteht ein gegenläufiger Zusammenhang, bewegen sich die Werte von r um null, so ist kein statistisch bedeutsamer linearer Zusammenhang vorhanden, bei positiven Werten fällt der Zusammenhang gleichgerichtet aus.

Um zufällige Korrelationen auszuschließen, berechnet man i. d. R. die Wahrscheinlichkeit, mit der eine Korrelation zufällig zustande gekommen sein kann. Das Maß für diese Wahrscheinlichkeit ist die Signifikanz p. Sie ist immer an eine Nullhypothese (in diesem Fall eine Korrelation von Null in der Population) gebunden. Ein nichtsignifikantes Ergebnis liegt vor, wenn die Nullhypothese beibehalten wird und kein Effekt vorliegt. Die statistische Signifikanz eines Effektes hängt vom Umfang der untersuchten Stichprobe ab. Da der Itemstichprobenumfang in dieser Arbeit nur eine moderate Größe aufweist, wird als Entscheidungsregel für ein signifikantes Ergebnis $p \leqslant 0.10$ angesetzt. So wird vermieden, einen tatsächlichen Effekt als nicht signifikant zu deklarieren. (Leonhart, 2008: 86ff)

Um die praktische Relevanz von signifikanten Ergebnissen darzustellen, wird deshalb auch die Effektstärke Cohen's d als Ergebnis des t-Tests angegeben. Bei einem t-Test wird die mittlere Itemschwierigkeit von Items, für die ein Code zutrifft, mit der mittleren Itemschwierigkeit aller übrigen Items verglichen, für die der Code nicht zutrifft. Die Effektstärke gibt an, wie stark der Einfluss einer Variablen auf eine abhängige Variable ist. Nach Bortz und Döring gilt bei t-Tests für unabhängige Stichproben eine Effektgröße von 0.2 als klein, ein Wert von 0.5 als mittlerer Effekt und ein Wert von 0.8 als großer Effekt.

3.6. Varianzanalysen

Varianzanalysen (analysis of variance, ANOVA) stellen ein statistisches Verfahren zur Auswertung von Vergleichen zwischen mehr als zwei Gruppen bezüglich einer abhängigen Variablen dar, indem die auf ein Treatment zurückzuführende Varianz in der Stichprobe mit der Gesamtvarianz in Beziehung gesetzt wird. Das Vorgehen untersucht, ob Unterschiede in den Ausprä-

gungen der abhängigen Variable (hier die Itemschwierigkeit) ursächlich auf die unabhängige(n) Variable(n) (hier die in den Einzelüberschriften benannten Merkmale) zurückzuführen sind. Es wird also z. B. getestet, ob die Varianz zwischen einzelnen Codestufen eines Einflussmerkmals größer ist als die auf Zufall zurückzuführende Varianz innerhalb dieses Codes. Auf diese Weise kann geprüft werden, ob die Codeeinteilung bedeutsam hinsichtlich der Varianzaufklärung der Itemschwierigkeit ist. Unterscheiden sich die Codes nicht signifikant voneinander, ist zu bezweifeln, dass in ihnen unterschiedliche Gesetzmäßigkeiten wirken. Das Verhältnis der Varianz zwischen und innerhalb der Gruppen wird durch den F-Wert F(k-1, n-k) ausgedrückt, wobei k die Anzahl der Faktorstufen und n die Anzahl der Messwerte sind. Diese Indizes bezeichnet man als Freiheitsgrade, da sie die Anzahl frei variierender Abweichungen angeben. Je höher F ist, desto mehr von der unabhängigen Variable erzeugte Varianz besteht zwischen den Gruppen. Im Fall von mehreren unabhängigen Variablen werden auch Interaktionseffekte zwischen den unabhängigen Variablen miteinander geprüft.

3.7. Lineare Regressionsanalysen

Lineare Regressionsanalysen sind ein auf Korrelationen basierendes statistisches Auswertungsverfahren (Cohen et al., 2003). Sie dienen dazu, lineare Beziehung zwischen einer abhängigen Variablen und mehreren unabhängigen Variablen zu ermitteln, die alle direkt gemessen werden. Mithilfe einer Regressionsanalyse kann berechnet werden, wie gut bestimmte Variablen ein bestimmtes Kriterium vorhersagen können und welche Variable einer Variablengruppe sich am besten dazu eignet, ein bestimmtes Ergebnis vorherzusagen. Auf der Grundlage der Regressionsanalyse können auch Aufgabenmerkmale kombiniert werden um eine bestmögliche Vorhersagekraft in Bezug auf die Itemschwierigkeit zu erhalten. Die Kombination einzelner Variablen lässt Rückschlüsse auf die von der Aufgabe erfassten Kompetenzen zu. Als Maßzahl wird der *Determinationskoeffizient* R^2 berechnet. Varianz- und Regressionsanalysen können mathematisch ineinander überführt werden. (Bortz & Döring, 2002: 503ff)

Freedle und Kostin (1996) berechneten mit 337 TOEFL-Hörverstehensitems eine schrittweise multiple Regression mit der Itemschwierigkeit als abhängige Variable und fanden dabei, dass 14 Itemmerkmale 35% der Varianz der Itemschwierigkeit über die 337 Items hinweg aufdeckten. Um die Varianzaufklärung zu erhöhen und methodische Verzerrung zu vermeiden, reanalysierten sie den Datensatz, wobei sie die Items in Gruppen (Main Idea Items, Explicit Statement Items, Inference Items) gliederten und nur jeweils eine Gruppe an Items für die Analysen verwendeten. Diese Analysen ergaben, dass unterschiedene Variablen bei unterschiedlichen Item-Typen eine Rolle spielen und die Varianzaufklärung konnte für alle drei Itemgruppen erhöht werden[9].

Aus diesem Grund werden auch in dieser Arbeit die Items, dort wo es sinnvoll ist, gruppiert und ein Teil der Analysen wird nicht mit dem gesamten Datensatz durchgeführt. Um die Stär-

9 Drei Variablen erklären bei den Main Idea Items 42%, bei den Explicit Statement Items erklären acht Variablen 54% und bei den Inference Items erklären sieben Variablen 63% der Varianz der Itemschwierigkeit (Freedle & Kostin, 1996)

ke des Einflusses der linguistischen Merkmale auf die empirische Itemschwierigkeit zu berechnen, werden die relevanten Variablen in dieser Arbeit einer additiven einfachen linearen Regressionsanalyse unterzogen. Grundannahme ist dabei, dass das Anforderungsprofil einer Aufgabe durch ein komplexes Zusammenwirken der einzelnen Merkmale geprägt ist, und der Einfluss der einzelnen Merkmale auf das Anforderungsprofil variieren kann. Die Aufgabenschwierigkeit ergibt sich aus der Summe der anforderungsrelevanten Merkmale. Die entsprechenden Variablen treten alle gleichzeitig in der Regressionsanalyse auf. Dabei wird jede Variable hinsichtlich ihrer Vorhersagekraft im Vergleich zu allen anderen Variablen analysiert. Auf diese Weise wird die Varianzaufklärung der einzelnen Variablen, aber auch der zu Gruppen zusammengefassten Variablen, erhoben. Unabhängige Variablen, die stärker als 0.9 korrelieren, werden nicht in die Regressionsanalyse miteinbezogen. Auch Variablen, die eine Kombination aus mehreren unabhängigen Variablen sind, werden aus den Analysen ausgeschlossen.

Eine elegantere Form sind linear-logistische Modelle, bei denen in einem gemeinsamen IRT-Modell die Schätzung von Schwierigkeit und Regression miteinander verbunden wird. Indem die Itemparameter in IRT-Modellen als Linearkombination einer geringeren Anzahl von Basisparametern betrachtet werden, können sie näher spezifiziert und inhaltlich erweitert werden. Eine Form linear-logistischer Modelle sind die Latent Trait Modelle. Sie gehen von quantitativen kontinuierlichen latenten Variablen aus, von denen die Wahrscheinlichkeit des manifesten Verhaltens des Probanden abhängt. Aus diesem Grund kann der Trait zur Erklärung von Verhaltensunterschieden dienen. (Moosbrugger, 2007: 220f) Das bekannteste Latent Trait Modell ist das *Linear Logistic Latent Trait Modell (LLTM)* (Fischer, 1995). Es wird verwendet um Items zu kalibrieren und Hypothesen über kognitive Komponenten zu testen, welche dem Itemlösungsprozess zugrunde liegen. Das LLTM fällt in die Klasse der *Component Latent Trait Modelle (CLTM)* (Embretson, 1984). CLTMs verbinden Antworten auf Items mit kognitiven Theorien durch eine Vorhersage der Itemschwierigkeit mittels mathematischer Modelle. Dadurch, dass der Einfluss kognitiver Komponenten auf die Itemschwierigkeit geschätzt wird, haben die Ergebnisse der CLTM direkten Einfluss auf die Konstruktvalidität und Items mit spezifischen kognitiven Komponenten können ausgewählt werden.

Das LLTM gehört insofern zur Familie der Rasch-Modelle als die Schwierigkeit den einzigen Unterschied zwischen den Items darstellt. Im LLTM wird die Itemschwierigkeit durch beurteilte Itemcharakteristika und Parameter zu ihrer Gewichtung ersetzt. Aus diesem Grund schätzt das LLTM den Einfluss von Design-Effekten auf die Itemschwierigkeit. Embretson (1999) verwendet eine Erweiterung des LLTM zur Vorhersage von Itemtrennschärfen durch Itemmerkmale. Es unterscheidet sich von regressionsanalytischen Ansätzen dadurch, dass im LLTM die Itemschwierigkeiten latent durch Itemmerkmale in einem einstufigen Prozess vorhergesagt werden. Im Regressionsmodell werden dagegen in einem ersten Schritt die Itemparameter geschätzt und in einem zweiten Schritt diese geschätzten, messfehlerbehafteten Itemschwierigkeiten als abhängige Variable im Regressionsmodell eingesetzt. Da Messfehler in der abhängigen Variablen enthalten sind, wird die Varianzaufklärung R^2 im Allgemeinen unterschätzt. Umgekehrt fallen die Standardfehler der Koeffizienten der Itemmerkmale im Vergleich zum LLTM kleiner aus, man erhält tendenziell also eher signifikante Effekte. Bei großen Itemstichproben sind die Unterschiede zwischen den beiden Vorgehen jedoch gering.

3.8. Methoden zur Trennung von Item- und Stimuluseffekten

Mehrebenenmodelle dienen dazu, hierarchisch strukturierte Daten zu modellieren, indem sie in linearen und verallgemeinerten linearen Regressionsmodellen für die Regressionskoeffizienten wiederum Regressionsmodelle spezifizieren (vgl. Raudenbush & Bryk, 2002). Mehrebenenmodelle (vgl. Ozuru et al., 2008) berücksichtigen im Gegensatz zur klassischen Regressionsanalyse, die von der stochastischen Unabhängigkeit der einzelnen Fälle ausgeht, in der Berechnung von Prüfgrößen und Effekten eine Clusterung der Datenstruktur (z. B. Schüler in Schulklassen oder der Items in Aufgaben mit einem Stimulus). Bei hinreichender Anzahl an vorhandenen Clustern wird so auch die Einführung von Prädiktoren für Merkmale auf der Individualebene möglich. Die separate Schätzung von Effekten auf der Individualebene und dem Mittelwert auf Gruppenebene zählt zu den wichtigsten Stärken der Mehrebenenmodelle. Da Mehrebenenmodelle die Variationsinformation auf der Individualebene mit einbeziehen und berücksichtigen, wie viel Varianz auf der individuellen Datenebene überhaupt auf Clusterzugehörigkeit zurückzuführen ist, sind sie im Verhältnis zur Aggregation von Individualdaten und der Korrelation dieser Aggregate genauer. Das Maß, das zur Berechnung der clusterspezifischen Varianz herangezogen wird, ist die *Intraklassenkorrelation (ICC)*. Sie gibt das Verhältnis von Klassenvarianz zur Gesamtvarianz an und verdeutlicht, in welchen Bereichen besonders große Unterschiede zwischen Klassen bestehen. (vgl. Robitzsch, 2009)

3.9. Dimensionsanalysen

I. d. R. werden zum Lösen eines Items mehrere Kompetenzen benötigt, das entsprechende Item misst also mehr als eine einzige Fähigkeit. Dimensionsanalysen untersuchen, ob sich Teilaspekte von Kompetenzen im Sinne von trennbaren Kompetenzen (Dimensionen) auch auf empirischer Ebene identifizieren lassen. Eindimensionalität bedeutet, dass ein Test eine einzelne homogene Fähigkeit abbildet. Dabei müssen die Items der Forderung nach lokaler stochastischer Unabhängigkeit genügen. Mehrdimensionalität impliziert, dass kein einzelnes dimensional homogenes Konstrukt operationalisiert wurde. (vgl. Böhme & Robitzsch, 2009) Dabei wird bei den Dimensionsanalysen im Gegensatz zu den DIF-Analysen, bei denen von Anfang an mehrere Personengruppen angenommen werden, davon ausgegangen, dass die Testpersonen in ihren Fähigkeiten homogen sind. Treten dennoch Leistungsunterschiede bei der Itembearbeitung in der Gesamtgruppe auf, weist dies auf die Mehrdimensionalität der Items hin. Fragen der Konstruktdimensionalität haben Auswirkungen auf die Operationalisierung des Konstrukts sowie die Überprüfung der Kompetenzausprägung, aber auch auf die Förderung des Konstrukts im Unterricht. Auch für die Dokumentation der Kompetenzstände in Kompetenzstufenmodellen sind diese Überlegungen relevant. (vgl. Böhme & Robitzsch, 2009: 261)

Im Rahmen von Large-Scale-Erhebungen von Schulleistungen gibt es im deutschen Sprachraum bislang wenige Analysen zur Dimensionalität von Zuhörkompetenz. Bei DESI (vgl. Kapitel *4.3.1. Nationale Studien*) wurden jedoch Analysen zur übergreifenden Struktur aller untersuchten sprachlichen Kompetenzen im Deutschen und im Englischen durchgeführt. (vgl. Jude et al., 2008) Im Rahmen von PISA 2000 und der Internationalen Grundschul-Lese-Untersuchung (IGLU) wurde die Dimensionalität der Leseverstehenskompetenz untersucht. (vgl. Artelt & Schlagmüller, 2004; Bos et al., 2007)

Um die Dimensionalität des Konstrukts Zuhören zu erfassen, werden explorative Faktorenanalysen sowie konfirmatorische Faktorenanalysen im Rahmen von Item-Response-Modellen nach Art der Stimuli (z. B. konzeptionell schriftlich oder mündlich) und nach Art der Informationsentnahme (Das Item verlangt z. B. globale Informationen, lokale Informationen oder eine Inferenz) durchgeführt. Bei der Aufgabenentwicklung wird häufig davon ausgegangen, dass jedes Item genau einer Dimension zuzuordnen ist (Between-Item-Dimensionality). Genauere inhaltliche Itemanalysen zeigen jedoch, dass z. T. mehrere notwendige Kompetenzen für die Itemlösung notwendig sind, sodass von einer Mehrfachladungsstruktur (Within-Item- Dimensionality) ausgegangen werden muss. (vgl. Winkelmann & Robitzsch, 2009)

Faktorenanalysen umfassen multivariate Analyseverfahren, die zur Reduktion größerer Datenmengen auf kleinere Mengen oder zur Überprüfung der Konstruktvalidität eines Tests eingesetzt wird. Dabei sollen latente Dimensionen (Faktoren) extrahiert werden, die die manifesten Daten strukturieren. Mithilfe der Faktorenanalyse kann untersucht werden, welche Variablen, die direkt gemessen werden, mit einem gemeinsamen latenten Konstrukt oder Faktor erklärt werden können, das/der nicht direkt gemessen wird. Die Faktorenanalyse dient also dazu, Gruppen innerhalb einer Variablenmenge zu identifizieren. Im Allgemeinen unterscheidet man die *exploratorische Faktorenanalyse (EFA)* als hypothesengenerierendes von der *konfirmatorischen Faktorenanalyse (CFA)* als hypothesenprüfendes Verfahren. Beide Methoden versuchen die Zusammenhänge zwischen den beobachteten Variablen zu erklären. Dabei ist die EFA im Gegensatz zur CFA ein struktursuchendes Verfahren (theoriefrei) und reduziert die komplexen Informationen in der beobachteten Korrelationsmatrix. So wird die inhaltliche Interpretation der Faktoren ermöglicht. Dagegen ist die CFA ein strukturüberprüfendes Verfahren (theoriegeleitet) und testet Hypothesen bezüglich der faktoriellen Struktur. Die inhaltliche Interpretation der Faktoren liegt hier bereits fest. (Moosbrugger & Schermelleh- Engel, 2007) Vor Beginn der EFA muss zunächst die Art der Extraktionsmethode ausgewählt werden, mit der die Faktoren extrahiert werden sollen. Bei der Extraktion werden auch die Faktorladungen, die Eigenwerte und die Kommunalitäten der Faktoren bestimmt. Dies kann beispielsweise mit der Hauptkomponenten- oder der Hauptachsenmethode erfolgen. Die Anzahl der Faktoren ergibt sich durch ein vorher festgelegtes Abbruchkriterium, wie den Scree-Test oder die Parallelanalyse. Der aus dem Scree-Test resultierende Scree-Plot ist eine grafische Darstellung der kumulierten erklärten Varianz je Faktorlösung. Für die Faktoren gilt die Faustregel, dass ein Faktor mindestens 5% Varianz erklären sollte. Zuletzt werden die Faktoren orthogonal (rechtwinklig) oder oblique (schiefwinklig) rotiert. (vgl. Moosbrugger & Schermelleh- Engel, 2007)

Im Unterschied zu EFA verlangt die CFA Vorannahmen bezüglich der Anzahl und Struktur der zu extrahierenden Faktoren. Sie prüft Annahmen über die Struktur, die den Daten zugrunde liegt. Dies geschieht häufig durch eine Prüfung mehrerer Modelle gegeneinander, um zu sehen, durch welches Modell die Daten am besten erklärt werden. Dabei werden zunächst latente Faktoren angenommen, auf welche die Ausprägungen in den gemessenen Variablen zurückzuführen sein sollen. In einem zweiten Schritt wird angegeben, auf welchen Faktoren welche Items laden und ob die Faktoren abhängig oder unabhängig von einander sind. In einem nächsten Schritt erfolgt die Berechnung der Fit-Indizes. Fit-Indizes bilden ab, mit welcher Genauigkeit empirische Daten, z. B. empirisch gewonnene Kovarianzen, durch Modellparame-

ter der Faktorenanalyse reproduziert werden können. Dabei gelten *Verhältnis Chi- Quadrat zu Freiheitsgraden*, der *Standardized Root Mean Square Residual (SRMR)*, der *Comparative Fit Index (CFI)* und der *Root Mean Square Error of Approximation (RMSEA*) als die vier wichtigsten (vgl. Moosbrugger & Schermelleh-Engel, 2007). Nach der Berechnung der Ladungen und der Fit-Indizes wird der Algorithmus i. d. R. durch Maximum- Likelihood-Schätzungen (ML) bestimmt.

3.10. Verfahren zur Erfassung der Beurteilerübereinstimmung

Kommen bei einem Test mindestens zwei geschulte, aber voneinander unabhängige Beurteiler zum Einsatz, empfiehlt es sich die Übereinstimmungen der verschiedenen Urteile zu prüfen. Dazu kann bei nominalskalierten Daten beispielsweise die absolute Übereinstimmung oder ein Objektivitätskoeffizient oder bei intervallskalierten Daten die Intraklassenkorrelation berechnet werden. Die absolute Übereinstimmung berechnet sich aus der Zahl der Übereinstimmungen dividiert durch die Anzahl der beobachteten Objekte. Da auch bei zufälliger Klassifizierung einige Beobachtungen übereinstimmen können, bieten sich stattdessen Maße an, die eine Zufallskorrektur der Übereinstimmung vornehmen. Eines dieser Maße ist Cohens Kappa. Dabei wird i. d. R. wird von Werten über 0.7 ausgegangen, um von hoher Übereinstimmung zu sprechen und eine adäquate Reliabilität des Tests zu gewährleisten. (vgl. Lienert & Raatz, 1998: 140) Bei intervallskalierten Daten bieten sich als Übereinstimmungsmaß die *Intraklassenkorrelation (ICC)* an, die als Reliabilität der Urteile eines beliebigen Beobachters zu interpretieren ist. Berechnet man die Intraklassenkorrelation ICC_k, werden die Reliabilität der Urteile über den Mittelwert der Beurteilungen von *k* Beurteilern zusammengefasst (vgl. Wirtz & Caspar, 2002: 157ff).

IV

Darstellung der Ergebnisse

IV Darstellung der Ergebnisse

Der vierte Teil widmet sich schließlich ausführlich der Beschreibung der durchgeführten Analysen und der Ergebnisse. In Kapitel 4.1. werden zunächst explorative Dimensionsanalysen der Item- und Stimulusmerkmale beschrieben, in Kapitel 4.2. erfolgt dann die Darstellung der erfolgten Zusammenhangsanalysen und Kapitel 4.3. schließt mit den Ergebnissen der durchgeführten Regressionsanalysen.

1. Explorative Dimensionsanalysen von Item- und Stimulusmerkmalen

Bevor mit den Einzelanalysen zu den identifizierten Merkmalen begonnen wird, wird mithilfe einer Korrelationsmatrix überprüft, welche Merkmale stark mit der Item- bzw. Aufgabenschwierigkeit zusammenhängen. Die Itemschwierigkeiten wurden über das Rasch-Modell gewonnen. Dabei wurde der Mittelwert der Schülerfähigkeiten der MSA-Schüler auf null gesetzt. Negative Itemschwierigkeiten bedeuten dann, dass die Items etwas leichter sind als Personen fähig sind. Die Aufgabenschwierigkeiten sind die aggregierten Itemschwierigkeiten.

In einem zweiten Schritt wird mittels einer Faktorenanalyse untersucht, welche Merkmale sich ggf. um einen gemeinsamen Faktor gruppieren. Erst dann sollen die einzelnen Merkmale weiteren Analysen unterzogen werden. Aufgrund der vorab gewonnenen Ergebnisse kann bei den Einzelanalysen ggf. nach sinnvollen Kriterien gefiltert werden, um den Einfluss einer einzelnen Variable auf die Item- bzw. Aufgabenschwierigkeit genauer zu erfassen. Dieser Schritt ist notwendig, da die Aufgaben nicht unter kontrollierten Bedingungen erstellt wurden und deshalb der Einfluss einzelner Merkmale auf die Schwierigkeit nicht isoliert überprüft werden kann. Bei den Aufgaben handelt es sich vielmehr um weitgehend authentische Stimuli und Items dazu, mit denen die sprachlichen Kompetenzen der Schüler getestet werden sollen.

Die folgenden Analysen wurden für Stimulus-, Item- und Personenmerkmale getrennt durchgeführt. Die mit dem Lehrerfragebogen erfassten Stimulusmerkmale und die am IQB gerateten Merkmale fanden separat Beachtung.

1.1. Stimulusmerkmale

1.1.1. Korrelation der Stimulusmerkmale aus den IQB-Ratings mit der Schwierigkeit

Die Korrelationen wurden mit den unkategorisierten Variablen berechnet. Dabei wurde zum einen die Korrelation eines Merkmals mit der Aufgabenschwierigkeit und zum anderen die mittlere Korrelation des Merkmals mit den einzelnen Itemschwierigkeiten berechnet. Insgesamt fallen die Korrelationen einzelner Merkmale auf Aufgabenebene deutlich höher als auf Itemebene aus. Als Grenzwerte für bedeutsame Korrelationen der Merkmale mit der Schwierigkeit wurde auf Item- und auf Aufgabenebene ±0.2 angesetzt. Werte, die größer bzw. kleiner als ±0.2 ausfallen und signifikant ($p < 0.1$) sind, wurden grau markiert (vgl. Anhang C, Tabelle C-1.). Die meisten Merkmale korrelieren sowohl mit den Item- als auch mit den Auf-

gabenschwierigkeiten, wenn sie auch höhere Korrelationen mit der Aufgabenschwierigkeit aufweisen. Die höheren Korrelationen auf Aufgabenebene hängen damit zusammen, dass bei der Aufgabenschwierigkeit die Schwierigkeit aller Items dieser Aufgabe gemittelt angegeben wird und sich dadurch insgesamt auch höhere Korrelationen ergeben. Die meisten Merkmale scheinen einen stärkeren Einfluss auf das Verständnis der Stimuli als auf die Beantwortung der Items zu haben. Die Merkmale „Inhaltswörter (IWH)", „Thema (THE)", die Relationstypen „Erklärung/Beweis/Ursache (REL4)" und „Ziel/Bedingung (REL6)" sowie die Merkmale „Neudeutsch/Anglizismen (RHE3)" und „Referenzen (REF)" korrelieren höher mit der Item- als mit der Aufgabenschwierigkeit. Sie scheinen für das Verständnis der Stimuli eine untergeordnete Rolle zu spielen, für die Beantwortung der Items jedoch relevant zu sein. Insgesamt kann man sagen, dass quantitativ auszählbare Merkmale zur Beschreibung der Stimuli einen recht hohen Einfluss auf die Aufgabenschwierigkeit haben.

Tabelle IV-1.1.1. gibt einen Überblick über die Stimulusvariablen, die einen positiven und negativen Zusammenhang von p = ±0.2 mit der Aufgabenschwierigkeit zeigen.

Tabelle IV-1.1.1: Zusammenfassung Korrelation Merkmal mit der Aufgabenschwierigkeit

Variable	Merkmal	r	Variable	Merkmal	r
PLW	Lange Wörter	0.42**	PEW	Einsilbige Wörter	-0.40**
WLS	Wortlänge	0.36*	HKO	Hörkontext	-0.39**
PMW	Mehrsilbige Wörter	0.34*	NEG	Negation	-0.37**
ASP	Anzahl Sprecher	0.33*	STR4	Anakoluthe	-0.32*
SLT	Literarischer Stimulus	0.31*	SGS	Sprechgeschwindigkeit	-0.27
WIE	Wiederaufnahmen	0.27	REL5	Reihenfolge/Aufzählung	-0.22
WEL	Hintergrundwissen	0.26			
REL1	Frage/Impuls/Themensetzung	0.24			
SUB	Substantive/Eigennamen/Appellative	0.22			
MLP	Mittlere Länge der Propositionen	0.22			
STR1	Ellipsen	0.21			

Anmerkung: r = Korrelation mit der Aufgabenschwierigkeit, * p < 0.10, ** p < 0.05

1.1.1.1. Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale

Die meisten Variablen beziehen sich auf sprachliche Merkmale („Einsilbige Wörter (PEW)", „Lange Wörter (PLW)", „Wortlänge (WLS)", „Mehrsilbige Wörter (PMW)", „Wiederaufnahmen (WIE)", „Substantive/Eigennamen/Appellative (SUB)", „Ellipsen (STR2)", Negationen (NEG) sowie Anakoluthe (STR4)). Die Merkmale PLW, WLS, PMW und PEW nehmen Bezug auf das Merkmal Wortlänge. Das Merkmal PLW (der Prozentanteil langer Wörter, d. h. mit mehr als sechs Buchstaben) hängt dabei stärker mit der Aufgabenschwierigkeit zusammen als das Merkmal PMW (der Prozentanteil mehrsilbiger Wörter, d. h. mit mehr als drei Silben). Erwartungsgemäß erhöhen größere Anteile der Variablen PMW, WLS und PLW die Aufgabenschwierigkeit. Das Merkmal PEW drückt erwartungskonform den umgekehrten Sachverhalt aus: Je höher die Anteile einsilbiger Wörter im Stimulus, desto geringer fällt die Aufgabenschwierigkeit aus.

Die Wortlänge als Ausdruck für die Aufgabenschwierigkeit kann mit der Arbeitsgedächtniskapazität zusammenhängen. Je mehr lange Wörter im Arbeitsgedächtnis aktiv gehalten werden müssen, desto eher ist dieses überlastet. Inhaltliche Kriterien, wie der Prozentanteil der Inhaltswörter[10] und die Worthäufigkeit[11] scheinen dabei eine untergeordnete Rolle zu spielen. Der Anteil der Substantive, Eigennamen und Appellative (SUB) korreliert hingegen mit der Aufgabenschwierigkeit. Wörter, die in diese Kategorie fallen, sind häufig komplexer und ihre Verarbeitung nimmt mehr Gedächtniskapazität in Anspruch.

Wiederaufnahmen (WIE) korrelieren erwartungskonform positiv mit der Aufgabenschwierigkeit. Sie stellen erhöhte Ansprüche an das Arbeitsgedächtnis, da es für ihr Verständnis notwendig ist, auch ältere Informationen aktiv zu halten.

Ellipsen (STR2) erschweren erwartungsgemäß das Stimulusverständnis, da der Zuhörer ausgesparte Informationen ergänzen müssen und Ellipsen auch indirekt dazu führen, dass in kürzerer Zeit mehr Informationen geäußert werden können und vom Arbeitsgedächtnis aufgenommen und verarbeitet werden müssen. Erwartungswidrig sind die Effekte der Merkmale „Negationen (NEG)" und „Anakoluthe (STR4)". Da beim Verstehen eines Negativsatzes zuerst die enthaltene Annahme extrahiert und dann die Negation verarbeitet wird, konnte in unterschiedlichen Studien (z. B. Freedle & Kostin, 1993a; Nissan et al., 1996) ein verstärkender Einfluss des Merkmals Negation auf die Itemschwierigkeit nachgewiesen werden. Bei den IQB-Aufgaben sinkt die Aufgabenschwierigkeit jedoch eher bei Stimuli, die einen hohen Anteil an Negationen aufweisen. Eine Prüfung der Stimuli mit einem hohen Anteil an Negationen ergab, dass es sich dabei um Diskurse handelt mit einer Häufung von negierten Ausdrücken, die häufig wechselseitig von den Gesprächsteilnehmern wiederholt und bestätigt werden.[12] Es handelt sich dabei also nicht um verschiedene, komplexe negierte Aussagen sondern um eher um redundante Verneinungen, die das Stimulusverständnis insgesamt nicht unbedingt erschweren. Im Fall der Anakoluthe wurde angenommen, dass die Unvollständigkeit der abgebrochenen Struktur eher Verständnis erschwerend wirkt. Im Gegenteil korreliert jedoch eine hohe Auftretenshäufigkeit dieses Merkmals eher mit einer geringeren Aufgabenschwierigkeit. Auch Stimuli, in denen viele Anakoluthe vorkommen, gehören eher zu den Diskursen und enthalten noch zahlreiche andere Merkmale gesprochener Sprache (z. B. Redundanzen), die Verständnis erleichternd wirken.

1.1.1.2. Merkmalsgruppe II: Präsentationsmerkmale

Negative Korrelationen treten auch zwischen der Aufgabenschwierigkeit und der mittleren Sprechgeschwindigkeit (SGS) auf. Es wurde erwartet, dass eine höhere Sprechgeschwindigkeit die Aufgabenschwierigkeit erhöht, da in kürzerer Zeit mehr Informationen auf den Zuhörer eintreffen und verarbeitet werden müssen. Insofern ist das erhaltene Ergebnis erwartungswidrig. Es wird vermutet, dass die Sprechgeschwindigkeit mit der Stimulusart (Text oder

10 IWH: r = 0.03 mit der Aufgabenschwierigkeit

11 GWS: r = -0.02 mit der Aufgabenschwierigkeit

12 Beispiel 1: „Ich wusste ja nichts, ich wusste ja nicht, wenn mich irgendwer gefragt hat, in welchen Abständen müssen die Hunde geimpft werden oder wie teuer ist das und das und ich saß dann immer da so, ähm, sorry ich bin Azubi, ich hab keine Ahnung, …";
Beispiel 2: „Nichts Falsches sagen!" - „Bitte?" - „Jetzt nichts Falsches sagen, Silke." „Nein, ich finde ihn sehr sehr lustig."

Diskurs) zusammenhängt und dass Diskurse schneller gesprochen werden, als vorgelesene Texte. Stimuli mit höherer Sprechgeschwindigkeit weisen demnach verstärkt Merkmale gesprochener Sprache auf und sind daher insgesamt einfacher zu verstehen.

Das Merkmal „Anzahl der Sprecher (ASP)" korreliert positiv mit der Aufgabenschwierigkeit. Dieses Ergebnis ist erwartungskonform. Je mehr unterschiedliche Sprecher der Zuhörer im Stimulus differenzieren muss, desto schwieriger.

1.1.1.3. Merkmalsgruppe III: Inhaltlich-thematische Merkmale

Merkmale, die sich auf inhaltlich-thematische Aspekte der Stimuli („Literarischer Stimulus (SLT)", „Hintergrundwissen (WEL)"), beziehen, korrelieren dagegen deutlich schwächer mit der Aufgabenschwierigkeit. Stimuli, für deren Verständnis Hintergrundwissen benötigt wird, sind schwieriger als Stimuli, die auch ohne Hintergrundwissen verstanden werden können. Dies ist plausibel, da das benötigte Hintergrundwissen u. U. nicht bei allen Schülern vorhanden ist. Insgesamt wurde zwar darauf geachtet, dass keine Stimuli zum Einsatz kamen, für deren Verständnis Fach- oder Hintergrundwissen notwendig ist. Bei den eingesetzten Stimuli gibt es jedoch einige Stellen (insbesondere bei den Diskursen) mit Andeutungen und (Wort-) Witzen, die leichter mit entsprechendem Weltwissen verstanden werden können. Beispielsweise spielt in einem der Stimuli eine Wurzelbehandlung beim Zahnarzt eine Rolle. Schüler, denen unbekannt ist, was bei diesem Eingriff geschieht und dass es sich dabei um eine sehr schmerzhafte Prozedur handelt, verstehen die entsprechende Stimulusstelle nur unvollständig. Dies kann einen Einfluss auf das gesamte Zuhörverhalten haben, da bei manchen Schülern u. U. Frustration oder Langeweile erzielt wird. Denkbar wäre auch, dass nachfolgende Items möglicherweise nicht korrekt bearbeitet werden, da der Gesamtzusammenhang des Stimulus nicht richtig erfasst wurde.

Nicht-literarische Stimuli sind schwieriger als literarische Stimuli. Der Einfluss der Variable „Literarischer Stimulus (SLT)" ist erwartungswidrig, da literarische Stimuli durch mehrdeutige und offene Stellen, die interpretiert und gefüllt werden müssen, häufig als schwieriger gelten. Im Fall der IQB-Aufgaben kann dieses Ergebnis jedoch mit der Auswahl der Stimuli und der dazugehörigen Items zusammenhängen. Es handelt sich dabei insgesamt nur um die folgenden fünf Aufgaben mit den angegebenen aggregierten Itemschwierigkeiten: zwei Gedichte (G104_Reklame: -0.58; G116_Mund; -1.23), ein kurzes Märchen (G118_Leila: -1.84), einen Song (G133_Aber sonst gesund: -1.87) und ein Hörspiel (G126_Bild im Ohr: 0.55). Bis auf die Aufgabe G126 handelt es sich also um extrem leichte Aufgaben.

Relativ hoch korreliert die Variable „Hörkontext (HKO)" mit der Aufgabenschwierigkeit. Die Korrelation könnte u. U. auf dem hohen Einfluss einzelner Stimuli beruhen. Insgesamt wurden 30 Aufgaben in sieben Kategorien eingestuft. Die Belegung der einzelnen Kategorien beläuft sich damit z. T. auf sehr wenige Fälle. Beispielsweise wurde die Kategorie „Hörspiel" mit nur zwei Stimuli belegt, die beide deutlich schwieriger sind, als die mittlere Aufgabenschwierigkeit der anderen Kategorien, wo sich Extreme ggf. leichter ausmitteln.

1.1.1.4. Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte
Stimuli, die häufig den Relationstyp 1 aufweisen, also viele Propositionen mit einer Frage, einem Impuls oder einer Themensetzung enthalten, sind schwieriger. Es wurde zunächst angenommen, dass die gliedernde Struktur dieses Relationstyps Verständnis erleichternd wird. Die Impulssetzungen und damit die Einleitung eines neuen Themenschwerpunkts führt jedoch auch dazu, dass mehr neue Informationen zu verarbeiten sind und dementsprechend Stimuli mit einem höheren Anteil der Variable „Relationstyp: Frage/Impuls/Themensetzung (REL1)" schwieriger zu verstehen sind. In diesem Sinn spielt auch die mittlere Länge von Propositionen (MLP) als Maß für die propositionale Dichte von Stimuli eine Rolle.

Der Relationstyp Reihenfolge oder Aufzählung (REL5) korreliert negativ mit der Aufgabenschwierigkeit. Dies ist insofern plausibel, als eine klare Gliederungsstruktur nach dem Schema erstens, zweitens, drittens, etc. dabei hilft, Inhalte zu erinnern. (vgl. Freedle & Kostin 1993b; 1996)

1.1.1.5. Merkmalsgruppe V: Globalurteil
Die Variable „Stimulusschwierigkeit (TSA)" zeigt keinen signifikanten Zusammenhang mit der Item- und der Aufgabenschwierigkeit.

1.1.2. Korrelation der Stimulusmerkmale aus den Lehrerfragebögen mit der Schwierigkeit
Die Variablen der Lehrerfragebogens korrelieren insgesamt sehr viel stärker mit der Aufgabenschwierigkeit (-0.55 bis 0.50) aber auch der Itemschwierigkeit (-0.29 bis 0.31) (vgl. Tabelle IV-1.1.2). Einzig die Werte der Variable „Kohärenz/Textzusammenhang (KOH)" weisen in beiden Bereichen sehr niedrige und nicht signifikante ($p > 0.1$) Korrelationen auf. Die liegt möglicherweise daran, dass bei manchen befragten Lehrkräften der Begriff „Kohärenz" unbekannt war oder unterschiedliche Konzepte des Begriffs vorlagen, obwohl der Begriff im Fragebogen in einer Fußnote erklärt wurde.

Die Variablen, die sowohl auf Item- als auch auf Aufgabenebene am stärksten signifikant ($p < 0.05$) mit Schwierigkeit korrelieren, sind die Merkmale „Wortschatz (WFA)" (-0.29 bzw. - 0.55), „Gesamteindruck (GIU)" (0.31 bzw. 0.56) und „Ton (TON)" (0.30 bzw. 0.50). Je stärker der Wortschatz eines Stimulus also von den Lehrkräften als fachspezifisch eingestuft wurde, und je uninteressanter der Gesamteindruck bzw. je unpersönlicher/sachlicher der Ton empfunden wurde, desto schwieriger war der Stimulus und waren die Items. Stimuli, deren Wortschatz als alltäglich eingestuft wurde, und die als interessant bzw. persönlich/ gefühlsbetont bewertet wurden, waren dagegen einfacher.

Die folgenden Variablen korrelieren signifikant mit $p < 0.1$ mit der Item- bzw. der Aufgabenschwierigkeit: „Vertrautheit mit dem Thema (VTH)" (0.26 bzw. 0.44), „Wortschatz (WAK)" (-0.29 bzw. -0.47), „Grammatik (GRA)" (-0.27 bzw. -0.43), „Ausdrucksweise (AGU)" (-0.29 bzw. -0.45) sowie „Wirkung (WEH)" (-0.23 bzw. -0.42). Demnach sind Stimuli, deren Thema den Schülern vertraut ist, einfacher als Stimuli, deren Thema als unvertraut eingestuft wurde. Wurde die Grammatik eines Stimulus als einfach eingestuft, der Gesamteindruck als eintönig und die Ausdrucksweise als umgangssprachlich, so handelte es sich um einfachere Stimuli als Stimuli,

bei denen die Grammatik mit schwierig, der Gesamteindruck mit abwechslungsreich und die Ausdrucksweise mit gewählt bewertet wurden. Der Einfluss der Variable AGU ist dabei nicht unbedingt erwartungskonform, da angenommen wird, dass abwechslungsreiche Stimuli für die Schüler interessanter und motivierender sind und deshalb einfacher. Es scheint jedoch so zu sein, dass abwechslungsreiche Stimuli in ihrer Machart komplizierter sind und auch mehr Informationen enthalten, insgesamt also schwieriger sind.

Tabelle IV-1.1.2.: Zusammenfassung Korrelation Lehrerfragebogen mit der Aufgabenschwierigkeit

Variable	r_{Item}	$r_{Aufgabe}$	Variable	r_{Item}	$r_{Aufgabe}$
Vertrautheit mit dem Thema (VTH): vertraut – gar nicht vertraut	0.26**	0.44*	Gesamteindruck (GAE): abwechslungsreich - eintönig	0.28**	0.48**
Wortschatz (WAK): abstrakt - konkret	-0.29**	-0.47*	Ton (TON): persönlich/gefühlsbetont – unpersönlich/sachlich	0.30**	0.50**
Wortschatz (WFA): fachspezifisch - alltäglich	-0.29**	-0.55**	Ausdrucksweise (AGU): gewählt - umgangssprachlich	-0.29**	-0.45*
Grammatik (GRA): schwierig - einfach	-0.27**	-0.43*	Ausdrucksweise (AKE): komplex/ausschweifend - einfach/knapp	-0.22**	-0.31*
Kohärenz/Textzusammenhang (KOH): stark ausgeprägt – wenig ausgeprägt	-0.09	-0.12	Informationsebene (INF): spitzfindig/tiefgründig – offensichtlich/oberflächlich	-0.21**	-0.28
Gesamteindruck (GIU): interessant - uninteressant	0.31**	0.56**	Wirkung (WNL): regt zum Nachdenken an – regt zum Lachen an	-0.20**	-0.32
Gesamteindruck (GEU): elegant - unbeholfen	-0.23**	-0.30	Wirkung (WEH): ernsthaft - humorvoll	-0.23**	-0.42*

Anmerkungen: $r_{Aufgabe}$ = Korrelation mit der Aufgabenschwierigkeit, r_{Item} = Korrelation mit der Itemschwierigkeit, * $p < 0.10$, ** $p < 0.05$

Insgesamt hängen also zahlreiche Stimulusmerkmale mit der Item- bzw. Aufgabenschwierigkeit zusammen. Im nächsten Schritt wird untersucht, ob zwischen den Merkmalen gemeinsame Faktoren angenommen werden können.

1.1.3. Faktorenanalysen der Stimulusmerkmale aus den IQB-Ratings

Für die Gruppierung der Stimulusmerkmale um Faktoren wurden nur Ladungen größer oder kleiner gleich ±0.4 berücksichtigt. Die meisten Merkmale fallen eindeutig einem Faktor zu. Merkmale, die mehr als einem Faktor zufallen, wurden bei dem Faktor angegeben, bei dem sie am stärksten laden. (vgl. Anhang D, Tabelle D-1.)

Der Scree-Plot der Faktorenanalyse der Stimulusmerkmale ergab insgesamt drei bzw. sechs auffällige Faktoren. Drei Faktoren erklären zusammen 46.3% der Varianz, sechs Faktoren erklären zusammen 65.5% Varianz. Insbesondere der erste Faktor tritt mit einem Eigenwert von 9.87 und einer Varianzaufklärung von 23.5% sehr deutlich in Erscheinung. Da die Faktoren 4 bis 6 zusammen deutlich weniger Varianz erklären als die Faktoren 1 bis 3 werden zwei Faktorenanalysen gerechnet, um die Verteilung der Merkmale auf drei bzw. sechs Faktoren zu prüfen. Tabelle IV-1.1.3a (Anhang D) fasst die beschriebenen Verteilungen zusammen.

Bei einer Faktorenanalyse mit drei Faktoren ergibt sich eine Gruppierung der folgenden Merkmale:

Faktor 1: sprachliche Merkmale zur quantitativen Beschreibung der Stimuli
„Wortlänge (WLS)", „Einsilbige Wörter (PEW)", „Lange Wörter (PLW)", „Mehrsilbige Wörter (PMW)", „Inhaltswörter (IWH)", „Worthäufigkeit (GWS)", „Adjazenzstrukturen (STR3)", „Anakoluthe (STR4)", „Deixis (DIE)", „Negationen (NEG)", „Substantive/ Eigennamen/Appellative (SUB)", „Hintergrundwissen (WEL)"

Faktor 2: sprachlich-inhaltliche Merkmale
„Länge in Minuten (LST)", „Wortzahl (AWS)", „Hörkontext (HKO)", „Thema (THE)", „Nähezeichen (STR5)", „Relationstyp Antwort (REL2)", „Uneigentliches Sprechen (RHE2)", „Jugendsprache/Umgangssprache (RHE4)", „Referenzen (REF)", „Anzahl der Propositionen (PRO)"

Faktor 3: überwiegend inhaltliche Merkmale
„Literarischer Stimulus (SLT)", „Funktion (TFU)", „Ellipsen (STR2)", „Verberststellung (STR6)", „Verben (VER)", „Länge der Propositionen (MLP)"

Die Merkmale, die sich um Faktor 1 gruppieren, können zusammenfassend als sprachliche Merkmale zur quantitativen Beschreibung der Stimuli bezeichnet werden. Demgegenüber stehen bei den Faktoren 2 und 3 stärker Merkmale, die auch inhaltliche Aspekte der Stimuli abdecken und auf die Bildhaftigkeit der Stimuli abzielen. Die Korrelationsmatrix der Faktoren verdeutlicht, wie die Faktoren miteinander zusammenhängen. Die Korrelationen der Faktoren untereinander sind bei drei Faktoren zu vernachlässigen, sodass davon auszugehen ist, dass tatsächlich alle drei Faktoren Unterschiedliches erfassen (vgl. Tabelle IV-1.1.3b., Anhang D).

Bei einer Faktorenanalyse mit sechs Faktoren ergibt sich eine Gruppierung der folgenden Merkmale:

Faktor 1: sprachliche Merkmale zur quantitativen Beschreibung der Stimuli
„Wortlänge (WLS)", „Einsilbige Wörter (PEW)", „Lange Wörter (PLW)", „Mehrsilbige Wörter (PMW)", „Inhaltswörter (IWH)", „Worthäufigkeit (GWS)", „Adjazenzstrukturen (STR3)", „Anakoluthe (STR4)", „Negationen (NEG)", „Substantive/Eigennamen/Appellative (SUB)", „Hintergrundwissen (WEL)"

Faktor 2: Merkmale zur Beschreibung der Länge und der Komplexität des Stimulus
„Länge in Minuten (LST)", „Wortzahl (AWS)", „Hörkontext (HKO)", „Thema (THE)", „Nähezeichen (STR5)", „Uneigentliches Sprechen (RHE2)", „Jugendsprache/Umgangssprache (RHE4)", „Referenzen (REF)", „Anzahl der Propositionen (PRO)"

Faktor 3: inhaltliche Merkmale
„Literarischer Stimulus (SLT)", „Funktion (TFU)", „Verberststellung (STR6)", „Verben (VER)", „Länge der Propositionen (MLP)"

Faktor 4: Merkmale gesprochener Sprache
„Sprechgeschwindigkeit (SGS)“, „Akzent/Dialekt/Aussprache (AST)“, „Referenz-Aussage- Strukturen (STR1)“, „Ellipsen (STR2)“, „Neudeutsch/Anglizismen (RHE3)“

Faktor 5: Merkmale zur Erfassung der literarischen Qualität
„Bildliche Darstellungsformen (RHE1)“, „Deixis (DEI)“, „Wiederaufnahmen (WIE)“, „Inferenzen (SFI)“

Faktor 6: Merkmale zur Erfassung des Diskurstyps
„Anzahl der Sprecher (ASP)“, „Relationstyp: Antwort (REL2)“, „Relationstyp: Spezifizierung (REL3)“

Die Faktoren 1 bis 3 sind mit ähnlichen Merkmalen wie bei der Faktorenanalyse mit drei Faktoren belegt. Faktor 4 scheint sich stark auf die Darbietungsmodalität und die Besonderheiten in der Präsentation durch unterschiedliche Sprecher zu beziehen. Faktor 5 bezieht sich auf die literarische Qualität der Stimuli und Faktor 6 gruppiert Merkmale zur Erfassung des Diskurstyps.

Die Korrelationsmatrix zeigt auch bei sechs Faktoren höchstens moderate Korrelationen zwischen den Faktoren an (vgl. Tabelle IV-1.1.3c., Anhang D), so dass angenommen werden kann, dass tatsächlich allen sechs Faktoren Unterschiedliches zugrunde liegt.

1.1.4. Faktorenanalysen der Stimulusmerkmale aus dem Lehrerfragebogen
Die Ergebnisse der Lehrereinschätzungen wurden sowohl auf Aufgabenebene als auch auf Ratingebene untersucht, um mögliche Effekte intraindividueller und interindividueller Strukturen auf einzelne Aufgaben zu identifizieren. Auf der intraindividuellen Ebene werden die Korrelationen der Angaben zu den einzelnen Fragebogenkategorien eines Raters über alle Aufgaben geprüft. Hier spielen demnach die Mittelwerte auf dem Fragebogen über alle Aufgaben eine Rolle. Auf der interindividuellen Ebene finden die Korrelationen der Angaben aller Rater zu einer Aufgabe hinweg Beachtung, es wird also mit den Mittelwerten über alle Ratings gearbeitet.

Mittels einer Faktorenanalyse wurden bei der interindividuellen Analysen auf Aufgabenebene und die der intraindividuellen Analyse auf Ratingebene jeweils drei dominante Faktoren identifiziert (vgl. Tabelle IV-1.1.4a., Anhang D). Diese drei Faktoren erklären auf Aufgabenebene zusammen 85% der Varianz, wobei der erste Faktor mit 59% Varianzaufklärung und einem Eigenwert von 8.32 deutlich dominiert. Auf Ratingebene wird insgesamt von den drei Faktoren nur 57% Varianz erklärt und auch der erste Faktor ist mit einem Eigenwert von 4.12 und einer Varianzaufklärung von 29% deutlich schwächer. Dieses Ergebnis kann mit der individuellen Fehlerstruktur der Ratings erklärt werden.

Für die Gruppierung der Merkmale um Faktoren wurden nur Ladungen größer oder kleiner gleich ±0.4 berücksichtigt. Merkmale, die mehr als einem Faktor zufallen, wurden bei dem Faktor angegeben, mit dem sie am stärksten korrelierten. Die Faktorstruktur der beiden Rotationstypen Varimax und Promax fällt auf Aufgabenebene relativ ähnlich, auf Ratingebene sogar identisch aus. Ein Vergleich der Ergebnisse der beiden Ebenen zeigt, dass auf Rating-

ebene weniger Mehrfachbelegungen auftreten. Insgesamt fallen die Ergebnisse für beide Analyseebenen jedoch unterschiedlich aus und können deshalb nicht aufeinander übertragen werden. (vgl. Anhang D, Tabelle D-2.). Es zeigt sich auf Aufgabenebene folgende Gruppierung der Merkmale zu Faktoren:

Faktor 1: Merkmale zur sprachlichen Gestaltung
„Vertrautheit mit dem Thema (VTH)", „Wortschatz: abstrakt – konkret (WAK)", „Grammatik (GRA)", „Gesamteindruck: elegant – unbeholfen (GEU)", „Ausdrucksweise: gewählt – umgangssprachlich (AGU)", „Ausdrucksweise: komplex/ausschweifend - einfach/knapp (AKE)", „Informationsebene (INF)"

Faktor 2: Merkmale zur Wirkung
„Wortschatz: fachspezifisch – alltäglich (WFA)", „Kohärenz/Textzusammenhang (KOH)", „Ton (TON)", „Wirkung: regt zum Nachdenken an – regt zum Lachen an (WNL)", „Wirkung: ernsthaft – humorvoll (WEH)"

Faktor 3: Gesamteindruck
„Gesamteindruck: interessant – uninteressant (GIU)", „Gesamteindruck: abwechslungsreich – eintönig (GAE)"

Auf Ratingebene gruppieren sich die Merkmale folgendermaßen:

Faktor 1: Merkmale zur sprachlichen Gestaltung
„Vertrautheit mit dem Thema (VTH)", „Wortschatz: abstrakt – konkret (WAK)", „Wortschatz: fachspezifisch – alltäglich (WFA)", „Grammatik (GRA)", „Ausdrucksweise: gewählt – umgangssprachlich (AGU)", „Ausdrucksweise: komplex/ausschweifend - einfach/knapp (AKE)", „Informationsebene (INF)"

Faktor 2: Merkmale zum Gesamteindruck
„Gesamteindruck: interessant – uninteressant (GIU)", „Gesamteindruck: abwechslungsreich – eintönig (GAE)", „Ton (TON)"

Faktor 3: Merkmale zur Wirkung „
Kohärenz (KOH)", „Gesamteindruck: elegant – unbeholfen (GEU)", „Wirkung: regt zum Nachdenken an – regt zum Lachen an (WNL)", „Wirkung: ernsthaft – humorvoll (WEH)"

Tabelle IV-1.1.4b., Anhang D zeigt die Korrelationen der einzelnen Faktoren untereinander auf Aufgaben- und auf Ratingebene. Auf Aufgabenebene korrelieren die Faktoren 1 und 2 stark ($r = 0.76$) und scheinen demnach ähnliche Inhalte zu haben. Faktor 3 scheint dagegen etwas anderes abzudecken. Inhaltlich sind dem Faktor 3 die Fragen zum Gesamteindruck zuzuordnen, bei den Faktoren 1 und 2 überschneiden sich hingegen Fragen zur sprachlichen Gestaltung sowie eher globalere Angaben zum Ton und zur Wirkung der Stimuli.

Auf Ratingebene fallen die Korrelationen der einzelnen Faktoren untereinander deutlich schwächer aus, was dafür spricht, dass alle Faktoren unterschiedliche Inhalte gruppieren. Betrachtet man die Merkmale, die sich auf Ratingebene um die drei Faktoren gruppieren, so wird auch eine inhaltliche Beschreibung der Faktoren einfacher als auf Aufgabenebene. Es handelt sich bei Faktor 1 im Wesentlichen um sprachliche Merkmale, bei Faktor 2 um den Gesamteindruck und bei Faktor 3 um die Wirkung.

Tabelle IV-1.1.4c. gibt einen Überblick darüber, welche Variablen sich auf Aufgaben- und Ratingebene überschneiden:

Tabelle IV-1.1.4c.: Vergleich Merkmalsgruppen Aufgabenebene und Ratingebene

Merkmalsgruppen	nur Aufgabenebene	Variablen gemeinsam	nur Ratingebene
I: Merkmale zur sprachlichen Gestaltung	GEU	VTH, WAK, GRA, AGU, AKE, INF	WFA
II: Merkmale zur Wirkung	WFA, TON	KOH, WNL, WEH	GEU
III: Gesamteindruck		GIU, GAE	TON

Die Merkmale „Gesamteindruck: elegant – unbeholfen (GEU)", „Wortschatz: fachspezifisch – alltäglich (WFA)" und „Ton (TON)" scheinen Eigenschaften beider Gruppen zu besitzen. Die Variablen TON und WFA gehören dabei zu den am stärksten mit der Item- und der Aufgabenschwierigkeit korrelierenden Variablen des Lehrerfragebogens (r = 0.50 bzw. r = -0.55).

1.2. Itemmerkmale

1.2.1. Korrelationsübersicht der Itemmerkmale

Die Ergebnisse der Korrelationsanalysen der Itemmerkmale mit der Itemschwierigkeit sind in Anhang C, Tabelle C-3. dargestellt. Bei den Korrelationsanalysen der Itemmerkmale mit der Itemschwierigkeit treten zwar höhere Werte auf als bei den Analysen mit den Stimulusvariablen zu beobachten waren, es korrelieren insgesamt aber nur sehr wenige Merkmale signifikant mit der Itemschwierigkeit. Als Grenzwert für einen bedeutsamen Wert wurde analog zu den Analysen der Stimulusmerkmale mit den Itemschwierigkeiten ±0.2 gewählt. Bemerkenswert ist dabei auch, dass von den insgesamt sieben Merkmalen vier dem Itemformat zuzurechnen sind und sich zwei auf die Plausibilität der Distraktoren, also einem Merkmal von MC-Aufgaben, beziehen. Inhaltliche Merkmale, wie die vom Item verlangte kognitive Operation oder die Art der gesuchten Information scheinen eine untergeordnete Rolle zu spielen.

1.2.1.1. Merkmalsgruppe I: Itemformat

Ankreuzformate wirken deutlich schwierigkeitssenkend, wohingegen Antwortformate, bei denen die Schüler selbst formulieren müssen, schwieriger sind. Dieser Befund wird von den beiden, zum Itemformat kodierten Variablengruppen bestätigt. Sowohl die Variablen IFK.GA (Geschlossen-Ankreuzen Item) bzw. IFK.01 (Einfache Antwort – 0/1-Kodierung), die sich auf die Kodierung der Items beziehen, als auch die Variablen IFA.HO (halboffenes Item) bzw. IFA.RF (Richtig-Falsch-Item), die auf dem Offenheitsgrad der Items fokussieren, korrelieren positiv bzw. negativ mit der Itemschwierigkeit und sind auch in der Höhe ihres Zusammenhangs sehr ähnlich.

1.2.1.2. Merkmalsgruppe II: Merkmale der Itempräsentation
Die Variable „Position des Items innerhalb der Aufgabe (PIA)“ korreliert weniger stark mit der Itemschwierigkeit als die Variable „Zeitpunkt der Itembearbeitung (ZIB)“. Werden die Items erst nach dem Zuhören beantwortet, so fällt die Itemschwierigkeit niedriger aus. Offenbar ist es schwieriger für die Schüler, mehrere Tätigkeiten gleichzeitig auszuführen (zuhören, nachdenken, schreiben), als zunächst den ganzen Stimulus anzuhören und ein mentales Modell des Gehörten aufzubauen.

1.2.1.3. Merkmalsgruppe III: Merkmale von MC-Items
Die höchste Korrelation mit der Itemschwierigkeit weist die Variable „Größte vorkommende Plausibilität der Distraktoren (PDI)“ auf. Je plausibler der beste Distraktor eines MC-Items eingeschätzt wurde, desto schwieriger ist das Item. Die Variable „Mittlere Plausibilität der Distraktoren (MPD)“ korreliert auch mit der Itemschwierigkeit, jedoch deutlich schwächer. Dieser Befund ist plausibel, da in der Regel eine sehr attraktive falsche Antwortoption von der richtigen ablenkt und das Mittel der Attraktivität der Distraktoren weniger aussagekräftig ist.

1.2.1.4. Merkmalsgruppe IV: Kognitive Anforderungen der Items
Insgesamt hängen die Merkmale der Gruppe IV eher schwach mit der Itemschwierigkeit zusammen. Geringfügige positive Korrelationen weisen die Variablen „Anforderungsbereich Items (AFB)“, „Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden (BS113)“,„Konkretheit der NI (TCO/TOR)“ auf.

1.2.1.5. Merkmalsgruppe V: Globalurteil
Die Variable „Eingeschätzte Itemschwierigkeit durch die Aufgabenentwickler (SEA)“ korreliert nur sehr schwach mit der Itemschwierigkeit.

1.2.2. Faktorenanalysen der Itemmerkmale
Für eine Faktorenanalyse der Itemmerkmale wurden die Variablen in drei Gruppen untersucht, wobei manche Variablen in mehreren Gruppen auftreten. In der ersten Gruppe treten hauptsächlich Merkmale auf, die allgemeine Itemmerkmale beschreiben. In die zweite Gruppe fallen Merkmale, die für Multiple-Choice-Items zutreffen. In der dritten Gruppe stehen Merkmale im Vordergrund, die die zur Itembeantwortung notwendige Information (NI) betreffen. Für die Faktorenanalyse der ersten Merkmalsgruppe wurde mit dem gesamten Itemdatensatz gearbeitet. Für die Analyse der zweiten Gruppe fand eine Filterung nach MC-Items statt und für die Analyse der dritten Gruppe, wurden nur die Fälle verwendet, in denen genau eine Information vom Item verlangt wurde (ANI = 1). Für die Gruppierung der Itemmerkmale um Faktoren wurden nur Ladungen größer oder kleiner gleich ±0.4 berücksichtigt (vgl. Anhang D, Tabelle D-3.). Merkmale, die mehr als einem Faktor zufallen, wurden stets bei dem Faktor angegeben, bei dem sie am stärksten laden.

1.2.2.1. Allgemeine Itemmerkmale
In der Faktorenanalyse wurden für die Itemmerkmale vier dominante Faktoren identifiziert. Diese vier Faktoren decken zusammen über 38% der Varianz auf (vgl. Tabelle IV-1.2.2.1a., Anhang D). Die Faktorstruktur der beiden Rotationstypen Varimax und Promax fällt sehr ähnlich aus. Es zeigt sich auf Aufgabenebene folgende Gruppierung der Merkmale zu Faktoren:

Faktor 1: Merkmale zur Beschreibung der kognitiven Anforderungen der Items
„Anforderungsbereich (AFB)", „Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben (BS142)", „Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln (BS143)", „Typ der NI: Detail (TNI2)", „Typ der NI: Stimmung/Einstellung des Sprechers (TNI4)", „Typ der NI: Geräusche/ Paraverbales (TNI9)"

Faktor 2: Merkmale zur Konkretheit der vom Item verlangten Information
„Konkretheit der NI (33-stufig) (TCO)" und „Konkretheit der NI (5-stufig) (TOR)", „Typ der NI: Sprecheridentifikation (TNI11)"

Faktor 3: Merkmale zum Itemformat
„Itemformat: Ankreuzen (MC+RF) (IFA.AK)", „Itemformat: Schreibitem (HO+OI) (IFA.S)"

Faktor 4: Merkmale zur Beschreibung der vom Item verlangten Information (NI)
„Typ der NI: Meinung des Zuhörers (TNI3)", „Typ der NI: Schlussfolgerung (TNI5)", „Typ der NI: Kommunikativer Zweck/Funktion/Wirkung des Stimulus (TNI6)"

Die Merkmale, die sich um Faktor 1 gruppieren, beschreiben die kognitiven Anforderungen der Items, wobei einerseits explizit Genanntes (z. B. Details), aber auch nonverbale Informationen (z. B. Stimmungen, Geräusche) angesprochen werden. Faktor 2 sammelt Merkmale zur Konkretheit der vom Item verlangten Information. Faktor 3 gruppiert Merkmale zum Itemformat, wobei zwischen geschlossenen und offenen Formattypen (Ankreuzen vs. Schreiben) unterschieden wird. Der vierte Faktor fokussiert auf der zur Itembeantwortung notwendigen Information, und zwar stehen hier Aspekte im Vordergrund, die sich auf implizite Informationen im Stimulus beziehen und eine Übertragungsleistung des Zuhörers erfordern. Die vier Faktoren korrelieren kaum miteinander, sodass davon auszugehen ist, dass sie sich inhaltlich wenig überschneiden. (vgl. Tabelle IV-1.2.2.1b., Anhang D)

1.2.2.2. Merkmale zur Beschreibung von MC-Items
Auch für die zweite Gruppe wurden vier Faktoren identifiziert, die gemeinsam dieser drei über 76% Varianz aufklären. Dabei fällt dem ersten Faktor bereits ein Anteil von 30% zu. (vgl. Tabelle IV-1.2.2.2a., Anhang D). Die Merkmale gruppieren sich um die folgenden vier Faktoren:

Faktor 1: Merkmale zur Einschätzung der Plausibilität der Distraktoren
„Größte vorkommende Plausibilität der Distraktoren (PDI)", „Mittlere Plausibilität der Distraktoren (MPD)", „Itemschwierigkeit (SEA)"

Faktor 2: Merkmale zu den Rahmenbedingungen der Items
„Zeitpunkt der Itembearbeitung (ZIB)", „Position des Items innerhalb der Aufgabe (PIA)", „Anforderungsbereich (AFB)"

Faktor 3: Merkmal Hintergrundwissen
„Hintergrundwissen (HGW)"

Faktor 4: Merkmal Position des Attraktors
„Position des Attraktors im MC-Item (PMC)“

Um Faktor 1 gruppieren sich Variablen, die eine Einschätzung der Plausibilität (und damit Schwierigkeit) der Distraktoren bzw. der Items enthalten. Um Faktor 2 gruppieren sich Merkmale, die objektivere Angaben zu den MC-Items machen und auf die Rahmenbedingungen der Items eingehen. Auf dem dritten und vierten Faktor lädt jeweils nur ein Merkmal, und zwar Hintergrundwissen auf Faktor 3 und die Variable PMC, die Angaben zum Attraktor macht, auf Faktor 4. Die Faktoren korrelieren kaum miteinander, wobei der höchste Zusammenhang zwischen dem zweiten und dem vierten Faktor auftritt (vgl. Tabelle IV-1.2.2.2b., Anhang D).

1.2.2.3. Merkmale zur Beschreibung der NI
Tabelle IV-1.2.2.3a., Anhang D gibt die Eigenwerte und die kumulative Varianzaufklärung der identifizierten Faktoren der dritten Merkmalsgruppe wieder. In der dritten Gruppe liegt die gemeinsame Varianzaufklärung der identifizierten drei Faktoren bei über 48%, wobei dem ersten Faktor der größte Anteil daran zufällt. Folgende Merkmale gruppieren sich um die drei Faktoren:

Faktor 1: Merkmale zur Beschreibung der vom Item verlangten Kompetenzen
„Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben (BS142)“, „Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln (BS143)“

Faktor 2: Merkmale zur Differenzierung der kognitiven Operationen
„Konkretheit der NI (5-stufig) (TOR)“, „Anforderungsbereich (AFB)“, „Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden (BS113)“

Faktor 3: Merkmal zur Beschreibung der verschiedenen Formen von Mündlichkeit
„Gesprächsbeiträge anderer verfolgen und aufnehmen (BS141)“

Faktor 1 gruppiert Variablen, die die zur Itembeantwortung benötigten Kompetenzen beschreiben. Um Faktor 2 gruppieren sich Merkmale, die die weiter die kognitiv verlangten Operationen differenzieren. Auf Faktor 3 lädt ein Merkmal, das auf die verschiedenen Formen der Mündlichkeit hin abzielt. Die drei Faktoren korrelieren kaum miteinander, sodass angenommen werden kann, dass sie tatsächlich unterschiedliche Inhalte versammeln (vgl. Tabelle IV-1.2.2.3b., Anhang D).

1.3. Personenmerkmale

1.3.1. Einschätzung der Stimuli durch die Schüler

Im Rahmen der Befragung der Schüler hinsichtlich ihrer Vertrautheit mit dem Thema, ihres Vorwissens und Interesses in Bezug auf die Stimuli sowie der wahrgenommenen Präsentationsqualität wird untersucht, ob diese personenbezogenen Angaben einen Zusammenhang mit den Leistungsdaten der Schüler zeigen. Es wird angenommen, dass motivierte Schüler

im Test besser abschneiden als Schüler, die die Zuhöraufgaben unmotiviert bearbeiten. Auch Schüler, die Vorwissen zu einem in den Stimuli angesprochenen Thema besitzen, oder mit einem der Stimuli bereits im Unterricht oder privat konfrontiert wurden, könnten einen Vorteil daraus ziehen. Nicht zuletzt kann auch die akustische Qualität der vorgespielten Stimuli Ausschlag für gute oder schlechte Leistungen geben. Ein Schüler, der im Klassenzimmer sehr weit vom Lautsprecher entfernt sitzt und aufgrund von Störgeräuschen im Raum oder in der Umgebung des Klassenzimmers (Mitschüler, die sich räuspern, Pausengong, Feueralarm, laute Schüler auf dem Gang, vorbeifahrender Verkehr, etc.) sowie aufgrund schlechter Tonqualität der Aufnahme (kleines Tonbandgerät mit schlechter Akustik, falsch eingestellte Lautstärke etc.) dem Hörbeitrag folgt, schneidet unter Umständen anders ab, als ein Schüler, der keinerlei Störungen oder Unterbrechungen des Hörbeitrags erlebt.

Insgesamt liegen für 15 Aufgaben zwischen 405 und 957 Schülereinschätzungen mit den Polen 1 = „trifft überhaupt nicht zu" und 5 = „trifft voll und ganz zu" vor. Die Anzahl der Antworten ist abhängig vom Testdesign. Cronbach's Alpha für die vier Variablen der Schülereinschätzungen beträgt 0.617. Die Mittelwerte der Schülerantworten liegen für die Variable „Vertrautheit mit dem Thema (VTS)" zwischen 1.75 und 2.99. Am bekanntesten wurde der Stimulus „G111_Vorstellungsgespräch" eingeschätzt, ein nicht-literarischer Stimulus über das korrekte Verhalten bei Vorstellungsgesprächen. Am wenigsten bekannt ist der Stimulus „G120_Ida Ehre spricht", bei dem eine Hamburger Schauspielerin im Mittelpunkt steht.

Bei der Variable „Motivation/Interesse (MOT)" liegen die Einschätzungen zwischen 2.07 und 2.89. Die Stimuli wurden also im Mittel als mäßig interessant eingestuft. Den höchsten Wert erzielt der Stimulus „G109_Torhüter", ein literarischer Beitrag über den Torhüter Lars Leese. Den niedrigsten Wert erzielte der Stimulus „G119_Installateur", bei dem ein Arbeitgeber seinen Angestellten verschiedene Arbeitsaufträge erteilt.

Die Mittelwerte der Variable „Bekanntheitsgrad des Stimulus (BEK)" liegen zwischen 1.35 und 1.66. Die eingesetzten Stimuli sind – wie gewünscht – also wenig bekannt. Da es sich bei den Stimuli im Wesentlichen um Radiobeiträge oder z. T. auch eigens für den Test geschriebene und vertonte Texte (z. B. „G119_Installateur") handelt, bleibt fraglich, ob es realistisch ist, dass einige Schüler, die eingesetzten Stimuli tatsächlich vor dem Test kannten. Dies ist zumindest im Fall der Aufgabe „G119_Installateur" nahezu unmöglich. Zu vermuten ist deshalb, dass viele Schüler in ihrem Ankreuzverhalten die Randbereiche der Skala ausgespart haben bzw. bei dieser Frage „Ich kannte den Text vorher schon sehr gut." die Modifizierung beantworteten. Die Itemformulierung wurde diesbezüglich schon im Vorfeld der Erhebung als kritisch betrachtet, jedoch beibehalten, da sie formell so in die fünfstufige Gesamtskala passte.

Die Mittelwerte bei der Variable „Verständlichkeit (VST)" liegen zwischen 2.60 und 3.87. Der niedrigste Wert wurde für den Stimulus „G105_Schulbesuch" erzielt. Dabei handelt es sich um eine Live-Reportage, bei der ein Reporter eine Klasse im Unterricht besucht. Die akustische Schwierigkeit liegt bei diesem Stimulus u. a. darin, dass viele unterschiedliche Personen sprechen und auch schulspezifische Nebengeräusche (z. B. eine Durchsage, Pausengong) auftreten. Erfreulicherweise liegen alle anderen Stimuli mit ihrem Mittelwert mindestens beim

Wert 3. Die Klangqualität der Stimuli wurde von den Schülern also als einigermaßen hinreichend empfunden.

Um mögliche Zusammenhänge der Schülereinschätzungen mit den Leistungsdaten zu bestimmen, werden die Schülerleistungen als Summenscores der Aufgaben, d. h. die Anzahl der richtig gelösten Items pro Aufgabe, mit den Einschätzungen korreliert. Bei einer ersten globalen Prüfung der Korrelationen der Aufgaben über alle Testhefte hinweg mit den Schülereinschätzungen dazu, ergeben sich keine signifikanten Zusammenhänge.

In einem nächsten Schritt werden die Zusammenhänge der Summenscores mit den Schülereinschätzungen für die Blöcke H8, H10, H4 und H1 ermittelt (vgl. Tabelle IV-1.3.1a. und Tabelle IV-1.3.1b.) Hier zeigen sich signifikante Korrelationen der Schülerangaben „Vertrautheit mit dem Thema", „Motivation/Interesse" und „Verständlichkeit" mit der mittleren Schwierigkeit der einzelnen Aufgaben. Durch die Spalten H1, H4 bzw. H8 und H10 wird die mittlere Korrelation des jeweiligen Blocks mit der Schwierigkeit ausgedrückt.

Tabelle IV-1.3.1a.: Korrelationsübersicht auf Blockebene: Summenscore – Schülereinschätzung (Block H1 und H4)

	Block H1					*Block H4*			
Aufgabe	**G105**	**G114**	**G118**	**G119**	**H1**	**G103**	**G111**	**G121**	**H4**
Vertrautheit mit dem Thema (VTS)	0.13*	0.12	0.03	0.00	0.07	0.11*	0.04	0.15*	0.10
Motivation/Interesse (MOT)	-0.02	0.11	0.01	0.04	0.03	0.08	0.01	0.12*	0.07
Bekanntheitsgrad des Stimulus (BEK)	-0.10	-0.07	-0.09	-0.05	-0.08	-0.05	-0.10	-0.05	-0.06
Verständlichkeit (VST)	0.06	0.06	0.06	0.18*	0.09	0.05	0.10	0.04	0.06

Anmerkung: * signifikante Zusammenhänge ($p < 0.01$)

Tabelle IV-1.3.1b.: Korrelationsübersicht auf Blockebene: Summenscore – Schülereinschätzung (Block H8 und H10)

	Block H8			*Block H10*		
Aufgabe	**G100**	**G120**	**H8**	**G107**	**G109**	**H10**
Vertrautheit mit dem Thema (VTS)	0.05	0.02	0.05	0.01	0.01	0.06
Motivation/Interesse (MOT)	0.04	0.17*	0.04	0.07	0.14*	0.11
Bekanntheitsgrad des Stimulus (BEK)	-0.14*	-0.11	-0.14	-0.19*	-0.12	-0.16
Verständlichkeit (VST)	0.05	0.03	0.05	0.10	0.09	0.09

Anmerkung: * signifikante Zusammenhänge ($p < 0.01$)

Ein systematischer Zusammenhang der Summenscores über die Blöcke und die Aufgaben hinweg mit den Schülereinschätzungen ist nicht zu erkennen. Zum Teil gibt es zwar signifikante ($p < 0.01$) Zusammenhänge, die insgesamt jedoch mit $r < 0.2$ sehr gering ausfallen. Die mittleren Korrelationen zeigen, dass kein nennenswerter Zusammenhang der Schülereinschätzungen mit den Leistungsdaten der Schüler auf Blockebene besteht.

Die Korrelationen zwischen den Summenscores der einzelnen Aufgaben liegen deutlich höher als die Korrelationen zwischen den Summenscores und den Schülereinschätzungen. Dies wird insbesondere bei Block H1 deutlich (vgl. Tabelle IV-1.3.1c.). Es liegen in diesem Block pro Aufgabe zwischen 452 und 507 Schülereinschätzungen vor.

Tabelle IV-1.3.1c.: Korrelationen der Summenscores mit den Schülereinschätzungen für Block H1

	GH105	**GH114**	**GH118**	**GH119**
GH105	1.00	0.52	0.40	0.41
GH114	0.52	1.00	0.41	0.39
GH118	0.40	0.41	1.00	0.32
GH119	0.41	0.39	0.32	1.00
GH105911a	**0.13**	0.08	0.02	0.10
GH105911b	**-0.02**	-0.09	-0.06	0.04
GH105911c	**-0.10**	-0.16	-0.08	0.00
GH105911d	**0.06**	-0.02	-0.07	0.00
GH114911a	0.05	**0.12**	0.09	0.03
GH114911b	0.07	**0.11**	0.03	0.15
GH114911c	-0.05	**-0.07**	-0.04	-0.06
GH114911d	0.05	**0.06**	0.04	0.00
GH118911a	-0.02	0.02	**0.03**	-0.03
GH118911b	0.04	0.02	**0.01**	0.02
GH118911c	-0.11	-0.14	**-0.09**	-0.07
GH118911d	0.14	0.11	**0.06**	0.07
GH119911a	-0.03	0.04	0.04	**0.00**
GH119911b	-0.05	-0.03	-0.04	**0.04**
GH119911c	-0.07	-0.10	-0.06	**-0.05**
GH119911d	0.08	0.06	-0.02	**0.18**

Anmerkung: fett gedruckt: Korrelationen zwischen den Summenscores der einzelnen Aufgaben und den Schülereinschätzungen

Ein Grund für diese geringen Korrelationen könnte sein, dass die Angaben über alle Schüler hinweg zu wenig aussagekräftig sind. Um zu überprüfen, ob die Ergebnisse für Subgruppen anders ausfallen, werden in einem nächsten Schritt die Berechnungen für die Gruppen „Hauptschüler" und „Gymnasiasten" wiederholt. Die Korrelationen fallen geringfügig höher aus, sind mit Werten bei $p < 0.2$ jedoch noch immer nicht signifikant.

Zusätzlich wird überprüft, ob sich Zusammenhänge zeigen, wenn nur diejenigen Schüler in die Berechnungen mit einbezogen werden, die in ihrem Ankreuzverhalten die Bandbreite der fünf Antwortmöglichkeiten ausschöpfen. Schüler, die für jede Frage beispielsweise nur den mittleren Antwortbereich wählen, also kaum Streuung in ihren Antworten haben, könnten die Zusammenhänge abschwächen. Für die Analysen werden in diesem Schritt nur die Schüler ausgewählt, die in ihrem Ankreuzverhalten eine Standardabweichung von 0.89 aufweisen. Die Analysen werden auf Blockebene für die beiden Gruppen „Hauptschüler" und „Gymnasiasten" wiederholt. Die Zusammenhänge zwischen den Schülerantworten und den Summenscores der Aufgaben bleiben unsystematisch und es sind keine signifikanten Korrelationen zu erkennen. Daraus muss gefolgert werden, dass die Merkmale „Vertrautheit mit dem Thema (VTS)", „Motivation/Interesse (MOT)", „Bekanntheitsgrad des Stimulus (BEK)" und „Verständlichkeit (VST)" keinen Einfluss auf die Schwierigkeit der Stimuli haben.

Das Merkmal „Interessantheit der Stimuli" wurde sowohl von den Schülern im Rahmen der beschriebenen Items als auch von den Lehrkräften im Rahmen des eingesetzten Lehrerfragebogens (vgl. Kapitel 2.2.2. *Stimulusmerkmale aus dem Lehrerfragebogen*) eingeschätzt. Daraus ergibt sich die Frage, ob beide Einschätzungen einander entsprechen. Die Korrelation der beiden Variablen beträgt $r = -0.27$. Der negative Wert ist dem Umstand geschuldet, dass die Skalen unterschiedlich gepolt sind. Die Lehrkräfte schätzten die Aufgaben auf einer Skala von 1 (interessant) bis 5 (uninteressant) ein, die Schüler schätzten die Aufgaben zur Frage „Ich finde den Text sehr interessant." auf einer Skala von 1 (trifft überhaupt nicht zu) bis 5 (trifft voll und ganz zu) ein. In Tabelle IV-1.3.1d. sind die Mittelwerte der Lehrer- bzw. Schülerangaben zum Interessantheitsgrad der Stimuli wiedergegeben. Bei vielen Aufgaben decken sich die Einschätzungen von Lehrern und Schülern im Mittel weitgehend, z. B. bei den Aufgaben „G118_Leila" und „G124_Dreck" (fett markiert). Auch die fünf Aufgaben am Pol „uninteressant" wurden von beiden Gruppen gleich ausgewählt, wenn auch geringfügige Unterschiede in der exakten Position auftreten. Bei manchen Aufgaben gab es jedoch stark unterschiedliche Einschätzungen. Beispielsweise schätzten die Lehrkräfte die Aufgabe „G109_Torhüter" (grau unterlegt) als eher mäßig interessant ein, die Schüler hielten diese Aufgabe jedoch für die interessanteste. Bei der Aufgabe „G105_Schulbesuch" (fett umrandet) verhielt es sich genau umgekehrt. Die Einschätzung der Lehrkräfte fällt hier deutlich positiver aus als das Urteil der Schüler.

Tabelle IV-1.3.1d.: Mittelwerte der Lehrer- und der Schülereinschätzungen zum Interessantheitsgrad der Stimuli

M			
Lehrereinschätzungen		**Schülereinschätzungen**	
uninteressant			
G119	4.00	G120	2.07
G103	3.83	G119	2.07
G120	3.60	G100	2.28
G131	2.83	G103	2.31
G100	2.67	G131	2.39
G104	2.60	G105	2.45
G118	**2.60**	**G118**	**2.46**
G109	2.50	G111	2.54
G122	2.50	G104	2.56
G121	2.40	G110	2.66
G110	2.33	G114	2.75
G111	2.33	G107	2.83
G105	1.83	G121	2.85
G107	1.80	G122	2.86
G124	**1.67**	**G124**	**2.87**
G114	1.50	G109	2.89
interessant			

Anmerkungen: fett gedruckt: die Einschätzungen von Lehrern und Schülern decken sich in etwa; grau unterlegt: stark unterschiedliche Einschätzungen (Schülerrating: interessant - Lehrerrating: uninteressant); fett umrandet: stark unterschiedliche Einschätzungen (Schülerrating: uninteressant – Lehrerrating: interessant)

1.3.2. Arbeitsgedächtnis

Die Ergebnisse des Tests zur Arbeitsgedächtniskapazität korrelieren dagegen etwas stärker mit den Hörleistungen. Der Test wurde von 1991 Schülern bearbeitet. Die Korrelation des Summenscores des Arbeitsgedächtnistests mit den zusammengefassten Ergebnissen des Leistungsteils zum Hörverstehen beträgt 0.21, Cronbach's Alpha beläuft sich auf 0.61.

In einem zweiten Schritt werden die aufgabenspezifischen Korrelationen mit dem Arbeitsgedächtnistest berechnet. Die Korrelationen liegen zwischen 0.02 und 0.44. Ein gewisser Zusammenhang der Höhe der Korrelation ist mit der Länge der Aufgaben erkennbar. Zu dem Stimulus, bei dem die Summenscores am wenigsten stark mit der Arbeitsgedächtniskapazität korrelieren („G118_Leila", r = 0.02), existieren nur vier Items und der Stimulus dauert 2,3 Minuten. Dagegen liegen die Korrelationen bei deutlich längeren Stimuli (z. B. „G100_Promiquiz": 9,5 Min oder „G114_Pinguin": 5,3 Min) bei 0.29 bzw. 0.20. Auch die empirische Item- bzw. Aufgabenschwierigkeit scheint einen leichten Einfluss auf diesen Zusammenhang zu haben. So handelt es sich bei der Aufgabe „G118_Leila" um eine sehr leichte Aufgabe mit einer mittleren

Aufgabenschwierigkeit von -1.84. Eine weitere sehr leichte Aufgabe („G121_Friseur": r = -1.47) liegt mit der Korrelation der Summenscores mit den Ergebnissen des Arbeitsgedächtniskapazitätstests ebenfalls im unteren Bereich bei 0.11. Die Summenscores der schwierigsten Aufgabe („G111_Vorstellungsgespräch": r = 0.77) korrelieren mit r = 0.25 mit dem Test zur Arbeitsgedächtniskapazität. Der Mittelwert der auftretenden Korrelationen für alle Aufgaben fällt mit 0.18 eher gering aus.

Eine getrennte Analyse nach Schulform („Hauptschule" vs. „Gymnasium") ergibt, dass die Arbeitsgedächtniskapazität in Abhängigkeit von der gestellten Aufgabe zu sehen ist (vgl. Tabelle IV-1.3.2.). Dabei ist zu berücksichtigen, dass nicht alle Aufgaben in Hauptschulen eingesetzt wurden. Ein Vergleich ist nur für die Aufgaben G103, G104, G105, G110, G111, G112, G114, G118, G119, G121 und G131 möglich. Die Aufgaben G100, G102, G107, G109, G120, G122 und G124 wurden nicht an Hauptschulen eingesetzt. Bei vielen Aufgaben fallen die Korrelationen sehr gering aus. Bei manchen Aufgaben ergeben sich jedoch schul-formspezifische Unterschiede (fett markierte Werte). Auffällig ist dabei, dass sich insbesondere im Bereich der Hauptschule auch häufig negative Korrelationen ergeben. Je besser also ein Schüler im Arbeitsgedächtnistest abschneidet, desto schlechtere Ergebnisse liegen im Bereich des Hörverstehenstests vor. Ein Beispiel dafür ist die Aufgabe „G104_Reklame", bei der es sich um einen lyrischen Stimulus handelt. Für die in den Gymnasien eingesetzten Aufgaben fallen die Korrelationen i. d. R. positiv aus. Ein leichter Zusammenhang mit dem Aufmerksamkeitstest liegt für die Aufgaben „G109_Torhüter", „G112_Wetterbericht" und „G119_Installateur" vor. Der Mittelwert der Korrelation aller Hauptschulaufgaben mit dem Aufmerksamkeitstest beträgt -0.06, der Mittelwert aller Gymnasialaufgaben 0.09.

Tabelle IV-1.3.2.: Korrelationsübersicht auf Aufgabenebene: Summenscores mit dem Test zur Arbeitsgedächtniskapazität

	Aufgabe	**G100**	**G102**	**G103**	**G104**	**G105**	**G107**	**G109**	**G110**	**G111**
Summenscore G999	alle Schüler	0.29	0.44	0.15	0.04	0.12	0.07	0.15	0.18	0.25
	nur HS			0.00	**-0.38**	-0.09			**-0.15**	0.10
	nur GYM	0.08	0.19	0.02	-0.03	0.08	0.10	**0.22**	-0.10	0.06
	Aufgabe	**G112**	**G114**	**G118**	**G119**	**G120**	**G121**	**G122**	**G124**	**G131**
Summenscore G999	alle Schüler	0.24	0.20	0.02	0.15	0.21	0.11	0.25	0.22	0.22
	nur HS	0.00	0.10	-0.14	-0.06		**-0.15**			0.05
	nur GYM	**0.19**	0.10	**0.18**	**0.24**	-0.03	0.01	0.08	0.12	0.04

Anmerkung: fett markiert: schulformspezifische Unterschiede des Aufmerksamkeitstests

1.3.3. Sprachkenntnisse

Angaben zu den beiden Fragen „Wie oft sprichst du zu Hause Deutsch?" und „Welche Sprache hast du in deiner Familie zuerst gelernt (Muttersprache)?" liegen jeweils von über 19 000 Schülern vor. 15 356 Schüler (79.1 %) gaben an, zu Hause immer Deutsch zu sprechen, 3 666 Schüler (18.9 %) sprechen zu Hause manchmal Deutsch und manchmal eine andere Sprache

und 379 Schüler (2.0 %) gaben an, zu Hause niemals Deutsch zu sprechen. 2 183 Schüler (11.3 %) gaben an, Deutsch nicht als Muttersprache gelernt zu haben. Die Ergebnisse fallen erwartungskonform aus. Die Testergebnisse im Bereich Zuhören fielen für Schüler niedriger aus, die angaben zu Hause nur manchmal oder niemals Deutsch zu sprechen. Dementsprechend hingen bessere Testleistungen mit der Angabe zusammen, deutsch als Muttersprache gelernt zu haben. Die Korrelationen fallen im Bereich der Hauptschule etwas stärker als im Gymnasium aus. Tabelle IV-1.3.3a. zeigt die Korrelationen der Angaben zum Sprachstand mit den Leistungsdaten des IQB-Hörverstehenstests.

Tabelle IV-1.3.3a.: Korrelation Angaben zum Sprachstand – Leistungsdaten Zuhören

	Wie oft sprichst du zu Hause Deutsch?	**Welche Sprache hast du in deiner Familie zuerst gelernt (Muttersprache)?**
alle Schüler	-0.22	0.22
nur Hauptschüler	-0.23	0.24
nur Gymnasiasten	-0.15	0.15

Eine Analyse auf Aufgabenebene zeigt stark unterschiedliche Zusammenhänge der Leistungsergebnisse mit den Angaben zum Sprachstand (vgl. Tabelle IV-1.3.3b.). Bei einigen Aufgaben, z. B. „G114_Pinguin", besteht kein beobachtbarer Zusammenhang zwischen den Testergebnissen und den im Fragebogen gemachten Aussagen. Bei anderen Aufgaben (fett markiert) fällt der Zusammenhang hingegen deutlicher aus. Insbesondere die Angabe, ob Deutsch als Muttersprache gelernt wurde, korreliert mit den Testergebnissen der Schüler. Ein Zusammenhang ist bei den Aufgaben G100, G103, G107, G110, G111, G120, G122 und G124 zu beobachten. Bis auf die Aufgaben „G122_Frauenfußball" und „G124_Dreck" handelt es sich bei diesen Aufgaben um eher schwierige Aufgaben. Die Aufgaben unterscheiden sich jedoch in der Anzahl ihrer Items, der Thematik und ihrer Machart, sodass keine zuverlässige Systematik erkannt werden kann.

Tabelle IV-1.3.3b.: Korrelation Angaben zum Sprachstand – Leistungsdaten Zuhören: Aufgabenspezifisch

	Aufgabe	**G100**	**G102**	**G103**	**G104**	**G105**	**G107**	**G109**	**G110**	**G111**
Summenscore G999	alle Schüler	0.29	0.44	0.15	0.04	0.12	0.07	0.15	0.18	0.25
	nur HS			0.00	**-0.38**	-0.09			**-0.15**	0.10
	nur GYM	0.08	0.19	0.02	-0.03	0.08	0.10	**0.22**	-0.10	0.06
	Aufgabe	**G112**	**G114**	**G118**	**G119**	**G120**	**G121**	**G122**	**G124**	**G131**
Summenscore G999	alle Schüler	0.24	0.20	0.02	0.15	0.21	0.11	0.25	0.22	0.22
	nur HS	0.00	0.10	-0.14	-0.06		-0.15			0.05
	nur GYM	**0.19**	0.10	**0.18**	**0.24**	-0.03	-0.01	0.08	0.12	0.04

Anmerkung: fett markiert: Korrelationen Angaben zum Sprachstand mit den Leistungsdaten Zuhören ⩾ 0.2

1.4. Zusammenfassung der explorativen Dimensionsanalysen

1.4.1. Zusammenfassung Stimulusmerkmale

Variablen, die die Komplexität des Wortschatzes und sprachliche Merkmale beschreiben (Gruppe I), zeigen den größten Zusammenhang mit der Aufgabenschwierigkeit. Dabei fallen insbesondere Auszählvariablen wie die Wortlänge ins Gewicht. Merkmale, die überwiegend in den Bereich gesprochener Sprache fallen und hauptsächlich die Diskurse betreffen (z. B. „Bildliche Darstellungsformen (RHE1)", „Uneigentliches Sprechen (RHE2)", „Neu-deutsch/Anglizismen (RHE3)", „Jugendsprache/Umgangssprache (RHE4)", „Referenz-Aussage-Strukturen (STR1)", etc.), korrelieren nicht signifikant mit der Aufgabenschwierigkeit.

Bei den Präsentationsmerkmalen (Gruppe II) zeigen lediglich die „Anzahl der Sprecher (ASP)" und die „Sprechgeschwindigkeit (SGS)" einen Zusammenhang mit der Schwierigkeit. Die Gesamtlänge des Stimulus scheint für die Aufgabenschwierigkeit kaum ins Gewicht zu fallen. Weder die Variable „Länge in Minuten (LST)" noch die Variable „Wortzahl (AWS)" korrelieren bedeutend mit der Item- bzw. der Aufgabenschwierigkeit. Dass das Merkmal LST kaum Zusammenhang mit der Schwierigkeit zeigt, erstaunt, da die Auszählmerkmale der Gruppe I (z. B. WLS, PLW und PMW) einen deutlichen Einfluss auf die Schwierigkeit aufwiesen und auch das Merkmal AWS mit der Schwierigkeit zusammenhängt. Offenbar spielt jedoch die Stimuluslänge eine weniger bedeutende Rolle für die Stimulusschwierigkeit und entscheidend ist vielmehr dessen sprachliche und inhaltliche Komplexität. Auch das Merkmal „Akzent/Dialekt/Aussprache (AST)" weist kaum Zusammenhang mit der Aufgabenschwierigkeit auf. Da bei der Auswahl der Stimuli jedoch darauf geachtet wurde, keine Materialien zu verwenden, die stärker dialektal gefärbt sind, ist dieses Ergebnis plausibel. So weisen die Stimuli höchstens leichte Akzentfärbung oder in Einzelfällen einen isolierte dialektale Ausdrücke auf. Das geratete Material war deshalb möglicherweise zu wenig umfangreich, um zu aussagefähigen Ergebnissen zu kommen.

Auch die Gruppe III mit den inhaltlich-thematischen Merkmalen „Literarischer Stimulus (SLT)", „Hörkontext (HKO)" und „Hintergrundwissen (WEL)" zeigt größere Zusammenhänge mit der Aufgabenschwierigkeit. Weder die Funktion (TFU) der Stimuli noch die Variable „Thema (THE)" korreliert mit der Aufgabenschwierigkeit. In der Testsituation scheint es also wenig bedeutend zu sein, mit welchem Thema Hörverstehen geprüft wird.

Aus der vierten Gruppe „Struktur der Stimuli und propositionale Dichte" korrelieren ebenfalls nur zwei Merkmale mit der Aufgabenschwierigkeit, und zwar die Relationstypen „Frage/Impuls/Themensetzung (REL1)" und „Reihenfolge/Aufzählung (REL5)". Keinen Zusammenhang mit der Stimulusschwierigkeit zeigen die Relationstypen „Antwort (REL2)", „Spezifizierung (REL3)", „Erklärung/Beweis/Ursache (REL4)" und „Ziel/Bedingung (REL6)". Auch die Merkmale „Schlussfolgerungen/Inferenzen (SFI)" und „Anzahl/Anteil der Propositionen (PRO/PRW)" hängen nicht mit der Schwierigkeit zusammen. Es wurde angenommen, dass die Anzahl der notwendigen Schlussfolgerungen (Merkmal „Schlussfolgerungen/Inferenzen (SFI)) einen Einfluss auf die Schwierigkeit hat. Diese Annahme konnte jedoch nicht bestätigt werden. Die Anzahl der notwendigen Schlussfolgerungen könnte eher als Itemmerkmal („Typ der NI (TNI)" – Code 5)

eine Rolle spielen. Ferner ist es nicht auszuschließen, dass sich viele Einschätzungen des Merkmals SFI mit dem Merkmal „Hintergrundwissen (WEL)“ überschneiden und zahlreiche Stimulusstellen eher dem Merkmal WEL zugeordnet wurden. Der mangelnde Einfluss der Variable SFI ist aus diesem Grund möglicherweise dem Ratingverfahren geschuldet. Die Stimuluslänge, ausgedrückt durch die Anzahl/den Anteil der Propositionen (PRO/PRW), ist als schwierigkeitsbeeinflussendes Merkmal weniger aussagekräftig als die Dichte des Stimulus, ausgedrückt durch die mittlere Länge der Propositionen (MLP).

In Tabelle IV-1.4.1a. werden die tatsächlichen Effekte noch einmal zusammenfassend dargestellt.

Tabelle IV-1.4.1a.: Übersicht über den vermuteten Einfluss der Stimulusmerkmale

Variable	**Stimulusmerkmale**	**Zusammenhang mit der Schwierigkeit**	
		angenommen	***tatsächlich***
Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale			
WLS	Wortlänge	+	+
PLW	Lange Wörter	+	+
PMW	Mehrsilbige Wörter	+	+
WIE	Wiederaufnahmen	+	+
STR4	Anakoluthe	-	+
SUB	Substantive/Eigennamen/ Appellative	+	+
STR2	Ellipsen	-	-
PEW	Einsilbige Wörter	-	-
NEG	Negationen	+	-
IWH	Inhaltswörter	+	0
GWS	Worthäufigkeit	-	0
STR1	Referenz-Aussage-Strukturen	-	0
STR3	Adjazenzstrukturen	-	0
STR5	Nähezeichen	-	0
STR6	Verberststellung	-	0
RHE1	Bildliche Darstellungsformen	+	0
RHE2	Uneigentliches Sprechen	+	0
RHE3	Neudeutsch/Anglizismen	+	0
RHE4	Jugendsprache/Umgangssprache	-	0
DEI	Deixis	+	0
REF	Referenzen	+	0
VER	Verben	-	0
Merkmalsgruppe II: Präsentationsmerkmale			
ASP	Anzahl der Sprecher	+	+
SPG	Sprechgeschwindigkeit	+	-
LST	Länge in Minuten	+	0
AWS	Wortzahl	+	0
AST	Akzent/Dialekt/Aussprache	+	0
AHO	Anzahl der Stimuluspräsentationen	-	0

Merkmalsgruppe III: Inhaltlich-thematische Merkmale			
WEL	Hintergrundwissen	+	+
HKO	Hörkontext – abstrakt	+	-
SLT	Literarischer Stimulus	+	-
TFU	Funktion – fern der Lebenswelt der Schüler	+	0
THE	Thema – abstrakt	+	0
Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte			
REL1	Relationstyp: Frage/ Impuls/Themensetzung	-	+
MLP	Länge der Propositionen	+	+
REL5	Relationstyp: Reihenfolge/Aufzählung	-	-
REL2	Relationstyp: Antwort	-	0
REL3	Relationstyp: Spezifizierung	-	0
REL4	Relationstyp: Erklärung/Beweis/Ursache	+	0
REL6	Relationstyp: Ziel/Bedingung	+	0
SFI	Schlussfolgerungen/Inferenzen	+	0
PRO/PRW	Anzahl/Anteil der Propositionen	+	0
Merkmalsgruppe V: Globalurteil			
TSA	Stimulusschwierigkeit	**+**	0

Anmerkungen: +: größere Anteile des Merkmals erhöhen die Schwierigkeit; - : größere Anteile des Merkmals senken die Schwierigkeit; 0: es konnte kein Effekt festgestellt werden

Bei der Faktorenanalyse der Stimulusmerkmale ergaben sich drei bzw. sechs Faktoren, um die sich die Merkmale gruppierten. Tabelle IV-1.4.1b. gibt einen Überblick, in welchen Bereichen sich die Variablengruppen auf der Grundlage didaktischer und sprachwissenschaftlicher Kriterien mit den durch die Faktorenanalyse gewonnenen Gruppen überschneiden. Während bei der ersten Merkmalsgruppe eine relativ hohe Zahl Variablen in beiden Gruppen vorkommt, gibt es bei den weiteren Merkmalsgruppen kaum noch Überschneidung, auch wenn sich die Gruppen auf einen ähnlichen Bereich beziehen[13]. Die Gegenüberstellung rechtfertigt eine Untersuchung sowohl der didaktisch, sprachwissenschaftlichen als auch der faktorenanalytischen Merkmalsgruppen, da angenommen werden muss, dass unterschiedliche, möglicherweise auch miteinander interagierende Aspekte der Stimuli von den Gruppen erfasst werden.

13 z. B. „Inhaltlich-thematische Merkmale" und „Überwiegend inhaltliche Merkmale"

Tabelle IV-1.4.1b.: Vergleich Merkmalsgruppen didaktisch/sprachwissenschaftlich und faktorenanalytisch

Beschreibung didaktisch/ sprachwissenschaftlich (MGD)	Variablen nur MGD	Gemeinsame Variablen	Variablen nur MGF	Beschreibung Faktoren-analyse (MGF)
Komplexität des Wortschatzes und sprachliche Merkmale	STR1, STR2, STR5, STR6, RHE, DIE, WIE, REF, VER	WLS, PEW, PLW, PMW, IWH, GWS, STR3, STR4, NEG, SUB	WEL	Sprachliche Merkmale zur quantitativen Beschreibung der Stimuli
Präsentationsmerkmale	ASP, AHO, LST, AWS	SGS, AST	STR1, STR2, RHE3	Merkmale gesprochener Sprache
Inhaltlich-thematische Merkmale	HKO, THE, WEL	SLT, TFU	STR6, VER, MLP	Überwiegend inhaltliche Merkmale
Struktur der Stimuli und propositionale Dichte	REL, SFI, PRW, MLP	PRO	HKO, THE, STR5, RHE2, RHE4, REF, LST, AWS	Sprachlich-inhaltliche Merkmale
Globalurteil	TSA			
			RHE1, DEI, WIE, SFI	Merkmale zur Erfassung der literarischen Qualität
			ASP, REL2, REL3	Merkmale zur Erfassung des Diskurstyps

Der Lehrerfragebogen eignet sich sehr gut, um die Aufgabenschwierigkeit vorherzusagen. Da er zum Teil mehrere Variablen zu einer Einschätzung aufweist[14], könnte der Fragebogen auf die am höchsten mit der Item- bzw. Aufgabenschwierigkeit korrelierenden Merkmale gekürzt werden. Die Variablen, die signifikant ($p < 0.10$) mit der Schwierigkeit korrelieren, sind: „Wortschatz (WFA)" mit den Ausprägungen „fachspezifisch" – „alltäglich",„Gesamteindruck (GIU)" mit den Ausprägungen „interessant" – „uninteressant",„Gesamteindruck (GAE)" mit den Ausprägungen „abwechslungsreich" – „eintönig",„Ton (TON)" mit den Ausprägungen „persönlich/gefühlsbetont" – „unpersönlich/sachlich",„Vertrautheit mit dem Thema (VTH)" mit den Ausprägungen „vertraut" – „gar nicht vertraut",„Grammatik (GRA)" mit den Ausprägungen „schwierig" – „einfach", „Wirkung (WEH)" mit den Ausprägungen „ernsthaft" – „humorvoll" sowie „Ausdrucksweise (AGU)" mit den Ausprägungen „gewählt" – „umgangssprachlich". Soll nur eine Variable zum Gesamteindruck im Fragebogen erscheinen, so bietet sich der „Gesamteindruck (GIU)" an, der mit $r = -0.56$ bei $p < 0.05$ das verlässlichere Merkmal gegenüber „Gesamteindruck (GAE)" mit $r = 0.48$ bei $p < 0.10$ ist.

1.4.2. Zusammenfassung Itemmerkmale

Die Variablen zum Itemformat (IFK/IFA) hängen, wie angenommen, insofern mit der Schwierigkeit zusammen, als geschlossene Itemformate, wie Ankreuzen, einfacher sind und offenere Formate, bei denen die Schüler stärker selbst formulieren müssen, schwieriger sind.

14 z. B. werden drei Einschätzungen zum Gesamteindruck mit den Adjektivpolen „interessant" – „uninteressant", „elegant" – „unbeholfen" und „abwechslungsreich" – „eintönig" erhoben

Das Bearbeiten der Items nach dem Anhören des Stimulus ist erwartungskonform einfacher als die Itembearbeitung während des Zuhörens. Die Konzentration auf mehrere Tätigkeiten gleichzeitig (Zuhören, Lesen, Schreiben) scheint mehr Arbeitsgedächtniskapazität in Anspruch zu nehmen als die Aktivhaltung des Gehörten. Die Variable „Position des Items innerhalb der Aufgabe (PIA)" korreliert dagegen nur sehr schwach mit der Itemschwierigkeit.

Aus der Merkmalsgruppe III (Merkmale von MC-Items) zeigen nur die Variablen zur Einschätzung der Plausibilität der Distraktoren einen Zusammenhang mit der Schwierigkeit. Erwartungsgemäß hängen hoch plausible Distraktoren mit einer höheren Itemschwierigkeit zusammen, wobei die größte vorkommende Plausibilität der Distraktoren (PDI) stärker ins Gewicht fällt als die mittlere Plausibilität der Distraktoren (MPD). An welcher Stelle im MC-Item der Attraktor steht („Position des Attraktors im MC-Item (PMC)") spielt jedoch für die Schwierigkeit keine Rolle.

Die Merkmale der Gruppen IV zeigen kaum Zusammenhang mit der Itemschwierigkeit. Es wurde angenommen, dass Items umso schwieriger werden, je stärker kognitive Operationen wie Schlussfolgern oder Transferieren gefordert sind, da diese Operationen das Arbeitsgedächtnis in verstärktem Maße beanspruchen. Dieser Aspekt wird durch die Variablen „Anforderungsbereich (AFB)", „Geprüfter Standard (BS)", „Anzahl der benötigten NI pro Item (ANI)", „Auftretenshäufigkeit der NI (ARN)", „Position der NI auf Stimulusebene (PST)", „Wortzahl der NI (WNI)", „Konkretheit der NI (TCO/TOR)", „Hintergrundwissen (HGW)" und „Typ der NI (TNI)" erfasst.

Auch die Variable der Gruppe V „Globalurteil" (Variable „Itemschwierigkeit (SEA)") hängt kaum mit der Schwierigkeit zusammen.

Tabelle IV-1.4.2a. gibt einen Überblick über die angenommenen und die tatsächlichen Zusammenhänge der einzelnen Itemmerkmale mit der Itemschwierigkeit.

Tabelle IV-1.4.2a.: Übersicht über den vermuteten und tatsächlichen Einfluss der Itemmerkmale

Variable	**Itemmerkmale**	**Zusammenhang mit der Schwierigkeit**	
		angenommen	***tatsächlich***
Merkmalsgruppe I: Itemformat			
IFK/IFA	Itemformat	je offener, desto schwieriger	je offener, desto schwieriger
Merkmalsgruppe II: Merkmale der Itempräsentation			
ZIB	Zeitpunkt der Itembearbeitung	einfacher nach dem Anhören	einfacher nach dem Anhören
PIA	Position des Items innerhalb der Aufgabe	je später, desto schwieriger	kein Effekt

Merkmalsgruppe III: Merkmale von MC-Items			
PMC	Position des Attraktors im MC-Item	je später, desto schwieriger	kein Effekt
PDI/MPD	Plausibilität der Distraktoren	je plausibler, desto schwieriger	je plausibler, desto schwieriger
Merkmalsgruppe IV: Kognitive Anforderungen der Items			
AFB	Anforderungsbereich	je höher, desto schwieriger	kein Effekt
BS	Geprüfter Standard	je kognitiv anspruchsvoller, desto schwieriger	kein Effekt
ANI	Anzahl der benötigten NI pro Item	je höher, desto schwieriger	kein Effekt
ARN	Auftretenshäufigkeit der NI	je höher, desto einfacher	kein Effekt
PST	Position der NI auf Stimulusebene	am Anfang und am Schluss einfacher	kein Effekt
WNI	Wortzahl der NI	je höher, desto schwieriger	kein Effekt
TCO/TOR	Konkretheit der NI (33-und 5-stufig)	je abstrakter, desto schwieriger	kein Effekt
HGW	Hintergrundwissen	je mehr erfordert, desto schwieriger	kein Effekt
TNI	Typ der NI	je kognitiv anspruchsvoller, desto schwieriger	kein Effekt
Merkmalsgruppe V: Globalurteil			
SEA	Itemschwierigkeit	je höher, desto schwieriger	kein Effekt

Für die faktorenanalytische Untersuchung der Itemmerkmale wurden drei Merkmalsgruppen gebildet. Die Merkmale gruppieren sich um die folgenden Faktoren:

Merkmalsgruppe: Allgemeine Itemmerkmale
Faktor 1: Merkmale zur Beschreibung der kognitiven Anforderungen der Items
Faktor 2: Merkmale zur Konkretheit der vom Item verlangten Information
Faktor 3: Merkmale zum Itemformat
Faktor 4: Merkmale zur Beschreibung der NI

Merkmalsgruppe: Merkmale zur Beschreibung von MC-Items
Faktor 1: Merkmale zur Einschätzung der Plausibilität der Distraktoren
Faktor 2: Merkmale zu den Rahmenbedingungen der Items
Faktor 3: Merkmal Hintergrundwissen
Faktor 4: Merkmal Position des Attraktors

Merkmalsgruppe: Merkmale zur Beschreibung der NI
Faktor 1: Merkmale zur Beschreibung der vom Item verlangten Kompetenzen
Faktor 2: Merkmale zur Differenzierung der kognitiven Operationen
Faktor 3: Merkmal zur Beschreibung der verschiedenen Formen von Mündlichkeit

Tabelle IV-1.4.2b. gibt einen Überblick über die Zugehörigkeit der einzelnen Variablen zu den Merkmals-und Faktorengruppen.

Tabelle IV-1.4.2b.: Überblick über die Zugehörigkeit der Itemvariablen zu Merkmals- und Faktorengruppen

Faktor-gruppe	Itemmerkmale	Merkmale zur Beschreibung	
		von MC-Items	*der NI*
1	Anforderungsbereich (AFB)	Größte vorkommende Plausibilität der Distraktoren (PDI)	
1	BS142	Mittlere Plausibilität der Distraktoren (MPD)	BS142
1	BS143	Itemschwierigkeit (SEA)	BS143
1	Typ der NI: Detail (TNI2)		
1	Typ der NI: Stimmung/ Einstellung des Sprechers (TNI4)		
1	Typ der NI: Geräusche/ Paraverbales (TNI9)		
2	Konkretheit der NI: 33-stufig (TCO) und 5-stufig (TOR)	Zeitpunkt der Itembearbeitung (ZIB)	Konkretheit der NI: 33-stufig (TCO) und 5-stufig (TOR)
2	Typ der NI: Sprecheridentifikation (TNI11)	Position des Items innerhalb der Aufgabe (PIA)	BS113
2		Anforderungsbereich (AFB)	Anforderungs-bereich (AFB)
3	Itemformat: Ankreuzen (MC+RF) (IFA.AK)	Hintergrundwissen (HGW)	BS141
3	Itemformat: Schreibitem (HO+OI) (IFA.S)		
4	Typ der NI: Meinung des Zuhörers (TNI3)	Position des Attraktors im MC-Item (PMC)	
4	Typ der NI: Schlussfolgerung (TNI5)		
4	Typ der NI: Kommunikativer Zweck/ Funktion/Wirkung des Stimulus (TNI6)		
	Typ der NI: Struktur des Stimulus/ Zusammenhang zwischen Teilen (TNI7)	Hintergrundwissen (HGW)	
	Typ der NI : Genre des Stimulus (TNI8)	Itemschwierigkeit (SEA)	
	Typ der NI: Leitidee/Kernaussage (TNI1)	Position der NI auf Stimulusebene (PST)	
	Typ der NI: Sprachliche Mittel (TNI10)	Wortzahl der NI (WNI)	
	BS141		
	BS113		
	Itemschwierigkeit (SEA)		
	Zeitpunkt der Itembearbeitung (ZIB)		
	Position des Items innerhalb der Aufgabe (PIA)		
	Itemformat: Offenes Item (IFA.OI)		
	Hintergrundwissen (HGW)		

Anmerkungen: BS142: Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben
BS143: Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln
BS113: Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden
BS141: Gesprächsbeiträge anderer verfolgen und aufnehmen

Die Merkmale gruppieren sich in den drei Merkmalsgruppen „Allgemeine Itemmerkmale", „Merkmale zur Beschreibung von MC-Items" und „Merkmale zur Beschreibung der NI" recht unterschiedlich um die einzelnen Faktoren. Es erscheint deshalb sinnvoll, die einzelnen Gruppierungen in einer Regressionsanalyse weiter zu untersuchen.

1.4.3. Zusammenfassung Personenmerkmale

Um den Einfluss personenbezogener Merkmale auf die Item- bzw. Aufgabenschwierigkeit zu prüfen, wurden die Schüler gebeten, die Stimuli hinsichtlich ihres Interessantheits-bzw. Be-kanntheitsgrades einzuschätzen und zusätzlich die Präsentationsqualität zu beurteilen. Ferner liegen Angaben zu den Sprachkenntnissen der Schüler vor sowie die Ergebnisse eines Testteils zur Arbeitsgedächtniskapazität. Alle personenbezogenen Merkmale wurden mit den Leistungsdaten korreliert. Im Fall der Einschätzung der Aufgaben durch die Schüler lässt sich für keine der vier Variablen „Vertrautheit mit dem Thema (VTS)", „Motivation/Interesse (MOT)", „Bekanntheitsgrad des Stimulus (BEK)" und „Verständlichkeit (VST)" ein systematischer signifikanter Zusammenhang mit den Leistungsdaten der Schüler erkennen. Die Schülereinschätzungen wurden mit den Summenscores für alle Schüler über alle Blöcke und alle Aufgaben hinweg korreliert. In Folgeanalysen wurden die Korrelationen zusätzlich für die Subpopulationen „Hauptschüler" und „Gymnasiasten" erhoben sowie in diesen Subpopulationen nur für die Schüler, die in ihrem Ankreuzverhalten die Bandbreite der fünf Antwortmöglichkeiten ausschöpfen. Die Schülereinschätzungen zum Bekannt- und Interessantheitsgrad der Stimuli sowie zu deren Verständlichkeit scheinen jedoch in keiner Weise mit den Schülerleistungen zusammenzuhängen. Daraus muss gefolgert werden, dass diese Merkmale keinen Einfluss auf die Item- bzw. die Aufgabenschwierigkeit besitzen.

Im Gegensatz zu den Variablen zur Einschätzung der Aufgaben durch die Schüler hängen die Ergebnisse des Tests zur Arbeitsgedächtniskapazität in allen Fällen signifikant (< 0.05) mit den Leistungsdaten zusammen. Die Korrelationen sind jedoch unabhängig von der Länge der Aufgaben und der empirischen Item- bzw. Aufgabenschwierigkeit. Analysen nach Schulform lassen stärkere Zusammenhänge im Bereich der Hauptschule erkennen, wobei die Korrelationen aufgabenbezogen sind. Eine Systematik bzgl. des aufgabenbezogenen Zusammenhangs lässt sich nicht ausmachen, es handelt sich dabei also nicht prinzipiell um Aufgaben eines bestimmten Typs, einer bestimmten Länge oder Schwierigkeit.

Auch die Angaben, wie häufig zu Hause deutsch gesprochen wird bzw. welche Sprache als Muttersprache gelernt wurde, korrelieren mit den Leistungsdaten der Schüler. Schüler, die angaben zu Hause nur manchmal oder niemals deutsch zu sprechen, zeigen geringere Leistungen im Hörverstehensteil des IQB-Deutschtests. Bessere Testleistungen hingen mit der Angabe zusammen, deutsch als Muttersprache gelernt zu haben. Die Korrelationen fallen im Bereich der Hauptschule etwas stärker als im Gymnasium aus. Auf Aufgabenebene lassen sich stark unterschiedliche Zusammenhänge der Leistungsergebnisse mit den Angaben zum Sprachstand erkennen.

2. Zusammenhangsanalysen

Die folgenden Analysen wurden in Abhängigkeit vom betrachteten Merkmal entweder auf Itemebene und/oder auf Aufgabenebene durchgeführt. Itemmerkmale wurden ausschließlich auf Itemebene analysiert, wohingegen Stimulusmerkmale sowohl auf Itemmerkmale als auch auf Aufgabenebene untersucht wurden. Dieses Vorgehen ermöglicht Aussagen über den Einfluss der Merkmale einerseits auf die Itemschwierigkeit und andererseits auf die Aufgabenschwierigkeit und lässt damit auch Rückschlüsse auf die Schwierigkeit der Stimuli zu. Für die Berechnungen auf Stimulusebene werden die aggregierten Itemschwierigkeiten eines Stimulus herangezogen, die als Aufgabenschwierigkeit interpretiert werden können. Da auf Stimulusebene nur insgesamt 30 Stimuli vorliegen, ist aufgrund dieser geringen Fallzahlen eher mit nicht signifikanten Ergebnissen und erhöhten Eta^2-Werten zu rechnen. I. d. R. gilt, dass der Signifikanz-Wert umso kleiner ist (man also eher signifikante Ergebnisse erhält), je größer die Stichprobe bei gleicher Varianzaufklärung (Eta^2) ausfällt. Die Einteilung der Code-Gruppen erfolgte unter dem Gesichtspunkt, das entsprechende Merkmal möglichst gut zwischen den einzelnen Code-Abstufungen zu streuen.

Zunächst werden die Mittelwerte der Itemschwierigkeiten auf den Faktorstufen der jeweiligen Variablen einer Zusammenhangsanalyse unterzogen. Da die Itemschwierigkeiten über das Rasch-Modell gewonnen wurden, wurde dabei der Mittelwert der Schülerfähigkeiten der MSA-Schüler auf null gesetzt. Dies bedeutet, dass auch negative Werte auftreten können. Negative Itemschwierigkeiten bedeuten, dass die Items etwas leichter sind als Personen im Mittel fähig sind. Unter der Annahme, dass eine Normalverteilung der Itemschwierigkeiten besteht, liegen etwa 68% der Schwierigkeiten im Bereich Mittelwert ± 1 Standardabweichung. Bei 2 Standardabweichungen beträgt der Anteil 95%.

2.1. Stimulusmerkmale

2.1.1. Mittelwertsvergleiche der Stimulusmerkmale aus den IQB-Ratings

Zu den Mittelwertsvergleichen wurden für dichotome Stimulusmerkmale auch Korrelationen berechnet. Die Werte können hier zwischen -1 und 1 liegen und sollten sich im Idealfall zwischen ± 0.1 -0.3 liegen. Je größer ein Wert ausfällt, desto stärker korreliert das entsprechende Merkmal positiv oder negativ mit der Item- bzw. Aufgabenschwierigkeit. Zusätzlich zur Korrelation wird noch die zweiseitige Signifikanz angegeben, bei der die Nullhypothese in zwei Richtungen getestet wird.

Zusätzlich wurden Partialkorrelationen für die Variablen berechnet, um den Zusammenhang weiterer Merkmale mit dem jeweiligen Merkmal und der Schwierigkeit auszuschließen. Die Merkmale, die bei der Berechnung der Partialkorrelationen kontrolliert werden, korrelieren stark mit den jeweiligen Variablen, für die ein Zusammenhang mit der Schwierigkeit berechnet wird.

2.1.1.1. Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale

Variablen „Wortlänge (WLS)", „Lange Wörter (PLW)", „Mehrsilbige Wörter (PMW)" und „Einsilbige Wörter (PEW)"

Die Variablen „Wortlänge (WLS)", „Lange Wörter (PLW)" und „Mehrsilbige Wörter (PMW)" verdeutlichen, dass Items umso leichter sind, je weniger lange Wörter (gemessen in Buchstaben und Silben) in den entsprechenden Stimuli vorkommen (vgl. Tabelle IV-2.1.1.1a., Anhang G). Dies ist damit zu erklären, dass kurze Wörter in der Regel einfacher sind, eher dem Grundwortschatz entsprechen und Funktionswörter darstellen. Die Ergebnisse für die drei Variablen sind statistisch signifikant ($p < 0.10$) und die Variablen haben auf Aufgabenebene einen hohen Anteil an der Varianzaufklärung. Ferner korrelieren die Variablen mit der Item- und Aufgabenschwierigkeit (WLS: $r = 0.11$ bzw. $r = 0.36$; PMW: $r = 0.14$ bzw. $r = 0.34$) und bestätigen das Resultat: Beträgt die mittlere Wortlänge mehr als fünf Buchstaben, bzw. je höher der Prozentanteil langer Wörter (> 20% bzw. 25%) bzw. mehrsilbiger Wörter ist (> 5%), desto schwieriger sind Items und Stimulus. Dieses Ergebnis ist für beide Variablen statistisch signifikant ($p < 0.10$). Hypothesenkonträr ist in diesem Zusammenhang, dass die Variablen „Worthäufigkeit (GWS)" und „Inhaltswörter (IWH)" keinen Effekt auf die Itemschwierigkeit hatten.

Die Variable „Einsilbige Wörter (PEW)" bestätigt die Ergebnisse der Variablen WLS, PLW und PMW. Je mehr Einsilbige Wörter in den Stimuli vorkommen, desto einfacher sind Items und Stimulus. Auch dieses Ergebnis ist sowohl auf Item- als auch auf Aufgabenebene statistisch signifikant ($p < 0.10$) und trägt auf beiden Ebenen mit einem großen Effekt zur Varianzaufklärung bei. Die Korrelationen mit der Item- bzw. der Aufgabenschwierigkeit betragen 0.17 bzw. 0.40 und sind in beiden Fällen statistisch signifikant. (vgl. Tabelle IV-2.1.1.1b., Anhang G)

Alle vier Merkmale korrelieren stark miteinander. Aus diesem Grund werden die Partialkorrelationen für diese Merkmale berechnet, indem jeweils die anderen drei Merkmale kontrolliert werden. Insgesamt liegen die Partialkorrelationen deutlich unter den unkontrollierten Werten, zeigen jedoch noch immer einen Effekt. (vgl. Tabelle IV-2.1.1.1c., Anhang G)

Variable „Inhaltswörter (IWH)"

Items und Stimuli, deren Anteil der Inhaltswörter in einem mittleren Bereich von 45-50% liegen sind schwieriger als Items und Stimuli, bei denen die Variable IWH weniger als 45% oder mehr als 50% beträgt. Die Ergebnisse sind jedoch statistisch nicht signifikant und die Variable IWH hat kaum Anteil an der Varianzaufklärung. Auffällig ist die geringe Fallzahlbelegung des dritten Codes auf Aufgabenebene von nur zwei Fällen, nach der anzunehmen ist, dass hier kein repräsentatives Ergebnis vorliegt (vgl. Tabelle IV-2.1.1.1d., Anhang G)

Die Partialkorrelation auf Aufgabenebene, bei der die Variablen „Worthäufigkeit (GWS)", „Ellipsen (STR.2)" und „Substantive/Eigennamen/Appellative (SUB)" kontrolliert wurden, beträgt -0.32 und liegt damit deutlich höher als die ursprünglich erhaltene Korrelation der Variable IWH mit der Stimulusschwierigkeit. Die Variable scheint also auf Aufgabenebene negativ mit der Schwierigkeit zusammenzuhängen. Dieser Zusammenhang wird jedoch durch das Zusammenspiel mehrerer weiterer Merkmale deutlich reduziert. Auf Itemebene ist der Effekt weniger

deutlich sichtbar. Hier beträgt die Partialkorrelation nur -0.06. Der Effekt ist erwartungswidrig. Es wurde angenommen, dass ein höherer Anteil an Inhaltswörtern im Stimulus eher mit einer höheren Item- bzw. Aufgabenschwierigkeit zusammenhängt, da mehr Propositionen zu verarbeiten sind. Es ist nicht auszuschließen, dass das Ergebnis auf Aufgabenebene durch die geringe Fallzahl stark von einzelnen Aufgaben beeinflusst wird.

Variable „Worthäufigkeit (GWS)"

Entgegen den Erwartungen zeigt sich, dass eine stärkere Überlappung der im Stimulus vorkommenden Wörter mit dem Grundwortschatz mit einer erhöhten Item- und Aufgabenschwierigkeit zusammenhängt. Das Ergebnis ist jedoch statistisch nicht signifikant. Die Variable GWS trägt auch nicht wesentlich zur Varianzaufklärung bei und sie korreliert kaum mit der Item- bzw. Aufgabenschwierigkeit (-0.01 bzw. -0.02). (vgl. Tabelle IV-2.1.1.1e., Anhang G)

Um zu prüfen, ob für dieses erwartungswidrige Ergebnis ggf. eine höhere Stimuluslänge in den Fällen mit stärkerer Überlappung verantwortlich ist, werden die Analysen nur für die Items und Stimuli wiederholt, die sich nicht in den Randbereichen der gezählten Wörter aufweisen, sondern eine Länge von 200 bis 600 Wörtern haben. (vgl. Tabelle IV-2.1.1.1f., Anhang G) An der Grundtendenz der Ergebnisse ändert sich auch bei Wiederholung der Analysen mit einem Teil der Daten nichts, das Resultat fällt stattdessen umso deutlicher aus. Je stärker der in den Stimuli vorkommende Wortschatz mit dem Grundwortschatz überlappt, desto schwieriger sind die Items und der Stimulus. Das Ergebnis ist auf Itemebene nun auch signifikant und deckt auf Item- und Aufgabenebene 3 bzw. 12% Varianz auf. Die Korrelation mit der Stimulusschwierigkeit beträgt 0.19. Die um die Variablen „Wortlänge (WLS)", „Inhaltswörter (IWH)" und „Substantive/Eigennamen/Appellative (SUB)" kontrollierte Partialkorrelation beträgt auf Aufgabenebene 0.25 und auf Itemebene 0.18. Die Variable GWS scheint also durchaus positiv mit der Schwierigkeit zusammenzuhängen. Dieser Zusammenhang relativiert sich jedoch durch das Zusammenwirken mehrerer anderer Merkmale. Da auf Aufgabenebene nur 17 Stimuli in die Analysen einflossen, ist nicht auszuschließen, dass der Effekt dem Einfluss einzelner Aufgaben geschuldet ist.

Variable „Strukturbestimmung (STR)"

Zusätzlich zu den hier aufgeführten Strukturmerkmalen wurden noch weitere Merkmale kodiert, und zwar: Operator-Skopus-Strukturen, Apokoinu-Strukturen, Konstruktionsübernahmen, Expansionen, abhängige Verbzweit-Konstruktionen, ursprüngliche Subjunktionen mit Verbzweitstellung, Dativ-Possessiv-Konstruktionen und Wiederaufnahmen. Für diese Fälle lagen jedoch nach Abschluss der Kodierung entweder zu wenige Fälle vor, um damit Analysen durchzuführen, oder die durchgeführten Analysen ergaben keinerlei Ergebnis. Dies kann u. U. auch auf Ungenauigkeiten bei der Kodierung zurückgeführt werden. Berichtet werden deshalb nur die Merkmale, die zumindest einen geringen Effekt auf die Schwierigkeit haben: „Referenz-Aussage-Strukturen (STR1)", „Ellipsen (STR2)", „Anakoluthe (STR4)", „Adjazenzstrukturen (STR3)", „Nähezeichen (STR5)" und „Verberststellung (STR6)".

Je höher die Anteile der Strukturtypen „Referenz-Aussage-Strukturen (STR1)", „Ellipsen (STR2)" und „Adjazenzstrukturen (STR3)" in den Stimuli sind, desto schwieriger sind die Items und der

Stimulus. Das Merkmal „Ellipsen (STR2)" liefert auf Item- und Aufgabenebene signifikante Ergebnisse ($p < 0.10$), bewegt sich hinsichtlich seiner Varianzaufklärung jedoch im kleinen bzw. mittleren Bereich (STR2: 5% auf Itemebene, 11% auf Aufgabenebene). Die Korrelation mit der Item- bzw. mit der Aufgabenschwierigkeit beträgt 0.15 bzw. 0.21. Für das Merkmal „Adjazenzstrukturen (STR3)" wurde nur auf Itemebene ein signifikantes ($p < 0.10$) Ergebnis erhalten, wobei die Variable kaum mit der Schwierigkeit korreliert und mit 2% auf Itemebene bzw. 4% auf Aufgabenebene auch kaum an der Varianzaufklärung beteiligt ist. Das Merkmal „Referenz-Aussage-Strukturen (STR1)" zeigt bei den Analysen mit dem Gesamtdatensatz keinen Einfluss auf die Schwierigkeit auf Item- bzw. Aufgabenebene. Tabelle IV-2.1.1.1g., Anhang G fasst die beschriebenen Ergebnisse zusammen.

Im Fall der Merkmale „Verberststellung (STR6)", „Anakoluthe (STR4)" und „Nähezeichen (STR5)" ergaben die Analysen umgekehrte Ergebnisse: Höhere Anteile der jeweiligen Merkmale in den Stimuli tendieren dazu, mit einfacheren Items bzw. Stimuli zusammenzuhängen. Dieser Effekt ist für alle drei Variablen nicht signifikant, deshalb ist bei diesen drei Variablen höchstens von einem tendenziellen Zusammenhang zu sprechen. Er ist auf Itemebene nur sehr schwach, auf Aufgabenebene aber etwas deutlicher zu beobachten. Die Ergebnisse der Korrelationsanalysen liegen für das Merkmal „Anakoluthe (STR4)" auf Itemebene bei -0.12 und auf Aufgabenebene bei -0.32. Je mehr Anakoluthe in einem Stimulus vorkommen, desto einfacher sind Items und Stimulus. (vgl. Tabelle IV-2.1.1.1h., Anhang G)

Die dargestellten Variablen sind alles Merkmale gesprochener Sprache. Ein weiteres Merkmal gesprochener Sprache bzw. Diskursen ist häufig eine erhöhte Sprechgeschwindigkeit im Vergleich zu vorgelesenen Texten. Die Analysen werden aus diesem Grund für einen Teil des Datensatzes wiederholt, der durch eine erhöhte Sprechgeschwindigkeit gekennzeichnet ist (vgl. Tabelle IV-2.1.1.1i., Anhang G).

Die Variable „Referenz-Aussage-Strukturen (STR1)" zeigt bei den Analysen mit dem eingeschränkten Datensatz einen leichten Effekt, der aber weder auf Item- noch auf Aufgabenebene signifikant ist. Je mehr Referenz-Aussage-Strukturen in einem Stimulus vorkommen, desto einfacher sind die Items und der Stimulus. Die Variable deckt auf Aufgabenebene lediglich 2% Varianz auf und das Ergebnis hat nur einen kleinen praktischen Effekt ($d = 0.12$ bzw. 0.33). Die Partialkorrelation für diese Variable mit der Aufgabenschwierigkeit steigt auf -0.42 bzw. auf -0.22 auf Itemebene, werden die Variablen „Relationstyp: Spezifizierung (REL3)", „Deixis (DEI)" und „Hintergrundwissen (WEL)" kontrolliert.

Die Ergebnisse der Analysen mit den Variablen „Ellipsen (STR2)" und „Adjazenzstrukturen (STR3)" replizieren quasi die Ergebnisse, die nach den Analysen mit dem gesamten Datensatz erhalten wurde. Je mehr Ellipsen bzw. Adjazenzstrukturen in einem Stimulus vorkommen, desto schwieriger sind Items und Stimulus. Die Varianzaufklärung fällt hier auf Aufgabenebene etwas höher aus als bei den vorhergehenden Analysen (13% vs. 11% für STR2 und 8% vs. 4% für STR3), was aber mit der geringeren Fallzahlbelegung zusammenhängen kann. Die Variable STR2 korreliert etwas mit der Aufgabenschwierigkeit, für die Variable STR3 sinkt die Korrelation im Vergleich zum Ergebnis, das mit dem gesamten Datensatz erhalten wurde.

Die Ergebnisse sind nur auf Itemebene signifikant ($p < 0.10$).

Die Partialkorrelation für die Variable STR2 mit der Aufgabenschwierigkeit beträgt bei Kontrolle der Variablen „Inhaltswörter (IWH)", „Worthäufigkeit (GWS)" und „Schlussfolgerungen/ Inferenzen (SFI)" 0.31. Auf Itemebene beträgt sie 0.24. Für die Variable STR3 liegt die Partialkorrelation mit der Aufgabenschwierigkeit bei -0.14, die Partialkorrelation mit der Itemschwierigkeit liegt bei 0.13. Kontrolliert wurden die Variablen „Worthäufigkeit (GWS)", „Relationstyp: Spezifizierung (REL3)" und „Relationstyp: Ziel/Bedingung (REL6)".

Insgesamt haben die Merkmale „Verberststellung (STR6)", „Anakoluthe (STR4)" und „Nähezeichen (STR5)" einen größeren Effekt auf Aufgabenebene als auf Itemebene. (vgl. Tabelle IV-2.1.1.1j., Anhang G) Insbesondere der Einfluss der Variable STR5 wird mit dem gefilterten Datensatz recht deutlich und drückt sich auch in einer stärkeren Korrelation mit der Aufgabenschwierigkeit aus. Je mehr Nähezeichen in einem Stimulus vorkommen, desto einfacher werden die Items und wird der Stimulus. Die Variable liefert auf Aufgabenebene 10% Varianzaufklärung, wenn auch das Ergebnis nicht statistisch signifikant ist ($p > 0.10$). Die Effektstärke liegt auf Aufgabenebene mit $d = 0.77$ recht hoch. Kontrolliert man bei der Berechnung der Partialkorrelation mit der Aufgabenschwierigkeit die Variablen „Relationstyp: Antwort (REL2)", „Referenzen (REF)" und „Anzahl der Propositionen (PRO)", ergibt sich ein Wert von -0.44. Auf Itemebene liegt die Partialkorrelation bei -0.05. Die Variable STR5 hat für sich genommen also einen Einfluss auf die Aufgabenschwierigkeit, wohingegen ihr Einfluss auf die Itemschwierigkeit zu vernachlässigen ist.

Das Merkmal „Anakoluthe (STR4)" zeigte mit dem Gesamtdatensatz auf Itemebene keinen Einfluss auf die Schwierigkeit, nun ist jedoch ein leichter Einfluss zu erkennen. Je mehr Anakoluthe im Stimulus vorkommen, desto einfacher sind die entsprechenden Items. Äußerungsabbrüchen sind im Rahmen gesprochener Sprache häufig und stellen i. d. R. für den Zuhörer keine größere Schwierigkeit dar. Gleichzeitig wird in Stimuli mit vielen Abbrüchen und Neuanfängen auch inhaltlich häufig weniger kommuniziert, was einen Einfluss auf die Schwierigkeit der Items haben kann. Auf Aufgabenebene hat sich im Vergleich zu den vorhergehenden Analysen wenig verändert. Die Variable korreliert nicht signifikant mit der Item- bzw. der Aufgabenschwierigkeit. Kontrolliert man die Variablen „Deixis (DEI)", „Negationen (NEG)" und „Substantive/Eigennamen/Appellative (SUB)", die hoch (0.55, 0.76 und 0.48) mit der Variable STR4 korrelieren beträgt die Partialkorrelation auf Aufgabenebene nur noch -0.02, auf Itemebene 0.01.

Ein signifikanter Einfluss der Variable „Verberststellung (STR6)" auf die Schwierigkeit ($p< 010$) lässt sich in diesem Analysedurchlauf nur noch auf Aufgabenebene feststellen. Korrelationen mit der Item- bzw. Aufgabenschwierigkeit liegen auch bei Berechnung der Partialkorrelation kaum vor. Kontrolliert wurden die am höchsten mit der Variable STR6 korrelierenden Merkmale „Literarischer Stimulus (SLT)", „Relationstyp: Reihenfolge/Aufzählung (REL5)" und „Verben (VER)". Die Partialkorrelation mit der Aufgabenschwierigkeit beträgt 0.03, mit der Itemschwierigkeit -0.03.

Bei den Einzelcodes der Variable STR haben nur die Merkmale „Referenz-Aussage-Strukturen (STR1)“, „Ellipsen (STR2)“ und „Nähezeichen (STR5)“ einen Einfluss auf die Schwierigkeit. STR1 und STR2 korrelieren zwar sowohl mit der Item- als auch mit der Aufgabenschwierigkeit, decken aber auf Aufgabenebene deutlich mehr Varianz auf. Merkmal STR5 korreliert nur mit der Aufgabenschwierigkeit und deckt auf Aufgabenebene 10% Varianz auf.

Variable „Rhetorische Mittel (RHE)”

Die Analysen zur Variable RHE ergeben, dass Items zu Stimuli umso schwieriger sind, je mehr bildliche Darstellungsformen und Formen uneigentlichen Sprechens sie enthalten. Metaphern, Bilder und Redewendungen sowie Ironie, Übertreibungen und Wortspiele erschweren das Verständnis eines Stimulus, da auf das tatsächlich Gemeinte geschlossen werden muss. Beide Merkmale („Bildliche Darstellungsformen (RHE1)“ und „Uneigentliches Sprechen (RHE2)“) haben mit statistisch signifikanten Ergebnissen ($p < 0.10$) einen kleinen bzw. mittleren Effekt auf die Varianzaufklärung (2% bzw. 10% für RHE1 und 2% bzw. 7% für RHE2), es liegen jedoch keine signifikanten Korrelationen mit der Item- oder der Aufgabenschwierigkeit vor. Tabelle IV-2.1.1.1k., Anhang G stellt die Ergebnisse der Mittelwertsanalysen für die Merkmale „Bildliche Darstellungsformen“ und „Uneigentliches Sprechen“ dar. Bei der Berechnung der Partialkorrelationen wurden für RHE1 die Merkmale „Schlussfolgerungen/Inferenzen (SFI)“ und „Hintergrundwissen (WEL)“ kontrolliert. Die Partialkorrelation auf Aufgabenebene beträgt 0.02, auf Itemebene -0.03. Für die Variable RHE2 werden die Merkmale „Wortzahl (AWS)“, „Referenzen (REF)“ und „Anzahl der Propositionen (PRO)“ kontrolliert. Die Partialkorrelation auf Aufgabenebene beläuft sich dann auf 0.17, auf Itemebene beträgt sie -0.01.

Im Fall des Merkmals „Neudeutsch/Anglizismen“ (RHE3) fallen die Ergebnisse auf Item- bzw. Aufgabenebene unterschiedlich aus: auf Itemebene hängen Stimuli mit einem höheren Anteil an neudeutschen Ausdrücken bzw. Anglizismen mit den einfacheren Items zusammen, auf Aufgabenebene sind diese Stimuli jedoch schwieriger. Dieses Ergebnis ist jedoch weder auf Item- noch auf Aufgabenebene signifikant ($p > 0.10$) und auch die Korrelationen mit der Item- und der Aufgabenschwierigkeit fallen sehr gering und nicht signifikant aus ($p > 0.10$). (vgl. Tabelle IV-2.1.1.1l., Anhang G) Bei der Berechnung der Partialkorrelationen auf Item- und Aufgabenebene werden für RHE3 die Merkmale „Anzahl der Sprecher (ASP)“, „Sprechgeschwindigkeit (SGS)“ und „Stimulusschwierigkeit (TSA)“ kontrolliert. Nach Kontrolle dieser Einflussfaktoren beträgt die Partialkorrelation für die Variable RHE3 auf Aufgabenebene 0.09, auf Itemebene 0.04.

Höhere Anteile des Merkmals „Jugendsprache/Umgangssprache (RHE4)“ tendieren eher dazu, dass Items zu diesen Stimuli leichter sind. Insgesamt muss jedoch darauf hingewiesen werden, dass dieser Effekt nur sehr schwach und statistisch nicht signifikant ist ($p > 0.10$). Auch Korrelationen mit der Item- bzw. der Aufgabenschwierigkeit sind kaum zu beobachten und besitzen keine statistische Signifikanz ($p > 0.10$). (vgl. Tabelle IV-2.1.1.1m., Anhang G) Die Partialkorrelation mit der Aufgabenschwierigkeit beträgt für die Variable RHE4 -0.09 und mit der Itemschwierigkeit 0.04. Dabei wurden die Merkmale „Nähezeichen (STR5)“, „Referenzen (REF)“ und „Anzahl der Propositionen (PRO)“ kontrolliert.

Insgesamt lässt sich kein Zusammenhang der Einzelcodes der Variable RHE mit der Schwierigkeit feststellen. Nur das Merkmal „Uneigentliches Sprechen (RHE2)“ korreliert leicht mit der Aufgabenschwierigkeit.

Variable „Deixis (DEI)“

Die Variable DEI beschreibt die prozentuale Auftretenshäufigkeit deiktischer Elemente in den Stimuli. Da die Schüler in der Testsituation nicht direkt an der Gesprächssituation beteiligt sind, könnten für sie deiktische Ausdrücke eher Verständnis erschwerend wirken. Die Item- und Aufgabenschwierigkeit tendiert dazu, mit höherer Auftretenshäufigkeit deiktischer Elemente (mehr als 10%) zu sinken. Auch die negativen Korrelationen mit der Item- und Aufgabenschwierigkeit weisen – obwohl statistisch nicht signifikant – mit -0.05 und -0.15 auf diese Tendenz. (vgl. Tabelle IV-2.1.1.1n., Anhang G) Bei den Stimuli mit einem hohen Anteil an deiktischen Elementen handelt es sich möglicherweise um Stimuli, die noch andere Merkmale gesprochener Sprache aufweisen, wie z. B. hohe Redundanz, und dadurch für Zuhörer leichter verständlich sind als Stimuli, die auditiv präsentierte Texte sind und überwiegend Merkmale geschriebener Sprache besitzen. Die beschriebenen Ergebnisse sind statistisch nicht signifikant und die Variable DEI hat nur einen geringen Anteil an der Varianzaufklärung. Bei der Berechnung der Partialkorrelationen auf Item- und Aufgabenebene wurden die Merkmale „Anakoluthe (STR4)“, „Negationen (NEG)“ und „Substantive/Eigennamen/Appellative (SUB)“ kontrolliert. Die Partialkorrelation der Variable DEI mit der Aufgabenschwierigkeit beträgt 0.05, mit der Itemschwierigkeit -0.01.

Variable „Wiederaufnahmen (WIE)“

Implizite und explizite Wiederaufnahmen wurden in Hinblick auf ihre prozentuale Auftretenshäufigkeit im Stimulus kategorisiert und hinsichtlich ihres Einflusses auf die Item- und die Stimulusschwierigkeit analysiert. Die Analysen ergeben, dass ein Stimulus mit vielen Wiederaufnahmen die Informationsentnahme stark erschwert und deshalb zu einer erhöhten Item- und Stimulusschwierigkeit beiträgt. Das Ergebnis ist auf beiden Analyseebenen statistisch signifikant und trägt mit 5% bzw. 25% mit einem mittleren bzw. starken Effekt zur Varianzaufklärung bei. (vgl. Tabelle IV-2.1.1.1o., Anhang G) Auch für die Variable WIE wurden Partialkorrelationen berechnet, wobei die Merkmale „Deixis (DEI)“ und „Schlussfolgerungen/Inferenzen (INF)“ kontrolliert wurden. Auf Aufgabenebene ergibt sich ein Wert von 0.26, auf Itemebene von 0.10.

Variable „Referenzen (REF)”

Tabelle IV-2.1.1.1p., Anhang G verdeutlicht, dass die Ergebnisse für die Variable REF nicht signifikant sind und die Variable auf Aufgabenebene keinen nachweisbaren Einfluss auf die Aufgabenschwierigkeit hat. Auf Itemebene ist ein geringer Effekt zu beobachten: Je mehr Referenzen in einem Stimulus vorkommen, desto einfacher sind die entsprechenden Items.

Der Einfluss der Variable REF auf die Itemschwierigkeit wird in einem nächsten Schritt nur für Stimuli untersucht sind, die in etwa gleich lang sind, um den Einfluss extremer Kürze oder Länge auf die Schwierigkeit auszuschließen. Ausgewählt wurden dafür die Stimuli, deren Länge sich im Rahmen von 200 – 600 Wörtern befindet. Im Aufgabenpool befinden sich Stimuli mit

einer Bandbreite von 63 bis 1728 Wörtern. Sowohl auf Item- als auch auf Aufgabenebene ist die Tendenz zu beobachten, dass die Schwierigkeit von Items und Stimuli umso geringer ausfällt, je mehr Referenzbezüge in einem Stimulus gezählt wurden. Dieses Ergebnis ist zwar nach wie nicht statistisch signifikant, jedoch klärt die Variable nun einen geringen Varianzanteil auf. (vgl. Tabelle IV-2.1.1.1q., Anhang G) Die Partialkorrelation mit der Aufgabenschwierigkeit beträgt -0.06, wobei der Einfluss der Variablen „Länge in Minuten (LST)", „Wortzahl (AWS)" und „Anzahl der Propositionen (PRO)" kontrolliert wurde. Auf Itemebene beträgt die Partialkorrelation bei Kontrolle derselben Merkmale -0.17.

Variable „Negation (NEG)"

Die Variable NEG drückt aus, wie hoch der Anteil an Negationen in den Stimuli ist. Die Ergebnisse sind auf Item- und Aufgabenebene nicht signifikant und nur sehr schwach. Ein hoher Anteil Negationen hängt auf Itemebene mit einer höheren Itemschwierigkeit zusammen, wohingegen ein hoher Anteil Negationen auf Aufgabenebene mit einer niedrigeren Itemschwierigkeit zusammenhängt. Die Varianzaufklärung beträgt auf Aufgabenebene 4% und auch Cohen's d weist mit 0.12 auf einen leichten Effekt hin. Die Korrelationen der Variable mit der Item- und der Aufgabenschwierigkeit liegen für die Itemschwierigkeit jedoch nur bei 0.03 und für die Aufgabenschwierigkeit bei -0.06. (vgl. Tabelle IV-2.1.1.1r., Anhang G) Die Variable korreliert hoch mit den Merkmalen „Anakoluthe (STR4)" (0.76) und „Substantive/Eigennamen/Appellative (SUB)" (-0.74), weshalb diese Variablen bei der Berechnung von Partialkorrelationen mit der Item- und Aufgabenschwierigkeit kontrolliert werden. Auf Aufgabenebene ergibt sich dann ein Wert von 0.17, auf Itemebene von 0.12. Dieses Ergebnis ist zwar nur schwach ausgeprägt, jedoch insofern plausibel als das Verständnis von Negationen mehr Arbeitsgedächtniskapazität in Anspruch nimmt und dementsprechend das Verständnis von Stimuli und Items bei einem höheren Anteil an Negationen schwieriger wird.

Variable „Substantive/Eigennamen/Appellative (SUB)"

Stimuli mit einem hohen Anteil an Substantiven, Eigennamen und Appellativen weisen schwierigere Items auf und sind auch selbst schwieriger zu verstehen. Die Ergebnisse zur Variable SUB sind sowohl auf Item- als auch auf Stimulusebene signifikant und haben auf Aufgabenebene einen mittleren Effekt hinsichtlich der Varianzaufklärung. Die Variable korreliert jedoch nicht signifikant ($p > 0.10$) mit der Item- und der Aufgabenschwierigkeit. (vgl. Tabelle IV-2.1.1.1s., Anhang G) Die Partialkorrelation mit der Aufgabenschwierigkeit liegt bei 0.08, wobei die Variablen „Inhaltswörter (IWH)", „Worthäufigkeit (GWS)" und „Negationen (NEG)" kontrolliert wurden. Auf Itemebene beträgt sie 0.05.

Variable „Verben (VER)"

Auch der Anteil der Verben in einem Stimulus hat einen ähnlichen Effekt: die entsprechenden Items und Stimuli sind tendenziell schwieriger, wenn in einem Stimulus mehr als 15% Verben vorkommen. Im Gegensatz zur Variable SUB sind die Ergebnisse der Analysen mit der Variable VER jedoch nicht signifikant und tragen mit 2% bzw. 3% auch kaum zur Varianzaufklärung bei. (vgl. Tabelle IV-2.1.1.1t., Anhang G) Die Korrelationen mit der Item- bzw. Aufgabenschwierigkeit liegen bei -0.06 bzw. -0.10, wobei in beiden Fällen $p > 0.10$ ausfällt. Bei der Berechnung der Partialkorrelation wurden die Merkmale „Inhaltswörter (IWH)", „Worthäufigkeit (GWS)" und

„Länge der Propositionen (MLP)“ kontrolliert. Dabei ergab sich eine Partialkorrelation mit der Aufgabenschwierigkeit von r > 0.01, mit der Itemschwierigkeit liegt sie bei -0.07.

2.1.1.2. Merkmalsgruppe II: Präsentationsmerkmale

Variable „Länge in Minuten (LST)“

Die Länge der Stimuli wurde in drei Kategorien zusammengefasst (vgl. Tabelle IV-2.1.1.2a., Anhang G). Die längeren Stimuli hängen mit den deutlich schwierigeren Items zusammen. Dies kann mit der erhöhten Inanspruchnahme des Arbeitsgedächtnisses zu tun haben, aber auch damit, dass zu längeren, komplexeren Stimuli tatsächlich häufig auch schwierigere Items existieren, die stärker kognitive Fähigkeiten wie „zusammenfassen“, „übertragen“ oder „interpretieren“ erfordern. Auf Aufgabenebene sind Stimuli mittlerer Länge am leichtesten. Dies kann zum einen mit den zu den Stimuli gehörenden Items zusammenhängen, zum anderen aber auch daran liegen, dass die Schüler oft eine gewisse Zeit zum Einhören und zum Aufmerksamkeitfokussieren benötigen. Bei sehr kurzen Stimuli kann es passieren, dass die Beiträge bereits abgespielt sind, während die Schüler sich noch auf die neue Zuhörsituation einstellen.

Da die Variable LST hoch mit den Variablen „Wortzahl (AWS)“, „Referenzen (REF)“ und „Anteil der Propositionen (PRW)“ korreliert, werden diese Merkmale bei der Berechnung einer Partialkorrelation kontrolliert. Die Partialkorrelation LST mit der Stimulusschwierigkeit beträgt -0.11, mit der Itemschwierigkeit 0.05. Die Variable LST scheint für sich genommen bei insgesamt nicht signifikanten Analyseergebnissen kaum Einfluss auf die Schwierigkeit zu haben.

Variable „Wortzahl (AWS)“

Die Ergebnisse auf Item- und auf Stimulusebene zeigen, dass sehr lange Stimuli dazu führen, dass sowohl die dazugehörigen Items als auch der Stimulus schwieriger sind. (vgl. Tabelle IV-2.1.1.2b., Anhang G) Dieses Ergebnis kann damit begründet sein, dass Ermüdungseffekte bei den Schülern auftreten. Meist werden zu langen Stimuli auch viele Items präsentiert, sodass die Schüler bei langen Stimuli aus Zeitmangel nicht alle Items beantworten, was ebenfalls zu einer erhöhten Aufgabenschwierigkeit führt. Am leichtesten sind die Items und die Stimuli wenn Stimuli mit mittlerer Länge (400 – 600 Wörter) vorliegen. Dieses Ergebnis ist auf Itemebene signifikant, die Variable AWS hat auf Itemebene jedoch nur einen kleinen Anteil an der Varianzaufklärung. Aufgrund der geringen Stichprobengröße auf Aufgabenebene ist nicht auszuschließen, dass das Ergebnis von einzelnen Aufgaben abhängt. Die um die Variablen „Länge in Minuten (LST)“, „Referenzen (REF)“ und „Anzahl der Propositionen (PRO)“ kontrollierte Partialkorrelation mit der Stimulusschwierigkeit beläuft sich auf 0.13, die Partialkorrelation mit der Itemschwierigkeit beträgt nur noch 0.02.

Variable „Anzahl der Sprecher (ASP)“

Aufgaben mit mehr als drei Sprechern sind deutlich schwieriger als Aufgaben mit zwei oder drei Sprechern (vgl. Tabelle IV-2.1.1.2c., Anhang G). Dieser Befund erstaunt insofern, als die Sprecher auch in Fällen, in denen mehr als drei auftauchen, nicht in einer Runde miteinander sprechen, sondern vielmehr nacheinander befragt werden. Allerdings handelt es sich bei diesen

Stimuli auch i. d. R. um längere Hörbeiträge, sodass ein Ermüdungseffekt für die erhöhte Itemschwierigkeit verantwortlich sein könnte. Die Befunde ähneln sich auf Item- und auf Stimulusebene und sind auf Itemebene signifikant. Durch die Kontrolle der Merkmale „Rela-tionstyp: Antwort (REL2)" und „Rhetorische Mittel: Uneigentliches Sprechen (RHE3)" ergibt sich eine Partialkorrelation auf Aufgabenebene von 0.39 und von 0.18 auf Itemebene.

Variable „Sprechgeschwindigkeit (SGS)"

Je schneller ein Stimulus gesprochen wird, desto einfacher ist er und sind die entsprechenden Items (vgl. Tabelle IV-2.1.1.2d., Anhang G). Je schneller einer der hier betrachteten Stimuli jedoch gesprochen wird, desto mehr Merkmale gesprochener Sprache hat er i. d. R. Bei den schnell gesprochenen Stimuli handelt es sich meist um Gespräche, die bestimmte Aspekte wiederholt behandeln und in denen neue Informationen nur in größeren Abständen auftreten. Gerade die Stimuli sind jedoch besonders sorgfältig und langsam gesprochen, die auf einer schriftlichen Grundlage basieren und Vorlesungscharakter haben. Es handelt sich dabei auch um inhaltlich schwierigere Stimuli. Eine Analyse der Korrelation des Merkmals SGS mit der Item- und der Aufgabenschwierigkeit ergab nur für die Itemschwierigkeit ein signifikantes Ergebnis ($p < 0.10$) und insgesamt eher geringe Korrelationen (-0.17 mit der Itemschwierigkeit und -0.27 mit der Aufgabenschwierigkeit).

Da anzunehmen ist, dass die Länge der Stimuli einen Effekt auf den Einfluss der Sprechgeschwindigkeit auf die Schwierigkeit hat, werden die Analysen nur für die Stimuli wiederholt, die in einem Bereich von 200 – 600 Wörtern liegen. Die Ergebnisse sind in Tabelle IV-2.1.1.2e., Anhang G dargestellt. Extrem lange und extrem kurze Stimuli sollen von den Analysen ausgeschlossen werden. Die Korrelation mit der Aufgabenschwierigkeit beträgt nun bei einem nicht signifikanten Ergebnis -0.41 und weist auf einen mittleren Einfluss der Variable hin. Dies wird auch durch Cohen's d von 0.53 bestätigt. Die Varianzaufklärung der Variable SGS liegt bei 7%. Die Partialkorrelation für diese Variable (kontrolliert um die Merkmale „Inhaltswörter (IWH)" und „Akzent/Dialekt/Aussprache (AST)" beträgt auf Aufgabenebene nun sogar -0.64 und auf Itemebene -0.34. Die Variable scheint also einen relativ starken Einfluss auf die Schwierigkeit zu haben.

Variable „Akzent/Dialekt/Aussprache (AST)"

Stimuli, die in Standardsprache gesprochen sind, sind tendenziell am Einfachsten zu verstehen (vgl. Tabelle IV-2.1.1.2f., Anhang G). Auch die entsprechenden Items tendieren dazu einfacher zu sein als bei Stimuli, die mit leichtem oder starkem regionalen Dialekt gesprochen sind. Zwei Stimuli weisen auch Teile ausländischen Akzents auf. Interessanterweise sind diese Stimuli besonders leicht zu verstehen und hängen auch mit den einfachsten Items zusammen. Dies kann daran liegen, dass diese Stimuli auch besonders langsam gesprochen sind und generell einfache Fragen dazu gestellt wurden. Die Ergebnisse auf beiden Analyseebenen sind statistisch nicht signifikant und tragen auch kaum zur Varianzaufklärung bei. Die Variable AST korreliert mit den Variablen „Inhaltswörter (IWH)" (-0.53), „Sprechgeschwindigkeit (SGS)" (0.52) und „Literarischer Stimulus (SLT)" (0.50). Bei der Berechnung der Partialkorrelation auf Item- und auf Aufgabenebene werden diese Variablen deshalb kontrolliert. Die Partialkorrelation auf Aufgabenebene beträgt 0.09 und auf Itemebene -0.01.

Variable „Anzahl der Stimuluspräsentationen (AHO)“
Mittelwertsanalysen und Korrelationsanalysen mit der Variable AHO zeigen, dass mehrmaliges Vorspielen die Aufgabenschwierigkeit deutlich senkt (vgl. Tabelle IV-2.1.1.2g., Anhang G). Dies hängt u. U. damit zusammen, dass die mentale Repräsentation des Stimulus und die gegebenen Antworten überprüft und ggf. korrigiert werden können. Die Varianzaufklärung der Variable AHO ist sowohl auf Item- als auch auf Aufgabenebene nicht signifikant. Auch die Ergebnisse der Korrelationsanalysen mit der Item- und der Aufgabenschwierigkeit sind in beiden Fällen nicht signifikant und fallen sehr gering aus. Die um die Variablen „Rhetorische Mittel: Bildliche Darstellungsformen (RHE1)“ und „Stimulusschwierigkeit (TSA)“ kontrollierte Partialkorrelation beträgt -0.30 auf Aufgabenebene und -0.12 auf Itemebene und liegt damit deutlich höher als die ursprünglich erhaltenen Werte.

2.1.1.3. Merkmalsgruppe III: Inhaltlich-thematische Merkmale

Variable „Literarischer Stimulus (SLT)“
Die Analysen mit der Variable SLT ergaben auf beiden Analyseebenen sowohl ein statistisch signifikantes Ergebnis als auch einen deutlich praktisch relevanten Effekt (vgl. Tabelle IV-2.1.1.2h., Anhang G). Die Varianzaufklärung beträgt auf Itemebene jedoch nur 1%, auf Aufgabenebene immerhin 10%. Das Ergebnis besagt, dass literarische Stimuli deutlich einfacher sind als nicht-literarische Stimuli. Dies kann zum einen an der Auswahl liegen (es liegen nur fünf literarische Stimuli vor), zum anderen wäre aber auch der Einfluss anderer Aspekte, wie der Länge denkbar. Die Korrelationen mit der Item- und der Aufgabenschwierigkeit sind eher gering (0.12 und 0.31).

Um den Einfluss der Länge auszuschließen, werden die Analysen nur für die Stimuli wiederholt, die sich in einem Bereich von 200 – 600 Wörtern bewegen. Die Ergebnisse sind in Tabelle IV-2.1.1.2i., Anhang G dargestellt. Sie unterscheiden sich kaum von denen mit dem gesamten Datensatz, lassen die beschriebenen Tendenzen jedoch noch deutlicher erkennen. Literarische Stimuli und die Items dazu sind deutlich einfacher als die eingesetzten nichtliterarischen Stimuli und Items. Auch die Korrelationen mit der Item- und Aufgabenschwierigkeit fallen mit dem gefilterten Datensatz etwas höher aus (0.16 und 0.38), auch wenn sie nach wie vor, insbesondere auf Itemebene, eher gering sind. Die Partialkorrelation auf Aufgabenebene, bei der die Variablen „Inhaltswörter (IWH)“, „Worthäufigkeit (GWS)“ und „Verben (VER)“ kontrolliert werden, beläuft sich auf 0.39. Auf Itemebene ergibt sich ein Wert von 0.15.

Variable „Hörkontext (HKO)“
Die Ergebnisse zur Variable HKO sind in Teilen nicht erwartungskonform. Die Stimuli, die die meisten Merkmale geschriebener Sprache aufweisen (wie vorgetragene Lyrik/Prosa und Hörspiele) zählen sowohl zu den schwierigsten (grau unterlegt) als auch zu den einfachsten (fett gedruckt) Stimuli (vgl. Tabelle IV-2.1.1.2j., Anhang G). Erklärbar wird dies durch die geringe Fallzahl der jeweils berücksichtigten Stimuli. Einzelne sehr schwierige oder sehr leichte Aufgaben haben so einen starken Einfluss auf das Gesamtergebnis. Insgesamt ist das Ergebnis auf Itemebene jedoch signifikant ($p < 0.10$) und trägt mit 8% bzw. 30% auch stark zur Varianzaufklärung bei. Die Korrelation mit der Aufgabenschwierigkeit beträgt -0.39 und erhöht sich

auf -0.50, wird um die Variablen „Rhetorische Mittel: Jugendsprache/Umgangssprache (RHE4)", „Nähezeichen (STR5)" und „Referenzen (REF)" kontrolliert. Der Hörkontext spielt also für sich genommen eine wichtige Rolle für die Aufgabenschwierigkeit. Auf Itemebene beträgt die Partialkorrelation bei Kontrolle dieser Variablen -0.23.

Variable „Stimulusfunktion (TFU)"

Überwiegend instruktive/appellative/informierende Stimuli, wie persönliche Anordnungen, Kurzdefinitionen oder Zusammenfassungen, sowie überwiegend beschreibende/erzählende Stimuli, wie Geschichten, impressionistische Beschreibungen oder Berichte, tendieren bei einem insgesamt nicht signifikanten Ergebnis dazu, einfacher als vorwiegend argumentative/erklärende Stimuli zu sein (vgl. Tabelle IV-2.1.1.2k., Anhang G). Die Grenzen zwischen diesen Stimulusfunktionen sind nicht immer trennscharf, die beschriebene Tendenz ist jedoch relativ deutlich zu beobachten. Das Ergebnis wird allerdings auf Itemebene generell und auf Aufgabenebene insbesondere bei Code 2 durch eine relativ hohe Streuung der Werte relativiert. Auch die z. T. recht geringe Fallzahlbelegung auf Aufgabenebene bei Code 2 und 4 lässt eine Verallgemeinerung der Ergebnisse nicht zu. Die Variable TFU leistet einen geringen bzw. mittleren Beitrag zur Varianzaufklärung, korreliert jedoch gering mit der Schwierigkeit auf Item- und auf Aufgabenebene. Die Partialkorrelation auf Aufgabenebene beträgt 0.24, wobei die Variablen „Rhetorische Mittel: Bildliche Darstellungsformen (RHE1)", „Rhetorische Mittel: Uneigentliches Sprechen (RHE2)" und „Länge der Propositionen (MLP)" kontrolliert wurden. Auf Itemebene beträgt die Partialkorrelation bei Kontrolle der gleichen Merkmale -0.09.

Variable „Thema (THE)"

Die Variable THE gibt Aufschluss darüber, welche Themen mit eher leichten oder eher schwierigen Aufgaben zusammenhängen (vgl. Tabelle IV-2.1.1.2l., Anhang G). Die Fallzahl pro Thema ist mit 1-5 sehr gering, sodass diese Ergebnisse vorsichtig zu interpretieren sind. Eher einfach sind Aufgaben zu den Themen „Gesundheit und Hygiene", „Menschliche Beziehungen" und „Tägliches Leben" (fett markiert). Sehr schwierig ist die Aufgabe zum Thema „Wohnen und Umwelt" (grau hinterlegt). Die Variable leistet zwar sowohl auf Item- als auch auf Aufgabenebene einen großen Beitrag zur Varianzaufklärung, jedoch fallen die Korrelationen mit der Item- bzw. der Aufgabenschwierigkeit sehr gering aus. Die um die Variablen „Rhetorische Mittel: Uneigentliches Sprechen (RHE2)", „Referenzen (REF)" und „Anteil der Propositionen (PRW)" kontrollierte Partialkorrelation beläuft sich auf -0.05 auf Aufgabenebene und auf -0.13 auf Itemebene.

Variable „Hintergrundwissen (WEL)"

Die Variable WEL gibt an, ob Hintergrundwissen zum Verständnis der Stimuli benötigt wird. Stimuli, bei denen Hintergrundwissen eine Rolle spielt, sind erwartungsgemäß schwieriger als Stimuli, die ohne Hintergrundwissen verstanden werden (vgl. Tabelle IV-2.1.1.2m., Anhang G). Im Fall der Variable WEL wird auf Aufgabenebene zwar 7% Varianz durch das Merkmal aufgeklärt, jedoch ist das Ergebnis nur auf Itemebene statistisch signifikant ($p < 0.10$).

Es ist anzunehmen, dass die Variable WEL mit der Variable „Inhaltswörter (IWH)" zusammenhängt, da Inhaltswörter i. d. R. auch die Wörter sind, für die ggf. Weltwissen benötigt wird.

Die Analysen werden aus diesem Grund für die Fälle wiederholt, bei denen der Anteil an Inhaltswörtern mindestens 45% beträgt (IWH ⩾ 45%). Die Ergebnisse sind in Tabelle IV-2.1.1.2n., Anhang G dargestellt. Die Ergebnisse fallen sowohl auf Item- als auch auf Aufgabenebene noch deutlicher aus: Je mehr Hintergrundwissen zum Verständnis eines Stimulus benötigt wird, desto schwieriger werden die Items und Stimuli. Bei den Analysen mit dem gefilterten Datensatz steigt die Varianzaufklärung der Variable WEL von 1% auf 3% auf Itemebene und von 7% auf 8% auf Aufgabenebene. Die Analysen ergeben ein auf Itemebene signifikantes Ergebnis ($p < 0.10$). Die Variable korreliert kaum mit der Itemschwierigkeit und leicht mit der Aufgabenschwierigkeit, wobei nur die Korrelationen mit der Itemschwierigkeit statistisch signifikant sind ($p < 0.10$). Zusätzlich werden für die Variable WEL Partialkorrelationen mit der Item- und der Aufgabenschwierigkeit berechnet, wobei die Merkmale „Worthäufigkeit (GWS)", „Substantive/Eigennamen/Appellative (SUB)" und „Schlussfolgerungen/Inferenzen (SFI)" kontrolliert werden. Auf Aufgabenebene ergibt sich so ein Wert von 0.06, auf Itemebene hat die Variable keinen Einfluss auf die Schwierigkeit. Die Partialkorrelation geht hier gegen null ($r < 0.01$).

2.1.1.4. Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte

Variable „Relationstypen (REL)"

Die Ergebnisse der Analysen mit der Variable REL sind in Tabelle IV-2.1.1.4a., Anhang G dargestellt. Stimuli, die höhere Sprechanteile der Relationstypen „Frage/Impuls/Themensetzung (REL1)" sind schwieriger zu verstehen. Dieses Ergebnis ist sowohl auf Item- als auch auf Aufgabenebene signifikant ($p < 0.10$) und deckt 3% bzw. 10% Varianz auf. Auch Cohen's d weist auf einen mittleren Effekt hin. Die Korrelation dieses Merkmals mit der Itemschwierigkeit liegt bei 0.16 und mit der Stimulusschwierigkeit bei 0.32. Bei der Berechnung der Partialkorrelation wurden für die Variable REL1 die am höchsten damit korrelierenden Merkmale „Adjazenzstrukturen (STR3)" und „Reihenfolge/Aufzählung (REL5)" kontrolliert. Auf Aufgabenebene beläuft sich die Partialkorrelation für REL1 auf 0.23, auf Itemebene beträgt sie 0.08.

Auf Itemebene hängt ein höherer Anteil des Relationstyps „Antwort (REL2)" eher mit leichteren Items zusammen. Auf Aufgabenebene ist kaum ein entsprechender Effekt zu beobachten. Korrelationen mit diesem Typ sind jedoch weder mit der Item- noch der Aufgabenschwierigkeit signifikant und fallen in beiden Fällen auch sehr gering aus (-0.03 bzw. -0.03). Für die Variable REL2 fand eine Kontrolle der Variablen „Nähezeichen (STR5)", „Referenzen (REF)" und „Anzahl der Propositionen (PRO)" bei der Berechnung der Partialkorrelation statt. Es ergab sich auf Aufgabenebene ein Wert von 0.08, der nur etwas höher als die unbereinigte Korrelation ausfällt. Auf Itemebene geht die Partialkorrelation dagegen gegen null ($r < -0.01$).

Die Variablen „Spezifizierung (REL3)" und „Ziel/Bedingung (REL6)" korrelieren nicht mit der Schwierigkeit. Bei der Berechnung der Partialkorrelation wurde für die Variable REL3 die Merkmale „Adjazenzstrukturen (STR3)", „Antwort (REL2)" und „Rhetorische Mittel: Jugendsprache/Umgangssprache (RHE4)" kontrolliert. Die Partialkorrelation beläuft sich auf Aufgabenebene auf 0.06 und auf Itemebene auf 0.11. Für die Variable REL6 fand eine Kontrolle der Variablen „Stimulusfunktion (TFU)" und „Adjazenzstrukturen (STR3)" statt. Die daraus resultierende Partialkorrelation beträgt 0.01 auf Aufgabenebene und 0.06 auf Itemebene. Obwohl sie

damit etwas höher als die unkontrollierte Korrelation mit der Aufgaben- und der Itemschwierigkeit ausfällt, ist kaum Einfluss der Variable REL6 auf die Schwierigkeit zu beobachten.

Die Typen „Reihenfolge/Aufzählung (REL5)" und „Erklärung/Beweis/Ursache (REL4)" zeigen die Tendenz, dass bei erhöhter Auftretenshäufigkeit die Stimuli einfacher zu verstehen sind. Jedoch weisen sowohl die Varianzaufklärung als auch die Effektstärke auf eher geringe Effekte hin. Das Merkmal „Reihenfolge/Aufzählung (REL5)" korreliert kaum mit der Item- und der Aufgabenschwierigkeit (-0.03 bzw. -0.05). Der Zusammenhang ist nicht signifikant ($p > 0.10$). Bei der Variable REL5 wurden zur Berechnung der Partialkorrelation die Merkmale „Verberststellung (STR6)", REL1 und REL4 kontrolliert. Die Partialkorrelation beträgt auf Aufgabenebene -0.10 und auf Itemebene -0.03.

Das Merkmal REL4 korreliert etwas höher mit der Schwierigkeit auf Itemebene (-0.14) und auf Aufgabenebene (-0.20). Je häufiger der Relationstyp „Erklärung/Beweis/Ursache" in einem Stimulus geratet wurde, desto einfacher sind Stimulus und Items. Das Ergebnis ist auf Itemebene signifikant. Bei der Berechnung der Partialkorrelation wurden für die Variable REL4 die Merkmale „Länge der Propositionen (MLP)" und REL5 kontrolliert. Die daraus resultierende Partialkorrelation beträgt auf Aufgabenebene -0.06 und auf Itemebene -0.07. Der Einfluss der Variable REL4 auf die Schwierigkeit ist also sehr gering.

Zusammenfassend kann gesagt werden, dass die Einzelcodes der Variable REL bis auf die Merkmale „Frage/Impuls/Themensetzung (REL1)" und „Erklärung/Beweis/Ursache (REL4)" kaum Einfluss auf die Schwierigkeit sowohl auf Item- als auch auf Aufgabenebene haben. Das Merkmal REL1 korreliert auf Aufgabenebene immerhin mit 0.32 mit der Schwierigkeit und deckt 10% Varianz auf, das Merkmal REL4 deckt 4% Varianz auf und korreliert mit 0.20 mit der Aufgabenschwierigkeit.

Das Merkmal Relationstypen ist abhängig von der Gesprächssituation. Es ist anzunehmen, dass einige Subtypen dieser Variable stärker in monologischen bzw. in dialogischen Gesprächssituationen Verwendung finden. Da die Anzahl der Sprecher in den Stimuli einen Einfluss auf die Variable REL haben könnte, werden die Analysen für die Stimuli wiederholt, in denen nur ein Sprecher vorkommt bzw. in denen nur zwei oder drei Sprecher vorkommen. Es wird angenommen, dass einige Relationstypen dann ggf. einen deutlicheren Einfluss auf die Schwierigkeit haben.

Analysen mit Stimuli mit nur einem Sprecher

Die Ergebnisse der Analysen mit monologischen Stimuli sind in Tabelle IV-2.1.1.4b., Anhang G zu sehen. Im Fall des Relationstyps „Frage/Impuls/Themensetzung (REL1)" kommt der Einfluss der Variablen bei Stimuli mit nur einem Sprecher nun etwas deutlicher zum Vorschein. Die Items und die Stimuli sind umso schwieriger, je mehr Sprechanteile im Stimulus dem Relationstyp 1 zugeordnet werden können. Die Varianzaufklärung dieser Variablen liegt bei 7% auf Itemebene bzw. 15% auf Aufgabenebene und auch Cohen's d weist mit Werten von -0.55 und -0.79 auf einen mittleren bzw. starken Effekt hin. Die Variable korreliert mit 0.26 mit der Itemschwierigkeit ($p < 0.10$). Auch mit der Aufgabenschwierigkeit lassen sich statistisch signifikante ($p < 0.10$) Korrelationen von 0.32 beobachten.

Der Einfluss der Variable „Antwort (REL2)“ zeigt einen schwachen Effekt auf Itemebene. Das Merkmal korreliert mit -0.11 mit der Itemschwierigkeit, wobei das Ergebnis nicht signifikant ($p > 0.10$) ist.

Die Variablen „Spezifizierung (REL3)“ und „Ziel/Bedingung (REL6)“ zeigen keinen Einfluss auf die Schwierigkeit. Insbesondere bei der Variable REL6 sind aufgrund der geringen Fallzahlbelegung der Kategorien (4 Fälle > 10% auf Itemebene und 1 Fall > 10% auf Aufgabenebene bei insgesamt 158 bzw. 14 Fällen) jedoch kaum Aussagen über den Einfluss der Variable auf die Schwierigkeit möglich.

Für die Variable „Erklärung/Beweis/Ursache (REL4)“ sinkt der Zusammenhang mit Schwierigkeit auf Stimulusebene. Die Varianzaufklärung liegt mit 3% auf Item- und Aufgabenebene im unteren Bereich und auch die Mittelwertsunterschiede der Schwierigkeiten haben sich im Vergleich zum ersten Analysedurchlauf mit dem kompletten Datensatz kaum verändert. Die Variable korreliert signifikant ($p < 0.10$), aber schwach negativ mit der Itemschwierigkeit (0.17) und nicht signifikant ($p > 0.10$) mit der Aufgabenschwierigkeit (-0.17).

Das Ergebnis für die Variable „Reihenfolge/Aufzählung (REL5)“ zeigt für die Fälle, bei denen nur ein Sprecher im Stimulus erscheint, die folgende Tendenz: Je mehr Relationen des Typs Reihenfolge oder Aufzählung geratet wurden, desto einfacher sind die entsprechenden Items und der Stimulus. Die Variable REL5 hat nun eine Varianzaufklärung von 6% auf Item- und von 17% auf Aufgabenebene, was Cohen's d mit Werten von 0.48 und 0.83 bestätigt. Die Variable korreliert mit einem Wert von -0.24 statistisch signifikant mit der Itemschwierigkeit ($p < 0.10$) und nicht signifikant mit der Aufgabenschwierigkeit ($p > 0.10$; $r = -0.41$).

Analysen mit Stimuli mit zwei oder drei Sprechern

Tabelle IV-2.1.1.4c., Anhang G gibt einen Überblick über die Ergebnisse der Analysen mit dialogischen Stimuli. Auf Aufgabenebene lagen bei den Analysen mit den Aufgaben, in denen nur zwei oder drei Personen sprechen, nur noch sieben Fälle vor. Die Variablen „Frage/Impuls/Themensetzung (REL1)“ und „Erklärung/Beweis/Ursache (REL4)“ zeigen bei den dialogischen Stimuli keinen Zusammenhang mit der Schwierigkeit.

Deutlicher als bei den Analysen mit dem gesamten Datensatz oder mit monologischen Stimuli wird bei dieser Wiederholung jedoch der Einfluss der Variable REL2 auf die Item- und die Stimulusschwierigkeit: Je mehr Stimulusanteile dem Relationstyp „Antwort (REL2)” zuzuordnen sind, desto einfacher werden die entsprechenden Items und der Stimulus. Die Variable korreliert mit -0.16 signifikant ($p < 0.10$) mit der Itemschwierigkeit und mit -0.51 mit der Stimulusschwierigkeit, wobei dieses Ergebnis nicht signifikant ($p > 0.10$) ist.

Ähnlich fallen auch die Analysen mit der Variable „Spezifizierung (REL3)“ aus. Sowohl auf Item- als auch auf Aufgabenebene korreliert dieses Merkmal positiv mit der Schwierigkeit. Je häufiger der Relationstyp „Spezifizierung“ also in einem Stimulus geratet wurde, desto schwieriger sind die entsprechenden Items. Das Ergebnis ist auf Itemebene signifikant ($p < 0.10$). Die Variable trägt auf Itemebene 11% und auf Aufgabenebene 43% zur Varianzaufklärung bei.

Die Variable „Reihenfolge/Aufzählung (REL5)“ korreliert auf Itemebene etwas geringer mit der Schwierigkeit als bei den Analysen mit den monologischen Stimuli (r = 0.17), auf Aufgabenebene hingegen etwas stärker (r = 0.49). Der Zusammenhang dieses Relationstyps mit der Schwierigkeit ist nun positiv. Hing die Variable REL5 bei den monologischen Stimuli noch mit den einfacheren Items zusammen, so hängt dieses Merkmal bei den dialogischen Stimuli eher mit einer höheren Item- bzw. Stimulusschwierigkeit zusammen. Die Variable trägt auf Aufgabenebene mit 24% zur Varianzaufklärung bei.

Die Variable „Ziel/Bedingung (REL6)“ zeigt bei den dialogischen Stimuli einen Effekt auf Aufgabenebene. Sie korreliert hier mit 0.29 mit der Schwierigkeit.

Tabelle IV-2.1.1.4d., Anhang G gibt einen Überblick über die Zusammenhänge der einzelnen Variablen mit der Item- bzw. der Stimulusschwierigkeit, bei den Analysen mit allen Stimuli, nur den monologischen und nur den dialogischen Stimuli. Die Subvariablen des Merkmals „Relationstyp“ korrelieren in ganz unterschiedlicher Weise mit der Schwierigkeit. Die Variablen „Frage/Themensetzung/Impuls (REL1)“ und „Erklärung/Beweis/Ursache (REL4)“ korrelieren bei den Analysen mit dem Gesamtdatensatz sowohl auf Item- als auch auf Stimulusebene mit der Schwierigkeit. Höhere Anteile der Variable REL1 korrelieren eher mit schwierigeren Items, höhere Anteile der Variable REL4 korrelieren eher mit einfacheren Items. Bei einer Wiederholung dieser Analysen nur mit monologischen Stimuli verstärkt sich das Ergebnis für die Variable REL1. Für die Variable REL4 fällt der Befund auf Itemebene geringfügig deutlicher aus, auf Aufgabenebene geringfügig schwächer. Einen deutlichen negativen Zusammenhang mit der Schwierigkeit zeigt nun auch die Variable „Reihenfolge/Aufzählung (REL5)“. Bei den Analysen nur mit dialogischen Stimuli kehrt sich der Zusammenhang der Variable REL5 ins Positive. Bei dialogischen Stimuli korreliert dieses Merkmal eher mit schwierigeren Items, wohingegen es bei monologischen Stimuli mit einfacheren Items korreliert. Bei den Stimuli mit zwei oder drei Sprechern korrelieren auch die Variablen „Antwort (REL2)“ und „Spezifizierung (REL3)“ mit der Schwierigkeit. Die Variable REL2 korreliert negativ mit der Schwierigkeit, die Variable REL3 korreliert eher mit schwierigeren Items. Die Variable „Ziel/Bedingung (REL6)“ korreliert nur auf Aufgabenebene leicht mit der Schwierigkeit.

Variable „Schlussfolgerungen/Inferenzen (SFI)“

Stimuli, bei denen zum Verständnis mehr als 2% Schlussfolgerungen notwendig sind, fallen sowohl durch erhöhte Itemschwierigkeit als auch durch erhöhte Aufgabenschwierigkeit auf (vgl. Tabelle IV-2.1.1.4e., Anhang G). Die Varianzaufklärung durch die Variable SFI bewegt sich jedoch eher im niedrigen Bereich und das Ergebnis ist nur auf Itemebene signifikant. Cohen's d von -0.22 bzw. -0.40 weist auf eher geringe Bedeutsamkeit hin. Die Ergebnisse der Korrelationsanalysen mit der Item- bzw. Aufgabenschwierigkeit sind nicht signifikant ($p > 0.10$) und liegen bei 0.01 bzw. 0.08. Für die Berechnung von Partialkorrelationen der Variable mit der Item- bzw. der Aufgabenschwierigkeit fand eine Kontrolle der Variablen „Worthäufigkeit (GWS)“, „Ellipsen (STR2)“ und „Anteil der Propositionen (PRW)“ statt. Die Partialkorrelation mit der Aufgabenschwierigkeit beträgt 0.15 und liegt damit deutlich über dem unkontrollierten Wert. Auf Itemebene ist kein Einfluss der Variable auf die Schwierigkeit zu beobachten. Die Partialkorrelation fällt < -0.01 aus.

Variablen „Anzahl Propositionen (PRO)" und „Anteil der Propositionen (PRW)"
Analysen zur Variable „Anzahl Propositionen (PRO)" lassen erkennen, dass Stimuli mit einer mittleren Propositionsdichte von 50 – 100 am einfachsten zu verstehen sind und diese auch mit den einfachsten Items zusammenhängen (vgl. Tabelle IV-2.1.1.4f., Anhang G). Wird ein Stimulus länger und weist deshalb mehr Propositionen auf oder verfügt er aufgrund seiner Kürze zwar insgesamt über weniger Propositionen, aber auch über eine höhere Propositionsdichte, so steigt die Itemschwierigkeit. Die Variable leistet auf Itemebene einen kleinen Beitrag zur Varianzaufklärung, auf Aufgabenebene liegt ihr Beitrag mit 15% deutlich höher. Allerdings ist nur das Ergebnis auf Itemebene statistisch signifikant. Die Korrelationen mit der Item- und der Aufgabenschwierigkeit sind insgesamt recht gering. Die Partialkorrelation geht auf Aufgabenebene sogar gegen null (< 0.01), auf Itemebene beträgt sie -0.03. Kontrolliert wurden dabei die Variablen „Länge in Minuten (LST)", „Wortzahl (AWS)" und „Referenzen (REF)". Es handelt sich bei den vorher beobachteten Effekten also um eine Scheinkorrelation.

Um den Einfluss der Stimuluslänge zu reduzieren und genauere Aussagen über den Einfluss der Propositionsdichte zu erhalten, werden die Analysen mit der Variable PRW wiederholt. Dabei ergeben sich für Item- und Aufgabenebene gegenläufige Ergebnisse, wie in Tabelle IV-2.1.1.4g., Anhang G dargestellt. Je höher der Anteil an Propositionen in einem Stimulus ist, desto schwieriger sind die entsprechenden Items, desto einfacher sind jedoch die Stimuli. Dies könnte dadurch zu erklären sein, dass bei einem Stimulus mit vielen Informationen auch nach diesen gefragt wird und die Schüler zur Beantwortung der Items viele Informationen gleichzeitig im Arbeitsgedächtnis halten müssen. Zu beachten ist jedoch, dass die Variable kaum mit der Schwierigkeit korreliert und zu hinterfragen ist, welchen Einfluss sie tatsächlich hat. Nach Kontrolle der Variablen „Verben (VER)", „Schlussfolgerungen/Inferenzen (SFI)" und „Länge der Propositionen (MLP)" beträgt die Partialkorrelation mit der Aufgabenschwierigkeit nur noch -0.02, mit der Itemschwierigkeit -0.01.

Variable „Länge der Propositionen (MLP)"
Die mittlere Länge der Propositionen hat keinen nennenswerten Einfluss auf die Item bzw. Aufgabenschwierigkeit und korreliert nur mit 0.06 mit der Item- und 0.22 mit der Aufgabenschwierigkeit. (vgl. Tabelle IV-2.1.1.4h.) Bei sehr langen Propositionen mit mehr als sieben Wörtern sind die Items deutlich schwieriger. Zu beachten ist dabei jedoch die geringe Fallzahlbelegung von Code 4. Der Beitrag der Variable zur Varianzaufklärung liegt mit 2% bzw. 10% im unteren bzw. mittleren Bereich. Für die Berechnung von Partialkorrelationen wurden die Merkmale „Literarischer Stimulus (SLT)" und „Verben (VER)" kontrolliert. Die Partialkorrelation mit der Aufgabenschwierigkeit beträgt 0.18 mit der Itemschwierigkeit nur noch 0.01.

Auch bei einer Dichotomisierung der Variable tritt kein deutlicheres Ergebnis ein (vgl. Tabelle IV-2.1.1.4i., Anhang G). Auf Itemebene scheint die Variable einen leichten Einfluss zu haben: die schwierigeren Items stehen mit einer höheren Propositionslänge in Verbindung. Das Ergebnis ist jedoch statistisch nicht signifikant. Auf Aufgabenebene ist kein Effekt zu beobachten und die Variable trägt weder auf Item- noch auf Aufgabenebene zur Varianzaufklärung bei.

Die Variable MLP hat gemäß der Faktorenanalyse einen gemeinsamen Faktor mit der Variable „Literarischer Stimulus (SLT)". Da angenommen wird, dass der Stimulustyp u. U. zusätzlichen Einfluss auf die Schwierigkeit von Items und Stimuli haben könnte, werden die Analysen mit der dichotomen Variable MLP nur für nicht-literarische Stimuli wiederholt (vgl. Tabelle IV-2.1.1.4j., Anhang G). Die Analysen mit dem gefilterten Datensatz zeigen den Einfluss der Variable auf Itemebene deutlicher. Längere Propositionen hängen mit den einfacheren Items zusammen. Auf Aufgabenebene ist kein Effekt zu beobachten. Die Partialkorrelation mit der Stimulusschwierigkeit liegt mit 0.26 etwas höher als die einfache Korrelation. Auf Itemebene beträgt die Partialkorrelation der Variable MLP mit der Schwierigkeit 0.03. Kontrolliert wurden die Variablen „Verben (VER)" und „Schlussfolgerungen/Inferenzen (SFI)".

2.1.1.5. Merkmalsgruppe V: Globalurteil

Variable „Stimulusschwierigkeit (TSA)"
Zu jedem Item liegt eine während der Itemerstellung gewonnene Einschätzung der Aufgabenentwickler hinsichtlich der vermuteten Stimulusschwierigkeiten vor. Diese Einschätzung entspricht recht gut den empirischen Aufgabenschwierigkeiten, wobei die Korrelationen mit der Schwierigkeit auf Item- und auf Aufgabenebene eher gering ausfallen (vgl. Tabelle IV2.1.1.5., Anhang G). Das Ergebnis ist auch nur auf Itemebene für die Variable TSA signifikant. Die Variable hat einen geringen Anteil an der Varianzaufklärung. Bei der Berechnung der Partialkorrelation wurden die Merkmale „Neudeutsch/Anglizismen (RHE3)", „Anzahl der Stimuluspräsentationen (AHO)" und „Negationen (NEG)" kontrolliert. Die Partialkorrelation mit der Itemschwierigkeit beträgt 0.11 und mit der Aufgabenschwierigkeit 0.12.

2.1.2. Mittelwertsvergleiche der Stimulusmerkmale aus dem Lehrerfragebogen
18 Stimuli wurden im Ländervergleich 2009 mittels eines fünfstufigen bipolaren Fragebogens von den Deutschlehrkräften der getesteten Klassen hinsichtlich unterschiedlicher Variablen eingeschätzt. Tabelle IV-2.1.2a., Anhang G gibt einen ersten Überblick über das Ergebnis der Ratings über alle Lehrkräfte und alle Aufgaben hinweg. Bei der Einschätzung der Aufgaben haben nicht alle Lehrer alle Kategorien bearbeitet. Beispielsweise liegen für die Variable GEU deutlich weniger Einschätzungen vor als für die Variable TON. Aufgrund des geringen Missing-Anteils von höchstens 5.5% ist jedoch nicht von starken Verzerrungen aufgrund fehlender Daten auszugehen, sodass in allen folgenden Analysen nur mit den verfügbaren Daten gearbeitet wird. Die folgenden Analysen werden nur auf Aufgabenebene durchgeführt.

Im Rahmen einer Reliabilitätsanalyse wurde zunächst ermittelt, wie viel Varianz durch die einzelnen Variablen aufgedeckt wird, hinsichtlich ihrer Einschätzung durch unterschiedliche Lehrer (Spalte L „Lehrer"), bei unterschiedlichen Aufgaben (Spalte A „Aufgabe") oder aufgrund weiterer nicht bekannter Faktoren (Spalte F „Sonstige Faktoren"). (vgl. Tabelle IV-2.1.2b., Anhang G) Dabei wird ein Varianzkomponentenmodell mit den beiden zufälligen Faktoren „Aufgabe" und „Lehrer" eingesetzt (Wirtz/Caspar, 2002) und es werden diese Varianzanteile bestimmt.

Zu wünschen ist, dass ein größtmöglicher Anteil systematischer Varianz durch die Aufgaben aufgedeckt wird und nicht von einem (zufälligen oder unsystematischen) Ankreuzverhalten der Lehrer oder unbekannten Faktoren abhängt. Eine hohe Varianzkomponente des Faktors „Lehrers" indiziert, dass über die Ratings aller Aufgaben hinweg systematische Strenge-Milde-Effekte bei Lehrern ausgeprägt sind. Insgesamt fallen die Varianzkomponenten hinsichtlich der Aufgaben und der Lehrer relativ gering aus. Dieser Effekt war bereits bei der Einschätzung der Plausibilität der Distraktoren (Variablen PDI und MPD) zu beobachten. Eine Varianzaufklärung von über 30% durch die Aufgaben kann deshalb bereits als gutes Ergebnis zählen. Dabei entfällt beispielsweise bei der Variable „Vertrautheit mit dem Thema (VTH)" ein Varianzanteil von 35.2% auf die Komponente „Aufgabe". Diese Maßzahl kann als Korrelation von 0.35 interpretiert werden, mit der zwei zufällig ausgewählte Lehrer die Variable einer Aufgabe beurteilen.

Besonders reliabel sind die grau unterlegten Variablen VTH, TON, AGU, WNL und WEH (siehe Tabelle IV-2.1.2b., Anhang G). Die fett markierten Variablen WAK und KOH erklären durch die Aufgaben kaum Varianz und erweisen sich damit als wenig reliabel. Diese Ergebnisse sollten auch bei den weiteren Analysen berücksichtigt werden.

Im Folgenden fließen die einzelnen Kategorien des Fragebogens als dichotom kategorisierte Variablen in die Analysen ein. Dargestellt werden die Mittelwerte aus allen Lehrereinschätzungen über die 18 Aufgaben hinweg.

Variable „Vertrautheit mit dem Thema (VTH)"
Bei dieser Variable schätzten die Lehrkräfte ein, wie vertraut die Schüler mit dem Thema des Stimulus waren. Bei sieben Stimuli wählten die Lehrer im Mittel die Stufen 4 oder 5, bei elf Stimuli wählten die Lehrer im Mittel die Stufen 1 bis 3. Die entsprechenden Stimuli weisen niedrigere Schwierigkeitswerte auf, wenn die Schüler nach Einschätzung der Lehrkräfte mit den Themen vertraut sind (vgl. Tabelle IV-2.1.2c., Anhang G). Die Variable VTH korreliert stark mit der Aufgabenschwierigkeit (0.44), wobei dieses Ergebnis statistisch signifikant ist ($p < 0.10$). Die Variable VTH erklärt 12% Varianz und auch Cohen's d von -0.68 zeigt die Bedeutsamkeit.

Variablen „Wortschatz ‚abstrakt – konkret' (WAK)" und „Wortschatz ‚fachspezifisch – alltäglich' (WFA)"
Stimuli zu eher abstrakten und fachspezifischen Themen sind gemäß den Erwartungen deutlich schwieriger als Stimuli zu konkreten und alltäglichen Themen. Beide Ergebnisse sind statistisch signifikant ($p < 0.10$) und tragen mit 21% bzw. 16% stark zur Varianzaufklärung bei. Die beiden Variablen korrelieren negativ mit der Aufgabenschwierigkeit (-0.47 für WAK und -0.55 für WFA) und bestätigen die Resultate: Je konkreter bzw. alltäglicher der Wortschatz eines Stimulus eingeschätzt wurde, desto einfacher ist der Stimulus. Die Ergebnisse der Korrelationsanalyse sind in beiden Fällen statistisch signifikant mit $p < 0.10$. Die Ergebnisse der Mittelwertsanalysen sind in Tabelle IV-2.1.2d., Anhang G dargestellt.

Variable „Grammatik (GRA)“
Stimuli, die nach Einschätzung der Lehrkräfte eher einfache grammatische Strukturen aufweisen, sind auch entsprechend den empirischen Aufgabenschwierigkeiten deutlich einfacher (vgl. Tabelle IV-2.1.2e., Anhang G). Die Variable GRA trägt mit 16% stark zur Varianzaufklärung bei. Cohen's d weist mit 0.91 auf einen starken praktischen Effekt. Die Korrelation mit der Aufgabenschwierigkeit beträgt -0.43 und ist statistisch signifikant ($p < 0.10$).

Variable „Kohärenz (KOH)“
Bei der Variable KOH ist nur eine schwache Tendenz zu erkennen: Je weniger kohärent die Stimuli sind, desto einfacher werden sie (vgl. Tabelle IV-2.1.2f., Anhang G). Dieses Ergebnis ist statistisch nicht signifikant ($p > 0.10$) und die Variable trägt mit 1% auch kaum zur Varianzaufklärung bei. Auch das Ergebnis einer Korrelationsanalyse ist statistisch nicht signifikant ($p > 0.10$), die Korrelation mit der Aufgabenschwierigkeit liegt bei -0.12.

Variablen „Gesamteindruck des Stimulus ‚interessant – uninteressant' (GIU)“, „Gesamteindruck des Stimulus ‚elegant – unbeholfen' (GEU)“ und „Gesamteindruck des Stimulus ‚abwechslungsreich – eintönig' (GAE)“
Die Ergebnisse der Mittelwertsanalysen mit den Variablen GIU, GEU und GAE sind in Tabelle IV-2.1.2g., Anhang G dargestellt. Stimuli, die gemäß der Lehrereinschätzung eher interessant sind (Variable GIU), sind einfacher als eher uninteressante Stimuli. Dieses Ergebnis ist plausibel, da Zuhören stark auf Prozessen der Aufmerksamkeitslenkung basiert. Zu hinterfragen ist bei der Variable GIU jedoch, ob die von den Lehrern als interessant eingestuften Stimuli auch von den Schülern als eher interessant empfunden werden. Dieser Frage wird in Kapitel *1.3.1. Einschätzung der Stimuli durch die Schüler* nachgegangen. Die Variable GIU korreliert mit 0.56 mit der Aufgabenschwierigkeit, wobei dieses Ergebnis auch statistisch signifikant ist ($p < 0.10$). Die Varianzaufklärung dieser Variable beträgt immerhin 15%. Bei der Einschätzung der Stimuli auf den Polen „elegant – unbeholfen“ und „abwechslungsreich – eintönig“ (Variable GEU) zeigt sich, dass Stimuli, die eher elegant und abwechslungsreich wirken, schwieriger sind als Stimuli, die als eher unbeholfen und eintönig eingestuft wurden. Es ist anzunehmen, dass ein eleganter oder abwechslungsreicher Gesamteindruck auf einer raffinierten und komplexen Machart der Stimuli basiert und es sich möglicherweise bei derartigen Stimuli auch um insgesamt schwierigere Stimuli handelt, als Stimuli, bei denen erwartungsgemäß linear Informationen gegeben werden und die dementsprechend als eintönig eingestuft werden. In Präpilotierungen mit den Aufgaben zeigte sich auch, dass gerade Stimuli, die die Schüler ansprachen, zu schlechteren Leistungen bei der Itembeantwortung einschließlich vielen nicht beantworteten Items führten, da hier das Hörerlebnis im Vordergrund stand.

Die Ergebnisse der Korrelationsanalysen sind nur für die Variable GAE signifikant ($p < 0.10$) und weist hier auf einen positiven Zusammenhang der Variablen mit der Aufgabenschwierigkeit hin (0.48). Die Korrelation der Variable GEU liegt bei -0.30. Beide Variablen haben mit 16% (GEU) und 24% (GAE) einen großen Anteil an der Varianzaufklärung.

Variable „Ton des Stimulus (TON)"
Eher persönliche und gefühlsbetonte Stimuli sind deutlich einfacher als eher unpersönliche und sachliche Stimuli. Dieses Fazit wird auch durch die Korrelationsanalyse bestätigt ($r = 0.50$), deren Ergebnis statistisch signifikant ist ($p < 0.10$). Die Variable TON hat mit 15% einen großen Anteil an der Varianzaufklärung. Die Ergebnisse der Mittelwertsanalysen sind in Tabelle IV-2.1.2h., Anhang G dargestellt.

Variablen „Ausdrucksweise ‚gewählt – umgangssprachlich' (AGU)" und „Ausdrucksweise ‚komplex/ ausschweifend – einfach/knapp' (AKE)"
Sehr umgangssprachliche bzw. einfache und knappe Stimuli sind einfacher als Stimuli, die eher als gewählt oder komplex und ausschweifend eingestuft wurden (vgl. Tabelle IV-2.1.2i., Anhang G). Das Ergebnis der Variable AGU ist statistisch signifikant ($p < 0.10$) und weist mit 39% einen sehr hohen Anteil an der Varianzaufklärung auf. Auch das Ergebnis der Korrelationsanalyse der Variable AGU mit der Aufgabenschwierigkeit ist statistisch signifikant ($p < 0.10$) und bestätigt den dargestellten Befund ($r = -0.45$). Dagegen sind die Ergebnisse die Variable AKE betreffend statistisch nicht signifikant ($p > 0.10$). Ihr Anteil an der Varianzaufklärung liegt mit 2% im unteren Bereich. Die Korrelation mit der Aufgabenschwierigkeit beträgt -0.31.

Variable „Informationsebene (INF)"
Sehr spitzfindige und tiefgründige Stimuli sind schwieriger zu verstehen als Stimuli, die als offensichtlich und oberflächlich eingestuft wurden. Das Ergebnis ist statistisch signifikant ($p < 0.10$) und trägt mit 20% stark zur Varianzaufklärung bei. Eine Korrelationsanalyse ergab kein signifikantes Ergebnis ($p > 0.10$) und eine Korrelation von -0.29. Die Ergebnisse werden in Tabelle IV-2.1.2j., Anhang G gezeigt.

Variablen „Wirkung des Stimulus ‚regt zum Nachdenken an – regt zum Lachen an' (WNL)" und „Wirkung des Stimulus ‚ernsthaft – humorvoll' (WEH)"
Die Einschätzung der Wirkung hängt bei beiden Variablen mit den Stimulusschwierigkeiten zusammen. Stimuli, die von den Lehrkräften als humorvoll empfunden wurden oder die zum Lachen anregen, sind einfacher als Stimuli, die als eher ernsthaft oder als zum Nachdenken anregend eingestuft wurden (vgl. Tabelle IV-2.1.2k., Anhang G). Die Variablen haben mit 7% (WNL) bzw. 17% (WEH) einen mittleren bis großen Anteil an der Varianzaufklärung. Korrelationen mit der Aufgabenschwierigkeit sind nur für die Variable WEH signifikant ($p < 0.10$) und liegen bei -0.32 (WNL) und -0.42 (WEH).

2.2. Itemmerkmale

2.2.1. Mittelwertsvergleiche der Itemmerkmale
Um den Einfluss der unterschiedlichen Ausprägungen der selektierten Variablen auf die Itemschwierigkeit zu erfassen, wurden deskriptive Mittelwertsvergleiche durchgeführt. Bei den Analysen auf Itemebene finden auch Variablen Beachtung, die auf einer Stimulus-Item-Interaktion beruhen. Grundlage für die Itemschwierigkeiten ist immer die mittlere Schwierigkeit der Items aus dem HSA- und dem MSA-Design. Für die Berechnung der Partialkorrelation werden die Drittmerkmale einbezogen, die mit dem zu untersuchenden Merkmal mindestens mit ±0.3 korrelieren.

2.2.1.1. Merkmalsgruppe I: Itemformat

Variable „Itemformat (IFK)" und „Itemformat (IFA)"
Ein Vergleich in einer Kreuztabelle verdeutlicht (Tabelle IV-2.2.1.1a.), dass es sich bei der Variable IFA im Wesentlichen um eine feinere Differenzierung der Variable IFK handelt. Multiple-Choice-Items (MC) und Richtig-Falsch-Items (RF) werden vollständig vom Code „Geschlossen-Ankreuzen (GA)", Reihenfolgeitems (RI) und Zuordnungsitems (ZO) vom Code „Geschlossen -Kodieren (GK)" und Offene Items (OI) vom Code 1+ „Antwort mit mehreren Codes" abgedeckt. Bei den Halboffenen Items (HO) kommen hingegen beide Codes, 0/1 „Einfache Antwort – 0/1 Kodierung" und 1+, zum Tragen. Dies ist darin begründet, dass es auch für Items, bei denen eine Kurzantwort verlangt wird (Code HO), Fälle gibt, bei denen mehrere Lösungsmöglichkeiten akzeptiert sein können.

Tabelle IV-2.2.1.1a.: Kreuztabelle – IFA/IFK

Code	GA	GK	0/1	1+	Gesamt
MC	87	0	0	0	87
RF	182	0	0	0	182
RI	0	5	0	0	5
ZO	0	4	0	0	4
HO	0	2	66	27	95
OI	0	0	0	11	11
Gesamt	**269**	**11**	**66**	**38**	**384**

Für beide Variablen wurden die Häufigkeitswerte berechnet (vgl. Tabelle IV-2.2.1.1b., Anhang G). Die Variable IFK klärt insgesamt 13% Varianz auf. Items, bei denen die Möglichkeit zur Antwort stark gelenkt wird (Code GA), fallen den Schülern dabei am leichtesten. Nicht erwartungskonform verhalten sich jedoch Items des Typs GK. Auch bei ihnen werden die Antworten der Schüler stark gelenkt, sie fallen jedoch relativ schwierig aus. Zu diesen Items gehören Formate wie Reihenfolge oder Zuordnung. Dies kann daran liegen, dass bei diesen Items i. d. R. die gesamte Antwort als falsch gewertet wird, wenn eine falsche Teilantwort gegeben wird, oder die Antwortformate den Schülern unvertraut sind. Aufgrund der sehr geringen Fallzahl von 11 sollten diese Ergebnisse mit Vorsicht interpretiert werden. Durch die genauere Differenzierung bei Variable IFA wird 19% Varianz aufgeklärt. Die Ergebnisse der Analysen bei Variable IFK werden bestätigt. Ankreuz-Items (Code MC und RF) erweisen sich als am einfachsten zu bearbeitendes Itemformat. Zuordnungsitems (ZO) und vor allem Reihenfolgeitems (RI) fallen überraschend schwierig aus, wobei die geringen Fallzahlen zu berücksichtigen sind. Für die halboffenen (HO) und die offenen Items (OI) ergibt sich ein erwartungskonformes Ergebnis. Je freier die Schüler in der Beantwortung eines Items sind, desto anspruchsvollere Leistungen werden i. d. R. erwartet (z. B. interpretieren, übertragen, analysieren) und desto schwieriger ist das Item.

In einem nächsten Schritt werden die Analysen für die einzelnen Codes der beiden Variablen wiederholt, um zu prüfen, ob ein bestimmtes Itemformat besonders großen Einfluss auf die

Schwierigkeit hat (vgl. Tabelle IV-2.2.1.1c., Anhang G). Dabei zeigt sich, dass insbesondere das Itemformat „Geschlossen – Ankreuzen" (IFK.GK) und das Itemformat „Einfache Antwort – 0/1-Kodierung" (IFK.01) mit 13% bzw. 6% einen mittleren bzw. großen Beitrag zur Varianzaufklärung leisten. Beide Formattypen korrelieren zudem signifikant mit der Itemschwierigkeit, wobei im Fall des Formats „Geschlossen – Ankreuzen" eine negative Korrelation vorliegt.

In einem nächsten Schritt wurden für die einzelnen Codes Partialkorrelationen berechnet, bei denen der Einfluss weiterer hoch mit den Codes korrelierender Faktoren ausgeschlossen wird. So wird bei dem Code „Geschlossen – Ankreuzen" der Einfluss der Formattypen „Einfache Antwort – 0/1-Kodierung",„Antwort mit mehreren Codes oder mit mehreren einfachen Variablen" und „Halboffen" kontrolliert, da die beiden Variablen hoch (mit -0.58, -0.53 und -0.68) mit dem Code „Geschlossen – Ankreuzen" korrelieren. Die so erhaltene Korrelation des Codes IFK.GA mit der Itemschwierigkeit beträgt -0.16.

Für die Variable „Geschlossen – Kodieren" beträgt die Korrelation mit der Variable „Geschlossen – Ankreuzen" -0.42. Rechnet man den Einfluss der Variable IFK.GA heraus, so verbleibt für die Variable IFK.GK nur noch eine Korrelation mit der Itemschwierigkeit von 0.03. Für die beiden Variablen IFK.01 und IFK.1P wurde jeweils der Einfluss der Codes IFK.GA und IFA.HO kontrolliert, da sie am höchsten (mit -0.57 und 0.56 bzw. mit -0.53 und 0.57) mit den Variablen IFK.01 und IFK.1P korrelieren. Für IFK.01 sind dann keine Korrelationen mehr zu beobachten.

Zusammenfassend wird durch die Berechnung der Partialkorrelationen für die Einzelcodes der Variable IFK deutlich, dass die einzelnen Formattypen an sich kaum Einfluss auf die Itemschwierigkeit besitzen und untereinander stark korrelieren. Das Merkmal „Formattyp" beeinflusst die Itemschwierigkeit jedoch weitaus stärker und trägt auch deutlich zur Varianzaufklärung bei.

Bei der Variable IFA weisen vor allem Items der Formattypen „Richtig-Falsch" und „Halboffen" signifikante Korrelationen mit der Itemschwierigkeit auf und haben mit 10% bzw. 16% einen starken Anteil an der Varianzaufklärung (vgl. Tabelle IV-2.2.1.1d., Anhang G). Im Wesentlichen deckt sich dieses Ergebnis mit den Resultaten der Analysen mit der Variable IFK, da Richtig-Falsch-Items eine Subgruppe des Formats „Geschlossen – Ankreuzen" darstellen und Items, die eine einfache Antwort im Sinne einer 0/1-Kodierung verlangen, i. d. R. vom Format „Halboffen" erfasst werden.

Die um den Einfluss des Formattyps „Richtig – Falsch" (IFA.RF) kontrollierte Korrelation des Codes „Multiple-Choice-Item" (IFA.MC) mit der Itemschwierigkeit beträgt -0.15 und fällt damit höher aus, als die unkontrollierte Korrelation von 0.09. Im Gegensatz dazu geht die Korrelation des Codes „Halboffen" (IFA.HO) gegen null (< -0.01), wenn sie um den Einfluss der Codes IFK.GA, IFK. 01, IFK.1P und IFA.RF kontrolliert wird, alles Variablen, mit denen IFA.HO mit über 0.5 korreliert. Die Partialkorrelation der Variable IFA.RF, die um den Einfluss der Codes IFA.MC und IFA.HO bereinigt ist, beträgt noch -0.24. Die Codes IFA.ZO, IFA.RI und IFA.OI korrelieren insgesamt deutlich geringer mit anderen Variablen als die Codes IFA.MC, IFA.RF und IFA.HO. Dies deutet darauf hin, dass diesen Codes tatsächlich stets etwas jeweils anderes zugrunde

liegt und andere Fähigkeiten von den Itemformaten erfasst werden. Die um den Einfluss des Codes IFK.GK bereinigten Partialkorrelationen von 0.06 für den Code IFA.ZO und von -0.03 für den Code IFA.RI sowie die Partialkorrelation von -0.01 für den Code IFA.OI (bereinigt um IFK.1P) fallen jedoch insgesamt sehr gering aus.

Die Formate „Multiple-Choice" und „Richtig-Falsch" scheinen demnach einen gewissen Einfluss auf die Itemschwierigkeit zu haben. Für alle anderen Formate gilt jedoch, dass die Partialkorrelationen eher darauf hinweisen, dass analog zur Variable IFK eher die gesamte Variable IFA als „Formatvariable" einen Einfluss auf die Schwierigkeit hat, als die einzelnen Formattypen.

2.2.1.2. Merkmalsgruppe II: Merkmale der Itempräsentation

Variable „Zeitpunkt der Itembearbeitung (ZIB)"

Um auszuschließen, dass die Häufigkeit, mit der der Stimulus vorgespielt wird, einen unbeabsichtigten Einfluss auf die Ergebnisse hat, werden die Analysen nur für die Fälle durchgeführt, in denen der Stimulus einmal vorgespielt wird. Die Itemschwierigkeit ist dann deutlich geringer, wenn die Items erst nach dem Hören gelesen und bearbeitet werden können. Die Präsentation und Aufnahme der Items während des Hörens stellt erhöhte Anforderungen an das Arbeitsgedächtnis, da durch die simultan ablaufenden Tätigkeiten des Zuhörens, Lesens und Ausfüllens (Ankreuzens oder Schreibens) eine erhöhte Beanspruchung des Arbeitsgedächtnisses anfällt. Dieses Ergebnis ist statistisch signifikant und hat mit 6% einen mittleren Anteil an der Varianzaufklärung. (vgl. Tabelle IV-2.2.1.2a., Anhang G) Die um den Einfluss des Itemformats „Einfache Antwort -0/1-Kodierung" (IFK.01) bereinigte Korrelation der Variable ZIB mit der Itemschwierigkeit beläuft sich auf -0.17.

Variable „Position des Items innerhalb der Aufgabe (PIA)"

Die Anordnung der Items erfolgt nicht zufällig, sondern ist von testtheoretischen und didaktischen Entscheidungen geleitet. So wird i. d. R. zu Beginn jeder Aufgabe ein sehr leichtes Item platziert, um den Schülern den Einstieg in die Aufgabe zu erleichtern und sie nicht zu entmutigen. Dieser Grundsatz spiegelt sich auch im Analyseergebnis wieder. Items, die eher zum Schluss einer Aufgabe stehen, verlangen i. d. R. Transferleistungen oder eine Interpretation oder Bewertung des Gehörten. Zum Teil reichen diese Items auch in andere Kompetenzbereiche hinein und haben beispielsweise Überschneidungen zum Kompetenzbereich Sprachreflexion oder Schreiben. Diese zusätzlichen Anforderungen erhöhen die Itemschwierigkeit. Zum Zeitpunkt der Bearbeitung eines Items ab neunter Stelle tritt häufig aber auch Ermüdung bei den Schülern ein. Die Variable PIA trägt nur mit 3% zur Varianzaufklärung bei und korreliert leicht positiv mit der Itemschwierigkeit ($p < 0.10$)(vgl. Tabelle IV-2.2.1.2b., Anhang G). Für die Berechnung der Partialkorrelation wurden die Merkmale „Hintergrundwissen (HGW)", „Position der NI auf Stimulusebene: am Anfang (PST1)" und „Zeitpunkt der Itembearbeitung (ZIB)" kontrolliert. Die Partialkorrelation mit der Itemschwierigkeit beläuft sich dann auf 0.10.

2.2.1.3. Merkmalsgruppe III: Merkmale von MC-Items
Der Einfluss der Variablen dieser Gruppe auf die Itemschwierigkeit wurde nur für MC-Items berechnet, da andere Itemformate nicht von den Variablen betroffen sind.

Variable „Position des Attraktors im MC-Item (PMC)"
Obwohl die Variable PMC nicht zur Varianzaufklärung beiträgt, zeigt das Analyseergebnis ein Absinken der Schwierigkeit von Code zu Code (vgl. Tabelle IV-2.2.1.3a., Anhang G). Ein Attraktor wird häufig erst dann als richtige Antwortmöglichkeit erkannt, wenn er mit den Distraktoren verglichen wird. Aufgrund der Testsituation kann es jedoch vorkommen, dass die Schüler nicht alle Antwortoptionen gründlich lesen, wenn sie die richtige Antwort in einer frühen Position vermuten und deshalb ein sorgfältiger Vergleich der Antwortoptionen nicht erfolgt. Eine weitere Erklärung könnte sein, dass für eine Prüfung des Attraktors aus einer frühen Position mit den anderen Ankreuzoptionen, dessen Aktivhaltung im Arbeitsgedächtnis erfordert. Steht der Attraktor jedoch an einer späteren Position, so ist es denkbar, dass die vorher stehenden Distraktoren abhängig von ihrer Plausibilität bereits verworfen wurden und nicht weiter im Arbeitsgedächtnis aktiv gehalten werden. Da die Variable PMC leicht mit den Variablen ANI ($r = -0.23$) und PST1 ($r = 0.21$) korreliert ($p > 0.1$), wird die um diese beiden Variablen kontrollierte Partialkorrelation berechnet. Sie ist mit -0.07 jedoch nach wie vor zu vernachlässigen.

Variablen „Größte vorkommende Plausibilität der Distraktoren (PDI)" und „Mittlere Plausibilität der Distraktoren (MPD)"
Items, bei denen der attraktivste Distraktor mit Code 4 bewertet wurde, sind für die Schüler am schwierigsten zu lösen (vgl. Tabelle IV-2.2.1.3b., Anhang G). Erwartungsgemäß weisen Items, bei denen alle Distraktoren unter Code 1 fallen, eine sehr geringe Itemschwierigkeit auf. Die Itemschwierigkeiten für Items mit Distraktoren des Codes 2 und 3 unterscheiden sich wie angenommen kaum voneinander. Das Ergebnis ist insgesamt signifikant und deckt mit $Eta^2 = 0.27$ viel Varianz auf. Im Rahmen des Ratingprozesses dürfte es bzgl. Code 1 und 3 Unschärfen gegeben haben und so ist auch das erzielte Ergebnis vorsichtig zu interpretieren. Es erscheint erwartungswidrig, dass offensichtlich abwegige Distraktoraussagen schwieriger sind als wenig plausible. Der Mittelwert beider Gruppen 1 und 3 zusammengefasst, beträgt 0.94. Damit sind plausible Distraktoraussagen, die nicht im Stimulus vorkommen, schwieriger, als nicht plausible bzw. offensichtlich abwegige Distraktoraussagen, die im Stimulus vorkommen.

Ein Ansteigen der Itemschwierigkeit mit zunehmender Code-Größe ist auch bei der Variable MPD zu beobachten. Allerdings fällt das Ergebnis weniger deutlich aus. Dies spricht für die Überlegung, dass nicht die durchschnittliche Attraktivität der Distraktoren vom Attraktor ablenkt, sondern i. d. R. einzelne besonders attraktive Distraktoren für eine Entscheidung den Attraktor nicht anzukreuzen verantwortlich sind. Die Korrelation mit der Itemschwierigkeit weist insbesondere bei der Variable PDI auf einen Einfluss dieses Merkmals auf die Schwierigkeit hin. Die Partialkorrelationen liegen sogar noch höher bei 0.53 für die Variable PDI und -0.41 für die Variable MPD. Kontrolliert wurden neben dem Einfluss der jeweils anderen Variable (PDI bzw. MPD) noch die Merkmale „Anforderungsbereich Items (AFB)" und „Itemschwierigkeit (SEA)". Dabei ist plausibel, dass zwischen der Variable PDI und der Schwierigkeit ein positiver Zusammenhang besteht, da angenommen wird, dass der attraktivste Distraktor von

der richtigen Lösung ablenkt. Nicht erwartungsgemäß ist jedoch der negative Zusammenhang der Variable MPD mit der Schwierigkeit. Der Grund dafür könnte darin liegen, dass eine hohe gemittelte Plausibilität nicht unbedingt auf eine hohe Plausibilität eines einzelnen Distraktors zurückzuführen sein muss, sondern auch durch mehrere mittlere Werte erreicht worden sein kann. Bei Items, für die dieser Fall zutrifft, kann deshalb u. U. trotz höherer mittlerer Plausibilität der Distraktoren der Attraktor ohne Schwierigkeiten zu erkennen sein. Insgesamt eignen sich jedoch beide Variablen sehr gut zur Erklärung der Itemschwierigkeit.

In einem zweiten Schritt wird die absolute Raterübereinstimmung der Einschätzungen von Rater 1 mit Rater 2 ermittelt (vgl. Tabelle IV-2.2.1.3c., Anhang G). Die absolute Raterübereinstimmung wird berechnet, indem die einzelnen Übereinstimmungswerte zwischen den beiden Ratern (grau hinterlegt) addiert werden. Die absolute Übereinstimmung der beiden Rater (Rater 1 und Rater 2) beläuft sich auf 38.7%. Dieser Wert ist relativ gering, was auf die Vierstufigkeit des Codes zurückgeführt werden kann. Codes, die in mehr als zwei Kategorien eingeschätzt werden müssen, führen i. d. R. zu geringeren Raterübereinstimmungen als dichotome Codes. Ein Vergleich mit einer IQB-Studie aus dem Bereich Englisch zeigt, dass bei derart komplexen Ratings meist keine höheren Übereinstimmungen zu erwarten sind. (vgl. Hartmann, 2008)

Eine Regressionsanalyse der relativen Distraktorwahlhäufigkeit für die beiden Variablen PDI und MPD ergab das erwartungskonforme Ergebnis, dass mit steigender Plausibilität alternativer Distraktoren die Distraktorwahlhäufigkeit eines betrachteten Distraktors sinkt. Insgesamt wurde eine Varianzaufklärung von 18% erreicht. Weitere Regressionsanalysen zeigten, dass die Erhöhung der abhängigen Variable „Relative Distraktorhäufigkeit (MSA)“ im Regressionsmodell um eine Einheit, eine Änderung der relativen Distraktorhäufigkeit um B stattfindet. Dieses Ergebnis kann so gedeutet werden, dass ein Distraktor umso häufiger gewählt wird, je plausibler er ist. Je attraktiver die anderen Distraktoren sind, desto weniger häufig wird ein Distraktor gewählt (vgl. Tabelle IV-2.2.1.3d., Anhang G).

2.2.1.4. Merkmalsgruppe IV: Kognitive Anforderungen der Items

Variable „Anforderungsbereich (AFB)“

Bei der Variable AFB fallen die Werte für Code 2 nicht erwartungskonform aus. Obwohl es in jedem Anforderungsbereich sowohl leichte als auch schwierigere Items gibt, wäre zu erwarten gewesen, dass die mittlere Itemschwierigkeit von Anforderungsbereich II unter der des Anforderungsbereichs III liegt. Für den Anforderungsbereich III liegen jedoch auch nur 70 Fälle vor im Gegensatz zu mindestens der doppelten Menge bei den Codes 1 und 2 (vgl. Tabelle IV-2.2.1.4a., Anhang G). Bei der Aufgabenentwicklung zeigte sich wiederholt, dass gerade die Erstellung von Items zum Anforderungsbereich III sehr schwierig fiel. Eine mögliche Erklärung für das vorliegende Ergebnis könnte also sein, dass die entwickelten Items insgesamt nur einen recht einfachen Teilbereich der Dimension „Reflektieren und Beurteilen“ abbilden. Die Variable AFB korreliert mit -0.35 mit der Variable „Mittlere Plausibilität der Distraktoren“ und der mit 0.35 mit der Variable „Konkretheit der NI (33-stufig)“. Die um diese beiden Variablen kontrollierte Partialkorrelation beträgt 0.09.

Variable „Geprüfter Standard (BS)“
Zunächst wurden die Ratings zur Variable BS insofern korrigiert, als zum Teil ein Haupt- und mehrere Nebenstandards von den Aufgabenentwicklern für jedes Item angegeben wurden. Nachträglich wurden die Angaben der Aufgabenentwickler überprüft und nur der jeweils überwiegende Standard für die Analysen miteinbezogen. Dabei ergab sich eine Verteilung wie in Tabelle IV-2.2.1.4b., Anhang G dargestellt. Dabei fällt auf, dass einige Standards nur sehr selten vergeben wurden und zum Teil auch Standards aus anderen Kompetenzbereichen (z. B. Schreiben oder Sprachreflexion) vorkommen. Für die weiteren Analysen wurden deshalb nur die grau unterlegten Standards BS113, BS141, BS142 und BS143 verwendet.

Die ausgewählten Standards verteilten sich wie in Tabelle IV-2.2.1.4c., Anhang G dargestellt auf die Items. Die Variable BS trägt 3% zur Varianzaufklärung bei. Von den einzelnen Codes erweist sich insbesondere Code BS113 als besonders schwierig. Die erhöhte Schwierigkeit liegt in den benötigten kognitiven Fähigkeiten. Etwas zu unterscheiden und anzuwenden sowie ggf. eine Wirkung zu beurteilen ist i. d. R. schwieriger als nur Informationen aufzunehmen und diese wiederzugeben. Zu beachten ist jedoch auch die eingeschränkte Aussagekraft dieses Werts aufgrund des recht kleinen Stichprobenumfangs.

Auf Codeebene wurden zusätzliche Analysen durchgeführt. Die Fallzahlen für jeden Code unterscheiden sich hier von der Codebelegung bei den zusammengefassten Kategorien, da nur die absolute Belegung der Codes für die Berechnung herangezogen wurde. In einem zweiten Schritt wurden die Analysen für die Einzelcodes wiederholt. Die Ergebnisse werden in Tabelle IV-2.2.1.4d., Anhang G zusammengefasst. Dabei zeigt sich, dass insbesondere die Codes BS141, BS142 und BS143 kaum mit der Itemschwierigkeit korrelieren, dafür aber stark untereinander. Der Code BS113 korreliert mit 0.15 etwas stärker mit der Itemschwierigkeit und weist keine nennenswerten Zusammenhänge mit den anderen drei Codes auf. Items, denen dieser Code zugeteilt wurde, scheinen also etwas anderes als die Items mit den Codes BS141, BS142 und BS143 zu erfassen. Für ihn wurden bei der Berechnung der Partialkorrelation die Itemformatvariablen „Multiple-Choice (IFA.MC)“ und „Zuordnung (IFA.ZO)“ kontrolliert. Die Partialkorrelation liegt bei 0.10. Die um den Code BS142 kontrollierte Partialkorrelationen für die Codes BS141 und BS143 betragen -0.02 und -0.06 und liegen damit in etwa im Bereich der unkontrollierten Korrelation der Codes mit der Itemschwierigkeit. Die Partialkorrelation für Code BS142 stieg nach Kontrolle der Codes BS141 und BS143 jedoch auf -0.13 und liegt damit etwas höher als der unkontrollierte Wert.

Variable „Anzahl der benötigten NI pro Item (ANI)“
An insgesamt 383 Items wurden die Mittelwerte für die Variable ANI untersucht. Dabei zeigt sich, dass etwa 68% der Itemschwierigkeiten im Bereich -2 bis 0 liegen. Unter der Annahme, dass eine Normalverteilung der Itemschwierigkeiten besteht, liegen etwa 68% der Schwierigkeiten im Bereich Mittelwert ± 1 Standardabweichung. Bei 2 Standardabweichungen ginge man per Definition von 95% aus. Die Mittelwerte in den ANI-Stufen unterscheiden sich nicht signifikant voneinander, es fällt jedoch auf, dass Items mit Code 0 besonders einfach sind (vgl. Tabelle IV-2.2.1.4e., Anhang G). Auch Items, bei denen die NI nur einmal auftritt (Code 1) oder global erfasst werden (Code 4), sind einfacher als Items, zu deren Lösung mehrere Infor-

mationen benötigt werden. Eine geringe praktische Relevanz des Effekts ist durch den Eta²-Wert von 0.02 gegeben. Daraus folgt, dass 2% der Itemschwierigkeiten durch die Variable ANI aufgeklärt werden. Da ANI mit den Variablen „Anforderungsbereich (AFB)" (r = 0.33) und „Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben (BS142)" (r = -0.30) korreliert, wird die Partialkorrelation um diese beiden Codes kontrolliert berechnet. Sie fällt mit 0.13 höher als der unkontrollierte Wert aus.

Variable „Auftretenshäufigkeit der NI (ARN)"

Der Code ARN hat mit 1% über alle Codes hinweg einen geringen Effekt auf die Varianzaufklärung und korreliert mit r = 0.03 mit der Itemschwierigkeit (vgl. Tabelle IV-2.2.1.4f., Anhang G). Um die Bedeutung der einzelnen Codestufen zu erfassen, werden deshalb die absoluten Belegungen für die einzelnen Codes berechnet. Dabei wurden die Codes 0 und 4 zusammengefasst, da beide Kategorien gemeinsam haben, dass die NI nicht konkret im Stimulus aufgezeigt werden kann. Der Code lautet nun: „Code 0/4: Redundanz der NI lässt sich nicht bestimmen, da es sich bei der NI um eine globale Einschätzung des Beitrags handelt oder die NI nicht im Stimulus vorkommt." Die angegebenen p- und d-Werte beziehen sich also auf diesen zusammengefassten Code 0/4. Die Ergebnisse der Analysen mit den Einzelcodes sind in Tabelle IV-2.2.1.4g., Anhang G dargestellt. Es wird deutlich, dass die Itemschwierigkeit sinkt, je häufiger eine NI im Stimulus auftritt, je redundanter sie also ist. Die Schwierigkeit für Items mit Code 0/4 steigt jedoch wieder etwas an. Dies erscheint plausibel: Einerseits taucht die NI in diesen Fällen i. d. R. mehrmals implizit auf, kann damit jedoch keiner genauen Stelle im Stimulus zugeordnet werden und erfordert häufig eine gewisse Abstraktionsleistung. Die Ergebnisse sind jedoch nur für Code 1 signifikant und die Einzelcodes leisten so gut wie keinen Beitrag zur Varianzaufklärung. Auch die Korrelation mit der Itemschwierigkeit ist in allen vier Fällen zu vernachlässigen.

In einem zweiten Schritt werden die Analysen für die Items wiederholt, bei denen gemäß der Variable ANI nur eine NI benötigt wird, um einen möglichst unverfälschten Einfluss der NI auf die Itemschwierigkeit zu prüfen (vgl. Tabelle IV-2.2.1.4h., Anhang G). Die Ergebnisse des zweiten Analyseschritts replizieren die der ersten Analyse, jedoch in deutlicherer Form. Die Resultate sind für Code 1 und Code 3 signifikant und haben einen kleinen Anteil an der Varianzaufklärung. Sie korrelieren nun auch etwas mit der Itemschwierigkeit. Generell lässt sich sagen, dass Items dann einfacher sind, wenn die NI häufiger als einmal auftritt.

Die Partialkorrelationen der Codes weisen auf einen nach wie vor schwachen Einfluss des Merkmals auf die Schwierigkeit. Kontrolliert wurden die Merkmale wie in Tabelle IV-2.2.1.4i., Anhang G zusammengefasst.

Dieses Ergebnis wird auch durch Tabelle IV-2.2.1.4j., Anhang G bestätigt. Die Variable ARN hat bei Fokussierung auf Items, für die nur eine NI benötigt wird, eine Varianzaufklärung von 4% und zeigt ein insgesamt signifikantes Ergebnis: Je häufiger die benötigte Information zur Beantwortung eines Items im Stimulus vorkommt, desto einfacher ist das Item.

Variable „Position der NI im Stimulus (PST)“
Ein Blick auf die Häufigkeitsverteilung der einzelnen Codes zeigt, dass in 192 Fällen eine Einschätzung für die Position der NI im Stimulus vorgenommen wurde, in insgesamt 50 Fällen jedoch zwei oder sogar drei Codes vergeben wurden. In zwei Fällen wurde der NI keine Position im Stimulus zugewiesen. Es ist naheliegend, die Analysen nur für die Fälle durchzuführen, in denen die NI genau einer Position im Stimulus zugewiesen wurde. Ferner wurden nur Fälle berücksichtigt, in denen für die Lösung eines Items genau eine NI benötigt wird (ANI = 1). Eine derartige Anpassung des Datensatzes führt zu nur mehr 150 Fällen. Die Ergebnisse fallen nun recht deutlich aus. Am leichtesten sind die Items, zu denen die gesuchte Information in der Mitte des Stimulus steht, die Items zu Informationen am Anfang oder am Ende der Stimuli sind in etwa gleich schwierig (vgl. Tabelle IV-2.2.1.4k., Anhang G). Dieses Ergebnis ist insofern erwartungswidrig, als Versuche mit Wortlisten aus der kognitiven Psychologie zeigten (vgl. Kapitel *5.3.5.2. Position der NI*), dass Informationen am Anfang und am Ende dieser Listen aufgrund ihrer markanten Position besser erinnert werden. Im Fall der IQB-Höraufgaben könnte dieses Ergebnis jedoch dadurch begründet werden, dass die Schüler eine gewisse Einhörzeit benötigen und zum Schluss der Stimuli bereits ein Ermüdungseffekt eintritt. Die Ergebnisse zu allen drei Variablen sind jedoch nicht signifikant und tragen praktisch nicht zur Varianzaufklärung bei.

Die drei Codes PST1, PST2 und PST 3 korrelieren stark miteinander (vgl. Tabelle IV-2.2.1.4l., Anhang G). Die Berechnung von Partialkorrelationen, bei denen immer jeweils die anderen beiden Codes kontrolliert werden, ergibt, dass die Partialkorrelationen für alle drei Codes etwas höher ausfallen, als die unkontrollierten Korrelationen mit der Itemschwierigkeit. Insgesamt sind die Werte aber nach wie vor recht gering. Dies ist ein Hinweis dafür, dass der Auftretensort der NI im Stimulus einen leichten Effekt auf die Itemschwierigkeit hat.

Variable „Wortzahl der NI (WNI)“
Die Ergebnisse der Mittelwertsanalysen für die Variable WNI werden in Tabelle IV-2.2.1.4m., Anhang G dargestellt. Die Itemschwierigkeit nimmt mit zunehmender Länge der zur Lösung des Items benötigten Information zu. Dies kann mit den erhöhten Ansprüchen an die Arbeitsgedächtniskapazität zusammenhängen.

Auch bei der Variable WNI werden in einem zweiten Analyseschritt nur die Fälle zu betrachtet, bei denen nur eine NI benötigt wird (ANI = 1). So soll der Einfluss der Variable WNI auf die Itemschwierigkeit möglichst unverfälscht beobachten werden können. (vgl. Tabelle IV-2.2.1.4n., Anhang G) Tatsächlich tritt der Einfluss der Variable WNI auf die Itemschwierigkeit jetzt noch deutlicher zu Tage: Je kürzer die NI ist, d. h. je weniger Wörter sie enthält, desto einfacher sind die entsprechenden Items. Die Varianzaufklärung liegt allerdings noch immer bei nur 1% und das Ergebnis ist nach wie vor nicht signifikant. Die um die Variable „Mittlere Plausibilität der Distraktoren (MPD)“ kontrollierte Partialkorrelation beläuft sich auf 0.11.

Variable „Konkretheit der NI – 33-stufig (TCO)" und Variable „ Konkretheit der NI – 5-stufig (TOR)"
Die TORI-Stufen beschreiben die Entnahme von Informationen unterschiedlichen Konkretheits- bzw. Abstraktionsgrades in aufsteigender Schwierigkeit. Demnach sollte es einfacher sein Items zu beantworten, die nach Personen oder Orten fragen, als Items richtig zu lösen, die eine Gesamteinschätzung des Beitrags erfordern. Die Mittelwerte der Itemschwierigkeiten stützen diese Annahme jedoch bedingt. Im Großen und Ganzen ist zwar eine Staffelung der Itemschwierigkeiten zu erkennen, jedoch tauchen auch immer wieder Werte auf, die deutlich über oder unter (± 0.5) dem erwarteten Stufenwert (Mittelwert der zur Stufe gehörigen Codes) liegen. Diese Werte sind fett markiert. Zum Teil sind diese Werte durch hohe Streuungen zu erklären, die im Fall des Codes 7 (Typen/Arten) bei 2 Standardabweichungen liegt. In anderen Fällen, wie bei Code 26 (Ähnlichkeiten) ist die Standardabweichung jedoch nicht auffällig. Eine weitere Rolle spielt die geringe Fallzahlbelegung der Codes. Einzelne Items haben so einen starken Einfluss auf den Gesamtwert. Die Analyse wurde zunächst für den gesamten Datensatz durchgeführt und in einem zweiten Schritt nur für die Items, die genau eine Information verlangen (ANI = 1) (vgl. Tabelle IV-2.2.1.4o., Anhang G).

Ein Vergleich der Ergebnisse ergibt sehr unterschiedliche Werte für die einzelnen Codes bei insgesamt höherer Varianzaufklärung der Variable TCO im angepassten Datensatz. Mit 25% Varianzaufklärung beschreibt die Variable TCO mit den differenzierten TORI-Stufen einen starken Effekt. Im Vergleich dazu leistet die Variable TOR mit 7% einen relativ geringen Beitrag zur Varianzaufklärung. (vgl. Tabelle IV-2.2.1.4p., Anhang G) Insgesamt haben beide Variablen diesbezüglich jedoch eine relativ hohe Aussagekraft. Die Ergebnisse der Analysen mit allen Items wiederholen sich im Wesentlichen mit dem angepassten Datensatz, fallen dabei jedoch etwas deutlicher aus. Am leichtesten sind Items, für die weniger konkrete Informationen wie Mengenangaben, Zeiten, Attribute, Aktionen/Versuche oder Angaben zur Art und Weise benötigt werden. Am schwierigsten sind Items, die nach hoch abstrakten Informationen wie dem Thema oder Äquivalenten/Paraphrasen fragen. Die mittlere Schwierigkeit der Codestufen 1, 3 und 4 entspricht in etwa dem Mittel der Itemschwierigkeit der verwendeten Items. Da die Variablen TCO und TOR das gleiche Konstrukt mit unterschiedlicher Genauigkeit erfassen, korrelieren sie erwartungsgemäß sehr hoch miteinander. Bei der Berechnung der Partialkorrelation für beide Variablen mit der Itemschwierigkeit wird demnach die jeweils andere Variable kontrolliert. Ferner findet in beiden Fällen eine Kontrolle der Variablen „Anforderungsbereich (AFB)" und „Anzahl der benötigten NI pro Item (ANI)" statt. Die Partialkorrelation lässt sich für die Variablen TCO und TOR dann praktisch nicht mehr erfassen.

Variable „Typ der NI (TNI)"
Auch bei der Variable TNI können Mehrfachbelegungen von Codes pro Item auftreten. In 358 Fällen wurde ein Code vergeben, in jeweils 13 Fällen wurden jedoch zwei oder sogar drei Codes für eine NI vergeben. Dies ist darin begründet, dass – wie bereits mit den Variablen ARN und ANI gezeigt wurde – für die Beantwortung eines Items mehrere Informationen aus dem Stimulus notwendig sind. Die Mehrfachbelegungen treten in insgesamt 26 Fällen auf. Zunächst wird der Einfluss der Variable TNI auf die Itemschwierigkeit insgesamt überprüft. Dabei werden nur für die Fälle berücksichtigt, für die genau ein Code vergeben wurde. Es zeigt sich, dass Items, bei denen nach der Meinung des Zuhörers gefragt wird, oder bei

denen ein Sprecher identifiziert werden muss, am leichtesten sind (fett markiert). Items, bei denen eine Schlussfolgerung verlangt wird oder die nach sprachlichen Mitteln fragen, sind am schwierigsten (grau hinterlegt). Die Variable TNI ist mit 4% an der Varianzaufklärung beteiligt, korreliert jedoch kaum mit der Itemschwierigkeit. Tabelle IV-2.2.1.4q., Anhang G fasst diese Ergebnisse zusammen.

Die Analysen werden wiederholt, für die Fälle bei denen vom Item genau eine Information verlangt wird (ANI = 1) sowie für die nur ein Code der Variable TNI vergeben wurde. Die Analysen mit dem angepassten Datensatz erfolgen nur auf der Grundlage von 203 Items im Gegensatz zum kompletten Datensatz mit 383 Items. Dabei zeigt sich, dass für einige Codes (z. B. Code 3 oder Code 7) die Fallzahlbelegung so gering ist, dass das Ergebnis von einzelnen Items abhängt und verallgemeinernde Aussagen nicht gemacht werden sollten. Die Ergebnisse der vorhergehenden Analyse bestätigen sich jedoch weitgehend (vgl. Tabelle IV-2.2.1.4r., Anhang G). Die Items bei denen nach der Meinung des Zuhörers bzw. der Stimmung oder der Einstellung des Sprechers fragen, sind am leichtesten (fett markiert), wohingegen die Items zum Genre des Stimulus oder zu Geräuschen oder paraverbalen Informationen am schwierigsten sind (grau hinterlegt). Die Variable TNI deckt bei dem gefilterten Datensatz 5% der Varianz auf, das Ergebnis ist jedoch nach wie vor nicht statistisch signifikant. Die Korrelation mit der Itemschwierigkeit beläuft sich hier nur noch auf 0.03 und sinkt weiter auf 0.02, berechnet man die Partialkorrelation indem man die am höchsten mit TNI korrelierende Variable „Position des Items innerhalb der Aufgabe (PIA)“ kontrolliert.

Um die Bedeutung der einzelnen Codes zu erfassen, werden für sie separat Analysen durchgeführt. Berücksichtigt werden die Fälle, bei denen vom Item nur genau eine Information verlangt wird (ANI = 1) sowie für die nur ein Code der Variable TNI vergeben wurde (vgl. Tabelle IV-2.2.1.4s., Anhang G). Insgesamt liegen bei den Ergebnissen mit dem angepassten Datensatz keine signifikanten p-Werte vor, Cohen's d weist jedoch in den meisten Fällen auf hohe praktische Relevanz hin. Eine Varianzaufklärung wird von den Variablen höchstens in sehr geringem Umfang geleistet. Die Tendenzen des Einflusses der Variablen auf die Itemschwierigkeit sind mit dem angepassten Datensatz deutlich zu beobachten, allerdings führen hier teilweise sehr geringe Codebelegungen dazu, dass die Ergebnisse sehr vorsichtig interpretiert werden müssen.

Die grau markierten Merkmale führen dazu, dass die entsprechenden Items leichter werden. Wird als zur Itembeantwortung notwendige Information also ein Detail, die Meinung des Zuhörers, die Stimmung bzw. Einstellung des Sprechers, die Angabe des kommunikativen Zwecks, der Funktion bzw. der Wirkung des Stimulus oder die Identifikation eines Sprechers verlangt, so sind die Items nicht nur einfacher als Items, von denen diese Informationen nicht gefordert wird, sondern diese Items sind auch einfacher als die mittlere Itemschwierigkeit der analysierten Itemstichprobe.

Müssen die Schüler jedoch eine Leitidee bzw. die Kernaussage des Stimulus erfassen, eine Schlussfolgerung ziehen, das Genre des Stimulus erfassen oder Fragen nach Geräuschen bzw. paraverbalen Merkmalen oder sprachlichen Mitteln beantworten, so sind die entsprechenden

Items schwieriger als im umgekehrten Fall und liegen in fast allen Fällen auch deutlich über der mittleren Itemschwierigkeit der Stichprobe. Die Merkmale, die mit einer höheren Itemschwierigkeit zusammenhängen, erscheinen sehr plausibel. Eine globale Aussage oder ein Genre zu erfassen ist schwieriger als auf bestimmte Details zu achten, da im ersten Fall das mentale Modell des Stimulus korrekt gebildet werden musste. Auch die Fähigkeit Schlussfolgerungen zu ziehen ist relativ elaboriert. Das Ergebnis entspricht dem Resultat der Analysen mit der Variable Schlussfolgerungen/Inferenzen (SFI), bei denen untersucht wird, ob sich die Stimulusschwierigkeit verändert, wenn zum Verständnis eines Stimulus viele Schlussfolgerungen gezogen werden müssen. Es wäre zu untersuchen, ob die Variable TNI5 (Schlussfolgerung) mit der Variable AFB, und zwar dem Code AFBIII (Reflektieren und Beurteilen) korreliert, da auch hier vom Item benötigte Transferleistungen erfasst werden. Fragen nach Geräuschen bzw. paraverbalen Merkmalen erfordern eine spezifische Aufmerksamkeitslenkung beim Zuhören. Wenn den Schülern diese Art Frage nicht vertraut ist, könnte gerade die Aufmerksamkeitslenkung weg vom Inhalt des Stimulus hin auf seine Machart die Schwierigkeit der Items begründen. Fragen nach sprachlichen Mitteln hingegen setzen häufig Metawissen voraus und können i. d. R. nicht beantwortet werden, wenn dieses Wissen zum Testzeitpunkt nicht erworben wurde.

Da die Art der vom Item geforderten Information auch den im Kompetenzstufenmodell beschriebenen kognitiven Fähigkeiten der Schüler entspricht, wird eine Einordnung der die dargestellten Merkmale tragenden Items ins Modell versucht (vgl. Tabelle IV-2.2.1.4t., Anhang G). In Abstimmung der einzelnen Codes mit ihrem Auftreten im Kompetenzstufenmodell (vgl. Kapitel *4.4.2. Kompetenzstufenmodelle Zuhören*) ist zwar eine gewisse Staffelung der mittleren Itemschwierigkeiten zu erkennen, jedoch fallen auch immer wieder Werte auf, die entgegen ihrer Einordnung auf bestimmten Kompetenzstufen deutlich einfacher oder schwieriger (± 0.5) sind (fett markiert). Dies lässt sich zumindest teilweise mit den geringen Fallzahlen erklären. Einzelne Items könnten das Gesamtbild hier deutlich verzerren.

Variable „Hintergrundwissen (HGW)"

Die Variable HGW gibt an, ob Hintergrundwissen zum Verständnis der Items benötigt wird. Items, bei denen Hintergrundwissen eine Rolle spielt, sind erwartungsgemäß deutlich schwieriger als Items, die ohne Hintergrundwissen verstanden werden (vgl. Tabelle IV-2.2.1.4u., Anhang G). Bei der Variable HGW ist das Ergebnis zwar statistisch signifikant, jedoch mit 1% Varianzaufklärung wenig aussagekräftig. Hintergrundwissen scheint also für die Bearbeitung der IQB-Items also keine Rolle zu spielen. Dies ist ein erfreuliches Ergebnis, da die IQB-Aufgaben Kompetenzen erfassen sollen, für die Hintergrundwissen keine Rolle spielt. Kontrolliert man bei der Berechnung von Partialkorrelationen den Einfluss der am höchsten mit der Variable HGW korrelierenden Merkmale „Position des Items innerhalb der Aufgabe (PIA)" (0.28) und „Itemformat: Zuordnung (IFA.ZO)" (0.23), so erhält man einen Wert von 0.02. Die Variable HGW scheint also praktisch keinen Einfluss auf die Itemschwierigkeit zu haben.

2.2.1.5. Merkmalsgruppe V: Globalurteil

Variable „Itemschwierigkeit (SEA)“
Zu jedem Item liegt eine während der Itemerstellung gewonnene Einschätzung der Aufgabenentwickler hinsichtlich der vermuteten Itemschwierigkeit vor. Die Aufgabenentwickler wurden gebeten anzugeben, wie schwierig die einzelnen Items einer Aufgabe wahrscheinlich für die Schüler sein werden. Dabei ist davon auszugehen, dass sich die Lehrkräfte nicht ausschließlich an den Vorgaben für den Hauptschulabschluss bzw. den Mittleren Schulabschluss orientierten, sondern insbesondere die eigenen Schülergruppen im Blick hatten. Diese Einschätzung der Aufgabenentwickler entspricht recht gut den empirischen Itemschwierigkeiten (vgl. Tabelle IV-2.2.1.5a., Anhang G), wobei die Korrelation insgesamt sehr schwach ausfällt. Die Variable SEA korreliert mit den Variablen PDI „Größte vorkommende Plausibilität der Distraktoren“, MPD „Mittlere Plausibilität der Distraktoren“ und AFB „Anforderungsbereich“. Auffallend dabei ist, dass alles drei Variablen sind, die auf Einschätzungen beruhen. Kontrolliert man den Einfluss dieser Variablen, erhält man eine Partialkorrelation der Variable SEA mit der Itemschwierigkeit von 0.20. Dieser Wert liegt deutlich über der unkontrollierten Korrelation der Variable mit der Itemschwierigkeit.

2.3. Zusammenfassung der Zusammenhangsanalysen

2.3.1. Zusammenfassung Stimulusmerkmale
Neben den Mittelwertsvergleichen wurden für dichotome Stimulusmerkmale auch Korrelationen berechnet. Dort wo es sich anbot, wurden außerdem Partialkorrelationen berechnet, um den Zusammenhang weiterer Merkmale mit dem jeweiligen Merkmal und der Schwierigkeit auszuschließen. Die Merkmale, die bei der Berechnung der Partialkorrelationen kontrolliert werden, korrelieren stark mit den jeweiligen Variablen, für die ein Zusammenhang mit der Schwierigkeit berechnet wird. In Tabelle IV-2.3.1a. werden die tatsächlichen Effekte noch einmal zusammenfassend dargestellt.

Tabelle IV-2.3.1a.: Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale Merkmalsgruppe I „Komplexität des Wortschatzes und sprachliche Merkmale“

Variable	Stimulusmerkmale	Zusammenhang mit der Schwierigkeit	
		angenommen	*tatsächlich*[15]
Merkmalsgruppe I: Komplexität des Wortschatzes und sprachliche Merkmale			
WLS	Wortlänge (< 5 Buchstaben; ⩾ 5 Buchstaben)	+	+
PLW	Lange Wörter (< 20%; 20% - 25%; > 25%)	+	+
PMW	Mehrsilbige Wörter (< 5%; ⩾ 5%	+	+
GWS	Worthäufigkeit (< 60%; ⩾ 60%))	-	+
STR2	Ellipsen (⩽ 15%; > 15%)	+	+
STR3	Adjazenzstrukturen (⩽ 1%; > 1%)	+	+
STR4	Anakoluthe (⩽ 1%; > 1%)	+	+
RHE2	Uneigentliches Sprechen (< 1%; ⩾ 1%)	+	+
WIE	Wiederaufnahmen (< 10%; 10% - 20%; > 20%)	+	+
NEG	Negationen (⩽ 2%; > 2%)	+	+
IWH	Inhaltswörter (< 45%; 45% - 50%; > 50%)	+	-
STR1	Referenz-Aussage-Strukturen (⩽ 1%; > 1%)	-	-
STR5	Nähezeichen (⩽ 1%; > 1%)	-	-
STR6	Verberststellung (⩽ 1%; > 1%)	-	-
REF	Referenzen (< 20; 20 – 50; > 50)	+	-
PEW	Einsilbige Wörter (⩽ 55%; > 55%)	-	-
RHE1	Bildliche Darstellungsformen (< 1%; ⩾ 1%)	+	**0**
RHE5	Neudeutsch/Anglizismen (< 1%; ⩾ 1%)	+	**0**
RHE6	Jugendsprache/Umgangssprache (< 1%; ⩾ 1%)	-	**0**
DEI	Deixis (⩽ 10%; > 10%)	+	**0**
SUB	Substantive/Eigennamen/Appellative (⩽ 20%; > 20%)	+	**0**
VER	Verben (⩽ 15%; > 15%)	-	**0**

Anmerkungen: +: größere Anteile des Merkmals erhöhen die Schwierigkeit;
-: größere Anteile des Merkmals senken die Schwierigkeit;
0: es konnte kein Effekt festgestellt werden

In der Merkmalsgruppe I „Komplexität des Wortschatzes und sprachliche Merkmale“ zeigten die Merkmale folgenden Einfluss auf die Schwierigkeit:

Bei den Variablen, die die Stimuli in ihrer Länge zu erfassen versuchen, lässt sich i. d. R. ein deutlicher Effekt erkennen, der verallgemeinert in „je länger, desto schwieriger“ zusammengefasst werden kann. Im Einzelnen haben die Variablen den folgenden Einfluss: Je größer die Anteile der Variablen „Lange Wörter (PLW)“, „Mehrsilbige Wörter (PMW)“ sind, desto schwieriger werden die Items bzw. wird die gesamte Aufgabe. Auch ein höherer Anteil an Wörtern im Stimulus, die länger als fünf Buchstaben sind (Variable „Wortlänge (WLS)“) hängt mit schwierigeren Items bzw. Stimuli zusammen. Je höher der Anteil der Variable „Einsilbige Wörter (PEW)“ ist, desto einfacher werden die Items bzw. wird die gesamte Aufgabe.

15 u. U. bei Analysen mit Teilstichproben

Bei den Variablen, die den Wortschatz der Stimuli abbilden, gab es nur sehr schwach ausgeprägte Ergebnisse, die im Fall der Variablen „Worthäufigkeit (GWS)“ und „Inhaltswörter (IWH)“ auch erwartungswidrig ausfielen. Die Variable GWS zeigt nach Berechnung der Partialkorrelation einen positiven Zusammenhang mit der Stimulusschwierigkeit. Ein höherer Anteil an Inhaltswörtern hängt nach Berechnung der Partialkorrelation negativ mit der Stimulusschwierigkeit zusammen. Dieser Effekt kann jedoch auf den Einfluss einzelner Aufgaben beruhen. Die Ausprägung der Variable „Substantive/Eigennamen/Appellative (SUB)“ hat nur einen geringen Einfluss auf die Schwierigkeit. Bei den Analysen mit dem Gesamtdatensatz zeigen sich auf Aufgabenebene Korrelationen mit der Schwierigkeit, die bei Kontrolle weiterer Variablen jedoch stark abnimmt. Auch die Wortart „Verben (VER)“ scheint das Verständnis von Stimuli und Items nicht zu beeinflussen. Die Variable „Negationen (NEG)“ zeigt erst bei der Berechnung der Partialkorrelation einen positiven Zusammenhang mit der Schwierigkeit. Da beim Aufbau mentaler Repräsentationen zunächst die positive Aussage der Proposition verarbeitet wird, bevor sie dann negiert wird, ist dieses Ergebnis plausibel.

Bei den Variablen, die die linguistische Struktur der Stimuli erfassen, sind kaum Zusammenhänge zu beobachten. Sie sind nach gezielten Analysen mit Teildatensätzen (z. B. nur für Stimuli gleicher Länge) und nach Kontrolle bestimmter hoch mit diesen Variablen korrelieren den Merkmalen jedoch deutlich stärker. Der Effekt fällt in allen Fällen auf Aufgabenebene höher als auf Itemebene auf. Positiv hängen mit der Item- bzw. der Stimulusschwierigkeit die folgenden Variablen zusammen: „Ellipsen (STR2)“, „Anakoluthe (STR4)“ und „Adjazenzstrukturen (STR3)“. Das Ergebnis ist in allen drei Fällen plausibel. Je mehr verkürzte oder abgebrochene Sätze bzw. Äußerungen im Stimulus vorstellen, desto mehr Leerstellen muss der Zuhörer füllen, desto mehr Arbeitsgedächtniskapazität wird benötigt und desto schwieriger sind entsprechend Items und Stimuli. Schwierig ist auch das Verständnis von „Adjazenzstrukturen (STR3)“, da sie die Aktivhaltung früherer Strukturen im Arbeitsgedächtnis erfordern. Insgesamt wurden nur sehr wenige Adjazenzstrukturen bei den Stimuli geratet. Es handelt sich dabei um typische Strukturmuster gesprochener Sprache. Es ist nicht auszuschließen, dass der Effekt dieser Variable nicht dem Merkmal, sondern den Aufgaben zu den entsprechenden Stimuli geschuldet ist, da die Fallzahlbelegung auf Aufgabenebene sehr gering ausfällt.

Ein negativer Zusammenhang zwischen der Schwierigkeit auf Item- bzw. Aufgabenebene besteht mit den folgenden Variablen: „Nähezeichen (STR5)“, „Verberststellung (STR6)“ und „Referenz-Aussage-Strukturen (STR1)“. Je größer die Ausprägungen dieser Merkmale in den Stimuli sind, desto einfacher werden die Items bzw. der Stimulus. Es handelt sich dabei um Besonderheiten gesprochener Sprache, die in der schriftsprachlichen Kommunikation, wohl aus stilistischen Gründen bzw. aufgrund einer klarer geregelten Kommunikationsabfolge kaum auftreten. Es ist anzunehmen, dass Stimuli mit einem hohen Anteil dieser Variablen insgesamt einfacher sind.

Alle vier Variablen zur Erfassung der rhetorischen Mittel eines Stimulus korrelieren auf Item- und auf Aufgabenebene kaum mit der Schwierigkeit. Auch die Berechnung der Partialkorrelation ergibt auf Aufgabenebene für die Variablen „Bildliche Darstellungsformen (RHE1)“, „Neudeutsch/Anglizismen (RHE3)“ und „Jugendsprache/Umgangssprache (RHE4)“ keinen

Zusammenhang mit der Schwierigkeit. Die Variable „Uneigentliches Sprechen (RHE2)" korreliert nach Kontrolle bestimmter hoch mit der Variable korrelierender Merkmale jedoch schwach positiv mit der Stimulusschwierigkeit. Insgesamt ist dieser Befund erwartungswidrig. Zwar zeigt die Variable RHE2 einen positiven Zusammenhang mit der Stimulusschwierigkeit, dieser Zusammenhang fällt aber sehr schwach aus. Es wäre zu erwarten gewesen, dass rhetorische Mittel einen größeren Einfluss auf das Verständnis der Stimuli haben und stark Verständnis erschwerend wirken. Eine Ausnahme wäre für die Variable „Jugendsprache/Umgangssprache (RHE4)" denkbar gewesen, die für die Schüler nicht Verständnis erschwerend wirken muss, da den Schülern eine umgangssprachliche Ebene vertrauter ist, als eine stärker formalisierte Sprachebene.

Bei den Variablen zur Kohärenz der Stimuli zeigt die Variable „Deixis (DEI)" einen schwachen Effekt auf Aufgabenebene auf die Schwierigkeit, der jedoch nach Berechnung der Partialkorrelation noch geringer wird. Deiktische Elemente scheinen für das Verständnis der Stimuli demnach keine Rolle zu spielen. Die Variable „Wiederaufnahmen (WIE)" korreliert auf Itemebene schwach und auf Aufgabenebene etwas stärker mit der Schwierigkeit. Je mehr Wiederaufnahmen in einem Stimulus geratet wurden, desto schwieriger sind die entsprechenden Items bzw. ist der Stimulus. Wiederaufnahmen erfordern vom Zuhörer die Erstellung einer mentalen Repräsentation der Inhalte, die im Arbeitsgedächtnis aktiv gehalten werden müssen. Bei der Wiederaufnahme eines Inhalts müssen dann die korrekten Inhalte aufgerufen werden, da es sonst zu Verständnisschwierigkeiten kommt. Für die Variable „Referenzen (REF)" fällt das Ergebnis gegenläufig aus. Sowohl auf Item- als auch auf Aufgabenebene hängt sie negativ mit der Schwierigkeit zusammen. Ein höherer Anteil an Referenzbezügen hängt also mit eher einfacheren Items bzw. Stimuli zusammen. Stimuli mit vielen Referenzbezügen sind meist auch kohärenter, was sich positiv auf das Verständnis auswirken kann.

Tabelle IV-2.3.1b.: Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale Merkmalsgruppe II „Präsentationsmerkmale"

Variable	Stimulusmerkmale	Zusammenhang mit der Schwierigkeit	
		angenommen	*tatsächlich*[16]
Merkmalsgruppe II: Präsentationsmerkmale			
AWS	Wortzahl (⩽200 Wörter; 201–400 Worte; 401–600 Wörter; ⩾600 Wörter)	+	+
ASP	Anzahl der Sprecher (ein Sprecher; zwei Sprecher; drei Sprecher; mehr als drei Sprecher)	+	+
SGS	Sprechgeschwindigkeit (<2 Wörter pro Sekunde; ⩾2 Wörter pro Sekunde)	+	-
AHO	Anzahl der Stimuluspräsentationen (einmal vorgespielt; zweimal vorgespielt)	-	-
AST	Akzent/Dialekt/Aussprache (Standardsprache; leichter regionaler Dialekt; starker regionaler Dialekt; ausländischer Akzent)	+	0
LST	Länge in Minuten (< 2 min; 2 – 4 min; > 4 min)	+	0

Anmerkungen: +: größere Anteile des Merkmals erhöhen die Schwierigkeit;
- : größere Anteile des Merkmals senken die Schwierigkeit;
0: es konnte kein Effekt festgestellt werden

16 u. U. bei Analysen mit Teilstichproben

In der Merkmalsgruppe II „Präsentationsmerkmale“ wird der Einfluss von Variablen zur Stimuluslänge, zur sprachlichen Qualität der Stimuli sowie zur Häufigkeit der Darbietung geprüft (vgl. Tabelle IV-2.3.1b.). Bei den Variablen zur Stimuluslänge eignet sich das Merkmal „Sprechgeschwindigkeit (SGS)“ am besten dazu, die Schwierigkeit von Items und Stimulus vorherzusagen. Längere Stimuli (Variable „Länge in Minuten (LST)“) hängen zwar mit einer höheren Item- bzw. Stimulusschwierigkeit zusammen, die Korrelationen sind jedoch nur sehr gering. Die Partialkorrelation wird auf Aufgabenebene negativ: Je länger die Stimuli sind, desto einfacher werden die Aufgaben. Dieses Ergebnis hängt sicher mit dem großen Einfluss einzelner Aufgaben zusammen, da die Fallzahlbelegung auf Aufgabenebene sehr gering ist. Insgesamt scheint die Variable LST nur wenig Einfluss auf die Schwierigkeit zu haben. Die Variable „Wortzahl (AWS)“ zeigt auf Stimulusebene einen geringen positiven Zusammenhang mit der Schwierigkeit. Für die Variable „Sprechgeschwindigkeit (SGS)“ wurden die Analysen nur mit Stimuli durchgeführt, deren Wortzahl zwischen 200 und 600 Wörtern liegen, um den Einfluss extremer Länge oder Kürze auszuschließen. Die Partialkorrelationen ergaben für diese Variable hohe negative Korrelationen auf Item- und auf Stimulusebene. Stimuli, bei denen die Sprechgeschwindigkeit über zwei Wörtern pro Sekunde liegt, hängen mit deutlich einfacheren Items und Stimuli zusammen. Da eine schnellere Sprechgeschwindigkeit an sich zu einer stärkeren Belastung des Arbeitsgedächtnisses führt, ist anzunehmen, dass die schnell gesprochenen Stimuli der Stichprobe insgesamt einfacher sind. Dies kann u. a. daran liegen, dass diese Stimuli andere Merkmale aufweisen, die Verständnis erleichternd wirken. Bei den langsamer gesprochenen Stimuli handelt es sich ferner in vielen Fällen um Stimuli mit z. T. sehr vielen Items.

Bei den Merkmalen zur sprachlichen Qualität der Stimuli zeigt die Variable „Akzent/Dialekt/Aussprache (AST)“ auf Aufgabenebene eine sehr geringe Partialkorrelation mit der Schwierigkeit. Kontrolliert wurden die Variablen „Inhaltswörter (IWH)“, „Sprechgeschwindigkeit (SGS)“ und „Literarischer Stimulus (SLT)“. Insgesamt scheint das Merkmal „Akzent/Dialekt/Aussprache (AST)“ jedoch kaum Einfluss auf die Item- bzw. die Stimulusschwierigkeit zu haben. Die Variable „Anzahl der Sprecher (ASP)“ korreliert sowohl auf Item- als auch auf Stimulusebene mit der Schwierigkeit. Je mehr Personen in einem Stimulus sprechen, desto schwieriger sind Items und Stimulus.

Die Variable „Anzahl der Stimuluspräsentationen (AHO)“ wirkt schwierigkeitssenkend. Wird ein Stimulus zweimal vorgespielt, so sind Items und Stimulus deutlich einfacher als bei einmaligem Vorspielen.

Tabelle IV-2.3.1c.: Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale Merkmalsgruppe III „Inhaltlich-thematische Merkmale“

Variable	Stimulusmerkmale	Zusammenhang mit der Schwierigkeit	
		angenommen	*tatsächlich*[17]
Merkmalsgruppe III: Inhaltlich-thematische Merkmale			
SLT	Literarischer Stimulus (Literarischer Stimulus; Nicht-literarischer Stimulus)	literarische Stimuli sind schwieriger	nicht-literarische Stimuli sind schwieriger
HKO	Hörkontext (Interview; Hörspiel; Reportage; Servicesendung; vorgetragene Lyrik/Prosa; nachgestellte Situation; Vortrag)	abstrakte, der Lebenswirklichkeit der Schüler wenig entsprechende Stimuli sind schwieriger	Kategorie „Hörspiel“ und „Interview“ sind am schwierigsten
TFU	Funktion (Vorwiegend argumentativ/ erklärend; Vorwiegend instruktiv/appellativ/informierend; Vorwiegend beschreibend/erzählend; Vorwiegend kommunikativ (phatisch))	abstrakte, der Lebenswirklichkeit der Schüler wenig entsprechende Stimuli sind schwieriger	Vorwiegend argumentative/ erklärende Stimuli sind am schwierigsten
THE	Thema (Informationen zur Person; Orte; Wohnen und Umwelt; Pflanzen und Tiere; Dienstleistungen; Gesundheit und Hygiene; Freizeit und Unterhaltung; Sprache; Ausbildung/ Beruf; Wetter; Menschliche Beziehungen; Tägliches Leben; Essen und Trinken)	abstrakte, der Lebenswirklichkeit der Schüler wenig entsprechende Stimuli sind schwieriger	Stimuli zum Thema „Wohnen und Umwelt“ sind am schwierigsten
WEL	Hintergrundwissen (Nicht zum Stimulusverständnis benötigt; Zum Stimulusverständnis benötigt)	Wird Hintergrundwissen benötigt, steigt die Schwierigkeit	kein Effekt

Die Variablen der Merkmalsgruppe III „Inhaltlich-thematische Merkmale“ eignen sich insgesamt eher moderat zur Vorhersage der Item- bzw. Stimulusschwierigkeit (vgl. Tabelle IV-2.3.1c.). Dies hängt jedoch auch mit der hohen Kategorienzahl einiger Variablen (z. B. „Thema (THE)“: 13 Kategorien) zusammen, die dazu führt, dass die einzelnen Kategorien nur sehr schwach belegt sind. Die beobachteten Effekte sind daher stark einzelnen Aufgaben geschuldet.

Nach Analysen mit einer Teilstichprobe (einbezogen wurden nur Stimuli, deren Wortzahl im Rahmen von 200 bis 600 Wörtern liegt), ergaben sich moderate Partialkorrelationen der Variable „Literarischer Stimulus (SLT)“ mit der Schwierigkeit auf Item- und auf Aufgabenebene. Erwartungswidrig hängen nicht-literarische Stimuli mit den schwierigeren Items zusammen. Es wird angenommen, dass dieses Ergebnis von der Qualität der literarischen Aufgaben abhängt. In der Teilstichprobe befinden sich nur noch vier literarische Stimuli, die einen großen Einfluss auf das Ergebnis haben.

Die Variable „Hörkontext (HKO)“ korreliert stark mit der Schwierigkeit auf Item- und auf Stimulusebene und trägt auch auf beiden Ebenen zur Varianzaufklärung bei. Am schwierigsten sind die Aufgaben, die in die Kategorien „Hörspiel“ und „Interview“ fallen. Von beiden

17 u. U. bei Analysen mit Teilstichproben

Kategorien ist nicht anzunehmen, dass sie der Lebenswirklichkeit der Schüler ferner sind, als beispielsweise die Kategorie „Vorgetragene Lyrik". Der Einfluss der Variable HKO ist nicht systematisch. Es ist zu vermuten, dass er stark von einzelnen Aufgaben abhängt, da die Stichprobe insgesamt – insbesondere auf Aufgabenebene – sehr gering ist und nur wenige Fälle (1 – 13) auf den einzelnen Kategorien liegen. Ähnlich unsystematisch sind die Ergebnisse der Analysen mit der Variable „Thema (THE)". Der Effekt zeigt sich deutlicher auf Item- als auf Aufgabenebene, ist jedoch aufgrund der geringen Fallzahlbelegung von einzelnen Aufgaben abhängig. Die Aufgaben zum Thema „Wohnen und Umwelt" und „Essen und Trinken" sind die schwierigsten Aufgaben. Die Stimuli sind nicht abstrakter bzw. der Lebenswelt der Schüler ferner als beispielsweise ein Stimulus der Kategorie „Sprache".

Die Variable „Stimulusfunktion (TFU)" zeigt eine Partialkorrelation mit der Schwierigkeit auf Aufgabenebene, nicht jedoch auf Itemebene. Vorwiegend argumentative bzw. erklärende Stimuli sind dabei am schwierigsten. Dieses Ergebnis ist insofern plausibel, als die auditive Erfassung einer Argumentstruktur schwieriger ist als beispielsweise dem Folgen einer linearen Erzählung.

Bei der Variable „Hintergrundwissen (WEL)" zeigt sich zwar, dass Stimuli, bei denen zum Verständnis Hintergrundwissen benötigt wird, mit den schwierigeren Items zusammenhängen. Insgesamt fallen die Korrelationen mit der Item- bzw. der Stimulusschwierigkeit jedoch eher gering aus und gehen bei der Berechnung der Partialkorrelation auf beiden Ebenen gegen null. Es scheinen demnach andere Faktoren die Schwierigkeit zu beeinflussen, der Einfluss der Variable WEL ist jedoch eher zu vernachlässigen.

Tabelle IV-2.3.1d.: Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale Merkmalsgruppe IV „Struktur der Stimuli und propositionale Dichte"

Variable	Stimulusmerkmale	Zusammenhang mit der Schwierigkeit	
		angenommen	*tatsächlich*[18]
Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte			
REL1	Relationstyp: Frage/Impuls/Themensetzung (⩽ 20%; > 20%)	-	+
REL3	Relationstyp: Spezifizierung (⩽ 30%; > 30%)	+	+
REL6	Relationstyp: Ziel/Bedingung (⩽ 10%; > 10%)	+	+
SFI	Schlussfolgerungen/Inferenzen (⩽ 2%; > 2%)	+	+
MLP	Länge der Propositionen (⩽ 6 Wörter; > 6 Wörter)	+	+
REL5	Relationstyp: Reihenfolge/Aufzählung (⩽ 20%; > 20%)	-	-/+
REL4	Relationstyp: Erklärung/Beweis/Ursache (⩽ 10%; > 10%)	+	-
REL2	Relationstyp: Antwort (⩽ 10%; > 10%)	-	-
PRO	Anzahl der Propositionen (< 50; 50 – 100; > 100)	+	0
PRW	Anteil der Propositionen (⩽ 15%; > 15%)	+	0

Anmerkungen: +: größere Anteile des Merkmals erhöhen die Schwierigkeit;
- : größere Anteile des Merkmals senken die Schwierigkeit;
0: es konnte kein Effekt festgestellt werden

18 u. U. bei Analysen mit Teilstichproben

Die vierte Merkmalsgruppe „Struktur der Stimuli und propositionale Dichte" versammelt Variablen zur Beschreibung der in den Stimuli auftretenden Relationstypen (REL1 -REL6), zur Erfassung der propositionalen Dichte (MLP, PRO/PRW) sowie zur Explizitheit der Stimuli (SFI) (vgl. Tabelle IV-2.3.1d.). Bei den Analysen mit dem Gesamtdatensatz zeigen die Variablen „Frage/Impuls/Themensetzung (REL1)" sowie „Erklärung/Beweis/Ursache (REL4)" auf Item- und auf Aufgabenebene einen Zusammenhang mit der Schwierigkeit. Items und Stimuli mit einem höheren Anteil des Relationstyps „Frage/Impuls/Themensetzung (REL1)" sind eher schwierig, wohingegen ein höherer Anteil des Relationstyps „Erklärung/Beweis/Ursache (REL4)" in den Stimuli eher mit einfacheren Items zusammenhängt. Die Variablen „Antwort (REL2)", „Spezifizierung (REL3)", „Reihenfolge/Aufzählung (REL5)" und „Ziel/Bedingung (REL6)" zeigen bei den Analysen mit allen Stimuli keinen Effekt auf die Schwierigkeit.

Die Analysen werden mit monologischen und dialogischen Stimuli wiederholt, da die Variable abhängig von der jeweiligen Gesprächssituation ist. Bei Stimuli, in denen nur eine Person spricht, zeigen die Variablen REL1, REL4 und REL5 einen Einfluss auf die Schwierigkeit. Der Relationstyp „Frage/Impuls/Themensetzung (REL1)" hängt noch immer positiv mit der Schwierigkeit zusammen. Dieses Ergebnis ist erwartungswidrig. Es wurde angenommen, dass dieser Relationstyp strukturell einfach ist und demnach eher mit einfacheren Items zusammenhängt. Ein Grund für die positive Korrelation mit der Schwierigkeit könnte sein, dass Stimuli, in denen viele Relationen des Typs 1 geratet wurden auch länger sind und durch die zahlreichen inhaltlichen Impulse auch eine stärkere Belastung des Arbeitsgedächtnisses mit sich bringen. Die Typen „Erklärung/Beweis/Ursache (REL4)" und „Reihenfolge/Aufzählung (REL5)" korrelieren hingegen negativ mit der Schwierigkeit. Auch das Ergebnis der Variable „Erklärung/Beweis/Ursache (REL4)" wurde so nicht erwartet. Denkbar wäre jedoch, dass Stimuli mit einem höheren Anteil dieses Relationstyps insgesamt kürzer sind und demnach weniger Arbeitsgedächtniskapazität beanspruchen.

Bei den Analysen mit Stimuli, in denen zwei oder drei Personen sprechen, korrelieren die Variablen REL2, REL3 und REL5 mit der Schwierigkeit. Der Relationstyp „Antwort (REL2)" hängt eher mit einfacheren Items zusammen. Die Relationstypen „Spezifizierung (REL3)" und „Reihenfolge/Aufzählung (REL5)" hängen eher mit schwierigeren Items zusammen. Dieses Ergebnis ist für das Merkmal REL3 erwartungsgemäß. Vor allem die Gegenläufigkeit des Merkmals REL5 bei mono- und dialogischen Stimuli ist jedoch erwartungskonträr. Denkbar wäre, dass eine aufzählende Struktur bei den dialogischen Stimuli aufgrund des Sprecherwechsels insgesamt deutlich schwieriger zu verfolgen ist als bei den monologischen Stimuli, wo es nur darum geht, die aufgezählten Inhalte zu behalten. Das Merkmal „Ziel/Bedingung (REL6)" korreliert auf Aufgabenebene leicht mit der Schwierigkeit.

Ein höherer Anteil der Variable „Schlussfolgerungen/Inferenzen (SFI)" hängt sowohl auf Itemals auch auf Aufgabenebene mit den schwierigeren Items zusammen, wobei die Variable nur bei Berechnung der Partialkorrelation mit der Aufgabenschwierigkeit positiv korreliert. Auf Itemebene ergeben sich keine signifikanten Partialkorrelationen. Dieses Ergebnis ist plausibel: Je mehr der Zuhörer in einem Stimulus schlussfolgern oder inferieren muss, desto schwieriger ist der Stimulus.

Die Partialkorrelation der Variablen „Anzahl Propositionen (PRO)“ und „Anteil der Propositionen (PRW)“ geht in beiden Fällen gegen null. Auch die Variable „Länge der Propositionen (MLP)“ zeigt kaum Einfluss auf die Schwierigkeit. Allerdings lässt sich bei sehr langen Propositionen mit mehr als sieben Wörtern ein Zusammenhang mit den schwierigeren Items beobachten. Analysen mit nur nicht-literarischen Stimuli sowie eine Dichotomisierung der Variable ergibt auf Aufgabenebene ein deutlicheres Ergebnis. Stimuli mit Propositionen, deren mittlere Länge sechs Buchstaben überschreitet, hängen dann mit einer höheren Schwierigkeit zusammen. Die Partialkorrelation der Variable erhöht sich auf Aufgabenebene sogar auf 0.26. Auf Itemebene zeigt die Variable keinen Effekt.

Tabelle IV-2.3.1e.: Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale Merkmalsgruppe V „Globalurteil“

Variable	Stimulusmerkmale	Zusammenhang mit der Schwierigkeit	
		angenommen	*tatsächlich*[19]
Merkmalsgruppe IV: Struktur der Stimuli und propositionale Dichte			
TSA	Stimulusschwierigkeit (leicht; mittel; schwierig)	+	+

Anmerkungen: +: größere Anteile des Merkmals erhöhen die Schwierigkeit

Obwohl die Einschätzung der Aufgabenentwickler recht gut mit der empirischen Schwierigkeit auf Item- und auf Aufgabenebene übereinstimmt, korreliert die Variable nur gering mit der Schwierigkeit. Auch die Berechnung der Partialkorrelation ergibt nur einen geringen Effekt (vgl. Tabelle IV-2.3.1e.).

2.3.2. Zusammenfassung Itemmerkmale

Zur Erfassung des Itemformats wurden die Items zum einen hinsichtlich des Offenheitsgrads der erwarteten Antwort (IFA) und zum anderen hinsichtlich ihres Kodieraufwands (IFK) eingestuft. Die Hypothese, dass offenere Itemformate auch mit den schwierigeren Items zusammenhängen, bestätigt sich nicht. (vgl. Tabelle IV-2.3.2a.) Bei der Einteilung nach Kodieraufwand liegen die schwierigsten Items in der Kategorie „Geschlossen-Kodieren“, bei der Einteilung nach dem Offenheitsgrad der erwarteten Antwort besitzen die schwierigsten Items das Format „Reihenfolge“ und „Zuordnung“. Alle drei Kategorien sind im Vergleich zu den übrigen Kategorien extrem schwach besetzt. Der erwartungswidrige Befund könnte deshalb durch den Einfluss einzelner Items erklärt werden. Ferner werden Items dieser Formate i. d. R. so kodiert, dass der Punkt nur gegeben wird, wenn das gesamte Item richtig gelöst wird. Möglicherweise sind Reihenfolge- bzw. Zuordnungsitems für die Schüler aber auch ungewohnte Formate und deshalb besonders schwierig. Insgesamt lässt sich jedoch die Tendenz beobachten, dass Items eher schwierig sind, wenn sie in einem offeneren Format vorliegen. Die Berechnung der Partialkorrelationen für die Einzelcodes der Variable IFK ergibt, dass die einzelnen Formattypen an sich kaum Einfluss auf die Itemschwierigkeit besitzen und untereinander stark korrelieren.

19 u. U. bei Analysen mit Teilstichproben
20 u. U. bei Analysen mit Teilstichproben

Das Merkmal „Formattyp" trägt hingegen deutlich zur Varianzaufklärung bei und korreliert mit der Schwierigkeit auf Itemebene. Bei der Variable IFA zeigen die Formate „Multiple-Choice" und „Richtig-Falsch" nach Berechnung der Partialkorrelation noch einen Zusammenhang mit der Schwierigkeit. Für alle anderen Formate dieser Variable weisen die Partialkorrelationen jedoch eher darauf hin, dass kaum Zusammenhänge mit der Schwierigkeit bestehen. Analog zur Variable IFK scheint also eher die gesamte Variable IFA als „Formatvariable" einen Einfluss auf die Schwierigkeit zu haben, als die einzelnen Formattypen.

Tabelle IV-2.3.2a.: Zusammenfassung Zusammenhangsanalysen Itemmerkmale Merkmalsgruppe I „Itemformat"

Variable	Stimulusmerkmale	Zusammenhang mit der Schwierigkeit	
		angenommen	***tatsächlich***[20]
Merkmalsgruppe I: Itemformat			
IFK/IFA	Itemformat	je offener, desto schwieriger	grob bestätigt

Tabelle IV-2.3.2b. fasst die Ergebnisse der Zusammenhangsanalysen für die Merkmalsgruppe II „Merkmale der Itempräsentation" zusammen. Je später die Items beantwortet werden (Variable „Position des Items innerhalb der Aufgabe (PIA)"), desto schwieriger sind sie, da die Informationen länger im Arbeitsgedächtnis aktiv gehalten werden müssen. Allerdings ist zu beachten, dass die Anordnung der Items von testtheoretischen und didaktischen Entscheidungen geleitet wird. Zu Beginn jeder Aufgabe wird meist ein sehr leichtes Item platziert, wohingegen Items, die eher zum Schluss einer Aufgabe stehen, i. d. R. Transferleistungen oder eine Interpretation oder Bewertung des Gehörten verlangen. Zum Teil weisen diese Items im Sinn einer Weiterführung auch Überschneidungen zum Kompetenzbereich Sprachreflexion oder Schreiben auf. Diese zusätzlichen Anforderungen erhöhen die Itemschwierigkeit. Ferner ist nicht auszuschließen, dass bei Aufgaben mit vielen Items zu einem späteren Zeitpunkt bei den Schülern auch Ermüdung eintritt. Nach Berechnung der Partialkorrelation zeigt die Variable PIA nur einen geringen Zusammenhang mit der Schwierigkeit.

Tabelle IV-2.3.2b.: Zusammenfassung Zusammenhangsanalysen Itemmerkmale Merkmalsgruppe II „Merkmale der Itempräsentation"

Variable	Stimulusmerkmale	Zusammenhang mit der Schwierigkeit	
		angenommen	***tatsächlich***[21]
Merkmalsgruppe II: Merkmale der Itempräsentation			
ZIB	Zeitpunkt der Itembearbeitung	einfacher nach dem Anhören	einfacher nach dem Anhören
PIA	Position des Items innerhalb der Aufgabe	je später, desto schwieriger	je später, desto schwieriger

21 u. U. bei Analysen mit Teilstichproben

Liegt der „Zeitpunkt der Itembearbeitung (ZIB)“ nach dem Anhören des Stimulus, wird die Aufgabe einfacher. Obwohl dann Informationen im Arbeitsgedächtnis aktiv gehalten werden müssen, wird dadurch insgesamt weniger Arbeitsgedächtniskapazität benötigt, als bei der Itembearbeitung während des Zuhörens. Die hier involvierten Tätigkeiten Zuhören, Lesen des Items, Nachdenken und Schreiben der Antwort erfordern insgesamt mehr Arbeitsgedächtniskapazität. Die Variable zeigt auch nach Berechnung der Partialkorrelation noch einen negativen Zusammenhang mit der Itemschwierigkeit.

Für die beiden Variablen „Größte vorkommende Plausibilität der Distraktoren (PDI)“ und „Mittlere Plausibilität der Distraktoren (MPD)“ zeigt sich, dass erwartungsgemäß die Itemschwierigkeit zunimmt, je plausibler die Distraktoren sind (vgl. Tabelle IV-2.3.2c.). Beide Variablen korrelieren deutlich positiv mit der Itemschwierigkeit. Bei der Berechnung der Partialkorrelationen ergibt sich für die Variable PDI ein positiver Zusammenhang von 0.53 mit der Schwierigkeit, die Variable MPD korreliert mit -0.41 negativ mit der Schwierigkeit. Dieses Ergebnis resultiert möglicherweise in der auf mehreren Distraktoren beruhenden Plausibilität mittleren Niveaus. Die Distraktoren weisen dann zwar im Mittel eine mittlere oder sogar höhere Plausibilität auf, lenken aber dennoch nicht allzu stark vom Distraktor ab. Zusammenfassend wird davon ausgegangen, dass i. d. R. einzelne besonders attraktive Distraktoren für eine Entscheidung den Attraktor nicht anzukreuzen verantwortlich sind und dass nicht die durchschnittliche Attraktivität der Distraktoren vom Attraktor ablenkt.

Tabelle IV-2.3.2c.: Zusammenfassung Zusammenhangsanalysen Itemmerkmale Merkmalsgruppe III „Merkmale von MC-Items“

Variable	**Stimulusmerkmale**	**Zusammenhang mit der Schwierigkeit**	
		angenommen	***tatsächlich***[22]
Merkmalsgruppe III: Merkmale von MC-Items			
PMC	Position des Attraktors im MC-Item	je später, desto schwieriger	es konnte kein Effekt festgestellt werden
PDI	Größte vorkommende Plausibilität der Distraktoren	je plausibler, desto schwieriger	je plausibler, desto schwieriger
MPD	Mittlere Plausibilität der Distraktoren	je plausibler, desto schwieriger	je plausibler, desto einfacher

Je später der Attraktor innerhalb des MC-Items erscheint, desto schwieriger ist er zu identifizieren, da auch mentale Repräsentationen der Distraktoren im Arbeitsgedächtnis aktiv gehalten werden müssen (Variable „Position des Attraktors im MC-Item (PMC)“). Diese Hypothese bestätigt sich zwar bei der Verteilung der Items auf die Codes, allerdings zeigt die Variable PMC auch nach Berechnung der Partialkorrelation keinen Zusammenhang mit der Schwierigkeit.

22 u. U. bei Analysen mit Teilstichproben

Tabelle IV-2.3.2d. fasst die Ergebnisse der Zusammenhangsanalysen für die Itemmerkmale der Gruppe IV „Kognitive Anforderungen der Items" zusammen.

Tabelle IV-2.3.2d.: Zusammenfassung Zusammenhangsanalysen Stimulusmerkmale Gruppe IV „Kognitive Anforderungen der Items"

Variable	**Stimulusmerkmale**	**Zusammenhang mit der Schwierigkeit**	
		angenommen	***tatsächlich***[23]
Merkmalsgruppe IV: Kognitive Anforderungen der Items			
AFB	Anforderungsbereich	je höher, desto schwieriger	je höher, desto schwieriger
BS	Geprüfter Standard	je kognitiv anspruchsvoller, desto schwieriger	
BS113	Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden	+	0
BS141	Gesprächsbeiträge anderer verfolgen und aufnehmen	+	0
BS142	Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben	+	0
BS143	Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln	+	0
ANI	Anzahl der benötigten NI pro Item	je mehr NIs benötigt, desto schwieriger	je mehr NIs benötigt, desto schwieriger
ARN	Auftretenshäufigkeit der NI	je höher die Auftretenshäufigkeit, desto einfacher	je höher die Auftretenshäufigkeit, desto einfacher
PST	Position der NI auf Stimulusebene	am Anfang und am Schluss einfacher	in der Mitte einfacher
WNI	Wortzahl der NI	je höher, desto schwieriger	je höher, desto schwieriger
TCO/ TOR	Konkretheit der NI – 33- und 5-stufig	je abstrakter, desto schwieriger	je abstrakter, desto schwieriger
HGW	Hintergrundwissen	je mehr erfordert, desto schwieriger	je mehr erfordert, desto schwieriger
TNI	Typ der NI	je kognitiv anspruchsvoller, desto schwieriger	je kognitiv anspruchsvoller, desto schwieriger

Anmerkungen: +: größere Anteile des Merkmals erhöhen die Schwierigkeit;
- : größere Anteile des Merkmals senken die Schwierigkeit;
0: es konnte kein Effekt festgestellt werden

23 u. U. bei Analysen mit Teilstichproben

Es wurde angenommen, dass Items umso schwieriger werden, je stärker kognitive Operationen wie Schlussfolgern oder Transferieren gefordert sind, da diese Operationen das Arbeitsgedächtnis in verstärktem Maße beanspruchen. Dieser Aspekt wird durch die Variablen „Anforderungsbereich (AFB)“,„Geprüfter Standard (STA)“ und „Typ der NI (TNI)“ erfasst. Die Variable AFB korreliert jedoch nur gering mit der Schwierigkeit (0.16), die Partialkorrelation liegt sogar noch darunter. Auch die Variable STA trägt nur wenig zur Varianzaufklärung bei. Die Subvariablen „BS141: Gesprächsbeiträge anderer verfolgen und aufnehmen“, „BS142: wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben“ und „BS143: Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln“ korrelieren stark untereinander, jedoch kaum mit der Itemschwierigkeit. Allerdings weist der Standard BS142 nach Kontrolle interferierender Variablen eine leicht negative Partialkorrelation mit der Schwierigkeit auf. Der Standard „BS113: verschiedene Formen mündlicher Darstellung unterscheiden und anwenden“ zeigt zunächst einen geringen Zusammenhang mit der Schwierigkeit, der bei Berechnung der Partialkorrelation jedoch verschwindet.

Auch bei der Variable TNI können Mehrfachbelegungen von Codes pro Item auftreten. Es werden jedoch nur die Fälle berücksichtigt, in denen genau ein Code vergeben wurde. Insgesamt korreliert die Variable kaum mit der Schwierigkeit und weist auch nach Berechnung der Partialkorrelation keinen stärkeren Zusammenhang mit der Schwierigkeit auf. Items, bei denen nach der Meinung des Zuhörers gefragt wird oder bei denen ein Sprecher identifiziert werden muss, sind am leichtesten. Items, bei denen geschlussfolgert werden muss oder bei denen Informationen zu sprachlichen Mitteln gegeben werden müssen, sind am schwierigsten. Die Variable TNI hat nur einen geringen Beitrag an der Varianzaufklärung und korreliert kaum mit der Itemschwierigkeit. Die Analysen werden wiederholt, für die Fälle bei denen vom Item genau eine Information verlangt wird sowie für die nur ein Code der Variable TNI vergeben wurde. Hier sind Items dann leicht, wenn sie nach der Meinung des Zuhörers bzw. der Stimmung oder der Einstellung des Sprechers fragen. Items zum Genre des Stimulus oder zu Geräuschen oder paraverbalen Informationen sind am schwierigsten. Zusätzlich werden für die einzelnen Codes Analysen durchgeführt, und zwar wieder nur für die Fälle, bei denen vom Item nur genau eine Information verlangt wird sowie für die nur ein Code der Variable TNI vergeben wurde. Insgesamt liegen keine signifikanten Ergebnisse vor, Cohen's d weist jedoch in den meisten Fällen auf hohe praktische Relevanz hin. Eine Varianzaufklärung wird von den Variablen höchstens in sehr geringem Umfang geleistet. Fragt ein Item nach einem Detail, der Meinung des Zuhörers, der Stimmung bzw. Einstellung des Sprechers, dem kommunikativen Zweck, der Funktion bzw. der Wirkung des Stimulus oder wird die Identifikation eines Sprechers verlangt, so sind die Items nicht nur einfacher als Items, die auf andere Informationen aus dem Stimulus abzielen. Diese Items sind auch einfacher als die mittlere Itemschwierigkeit der analysierten Itemstichprobe. Wird jedoch nach der Leitidee bzw. der Kernaussage des Stimulus, dessen Genre, nach Geräuschen bzw. paraverbalen Merkmalen oder sprachlichen Mitteln gefragt oder eine Schlussfolgerung verlangt, so sind die entsprechenden Items schwieriger und liegen in fast allen Fällen auch deutlich über der mittleren Itemschwierigkeit der Stichprobe. Die Merkmale, die mit einer höheren Itemschwierigkeit zusammenhängen, erscheinen sehr plausibel. Eine globale Aussage oder ein Genre zu erfassen ist schwieriger als auf bestimmte Details zu achten, da im ersten Fall das mentale Modell des Stimulus korrekt

gebildet werden musste. Auch die Fähigkeit Schlussfolgerungen zu ziehen ist relativ elaboriert. Bei einer Einordnung der den Codestufen zugefallenen Items in die Stufen des Kompetenzstufenmodells ergibt sich mit einigen Ausnahmen eine den Cutoffs der Kompetenzstufen entsprechende Staffelung der mittleren Itemschwierigkeiten.

Ein Anstieg der Itemschwierigkeit wird auch angenommen, wenn mehrere Informationen zur Lösung benötigt werden, die an unterschiedlichen Stellen im Stimulus auftreten und nur einmal genannt werden. Die dafür entwickelten Variablen lauten „Anzahl der benötigten NI pro Item (ANI)", „Position der NI auf Stimulusebene (PST)" und „Auftretenshäufigkeit der NI (ARN)". Die Variable ANI zeigt einen leichten positiven Zusammenhang mit der Schwierigkeit nach Berechnung der Partialkorrelation. Dabei hängen die einfachsten Items mit der Einschätzung „Es wird 1 NI benötigt, die weder im Stimulus noch im Item auftaucht." zusammen. Es wird davon ausgegangen, dass diese Items aufgrund von Weltwissen gelöst werden können und deshalb einfacher sind.

Bei der Variable PST wurden nur Fälle berücksichtigt, in denen die NI genau einer Position im Stimulus zugewiesen wurde sowie in denen für die Lösung eines Items genau eine NI benötigt wird. Obwohl weder Korrelationen noch Partialkorrelationen der Variable mit der Itemschwierigkeit auftreten, zeigt die Mittelwertsanalyse, dass Items am leichtesten sind, zu denen die gesuchte Information in der Mitte des Stimulus steht. Dieses erwartungswidrige Ergebnis könnte dadurch entstehen, dass die Schüler eine gewisse Einhörzeit benötigen und anfangs leicht Informationen überhören und zum Schluss der Stimuli bereits ein Ermüdungseffekt eintritt, die Informationen aus der Stimulusmitte also am besten behalten werden.

Bei der Variable ARN wurden die Codes 0 und 4 in den Code 0/4 „Redundanz der NI lässt sich nicht bestimmen, da es sich bei der NI um eine globale Einschätzung des Beitrags handelt oder die NI nicht im Stimulus vorkommt." zusammengefasst. Je redundanter eine NI dadurch ist, dass sie häufiger im Stimulus auftritt, desto geringer ist die Itemschwierigkeit. Analysen mit den Einzelcodes ergeben kaum Korrelationen mit der Itemschwierigkeit und auch die Varianzaufklärung der Einzelcodes ist zu vernachlässigen. Bei einer Wiederholung der Analysen für die Items, bei denen gemäß der Variable ANI nur eine NI benötigt wird, ergeben sich ein leichter positiver Zusammenhang mit der Schwierigkeit von Items, bei denen die NI nur einmal im Stimulus auftritt und ein leichter negativer Zusammenhang von Items, bei denen die NI dreimal oder öfter im Stimulus vorkommt. Bei Kontrolle interferierender Variablen ergibt sich für Code 1 nach wie vor ein leicht positiver Zusammenhang mit der Schwierigkeit. Ein leicht negativer Zusammenhang ist hingegen nur noch für den Code 2 („NI kommt 2x vor.") zu erkennen. Insgesamt ist das Ergebnis der Variable ARN jedoch erwartungsgemäß.

Ferner wird angenommen, dass der Abstraktheits- bzw. Konkretheitsgrad (Variable „Konkretheit der NI (TCO/TOR)"), der Umfang der gesuchten Information (Variable „Wortzahl der NI (WNI)") und die Notwendigkeit Hintergrundwissen zum Verständnis der Items zu aktivieren (Variable „Hintergrundwissen (HGW)") zur erhöhten Schwierigkeit der Items beitragen. Je länger die zur Lösung des Items benötigte Information ist (Variable WNI), desto schwieriger sind die Items. Dieses Ergebnis hängt wahrscheinlich mit der erhöhten Beanspruchung des Ar-

beitsgedächtnisses zusammen. Ein zweiter Analysedurchlauf mit ausschließlich den Fällen, bei denen gemäß der Variable ANI nur eine NI benötigt wird, bestätigt das Ergebnis. Die Partialkorrelation der Variable WNI mit der Schwierigkeit beläuft sich auf 0.11.

Bei den Variablen zur Beschreibung der Entnahme von Informationen unterschiedlichen Konkretheits-bzw. Abstraktionsgrades in aufsteigender Schwierigkeit TCO bzw. TOR werden die Hypothesen nur bedingt bestätigt. Zwar ist eine grobe Staffelung der Itemschwierigkeiten zu erkennen, allerdings liegen zahlreiche Subcodes in ihren Mittelwerten deutlich über oder unter dem erwarteten Wert. Dabei treten z. T. sehr hohe Standardabweichungen bzw. nur geringe Fallzahlbelegungen auf. Eine Wiederholung der Analysen mit nur den Items, die genau eine Information verlangen, ergibt eine insgesamt höhere Varianzaufklärung der Variablen TCO und TOR. Items sind am leichtesten, wenn für sie weniger konkrete Informationen wie Mengenangaben, Zeiten, Attribute, Aktionen/Versuche oder Angaben zur Art und Weise benötigt werden. Items, die nach hoch abstrakten Informationen wie dem Thema oder Äquivalenten/Paraphrasen fragen, sind am schwierigsten. Allerdings liegen leichtere Items auf TOR-Stufe 4 (sehr abstrakte Informationen) als auf Stufe 3 (abstrakte Informationen) oder 1 (hoch konkrete Informationen), und leichtere Items auf Stufe 2 (weniger konkrete Informationen) als auf Stufe 1. Obwohl beide Variablen leicht mit der Schwierigkeit korrelieren, gehen die Partialkorrelationen für beide Variablen gegen null.

Die Analysen mit der Variable HGW ergeben, dass Items, bei denen Hintergrundwissen zur Beantwortung benötigt wird, erwartungsgemäß deutlich schwieriger sind als Items, die ohne Hintergrundwissen verstanden werden. Allerdings korreliert die Variable nur gering mit der Schwierigkeit. Dieser Zusammenhang verschwindet beinahe vollständig bei der Berechnung der Partialkorrelation.

Tabelle IV-2.3.2e.: Zusammenfassung Zusammenhangsanalysen Itemmerkmale Merkmalsgruppe V „Globalurteil“

Variable	**Stimulusmerkmale**	**Zusammenhang mit der Schwierigkeit**	
		angenommen	***tatsächlich*[24]**
Merkmalsgruppe V: Globalurteil			
SEA	Itemschwierigkeit	je höher, desto schwieriger	je höher, desto schwieriger

Die zu jedem Item vorliegende Einschätzung der Aufgabenentwickler hinsichtlich der vermuteten Itemschwierigkeit korreliert leicht mit den empirischen Itemschwierigkeiten. Die Partialkorrelation der Variable SEA mit der Itemschwierigkeit beträgt 0.20.

Zusammenfassend lässt sich sagen, dass durch die Kontrolle hoch mit den Variablen korrelierender Merkmale die Partialkorrelation zwar häufig höher als die unkontrollierte Korrelation der Merkmale mit der Schwierigkeit ausfällt, jedoch oft noch immer sehr gering ist. Nur die

24 u. U. bei Analysen mit Teilstichproben

folgenden Merkmale haben einen signifikanten Einfluss ($p < 0.10$) von mindestens $r = 0.20$ auf die Aufgabenschwierigkeit: „Wortlänge (WLS)", „Lange Wörter (PLW)", „Mehrsilbige Wörter (PMW)", „Einsilbige Wörter (PEW)", „Anzahl Sprecher (ASP)", „Literarischer Stimulus (SLT)", „Hörkontext (HKO)", „Negationen (NEG)" und „Anakoluthe (STR4)".

Bei den Variablen des Fragebogens zur Einschätzung der Stimuli durch Lehrkräfte treten sehr hohe Korrelationen zwischen fast allen Variablen auf. Dies lässt darauf schließen, dass sich der Fragebogen – obwohl er versucht unterschiedliche Facetten der Stimuli zu erfassen – doch im Wesentlichen auf eine globale Einschätzung der Stimuli bezieht. Auf eine Berechnung von Partialkorrelationen wird aus diesem Grund verzichtet. Bis auf die Variable „Kohärenz/Textzusammenhang (KOH)" korrelieren alle Merkmale sowohl mit der Item- als auch der Aufgabenschwierigkeit.

Bei den Itemmerkmalen hängen nach Kontrolle hoch korrelierender Merkmale nur noch wenige Variablen mit der Schwierigkeit zusammen. Signifikant ($p < 0.05$) korrelieren folgende Variablen mit der Itemschwierigkeit: „Größte vorkommende Plausibilität der Distraktoren (PDI)", „Mittlere Plausibilität der Distraktoren (MPD)", „Itemformat: Richtig – Falsch (IFA.RF)", „Itemformat: Geschlossen – Ankreuzen (IFA:GA)", „Itemformat: Halboffen (IFA.HO)", „Itemformat: Einfache Antwort – 0/1 Kodieren (IFK.01) und „Zeitpunkt der Itembearbeitung (ZIB)" mit der Itemschwierigkeit zusammen. Dabei fällt auf, dass alle genannten Merkmale (bis auf IFA.RF) auf einer Einschätzung hinsichtlich Plausibilität oder Schwierigkeit beruhen und keine objektiven Auszählvariablen sind. Die Güte dieser Merkmale ist also abhängig von der Beurteilergruppe, hier die Aufgabenentwickler und die Autorin dieser Arbeit.

3. Regressionsanalysen

Zur gleichzeitigen Analyse mehrerer Prädiktoren werden regressionsanalytische Modelle eingesetzt. Bei festen Faktoren wird im Regressionsmodell für jede Faktorstufe des Prädiktors ein Regressionsparameter bestimmt, z. B. werden für die Variable „Anforderungsbereich" zwei Parameter für die drei Ausprägungen bestimmt, denn es wird angenommen, dass der Achsenabschnitt bereits einer Kategorie entspricht. Bei Behandlung eines Prädiktors als Kovariate wird nur ein Parameter geschätzt. Dies entspricht einer Partialkorrelation, bei der alle anderen Prädiktoren im Regressionsmodell kontrolliert werden.

3.1. Aufgabenschwierigkeiten und Stimulusmerkmale

Gemäß den Ergebnissen der Faktorenanalysen zu den Stimulusmerkmalen wurden zunächst die Merkmale, die auf einem gemeinsamen Faktor luden, in einem allgemeinen linearen Modell und dann in einem Regressionsmodell berechnet. Am Regressionskoeffizienten B lässt sich der Beitrag der unabhängigen Variablen („Inhaltswörter (IWH)", „Lange Wörter (PLW)" und „Negationen (NEG)") für die Erklärung der abhängigen Variable (Aufgabenschwierigkeit) ablesen. Je höher B ausfällt, desto größer ist der Einfluss der jeweiligen unabhängigen Variable auf die Aufgabenschwierigkeit. Die Güte einer Schätzung lässt sich mithilfe des Bestimmt-

heitsmaßes R^2 erkennen. R^2 drückt den Anteil der durch das Schätzmodell erklärten Varianz an der Gesamtvarianz der abhängigen Variablen aus.

3.1.1. Die am höchsten mit der Aufgabenschwierigkeit korrelierenden Merkmale

In einem ersten Schritt werden alle Merkmale ins Modell gegeben, die bei den vorhergehenden Analysen einen signifikanten Einfluss auf die Aufgabenschwierigkeit hatten. Dies trifft auf die Merkmale WLS, PLW, PMW, IWH, GWS, ASP, SGS, AHO, SLT, HKO, TFU, STR1, STR2, STR3, STR4, STR5, REL1, REL5, WIE, NEG, SUB, WEL und MLP zu. Aus dem Modell wurden dann schrittweise die am wenigsten aussagekräftigen Merkmale entfernt. Es verblieben so noch zwölf Merkmale, die zusammen 87.9% Varianz erklären, wobei insbesondere die Merkmale GWS, STR.1 und NEG daran einen großen Anteil haben. (vgl. Tabelle IV-3.1.1a., Anhang E)

3.1.2. Faktorenanalytisch bestimmte Merkmalsgruppen

In einem nächsten Schritt soll überprüft werden, ob die Ergebnisse der Faktorenanalyse zu Variablengruppen führen, die sich dazu eignen, besonders viel Varianz in der Aufgabenschwierigkeit zu erklären. Zur Erinnerung: Es wurden für die Stimulusvariablen in zwei Analyseschritten einmal drei und einmal sechs Faktoren identifiziert, wobei sich die Merkmale, die jeweils auf den ersten drei Faktoren liegen, in beiden Fällen decken. Die Variablen um den ersten Faktor **„Sprachliche Merkmale zur quantitativen Beschreibung der Stimuli"** sind WLS, PEW, PLW, PMW, IWH, GWS, STR3, STR4, NEG, SUB, WEL.

Da die Variablen WLS, PEW, PLW und PMW sowie die Variablen IWH und GWS stark miteinander korrelieren, wurde aus beiden Gruppen nur das jeweils am besten funktionierende Merkmal im Modell behalten, und zwar die Variable PLW und die Variable GWS. Ferner wurden wie in den vorhergehenden Analysen alle Variablen aus dem Modell entfernt, deren p-Werte > 0.10 betrugen. Die verbleibenden drei Variablen erklären bei insgesamt signifikanten Ergebnissen zusammen 23.3% der Varianz. Die Ergebnisse sind in Tabelle IV-3.1.2a., Anhang E dargestellt.

Die Variablen um Faktor 2 **„Merkmale zur Beschreibung der Länge und Komplexität der Stimuli"** sind LST, AWS, HKO, THE, STR5, RHE2, RHE4, REF, PRO. Bei den Variablen des zweiten Faktors eignen sich die Merkmale „Hörkontext (HKO)", „Länge des Stimulus (LST)" und „Jugendsprache/Umgangssprache (RHE4)" gut zur Vorhersage der Schwierigkeit. Die Merkmale des zweiten Faktors erklären zusammen 30.8% Varianz (vgl. Tabelle IV-3.1.2b., Anhang E). Die Variablen AWS und REF wurden aufgrund schlechter Passung aus dem Modell entfernt.

Um den dritten Faktor **„Inhaltliche Merkmale"** gruppieren sich die Variablen SLT, TFU, STR6, VER, MLP. Von den Variablen des dritten Faktors passte nur die Variable VER nicht ins Modell. Die verbleibenden vier Variablen leisten eine gemeinsame Varianzaufklärung von 10.4%. Den größeren Einfluss hat dabei die Variable SLT. (vgl. Tabelle IV-3.1.2c., Anhang E)

Von den Merkmalen auf dem vierten Faktor **„Merkmale gesprochener Sprache"** SGS, AST, STR1, STR2, RHE3 zeigt lediglich die Variable RHE3 keine Modellpassung. Sie wurde deshalb entfernt. Die verbleibenden vier Merkmale erklären gemeinsam 20.9% Varianz bei insgesamt signifikanten Ergebnissen. Den größten Beitrag leisten die Merkmale SGS und AST (vgl. Tabelle IV-3.1.2d., Anhang E).

Auf Faktor 5 **„Merkmale zur Erfassung der literarischen Qualität“** liegen die Variablen RHE1, DEI, WIE, SFI. Von diesen Merkmalen funktionieren nur die Variablen RHE1 und WIE im Modell. Die beiden Merkmale erklären zusammen 6.3% Varianz. (vgl. Tabelle IV-3.1.2e., Anhang E).

Um Faktor 6 **„Merkmale zur Erfassung des Diskurstyps“** gruppieren sich die Merkmale ASP, REL2, REL3. Die Merkmale des sechsten Faktors erklären zusammen 16.3% Varianz. Das Merkmal REL2 zeigte keine Modellpassung und wurde für die Analysen entfernt. Die Ergebnisse sind in Tabelle IV-3.1.2f., Anhang E dargestellt.

Zusammenfassend kann festgestellt werden, dass die Variablengruppe um den ersten Faktor recht aussagestark hinsichtlich der Aufgabenschwierigkeit ist. Drei Merkmale erklären zusammen 37.7% Varianz. Die durch die Faktorenanalyse bestimmten Merkmalsgruppen um die Faktoren 2 und 3 haben im Vergleich dazu weniger Einfluss auf die Aufgabenschwierigkeit: jeweils zwei Merkmale decken zusammen 25.4% bzw. 18.2% Varianz auf. Die Merkmale auf den Faktoren 4, 5 und 6 eignen sich als Gruppe weniger gut dazu, die Aufgabenschwierigkeit vorherzusagen. Die Variablen erklären bei insgesamt nicht signifikanten Werten zusammen 16.9% (Faktor 4), 12.1% (Faktor 5) und 17.9% (Faktor 6). Die den Merkmalen zugrunde liegenden gemeinsamen Eigenschaften/Faktoren weisen also weniger gute Vorhersagekraft bezüglich der Schwierigkeit auf.

In einem nächsten Schritt sollen aus diesem Grund Variablengruppen untersucht werden, die nach thematischen Gesichtspunkten für die Formulierung der Arbeitshypothesen dieser Arbeit gebildet wurden.

3.1.3. Thematisch gebildete Variablengruppen

Für diese Gruppen wird in einer Regressionsanalyse ihr Einfluss auf die Aufgabenschwierigkeit ermittelt. In Gruppe I **„Komplexität des Wortschatzes und sprachliche Merkmale**“ liegen die Variablen WLS, PEW, PLW, PMW, IWH, GWS, STR, RHE, DEI, WIE, REF, NEG, SUB, VER. Acht Merkmale klären zusammen 63% der Varianz auf. Besonders hohe Vorhersagekraft haben die Merkmale PLW, WIE und VER. Die übrigen Merkmale der Gruppe wurden aufgrund schlechter Eignung aus dem Modell entfernt. Die Ergebnisse der Regressionsanalyse mit den Variablen der Gruppe I sind in Tabelle IV-3.1.3a., Anhang E dargestellt.

In der Gruppe II **„Präsentationsmerkmale“** liegen die Merkmale LST, AWS, ASP, SGS, AST, AHO. Drei Merkmale erklären gemeinsam 23.5% Varianz, wobei dem Merkmal ASP die größte Bedeutung zukommt. (vgl. Tabelle IV-3.1.3b., Anhang E)

In der Gruppe III **„Thematische Merkmale“** liegen die Merkmale SLT, HKO, TFU, THE, WEL. Die ausgewählten drei Variablen „Hörkontext (HKO)“, „Stimulusfunktion (TFU)“ und „Hintergrundwissen (WEL)“ leisten bei insgesamt signifikanten Ergebnissen zusammen einen Beitrag von 30.5% zur Varianzaufklärung, wobei die Variable HKO etwas stärker an der Varianzaufklärung beteiligt ist. (vgl. Tabelle IV-3.1.3c., Anhang E)

In der vierten Gruppe „Struktur der Stimuli und propositionale Dichte“ liegen die Merkmale REL, SFI, PRO, PRW, MLP. Die Merkmale der Gruppe IV **„Struktur der Stimuli und propositionale Dichte“** leisteten zusammen bei nicht signifikanten Werten kaum einen Einfluss auf die Varianzaufklärung und auch der Ausschluss der problematischsten Variablen aus dem Modell brachte keine Verbesserung. Die Variablengruppe „Struktur der Stimuli und propositionale Dichte“ hat also kaum Einfluss auf die Aufgabenschwierigkeit.

In der Gruppe V **„Globalurteil“** liegt nur das Merkmal „Stimulusschwierigkeit (TSA)“. Eine Berechnung von Regressionsanalysen macht in diesem Fall also keinen Sinn.

Zusammenfassend kann gesagt werden, dass die nach inhaltlichen Gesichtspunkten gebildeten Merkmalsgruppen recht gute Vorhersagekraft hinsichtlich der Aufgabenschwierigkeit besitzen. Nur für die Merkmale der vierten Gruppe konnte kein gemeinsamer Einfluss festgestellt werden. Bei der ersten Gruppe „Komplexität des Wortschatzes und sprachliche Merkmale“ decken acht Merkmale zusammen 63.2% Varianz auf, bei der zweiten Gruppe „Präsentationsmerkmale“ klären drei Merkmale immerhin 23.5% auf und in der dritten Gruppe „Thematische Merkmale“ werden von drei Merkmalen sogar 30.5% aufgeklärt.

3.1.4. Ausgewählte Stimulusmerkmale der vorhergehenden Analysen

In einem letzten Schritt werden alle Variablen, die sich in den vorherigen Durchläufen als relevant erwiesen haben, ins Modell gegeben. Die in Tabelle IV-3.1.4a., Anhang E dargestellten Merkmale erwiesen sich dabei als die besten Prädiktoren. Zwölf Merkmale erklären zusammen 90% Varianz, wobei die Merkmale GWS und PLW den größten Anteil daran haben.

3.2. Itemschwierigkeiten und Stimulusmerkmale

In einem nächsten Schritt wird geprüft, ob die untersuchten Stimulusmerkmale auch einen Einfluss auf die Itemschwierigkeit haben. Die Analysen werden wieder in den eben beschriebenen Gruppierungen vorgenommen.

3.2.1. Die am höchsten mit der Itemschwierigkeit korrelierenden Merkmale

Zunächst werden wieder alle Variablen ins Modell gegeben, die in den vorhergehenden Analysen signifikant mit der Itemschwierigkeit korrelierten, und zwar die Variablen SGS, HKO, STR1, STR2. Aus dem Modell wurden dann schrittweise all die Merkmale entfernt, die einen p-Wert > 0.10 aufwiesen. Insgesamt fällt auf, dass wesentlich weniger Stimulusmerkmale einen signifikanten Einfluss auf die Itemschwierigkeit aufweisen und die Anzahl dieser Merkmale im Rahmen der Regressionsanalyse weiter auf drei Variablen reduziert werden mussten. Die verbleibenden drei Merkmale erklären zusammen 8.1% Varianz. Der Variable HKO fällt im Rahmen der Varianzaufklärung der größte Beitrag zu, wobei die Variable allein ebenso 8% erklärt. Die Ergebnisse der Regressionsanalyse mit den ausgewählten stärksten Stimulusmerkmalen sind in Tabelle IV-3.2.1a., Anhang E dargestellt.

3.2.2. Faktorenanalytisch bestimmte Merkmalsgruppen

Es wird dann überprüft, ob die Variablengruppen, die sich um die faktorenanalytisch identifizierten Faktoren gruppieren, dazu eignen, Varianz in der Itemschwierigkeit zu erklären. Um

den ersten Faktor **„Sprachliche Merkmale zur quantitativen Beschreibung der Stimuli“** gruppieren sich die folgenden Merkmale: WLS, PEW, PLW, PMW, IWH, GWS, STR3, STR4, NEG, SUB und WEL. Wie in den vorhergehenden Analysen wurden wieder alle Variablen aus dem Modell entfernt, deren p-Werte > 0.10 betrugen. Wie bei den Analysen auf Aufgabenebene spielen auch hier die Merkmale PLW und GWS eine wichtige Rolle. Zusätzlich passen nun jedoch auch die Variablen WLS, IWH, STR4 und WEL ins Modell. Die sechs Variablen erklären zusammen 8% Varianz. (vgl. Tabelle IV-3.2.2a., Anhang E)

Um den zweiten Faktor **„Merkmale zur Beschreibung der Länge und der Komplexität des Stimulus“** gruppieren sich die Variablen LST, AWS, HKO, THE, STR5, RHE2, RHE4, REF, PRO. Die Variablen auf dem zweiten Faktor erklären auf Itemebene zusammen 6.7% Varianz. Zusätzlich zu den Variablen HKO und RHE4 passt auf Itemebene auch die Variable LST ins Modell. (vgl. Tabelle IV-3.2.2b., Anhang E)

Auf dem dritten Faktor **„Inhaltliche Merkmale“** (SLT, TFU, STR6, VER und MLP) eignen sich – ebenso wie bei den Analysen auf Stimulusebene – die Variablen SLT und STR6 dazu, die Itemschwierigkeit vorherzusagen. Auf Itemebene erklären die beiden Merkmale jedoch nur 2.2% Varianz (im Vergleich zu 18.2% auf Stimulusebene). Die diesen beiden Merkmalen zugrunde liegenden Eigenschaften scheinen demnach kaum Einfluss auf die Itemschwierigkeit, wohl aber auf die Stimulusschwierigkeit, zu haben. (vgl. Tabelle IV-3.2.2c., Anhang E)

Um Faktor 4 **„Merkmale gesprochener Sprache“** gruppieren sich die Merkmale SGS, AST, STR1, STR2 und RHE3. Im Gegensatz zu den Analysen auf Aufgabenebene eignen sich auch drei Merkmale des vierten Faktors dazu, die Schwierigkeit auf Itemebene vorherzusagen. Die drei Merkmale erklären zusammen jedoch nur 5.1% Varianz. Den größten Einfluss hat das Merkmal SGS. (vgl. Tabelle IV-3.2.2d., Anhang E)

Die Variablen, die sich um den Faktor 5 **„Merkmale zur Erfassung der literarischen Qualität“** (RHE1, DEI, WIE, SFI) gruppieren, haben zusammen keinen Einfluss auf die Itemschwierigkeit.

Von den Merkmalen um den sechsten Faktor **„Merkmale zur Erfassung des Diskurstyps“** (ASP, REL2, REL3) passen zwei Merkmale („Anzahl der Sprecher (ASP)“ und „Spezifizierung (REL3)“ gut ins Modell und erklären gemeinsam 4.4% Varianz. Tabelle IV-3.2.2e., Anhang E zeigt die Ergebnisse der Regressionsanalyse mit den beiden Merkmalen.

Insgesamt zeigen die faktorenanalytisch gebildeten Gruppen der Stimulusmerkmale auf Itemebene deutlich weniger Einfluss auf die Schwierigkeit als auf Stimulusebene. In einem nächsten Schritt wird überprüft, ob sich die unter inhaltlichen Gesichtspunkten gebildeten Gruppen vielleicht besser zur Vorhersage der Itemschwierigkeit eignen.

3.2.3. Thematisch gebildete Variablengruppen

Auch für die thematisch gebildeten Variablengruppen werden Regressionsanalysen berechnet, um ihren Einfluss auf die Itemschwierigkeit zu ermitteln. In der ersten Gruppe **„Komplexität des Wortschatzes und sprachliche Merkmale“** befinden sich die Merkmale WLS, PEW, PLW,

PMW, IWH, GWS, STR, RHE, DEI, WIE, REF, NEG, SUB und VER. Neun Merkmale klären zusammen 15.8% der Varianz auf. Besonders hohe Vorhersagekraft haben die Merkmale WIE, SUB und VER. Die übrigen Merkmale der Gruppe wurden aufgrund schlechter Eignung aus dem Modell entfernt. Im Vergleich zu den Ergebnissen auf Stimulusebene, wo acht Merkmale beispielsweise 63.2% Varianz aufklärten, erscheint der Einfluss der Stimulusvariablen auf Itemebene eher gering. (vgl. Tabelle IV-3.2.3a., Anhang E)

In der Gruppe II **„Präsentationsmerkmale"** liegen die Merkmale LST, AWS, ASP, SGS, AST und AHO. Genau wie bei den Analysen mit der Aufgabenschwierigkeit erklären auch jetzt die drei Merkmale „Anzahl der Sprecher (ASP)", „Sprechgeschwindigkeit (SGS)" und „Anzahl der Stimuluspräsentationen (AHO)" gemeinsam am meisten Varianz, wenn auch die gemeinsame Varianzaufklärung mit 6.4% deutlich geringer ist, als bei den Analysen auf Aufgabenebene (23.5%). (vgl. Tabelle IV-3.2.3b., Anhang E)

Auch in der dritten Gruppe **„Thematische Merkmale"** (SLT, HKO, TFU, THE und WEL) erweisen sich die gleichen Merkmale als relevant für die Varianzaufklärung wie in den Analysen auf Aufgabenebene. Allerdings fällt auch hier die geleistete Varianzaufklärung mit nur 6.7% deutlich geringer aus. (vgl. Tabelle IV-3.2.3c., Anhang E)

Obwohl die ausgewählten Merkmale der Gruppe IV **„Struktur der Stimuli und propositionale Dichte"** einigermaßen ins Modell zu passen scheinen, tragen sie zusammen doch kaum zur Varianzaufklärung bei. In der Gruppe IV liegen die Merkmale REL, SFI, PRO, PRW, MLP. Insgesamt werden nur 3.6% Varianz von sechs Merkmalen erklärt. Den größten Beitrag leistet die Variable „Relationstyp: Reihenfolge/Aufzählung (REL5)" (vgl. Tabelle IV-3.2.3d., Anhang E).

Das Merkmal **„Stimulusschwierigkeit (TSA)"** der Gruppe V „Globalurteil" hat auf Itemebene einen Anteil von 1% an der Varianzaufklärung und ist als schwierigkeitsbeeinflussendes Merkmal zu vernachlässigen.

3.2.4. Ausgewählte Stimulusmerkmale der vorhergehenden Analysen

Auf Itemebene erklären die vorhersagekräftigsten Stimulusmerkmale deutlich weniger Varianz als auf Aufgabenebene. Von elf Merkmalen wird 22.6% Varianz aufgeklärt, wobei die stärksten Variablen die Merkmale „Lange Wörter (PLW)" und „Verben (VER)" sind. Die Merkmale „Sprechgeschwindigkeit (SGS)", „Lange Wörter (PLW)", „Referenz-Aussage-Strukturen (STR1)", „Wiederaufnahmen (WIE)", „Verben (VER)", „Anzahl der Sprecher (ASP)" und „Länge der Propositionen (MLP)" sind dabei sowohl auf Item- als auch auf Aufgabenebene relevant. Zusätzlich tragen dazu auf Itemebene die Merkmale „Ellipsen (STR2)", „Referenzen (REF)", „Relationstyp: Erklärung/Beweis/Ursache (REL4)" und „Relationstyp: Reihenfolge/Aufzählung (REL5)" zur Varianzaufklärung bei. Auf Aufgabenebene spielen hingegen die Merkmale „Worthäufigkeit (GWS)", „Adjazenzstrukturen (STR3)", „Relationstyp: Frage/Impuls/Themensetzung (REL1)", „Negationen (NEG)" und „Deixis (DEI)" eine Rolle. Die Ergebnisse der Regressionsanalyse mit den ausgewählten Stimulusmerkmalen sind in Tabelle IV-3.2.4a., Anhang E zusammengefasst.

3.3. Itemschwierigkeiten und Itemmerkmale

Wie bei den Stimulusmerkmalen sollen zunächst die am höchsten mit der Itemschwierigkeit korrelierenden Variablen gemeinsam in einem Regressionsmodell untersucht werden, in einem zweiten Schritt sollen dann Gruppen, die sich aufgrund faktorenanalytischer Verfahren ergeben haben untersucht werden und in einem dritten Schritt werden Merkmalsgruppen, die sich aufgrund inhaltlicher Überlegungen ergaben, analysiert.

3.3.1. Die am höchsten mit der Itemschwierigkeit korrelierenden Merkmale

Aufgrund der vorhergehenden Analyseergebnisse werden folgende Variablen gemeinsam ins Regressionsmodell gegeben: PDI, MPD, ZIB, IFK.GA, IFK.01, IFA.RF, IFA.HO, SEA. Die drei gut im Modell funktionierenden Merkmale „Größte vorkommende Plausibilität der Distraktoren (PDI)", „Zeitpunkt der Itembearbeitung (ZIB)" und „Itemformat: Geschlossen – Ankreuzen (IFK.GA)" erklären zusammen 41.3% Varianz, wobei den größten Anteil daran die Variable PDI hat (vgl. Tabelle IV-3.3.1a., Anhang E)

3.3.2. Faktorenanalytisch bestimmte Merkmalsgruppen

Im Bereich der Itemmerkmale wurden die Merkmale in drei Gruppen faktorenanalytisch untersucht. Dabei ergaben sich die Faktoren, wie in Tabelle IV-3.3.2a. dargestellt.

Tabelle IV-3.3.2a.: Übersicht über die faktorenanalytischen Ergebnisse bei den Itemmerkmalen

	Gruppe I **Allgemeine Itemmerkmale**	**Gruppe II** **Merkmale zur Beschreibung von MC-Items**	**Gruppe III** **Merkmale zur Beschreibung der NI**
Faktor 1	AFB, BS143, BS113, TNI2, TNI4, TNI9	PDI, MPD, SEA	BS143, BS113
Faktor 2	TCO, TOR, TNI11	ZIB, PIA, AFB	TOR, AFB, BS141
Faktor 3	IFA.AK, IFA.S	HGW	BS142
Faktor 4	TNI3, TNI5, TNI6	PMC	

Gruppe I: allgemeine Itemmerkmale

Obwohl auf dem ersten Faktor **„Merkmale zur Beschreibung der kognitiven Anforderungen der Items"** mehrere Merkmale (AFB, BS143, BS113, TNI2, TNI4, TNI9) liegen, funktionieren nur jeweils zwei davon im Regressionsmodell. Die beiden Merkmale „Anforderungsbereich (AFB)" und „Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln (BS143)" tragen zusammen mit 3.4% jedoch nur wenig zur Varianzaufklärung bei (vgl. Tabelle IV-3.3.2b., Anhang E). Die beiden Merkmale auf dem dritten Faktor **„Merkmale zum Itemformat"** „Itemformat Ankreuzen (MC+RF) (IFA.AK)" und „Itemformat Schreiben (HO+OI) (IFA.S)" erklären zusammen 12.1% Varianz (vgl. Tabelle IV-3.3.2c., Anhang E). Die Merkmale auf dem zweiten Faktor „Merkmale zur Konkretheit der NI" (TCO, TOR, TNI11) zeigen keine gute Modellpassung. Auch die Merkmale, die auf dem vierten Faktor **„Merkmale zur Beschreibung der NI"** liegen (TNI3, TNI5, TNI6), tragen nicht dazu bei, die Itemschwierigkeit zu erklären.

Gruppe II: Merkmale zur Beschreibung von MC-Items

Auf dem ersten Faktor **„Merkmale zur Einschätzung der Plausibilität der Distraktoren"** liegen die Variablen PDI, MPD und SEA. Die beiden Merkmale „Größte vorkommende Plausibilität der Distraktoren (PDI)" und „Itemschwierigkeit (SEA)" sind mit 24.1% deutlich stärker an der Varianzaufklärung beteiligt als die Merkmale auf dem zweiten Faktor „Merkmale zu den Rahmenbedingungen der Items" (8.6%). Insbesondere die Variable PDI hat diesbezüglich eine hohe Voraussagekraft. Allerdings ist zu beachten, dass die Variable PDI für sich genommen bereits mit 27% an der Varianzaufklärung beteiligt ist und ihr Einfluss durch die Variable SEA demnach eingeschränkt wird (vgl. Tabelle IV-3.3.2d., Anhang E). Die Merkmale „Zeitpunkt der Itembearbeitung (ZIB)", „Position des Items innerhalb der Aufgabe (PIA)" und „Anforderungsbereich (AFB)" um den Faktor 2 **„Merkmale zu den Rahmenbedingungen der Items"** erklären gemeinsam 8.6% Varianz. Die Variable ZIB hat dabei einen geringfügig größeren Anteil als die Variablen PIA und AFB (vgl. Tabelle IV-3.3.2e., Anhang E). Da auf den Faktoren 3 **„Hintergrundwissen"** und 4 **„Position des Attraktors"** jeweils nur ein Merkmal gleichen Namens liegt, wird auf die Berechnung einer Regressionsanalyse verzichtet.

Gruppe III: Merkmale zur Beschreibung der NI

Die Merkmale, die auf Faktor 1 **„Zur Itembeantwortung benötigte Kompetenzen**" (BS142, BS143) liegen, ergaben keine gute Passung im Regressionsmodell. Sie eignen sich offensichtlich nicht, die Itemschwierigkeit vorherzusagen. Von den Merkmalen auf Faktor 2 **„Merkmale zur Differenzierung der kognitiven Operationen"** (TOR, AFB, BS113) eignen sich die Variablen „Konkretheit der NI – 5-stufig (TOR)" und „Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden (BS113)" dazu die Schwierigkeit vorherzusagen. Die beiden Variablen TOR und STA1 decken bei einem insgesamt signifikanten Ergebnis zusammen 4.5% Varianz auf. (vgl. Tabelle IV-3.3.2f., Anhang E) Auf dem dritten Faktor **„Merkmal zur Beschreibung der verschiedenen Formen von Mündlichkeit"** liegt nur die Variable „Verschiedene Formen mündlicher Darstellung unterscheiden und anwenden (BS141)". Auf eine Analyse im Regressionsmodell wird hier verzichtet. Von den für die einzelnen Gruppen durchgeführten faktorenanalytischen Untersuchungen eigneten sich nur jeweils wenige Merkmalsgruppen als Prädiktoren für die Itemschwierigkeit. Relativ viel Varianz wird von den beiden Merkmalen auf dem ersten Faktor der Gruppe II (Merkmale zur Beschreibung von MC-Items) mit 24.1% geleistet. Auch die Merkmale zum Itemformat (Merkmale auf dem dritten Faktor der Gruppe I: allgemeine Itemmerkmale) klären mit 12.1% im Verhältnis zu den anderen Merkmalsgruppen noch recht viel Varianz auf.

3.3.3. Thematisch gebildete Variablengruppen

Da die Merkmale der ersten Gruppe **„Itemformat"** (IFA und IFK) sehr hoch miteinander korrelieren, werden sie nicht zusammen im Regressionsmodell analysiert. Die beiden Merkmale der Gruppe II **„Merkmale der Itempräsentation"** „Zeitpunkt der Itembearbeitung (ZIB)" und „Position des Items innerhalb der Aufgabe (PIA)" erklären bei einem signifikanten Ergebnis zusammen 7.3% Varianz. (vgl. Tabelle IV-3.3.3a., Anhang E) Die Merkmale der Gruppe III **„Merkmale von MC-Items"** PMC, PDI und MPD haben keine gute Modellpassung und tragen kaum dazu bei, als Gruppe die Itemschwierigkeit zu erklären. In der Gruppe IV **„Kognitive Anforderungen der Items"** befinden sich die Variablen AFB, BS, ANI, ARN, PST, WNI, TCO/TOR, HGW

und TNI. Von diesen Variablen erweisen sich nur die Merkmale „Gesprächsbeiträge anderer verfolgen und aufnehmen (BS141)“, „Wesentliche Aussagen aus umfangreichen gesprochenen Stimuli verstehen, diese Informationen sichern und wiedergeben (BS142)“, „Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln (BS143)“, „Wortzahl der NI (WNI)“ und „Konkretheit der NI (5-stufig) (TOR)“ als Modell passend. Zusammen erklären sie 8.5% Varianz (vgl. Tabelle IV-3.3.3b., Anhang E). Da sich in der vierten Gruppe „Globalurteil“ nur ein Merkmal („Itemschwierigkeit (SEA)“) befindet, wird hier auf die Berechnung einer Regressionsanalyse verzichtet. Im Vergleich zu den Ergebnissen der thematischen Gruppen der Stimulusmerkmale fallen die Ergebnisse der Itemmerkmalsgruppen deutlich geringer aus. Zwei bzw. fünf Merkmale erklären gemeinsam nur 7.3% bzw. 8.5% Varianz.

3.3.4. Ausgewählte Itemmerkmale der vorhergehenden Analysen

Alle Variablen, die sich in den vorherigen Durchläufen als relevant erwiesen haben, werden wieder gemeinsam ins Modell gegeben, wobei nur die stärksten Merkmale behalten werden. Die Merkmale in Tabelle IV-3.3.4a., Anhang E erwiesen sich dabei als die besten Prädiktoren. Vier Merkmale erklären zusammen 45% Varianz, wobei die Merkmale „Größte vorkommende Plausibilität der Distraktoren (PDI)“ und „Zeitpunkt der Itembearbeitung (ZIB)“ den größten Anteil daran haben. Bei der inhaltlichen Betrachtung dieser Itemmerkmale fällt auf, dass sich zwei Variablen auf das Itemformat, insbesondere Ankreuz-Items bzw. Multiple-Choice-Items, beziehen. Auch die Variable ZIB beschreibt kein inhaltliches Merkmal von Items, sondern die Rahmenbedingungen der Testadministration. Ein Großteil der Itemschwierigkeit scheint demnach also stark vom Testformat und der Testadministration abzuhängen.

3.4. Zusammenfassung der Regressionsanalysen

3.4.1. Zusammenfassung: Einfluss Stimulusmerkmale auf die Aufgabenschwierigkeit

Die unterschiedlichen Arten der Gruppierung ergaben verschiedene Ergebnisse. Zwölf Merkmale der Gruppe „Die am höchsten mit der Aufgabenschwierigkeit korrelierenden Merkmale“ erklärten mit 87.9% einen Großteil der Varianz. Es handelt sich dabei um die Merkmale „Worthäufigkeit (GWS)“, „Sprechgeschwindigkeit (SGS)“, „Anzahl der Stimuluspräsentationen (AHO)“, „Hörkontext (HKO)“, „Stimulusfunktion (TFU)“, „Referenz-Aussage-Strukturen (STR1)“, „Ellipsen (STR2)“, „Adjazenzstrukturen (STR3)“, „Wiederaufnahmen (WIE)“, „Negationen (NEG)“, „Hintergrundwissen (WEL)“ und „Lange Wörter (PLW)“.

Bei den faktorenanalytischen Merkmalsgruppen erklärten sieben Merkmale um die Faktorgruppe „Merkmale zur Beschreibung der Länge und der Komplexität des Stimulus“ 30.8% Varianz. Die Faktorgruppen „Merkmale zur Erfassung der literarischen Qualität“ (zwei Merkmale erklären 6.3%) sowie „Inhaltliche Merkmale“ (vier Merkmale erklären 10.4%) als weniger geeignet, um die Aufgabenschwierigkeit vorherzusagen. Insbesondere die Gruppe „Sprachliche Merkmale zur quantitativen Beschreibung der Stimuli“ eignet sich mit den Variablen „Lange Wörter (PLW)“, „Worthäufigkeit (GWS)“, „Anakoluthe (STR4)“ und „Hintergrundwissen (WEL)“ hingegen besser zur Vorhersage der Schwierigkeit. (R^2 = 23.3%) Die Gruppen „Merkmale zur Erfassung des Diskurstyps“ und „Merkmale gesprochener Sprache“ erklären mit zwei bzw. vier Variablen immerhin 16.3% bzw. 20.9% Varianz.

Bei den thematisch gebildeten Variablengruppen erwies sich die Gruppe „Komplexität des Wortschatzes und sprachliche Merkmale“ mit acht Merkmalen mit 63.2% Varianzaufklärung als vorhersagestark. Aber auch die Gruppen „Präsentationsmerkmale“ und „Thematische Merkmale“ erklären mit jeweils drei Variablen 23.5 bzw. 30.5% Varianz.

Das am häufigsten in den Gruppen auftretende Merkmal ist die Variable „Sprechgeschwindigkeit (SGS)“, die allein nur 2% Varianz auf Aufgabenebene aufklärt, in der Interaktion mit anderen Merkmalen jedoch eine wichtige Rolle zu spielen scheint. Ferner tauchten die Variablen „Worthäufigkeit (GWS)“, „Anzahl der Stimuluspräsentationen (AHO)“, „Hörkontext (HKO)“, „Stimulusfunktion (TFU)“, „Referenz-Aussage-Strukturen (STR.1)“, „Wiederaufnahmen (WIE)“, „Negationen (NEG)“, „Hintergrundwissen (WEL)“, „Lange Wörter (PLW)“ und „Verben (VER)“ als wichtigste Merkmale in mehreren Gruppen auf. Sie scheinen demnach für die Vorhersage der Aufgabenschwierigkeit von großer Bedeutung zu sein. Das finale Modell mit den Merkmalen, die in den vorhergehenden Analysen am meisten Varianz erklärten, eignet sich am besten dazu, die Aufgabenschwierigkeit vorherzusagen. Zwölf Merkmale erklären zusammen 90% Varianz.

Anhang E, Tabelle E-1. gibt einen Überblick über die Ergebnisse der regressionsanalytischen Untersuchungen mit den Stimulusmerkmalen in Bezug auf die Aufgabenschwierigkeit.

3.4.2. Zusammenfassung: Einfluss Stimulusmerkmale auf die Itemschwierigkeit

In allen Fällen fiel die geleistete Varianzaufklärung durch die Stimulusmerkmale auf Itemebene deutlich geringer aus, als auf Aufgabenebene. Dabei ist jedoch zu berücksichtigen, dass auch die Stimulusmerkmale an sich auf Itemebene deutlich geringere Eta²-Werte aufweisen. Stimulusmerkmale eignen sich offensichtlich weniger gut dazu, die Itemschwierigkeit vorauszusagen. Anhang E, Tabelle E-2. gibt einen Überblick über die Ergebnisse der regressionsanalytischen Untersuchungen mit den Stimulusmerkmalen in Bezug auf die Itemschwierigkeit.

Im Bereich der Gruppe „Die am höchsten mit der Itemschwierigkeit korrelierenden Merkmale“ eignen sich die Variablen „Sprechgeschwindigkeit (SGS)“, „Hörkontext (HKO)“ und „Ellipsen (STR2)“ am besten dazu, die Itemschwierigkeit vorherzusagen. Dennoch liegt die Varianzaufklärung mit 8% vergleichsweise niedrig.

Bei den faktorenanalytisch bestimmten Merkmalsgruppen erweisen sich nur einige Variablen der Gruppen 1 „Sprachliche Merkmale zur quantitativen Beschreibung der Stimuli“ und 2 „Merkmale zur Beschreibung der Länge und der Komplexität des Stimulus“ als geeignet, die Itemschwierigkeit vorherzusagen. Die Merkmale der Gruppen um die Faktoren 3 „Inhaltliche Merkmale“, 4 „Merkmale gesprochener Sprache“, 5 „Merkmale zur Erfassung der literarischen Qualität“ und 6 „Merkmale zur Erfassung des Diskurstyps“ zeigen hingegen keinen Effekt auf die Itemschwierigkeit.

Die thematisch gebildeten Variablengruppen eignen sich besser dazu, die Itemschwierigkeit vorherzusagen. Durch die Merkmale der Gruppe I „Komplexität des Wortschatzes und sprachliche Mittel“ werden 15.8% Varianz erklärt. Jeweils drei Merkmale der Gruppen „Präsentations-

merkmale“ und „Thematische Merkmale“ erklären 6.4 bzw. 6.7% Varianz. Im Gegensatz zu den Analysen mit den Stimulusmerkmalen und der Aufgabenschwierigkeit zeigt nun auch die Gruppe IV „Struktur der Stimuli und propositionale Dichte“ einen – wenn auch schwachen – Einfluss auf die Itemschwierigkeit. Sechs Merkmale erklären 3.6% Varianz.

Aus diesen Analysen werden die einflussreichsten Variablen ermittelt und erneut in ein Regressionsmodell gegeben. Es zeigt sich, dass elf Merkmale 22.6% Varianz erklären. Die wichtigsten Stimulusmerkmale für die Varianzaufklärung auf Itemebene sind die Variablen „Sprechgeschwindigkeit (SGS)“, „Hörkontext (HKO)“, „Ellipsen (STR2)“, „Hintergrundwissen (WEL)“, „Lange Wörter (PLW)“ und „Nähezeichen (STR5)“.

3.4.3. Zusammenfassung: Einfluss Itemmerkmale auf die Itemschwierigkeit

Zusammenfassend fällt auf, dass nicht nur der Einfluss der Stimulusmerkmale auf die Itemschwierigkeit sondern auch der Einfluss der Itemmerkmale darauf deutlich geringer ist, als die Stimulusmerkmale die Aufgabenschwierigkeit zu beeinflussen scheinen (vgl. Anhang E, Tabelle E-3.). Relativ starke Vorhersagekraft fällt der Gruppe „Die am stärksten mit der Itemschwierigkeit korrelierenden Merkmale“ mit den Merkmalen „Größte vorkommende Plausibilität der Distraktoren (PDI)“, „Zeitpunkt der Itembearbeitung (ZIB)“ und „Itemformat: Geschlossen – Ankreuzen (IFK. GA)“ mit 41.3% Varianzaufklärung zu.

Die Itemmerkmale wurden im Rahmen der faktorenanalytischen Untersuchungen in drei Gruppen eingeteilt: Gruppe I „Allgemeine Itemmerkmale“, Gruppe II „Merkmale zur Beschreibung von MC-Items“ und Gruppe III „Merkmale zur Beschreibung der NI“. Bei diesen faktorenanalytisch bestimmten Merkmalsgruppen zeigen sich je Gruppe stets nur wenige Variablen als vorhersagestark, was die Itemschwierigkeit betrifft. Bei der Gruppe I sind dies um den ersten Faktor „Merkmale zur Beschreibung der kognitiven Anforderungen der Items“ die Variablen „Anforderungsbereich (AFB)“ und „BS143: Aufmerksamkeit für verbale und non-verbale Äußerungen entwickeln“ sowie um den dritten Faktor „Merkmale zum Itemformat“ die Variablen „Ankreuzen (MC/RF)“ und „Schreiben (HO/OI)“. Bei der zweiten Gruppe um den Faktor „Merkmale zur Einschätzung der Plausibilität der Distraktoren“ zeigen die Variablen „Größte vorkommende Plausibilität der Distraktoren (PDI)“ und „Itemschwierigkeit (SEA)“ sowie die Variablen „Zeitpunkt der Itembearbeitung (ZIB)“, „Position des Items innerhalb der Aufgabe (PIA)“ und „Anforderungsbereich (AFB)“ um den Faktor 2 „Merkmale zu den Rahmenbedingungen der Items“ den größten Einfluss auf die Itemschwierigkeit. In der dritten Gruppe erweisen sich lediglich die Variablen „Type of requested information (5-stufig) (TOR)“ und „BS113: verschiedene Formen mündlicher Darstellung unterscheiden und anwenden“ um den zweiten Faktor „Merkmale zur Differenzierung der kognitiven Operationen“ als aussagekräftig bezüglich der Itemschwierigkeit.

Aus der Gruppe der thematisch gebildeten Variablengruppen zeigt die Gruppe 3 „Thematische Merkmale“ mit den Variablen „Hörkontext (HKO)“, „Textfunktion (TFU)“ und „Hintergrundwissen (WEL)“ den größten Effekt. Die genannten drei Merkmale erklären zusammen 6.7% Varianz. Aus der Gruppe „Präsentationsmerkmale“ erklären drei Merkmale 6.4% Varianz und aus der Gruppe „Komplexität des Wortschatzes und sprachliche Mittel“ erklären neun Merkmale 15.8% Varianz.

Die aus den vorhergehenden Regressionsanalysen ausgewählten Variablen eignen sich am besten dazu, die Itemschwierigkeit vorherzusagen. Vier Merkmale erklären hier 45% der Gesamtvarianz. Ansonsten liegt die geleistete Varianzaufklärung der Itemmerkmale bei zwei oder drei Merkmalen je Gruppe meist im einstelligen Bereich. Nur die Merkmale „Größte vorkommende Plausibilität der Distraktoren (PDI)“ und „Itemschwierigkeit (SEA)“ sowie die Merkmale zum Itemformat „Ankreuzen (IFA.AK)“ und „Schreiben (IFA.S)“ tragen mit 24.1% bzw. 12.1% etwas mehr zur Varianzaufklärung bei. Die am häufigsten in den einzelnen Gruppen auftretenden Merkmale sind die Variablen „Größte vorkommende Plausibilität der Distraktoren (PDI)“, „Anforderungsbereich (AFB)“, „BS143: Aufmerksamkeit für verbale und nonverbale Äußerungen entwickeln“, „Zeitpunkt der Itembearbeitung (ZIB)“, „Position des Items innerhalb der Aufgabe (PIA)“ und „Type of requested Information (5-stufig) (TOR)“.

V

Zusammenfassung und Diskussion

VI Zusammenfassung und Diskussion

Ziel dieser Arbeit war es, die Bedeutung ausgewählter Merkmale im Kompetenzbereich „Zuhören" besser zu verstehen und ihren Einfluss auf die Schwierigkeit von Items und Stimuli zu bestimmen. Untersucht wurden korpusanalytisch ermittelte Merkmale der Stimuli, der Items sowie Merkmale, die aus der Interaktion der Items mit dem entsprechenden Stimulus resultieren. Außerdem sollte die Übereinstimmung der Einschätzung der Aufgabenentwickler bezüglich der Schwierigkeit von Items und Stimuli mit den empirischen Schwierigkeiten überprüft werden. Zu den Stimuli liegen außerdem Einschätzungen von Lehrkräften auf der Grundlage einer Ratingskala vor. Zusätzlich werden Einschätzungen der Testpersonen zu den Stimuli und weitere Merkmale der Testpersonen in die Analysen miteinbezogen.

Die Kenntnis schwierigkeitsbeeinflussender Merkmale ist sowohl aus wissenschaftlicher als auch aus didaktischer Perspektive wünschenswert. In einem wissenschaftlichen Rahmen kann Wissen über Prädiktoren der Itemschwierigkeit zur Optimierung der Aufgabenentwicklung, aber auch bei der Erstellung von Kompetenzstufenmodellen eingesetzt werden. Für die theoriegeleitete Aufgabenkonstruktion ist dabei zunächst die Kenntnis von Stimulusmerkmalen wichtig, um geeignete, d. h. in ihrer Schwierigkeit für die Zielgruppe passende Stimuli zu finden. Bei der Itemgenerierung hilft die Kenntnis schwierigkeitsbeeinflussender Merkmale, Items systematisch zu manipulieren und in ihrer Schwierigkeit gezielt für bestimmte Zielgruppen anzupassen. Im Rahmen der Itemrevision können dann Unterschiede in den Schwierigkeiten und Trennschärfen der Items besser erklärt werden. Wissen über derartige Prädiktoren könnte in diesem Sinne auch für die Beschreibung der einzelnen Stufen in einem Kompetenzstufenmodell und zur Validierung des untersuchten Konstrukts genutzt werden. Im pädagogischen Kontext ist die Kenntnis von Prädiktoren der Itemschwierigkeit ausschlaggebend für die Passung der Unterrichtsmaterialien für die Zielgruppe einerseits durch die Auswahl geeigneter, bereits vorhandener Materialien und andererseits durch die Modifikation der Materialien beispielsweise für die Binnendifferenzierung im Unterricht. Auf diese Weise generierte Items ermöglichen gleichzeitig auch genauere Aussagen über die Schülerfähigkeiten, sodass leichter individuelle Fähigkeitsprofile erstellt werden können und die Qualität der diagnostischen Rückmeldungen optimiert wird.

Die Untersuchung schwierigkeitsbeeinflussender Merkmale hat gerade im englischsprachigen Raum für den Kompetenzbereich „Leseverstehen" eine längere Tradition (z. B. Freedle & Kostin, 1993; Evetts & Gauthier, 2005, Chalifour & Powers, 1989; Nold & Rossa, 2007) Für den Bereich „Zuhören" gibt es hingegen weniger Untersuchungen (z. B. Freedle & Kostin, 1996; Bae & Bachman, 1998). Für die Untersuchung des Bereichs „Zuhören" im Deutschen liegen Untersuchungen für den Primarbereich von Bremerich-Vos et al. (2009) und Böhme et al. (2010) vor, für den Sekundarbereich gab es mit Blick auf o. g. Fragestellungen jedoch noch keine Überprüfung, wie sie in der vorliegenden Arbeit vorgenommen wurde. Auch spezielle empirisch validierte Fragebögen, mit deren Hilfe Lehrkräfte die Schwierigkeit der einzusetzenden Hörstimuli beurteilen können, fehlen derzeit.

Die vorliegende Arbeit stellte in Kapitel II.1. zunächst die Eigenart von gesprochener Sprache und mündlicher Kommunikation dar, um zu zeigen, vor welchen Herausforderungen die Kompetenzdiagnostik im Bereich des Hörverstehens in der deutschen Sprache steht. Obwohl mündliche Kommunikation menschheitsgeschichtlich älter als Schriftsprache ist, wird der gesprochenen Sprache aufgrund ihrer Flüchtigkeit weniger gesellschaftliche Bedeutung beigemessen (Thaler, 2007). Die besonderen Umstände in Art (z. B. Übertragung durch Schallwellen) und Produktion (z. B. in Echtzeit durch mehrere Gesprächspartner) gesprochener Sprache resultieren in bestimmten Bearbeitungsprozessen durch die Zuhörer und Sprecher. Es wurde vermutet, dass gerade Prozesse, die im Sinne von Formulierungsabbrüchen oder Konstruktionsmischungen Ausdruck von Schwierigkeiten bei der Versprachlichung sind, auch für den Zuhörer schwieriger zu erfassen sein würden. Die identifizierten Merkmale sollen im Rahmen korpusanalytischer Methoden untersucht werden und sind überwiegend auf lexikalischer und syntaktischer Ebene zu finden. Aus diesem Grund lag bei den Analysen für diese Arbeit der Schwerpunkt auf Lexik und Syntax (vgl. Fiehler, 2005).

Aus psycholinguistischer Perspektive wurde in Kapitel II.2. beschrieben, wie Sprachverstehen im Sinne von Satz- und Textverstehen erfolgt. Dabei wurde beispielhaft das Modell des Textverstehens von Kintsch und van Dijk (1978) vorgestellt. Verschiedene Theorien und Modelle für Sprach-, Gesprächs- und Hörverstehenskompetenz sollten verdeutlichen werden, welche Anforderungen an die Schüler im Bereich des Hörverstehens gestellt werden und inwieweit Schülerfähigkeiten deshalb ggf. von den Itemmerkmalen abhängen.

Anschließend wurde in Kapitel II.3. auf den Stand der Forschung zur Hörverstehensdiagnostik in der Psychologie eingegangen. Es wurde dargestellt, in wie weit Wahrnehmung, Aufmerksamkeit und Informationsverarbeitung für das Zuhören und Hörverstehen relevant sind und sich die Verarbeitung akustischen Inputs von der Rezeption geschriebener Sprache unterscheidet. Aufgrund der hohen Präsentationsgeschwindigkeit können auditiv z. B. häufig nicht alle Informationen erfasst werden. Es wurden Theorien vorgestellt, die Aufschluss darüber geben, nach welchen Kriterien die Informationsauswahl erfolgt. Ferner wurde auf die Organisation und Speicherung der Informationen im Langzeitgedächtnis eingegangen. Für die Speicherung von Wissen ist zunächst die Wahrnehmung bestimmter Informationen aus dem Lautfluss ausschlaggebend, aber auch der Umfang dessen, wie stark die neuen Informationen an bereits vorhandenes Wissen angegliedert werden können. Die ausgewählten Informationen werden im Gedächtnis enkodiert und gespeichert, wobei auf Bedingungen eingegangen wird, die für das Behalten und den Abruf dieser Informationen eine Rolle spielen (vgl. Anderson, 2001).

Relevant für diese Arbeit war insbesondere die Integration der Überlegungen in den aktuellen Stand der Forschung in der Deutschdidaktik. Hier wurde der Fokus erst in neuerer Zeit dem Kompetenzbereich „Zuhören" zugewendet (z. B. Bernius & Imhof, 2010). Meist wird er noch als Teilbereich von Unterrichtsgesprächen, Diskussionen und Debatten betrachtet (vgl. Vogt, 2002), findet vereinzelt aber auch im Sinne einer Hörerziehung Beachtung (z. B. Kinder- /Jugendliteratur und Medien, in Forschung, Schule und Bibliothek/Kjl&m, 2008). Um die Bedeutung des Hörverstehens für den Unterricht zu erfassen, wurden in Kapitel II.4. zunächst

systematische bundesländerspezifische Lehrplanvorgaben mit den länderübergreifenden Bildungsstandards verglichen. Unterschiede zu den Ansätzen des Hörverstehens in den Fremdsprachen müssen im Vergleich zu den herausgearbeiteten Systematisierungen im muttersprachlichen deutschen Bereich gebracht werden. Ein Überblick über Arbeiten im Rahmen der Literalitätsforschung und großer Schulleistungsstudien, wie der DESI-Studie, macht deutlich, an welcher Stelle die Arbeiten des IQB ansetzen. Dabei dienen die bei DESI verwendeten dreistufigen Kategorien der Textschwierigkeit Lexik, Syntax und Vertextung (Klieme et al., 2003: 38ff) zum Kompetenzbereich „Lesen" unter anderem als Basis für die Untersuchungen im Bereich des Hörverstehens, mit dem Ziel geeignete Merkmale zu übernehmen, ggf. zu adaptieren und zu ergänzen.

Aus der Analyse der besonderen Eigenarten der gesprochenen Sprache sowie der psychologischen, psycholinguistischen und didaktischen Grundlagen des Hörverstehens ergaben sich bestimmte Merkmale in den IQB-Hörstimuli und den Items, von welchen vermutet wurde, dass sie sich als Prädiktoren für die Itemschwierigkeit erweisen. (vgl. Kapitel II.5.) Unterstützt wurden diese Vermutungen durch Studien im Bereich des fremdsprachlichen Leseverstehens (z. B. Katz et al. 1990; Keshavarz et al., 2007), des fremdsprachlichen Hörverstehens (z. B. Nissan et al., 1996; Pallier et al., 1997) und im Leseverstehen in der Erstsprache (Willenberg, 1995, 2007; Artelt & Schlagmüller, 2004). Untersuchungen an den IQB-Hörverstehensaufgaben aus dem Primarbereich Deutsch ergänzen die Ausführungen (vgl. Bremerich-Vos et al., 2009; Böhme et al. 2010).

Die Analysen wurden an Aufgaben durchgeführt, die auf der Grundlage der Bildungsstandards zur Überprüfung der Hörverstehenskompetenz im Fach Deutsch von 16 Fachlehrkräften aus allen Bundesländern unter der Leitung des IQB entwickelt wurden. Die Aufgaben für den Kompetenzbereich „Zuhören" operationalisieren Aspekte des Hörverstehens, wie sie von den Lehrplänen der Länder und den Bildungsstandards für den Hauptschulabschluss (Jahrgangsstufe 9) und den Mittleren Schulabschluss (Jahrgangsstufe 10) gefordert werden. Die Bildungsstandards wurden 2003 und 2004 von der Kultusministerkonferenz eingeführt und sind seitdem für alle Schularten verbindlich. (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, 2004) Das gesamte Itemmaterial besteht im Bereich Hörverstehen aus 30 Aufgabenstämmen mit insgesamt 195 Items. Jede Aufgabe besteht aus einem Stimulus, der ggf. auch in mehreren Teilen präsentiert wird, zu dem Items zu bearbeiten sind. Die Stimuli lagen zunächst nur phonisch vor und wurden dann transkribiert. Sie sind zum Teil konzeptionell schriftlich, d. h. sie weisen mehr Merkmale geschriebener Sprache als gesprochener Sprache auf.

Die Analyse der Daten gliederte sich entsprechend den Zielen dieser Arbeit in vier Bereiche: Analysen bezogen erstens auf die einzelnen Merkmale (Kapitel IV-1. „Explorative Dimensionsanalysen von Item- und Stimulusmerkmalen" und Kapitel IV-2. „Zusammenhangsanalysen") und zweitens auf Merkmalsgruppen (Kapitel IV-3. „Regressionsanalysen"), drittens Analysen in Bezug auf eine Ratingskala und viertens Analysen mit den Angaben der Probanden.

Hypothese 1: Einfluss einzelner Item- und Stimulusmerkmale auf die Item- bzw. Aufgabenschwierigkeit

Die Aufgaben wurden auf Item- und auf Stimulus-Ebene nach linguistischen Merkmalen, aber auch unter inhaltlichen und kognitionspsychologischen Gesichtspunkten analysiert. Die grammatische Struktur wurde an Stimulussegmenten untersucht, für die Untersuchung des Wortschatzes wurden Kategorien wie Wortarten, Inhalts- bzw. Funktionswörter, Negationen und Wiederaufnahmen berücksichtigt. Inhaltlich wurden Merkmale wie Thema und Aufbau der Stimuli erfasst. Zu den untersuchten kognitionspsychologischen Merkmalen gehören Faktoren wie die zur Itembeantwortung notwendigen kognitiven Operationen. Aufgrund ihrer großen Anzahl wurden die Merkmale in einer geringen Anzahl von Prädiktorgruppen angeordnet. Dabei wurden einerseits inhaltliche Erwägungen berücksichtigt und andererseits faktorenanalytische Verfahren eingesetzt. Einige der Merkmale (z. B. „Notwendigkeit zur Inferenzbildung" oder „Plausibilität der Distraktoren") wurden von mehreren Ratern unabhängig voneinander ermittelt. Für diese Merkmale wurde die Raterübereinstimmung berechnet, um die Verlässlichkeit der Einschätzung zu prüfen. Obwohl die Rater vorab die Vorgehensweise beim Rating im Rahmen einer IQB-internen Schulung besprachen und genaue Kriterien zum Einstufen der Stimuli vorlagen, gab es zum Teil unterschiedliche Einstufungsergebnisse. Für noch stärker abgesicherte Ergebnisse wären Mehrfachratings für alle untersuchten Merkmale demnach unumgänglich. Im Idealfall könnte die Analyse der Stimuli sogar automatisiert mittels eines Computerprogramms erfolgen, um größtmögliche Objektivität zu gewährleisten. Derartige Untersuchungen wurden bereits mit englischen Korpora von Graesser et al. (2004) oder Crossley et al. (2006) mit dem Programm Coh-Metrix durchgeführt.

Auf Korrelationen basierende Methoden gaben Aufschluss über den Zusammenhang der verschiedenen linguistischen Variablen (vgl. Kapitel IV-1. „Explorative Dimensionsanalysen von Item- und Stimulusmerkmalen" und Kapitel IV-2. „Zusammenhangsanalysen"). Die explorativen Dimensionsanalysen wurden für die Stimulus- und die Itemmerkmale getrennt vollzogen. Die Korrelationen wurden mit den unkategorisierten und den kategorisierten Variablen berechnet. Um den Zusammenhang hoch mit einzelnen Merkmalen korrelierender Variablen auszuschließen, wurden zusätzlich Partialkorrelationen berechnet. Dabei zeigten sich die Effekte u. U. erst bei Analysen mit Teilstichproben. Die Mehrzahl der Merkmale erwies sich dabei nicht als signifikanter Prädiktor. Kritisch ist hierzu anzumerken, dass bei der Analyse der Teilstichproben zum Teil nur noch eine sehr geringe Anzahl von Items vorlag. In weiterführenden Untersuchungen wäre es daher lohnenswert, die entsprechenden Merkmale an einer größeren Itemstichprobe zu analysieren um die Aussagekraft der Ergebnisse zu erhöhen.

Insgesamt fiel auf, dass nicht nur der Einfluss der Stimulusmerkmale auf die Itemschwierigkeit, sondern auch der Einfluss der Itemmerkmale darauf deutlich geringer ausfiel, als der Einfluss der der Stimulusmerkmale auf die Aufgabenschwierigkeit. Von den insgesamt 43 untersuchten Stimulusmerkmalen zeigten nur 23 einen Zusammenhang mit der Schwierigkeit. Die meisten Merkmale lassen sich der Kategorie „Komplexität des Wortschatzes und sprachliche Mittel" zuordnen. Die Schwierigkeit eines Hörbeitrags scheint also stark von seiner sprachlichen Gestaltung beeinflusst zu werden. Andererseits waren die sprachlichen Variablen im Gegensatz zu stärker inhaltlichen Kriterien aber auch trennschärfer zu erfassen und relativ

einfach nachzuvollziehen. Dass auch inhaltliche Kriterien einen Einfluss auf die Schwierigkeit eines Hörbeitrags haben, zeigt sich bei der Betrachtung der Merkmale, die am höchsten mit der Schwierigkeit korrelieren. Von diesen insgesamt zwölf Merkmalen lassen sich drei der Kategorie „Inhaltlich-thematische Merkmale" zuordnen. Von den 18 untersuchten Itemmerkmalen korrelierten sieben signifikant mit der Schwierigkeit. Die meisten dieser Merkmale fielen in die Kategorie „Itemformat". Insbesondere das Format „Geschlossen- Ankreuzen" hing mit einer geringeren Itemschwierigkeit zusammen. Entscheidend war ferner der „Zeitpunkt der Itembearbeitung (ZIB)" und bei Multiple-Choice-Items die Plausibilität der Distraktoren. Für die kognitiven Anforderungen der Items konnten hingegen kaum Zusammenhänge mit der Schwierigkeit nachgewiesen werden.

Auf die Analysen wirkte erschwerend, dass nicht alle identifizierten Merkmale für alle untersuchten Stimuli und Items gleichermaßen relevant waren. So zählen zu den Stimuli beispielsweise sowohl Texte als auch Diskurse. Manche Merkmale, wie z. B. „Jugendsprache/ Umgangssprache (RHE4)", treten jedoch fast ausschließlich in Diskursen auf, und haben in Bezug auf Texte kaum Aussagekraft. Ähnliche Beispiele lassen sich für literarische und nicht-literarische Stimuli finden. Obwohl auch in dieser Arbeit die Stimuli nach entsprechenden Kriterien für die Analysen gruppiert wurden, wurden häufig aufgrund der insgesamt sehr geringen Stimulusanzahl Grenzen erreicht. Lohnenswert könnte daher in weiterführenden Arbeiten die Analyse einzelner Merkmale in Bezug auf größere Gruppen ähnlicher Stimulustypen (z. B. nur Diskurse, nur literarische Texte, etc.) sein.

Im Rahmen derartiger weiterführender Arbeiten wäre es auch sinnvoll, den Einfluss einzelner Merkmale genauer in kontrollierten Versuchsanordnungen zu untersuchen. Um den wechselseitigen Einfluss verschiedener Merkmale auf die Schwierigkeit auszuschließen, müsste an Items oder Stimuli ausschließlich ein Merkmal manipuliert werden und alle anderen Aspekte müssten konstant gehalten werden. In diesem Zusammenhang muss auch kritisch angemerkt werden, dass manche Merkmale, wie beispielsweise die Itemmerkmale „Hintergrundwissen (HGW)" oder „Typ der NI (TNI)" z. B. mit dem Subcode „Schlussfolgerungen" mehrere Subfacetten aufweisen, die jedoch für die Ratings in dieser Arbeit nicht genauer identifiziert werden konnten. So können Schlussfolgerungen notwendig sein um referenzielle Kohärenz herzustellen, kausale Zusammenhänge zu erkennen oder um die emotionalen Reaktionen im Stimulus auftretender Personen zu verstehen. Die Schlussfolgerungen können sich auf lokale oder globale Kohärenz beziehen. (vgl. Graesser et al., 1994; Long & Chong, 2001) Bei den Ratings in dieser Arbeit wurden die Schlussfolgerungen jedoch nicht genauer klassifiziert, sondern lediglich als solche eingestuft. Für genauere Analysen war das vorliegende Stimulusmaterial nicht ergiebig genug. Dennoch tritt so eine Unschärfe ein, da die Schwierigkeit der Stimuli durch ein Merkmal erklärt wird, das möglicherweise unterschiedliche Ausprägungen hat. In weiterführenden Arbeiten wären daher die von dieser Problematik betroffenen Merkmale nochmals genauer einzugrenzen und an geeignetem Stimulusmaterial vertieft zu untersuchen.

Hypothese 1 wird insofern bestätigt, als von den untersuchten 59 Merkmalen lediglich 14 keinen Effekt auf die Item- oder die Stimulusschwierigkeit zeigten. Hinsichtlich der Richtung ihres Einflusses wurde für 13 Merkmale eine nicht zutreffende Prognose formuliert.

Hypothese 2: Einfluss von Merkmalsgruppen auf die Item- und Aufgabenschwierigkeit
Um die Stärke der Merkmalseinflüsse auf die empirische Itemschwierigkeit zu berechnen, wurden die Merkmale gruppiert und einer linearen Regressionsanalyse unterzogen (vgl. Kapitel IV-3. „Regressionsanalysen"). Dabei wurden die folgenden vier Merkmalsgruppen für die Stimulus- und die Itemvariablen betrachtet:

1. „Die am höchsten mit der Aufgabenschwierigkeit korrelierenden Merkmale",
2. „Faktorenanalytisch bestimmte Merkmalsgruppen",
3. „Thematisch gebildete Variablengruppen" und
4. „Ausgewählte Stimulusmerkmale den Zusammenhangsanalysen".

Die Stimulusmerkmale wurden in Bezug auf die Stimulus- und die Itemschwierigkeit untersucht, die Itemmerkmale nur in Bezug auf die Itemschwierigkeit. In allen Fällen fiel die geleistete Varianzaufklärung durch die Stimulusmerkmale auf Itemebene deutlich geringer aus als auf Aufgabenebene, wobei jedoch die Stimulusmerkmale auf Itemebene deutlich geringere Eta^2-Werte aufweisen als die Itemmerkmale auf Itemebene. Nach Böhme et al. (2010: 100) ist bei der Untersuchung einer Vielzahl an Merkmalen an stark heterogenen Aufgaben eine Varianzaufklärung von mehr als 50% kaum zu erwarten. Die enorm hohe Varianzaufklärung von 87.9 bzw. 90.1% in dieser Arbeit muss relativiert gesehen werden, da bei vielen Variablen und relativ wenigen Fällen (33 Aufgaben) Overfitting eintreten kann. Das bedeutet, dass nur wenige Variablen signifikante Ausprägungen zeigen, die Varianzaufklärung insgesamt jedoch sehr hoch ausfällt. Dabei handelt es sich in gewisser Weise um einen künstlichen Effekt und der Einfluss separater starker Variablen sollte höher gewertet werden, als der Einfluss einer Gruppe. Aus diesem Grund ist es sinnvoll, sich die Einzelmerkmale der beiden Gruppen noch einmal genauer anzusehen.

Die folgenden Ergebnisse wurden im Rahmen dieser Arbeit erzielt:
1. Die Durchführung einer linearen Regressionsanalyse für die am höchsten mit der Schwierigkeit korrelierenden Merkmale ergab für die zwölf ausgewählten Stimulusmerkmale einen Anteil an der Gesamtvarianzaufklärung von 87.9%. Die drei am höchsten mit der Schwierigkeit korrelierenden Itemmerkmale erklärten zusammen 41.3%.

2. Eine Gruppierung der Merkmale aufgrund faktorenanalytischer Kriterien führte nicht zur Identifikation besonders vorhersagestarker Variablengruppen. Die faktorenanalytisch bestimmten Merkmalsgruppen liegen mit einer Varianzaufklärung von maximal 30.8% im moderaten Bereich. Am aussagekräftigsten sind hier für die Stimulusmerkmale die Merkmalsgruppen „Merkmale zur Beschreibung der Länge und Komplexität des Stimulus" ($r^2 = 30.8$) und „Sprachliche Merkmale zur quantitativen Beschreibung der Stimuli" ($r^2 = 23.3$). Bei den Itemmerkmalen erklären die „Merkmale zur Einschätzung der Plausibilität der Distraktoren" den größten Varianzanteil (24.1%).

3. Eine nach thematischen Überlegungen vorgenommene Gruppierung der Merkmale führte zu einer höheren Varianzaufklärung im Bereich der Stimulusmerkmale. Hier erklärt die Gruppe

„Komplexität des Wortschatzes und sprachliche Merkmale“ 63.2%, die Gruppe „Thematische Merkmale“ 30.5%. Im Bereich der Itemmerkmale liegt die Varianzaufklärung bei 6.7% (Gruppe „Thematische Merkmale).

4. Die aus den vorhergehenden Regressionsanalysen ausgewählten Itemmerkmale eigneten sich im Vergleich zu den anderen Gruppen der Itemmerkmale am besten dazu, die Itemschwierigkeit vorherzusagen. Zwölf Stimulusmerkmale erklären hier zusammen 90.1% Varianz. Es handelt sich dabei im Wesentlichen um sprachliche Merkmale bzw. um Merkmale, die für Diskurse typisch sind. Bei den Itemmerkmalen erklären die vier Merkmale 45% der Gesamtvarianz. Die größte Aussagekraft haben im Bereich der Itemmerkmale Formatvorgaben und die Varianz des Bearbeitungszeitpunktes. Dies kann damit zusammenhängen, dass die stärker inhaltlich motivierten Aspekte eher durch ein stärker kontrolliertes Versuchssetting zu beobachten gewesen wären. Da die Stimuli in Länge und Art stark variieren und zu jedem Stimulus sehr unterschiedliche Items formuliert wurden, lassen sich differenzierte Aussagen zum Einfluss bestimmter, insbesondere stärker inhaltlich verankerter Merkmale auf die Schwierigkeit kaum generalisierend machen.

Hypothese 2 wurde demnach bestätigt. Statt einer extensiven Analyse der Stimuli mittels einer Vielzahl von Variablen bietet sich eine gezielte Untersuchung nach einigen wenigen Merkmalen an. Die vorhersagestärksten Gruppen[25] beinhalten beide die Merkmale „Worthäufigkeit (GWS)“, „Sprechgeschwindigkeit (SGS)“, „Referenz-Aussage-Konstruktionen (STR1)“, „Adjazenzstrukturen (STR3)“, „Frage/Impuls/Themensetzung (REL1)“, „Wiederaufnahmen (WIE)“ und „Negationen (NEG)“. Während die erste Gruppe jedoch durch die Merkmale „Anzahl der Stimuluspräsentationen (AHO)“, „Hörkontext (HKO)“, „Stimulusfunktion (TFU)“ und „Hintergrundwissen (WEL)“ zusätzlich den Schwerpunkt auf den inhaltlichen Rahmen der Beiträge legt, liegt der Fokus der zweiten Gruppe durch die Variablen „Lange Wörter (PLW)“, „Verben (VER)“, „Anteil der Propositionen (PRW)“, „Deixis (DEI)“ und „Anzahl der Sprecher (ASP)“ eher auf sprachlichen Besonderheiten der Hörbeiträge. Bis auf wenige Ausnahmen korrespondieren die Korrelationen der einzelnen Merkmale mit ihren beta-Gewichten. Die Merkmale zeigen also einzeln einen ähnlichen Einfluss auf die Schwierigkeit wie in der Gruppe.

Zu prüfen wäre in weiteren Arbeiten, wie praktikabel diese Merkmale für ungeschulte Personen sind, um die Schwierigkeit von Hörtexten einzuschätzen. Gerade die Einschätzung eines Merkmals wie der „Worthäufigkeit“ oder dem „Anteil der Propositionen“ ist mit relativem Aufwand verbunden und eine reliable Erfassung dieser Merkmale muss möglicherweise vorab geschult werden. Hingegen lässt sich die mittlere Sprechgeschwindigkeit eines Hörstimulus relativ einfach ermitteln, wenn das Transkript dazu vorliegt, die Wortzahl also bekannt ist. Zu prüfen wäre weiterhin, ob die Untersuchung der aufgezählten Merkmale bei Lehrkräften auf Akzeptanz stößt, da nur punktuelle, überwiegend sprachliche Eigenarten der Stimuli erfasst werden und inhaltliche Faktoren oder Kohärenzaspekte kaum eine Rolle spielen. Im Bereich der Lesbarkeitsformeln führte genau diese Problematik zum Vorwurf der mangelnden Face-Validität (vgl. Klein-Braley, 1994: 189).

25 „Die am höchsten mit der Schwierigkeit korrelierenden Merkmale“ und „Ausgewählte Merkmale aus den vorhergehenden Analysen“

Hypothese 3: Globalurteil und Einschätzung der Stimuli mittels einer Ratingskala
Das Globalurteil der Aufgabenentwickler hinsichtlich der Stimulus- und der Itemschwierigkeit ist weniger dazu geeignet, die empirische Schwierigkeit vorherzusagen. Die Aufgabenentwickler wurden gebeten anzugeben, wie schwierig die jeweiligen Stimuli wahrscheinlich für die Schüler sein werden. Dabei orientierte sich jeder Aufgabenentwickler sicherlich implizit an den ihm bekannten Schülergruppen. Aufgabenentwickler, die an einer Realschule unterrichten, stuften die Stimuli also für „ihre" Schüler ein, auch wenn der Stimulus für den Mittleren Schulabschluss vorgesehen war und damit auch für Schüler der Gymnasien. Auf der Grundlage dieses Ergebnisses ist von einer einfachen Globaleinschätzung mit den Stufen „leicht", „mittel", „schwierig" zur Beurteilung der Schwierigkeit eines Hörstimulus abzuraten. Dieses Ergebnis kann damit zusammenhängen, dass jeder Aufgabenentwickler die Schwierigkeit zwar gut für die eigene Referenzgruppe vorhergesagt hätte, jedoch weniger gut für eine unbekannte Gruppe, deren Leistungsstand mittels den Vorgaben der Bildungsstandards eingestuft werden sollte. Es wäre zukünftig zu untersuchen, ob Lehrer für eine ihnen bekannte Testgruppe die Schwierigkeit besser vorhersagen können.

Für eine genauere Einschätzung der Stimuli wurde ein entsprechender Fragebogen mit Kriterien im Sinne subjektiver Ratings entwickelt. Der Fragebogen fokussiert einerseits auf der Vertrautheit der Schüler mit den Stimuli („Vertrautheit mit dem Thema") und andererseits auf Kategorien, zu denen auch quantitative Analysen durchgeführt werden, wie Wortschatz, Grammatik, Kohäsion und Kohärenz. In Anlehnung an die Studien von Carroll (1964) wurden Merkmale mit mindestens einem Adjektivpaar zur Beschreibung in den IQB-Fragebogen übernommen. Für alle Ratings wurde die Beurteilerreliabilität ermittelt. Ferner wurden Mittelwertsvergleiche durchgeführt und eine Reliabilitätsanalyse vorgenommen.

Die Vorhersage der Stimulusschwierigkeit mittels der durch den Lehrerfragebogen erhaltenen Einschätzungen funktionierte für die meisten Variablen sehr gut, mit einer mittleren Varianzaufklärung von 16% und einem Range von 38. Der Fragebogen bezieht sich mehrfach mit unterschiedlichen Variablen auf das gleiche Merkmal[26] und könnte deshalb auf die am höchsten mit der Stimulusschwierigkeit korrelierenden Merkmale gekürzt werden. Für eine Kurzversion sollten die Variablen „Wortschatz (WFA)" mit den Ausprägungen „fachspezifisch" – „alltäglich", „Gesamteindruck (GIU)" mit den Ausprägungen „interessant" – „uninteressant" und „Ton (TON)" mit den Ausprägungen „persönlich/gefühlsbetont" – „unpersönlich/ sachlich" eingesetzt werden. Die Langversion des Fragebogens sollte zusätzlich die Merkmale „Vertrautheit mit dem Thema (VTH)" mit den Ausprägungen „vertraut" – „gar nicht vertraut", „Grammatik (GRA)" mit den Ausprägungen „schwierig" – „einfach", „Gesamteindruck (GEU)" mit den Ausprägungen „abwechslungsreich" – „eintönig" sowie „Ausdrucksweise (AGU)" mit den Ausprägungen „gewählt" – „umgangssprachlich" einschätzen lassen.

Eine Vorabanalyse der im Unterricht oder für die Testkonstruktion einzusetzenden Hörbeiträge mittels des Lehrerfragebogens erscheint praktikabel und führt zu einer ersten Schwierigkeitseinschätzung der Stimuli. In weiterführenden Arbeiten müsste dieser Fragebogen an

26 z. B. Merkmal „Gesamteindruck" mit den Adjektivpolen „interessant" – „uninteressant", „elegant" – „unbeholfen" und „abwechslungsreich" – „eintönig"

weiteren Stimuli nochmals validiert werden. Zu prüfen wäre auch, ob andere Adjektivpole als die gewählten (z. B. Gesamteindruck – „langweilig" vs. „spannend") ggf. zu höherer Varianzaufklärung führen. Wünschenswert wäre ferner eine Schärfung des Fragebogens für unterschiedliche Stimulusarten. Denkbar wäre es beispielsweise, Merkmale bzw. Adjektivpole zu integrieren, die die literarische Qualität oder den Grad der Mündlichkeit von Stimuli erfassen.

Hypothese 3 wurde insofern bestätigt, als eine Vorhersage der Schwierigkeit mittels eines Fragebogens zur Befragung von Lehrkräften zu guten Ergebnissen führte. Widerlegt wurde jedoch die Annahme, auch eine Globaleinschätzung der Aufgabenentwickler könnte die Schwierigkeit der Stimuli vorhersagen.

Hypothese 4: Personenmerkmale
Untersucht wurde, in welcher Weise personenbezogene Merkmale die Testleistung und damit die Schwierigkeit von Items und Stimuli beeinflussen. Beurteilt wurden von den Schülern der Interessantheits- bzw. Bekanntheitsgrad der Stimuli sowie die Präsentationsqualität. Zusätzlich liegen Angaben zu den Sprachkenntnissen der Schüler und die Ergebnisse eines Testteils zur Arbeitsgedächtniskapazität vor.

Hypothese 4 wurde widerlegt, da für keine der vier Variablen „Vertrautheit mit dem Thema (VTS)", „Motivation/Interesse (MOT)", „Bekanntheitsgrad des Stimulus (BEK)" und „Verständlichkeit (VST)" ein systematischer signifikanter Zusammenhang mit den Leistungsdaten der Schüler in allen Schulformen bzw. für die Subpopulationen „Hauptschüler" und „Gymnasiasten" nachgewiesen werden konnte. Die Schülereinschätzungen zum Bekannt- und Interessantheitsgrad der Stimuli sowie zu deren Verständlichkeit eigneten sich bei den vorliegenden Aufgaben nicht dazu, die Schwierigkeit vorherzusagen. Dies kann damit zusammenhängen, dass den Schülern die Testsituation bewusst war und sie unabhängig davon, wie interessant sie die Aufgaben fanden, versuchten gut abzuschneiden. Dass auch für die Variable „Verständlichkeit (VST)" kein Zusammenhang mit der Schwierigkeit festgestellt werden konnte, kann daran liegen, dass die Hörstimuli offenbar relativ gut für alle teilnehmenden Schüler zu verstehen waren. In weiterführenden Studien müsste nochmals der Einfluss dieser Variablen, beispielsweise auch auf die Mitarbeit der Schüler im Unterricht überprüft werden.

Sowohl für die genannten vier Variablen als auch für den Lehrerfragebogen könnte es lohnenswert sein, das Antwortverhalten der Schüler bzw. Lehrer im Umgang mit den Ratingskalen genauer zu untersuchen. Systematische Tendenzen Ratingskalen zu verwenden, indem beispielsweise immer die Extrempole ausgewählt werden, eine Tendenz zur Mitte zu beobachten ist oder die gegebenen Antworten stets um Zustimmung bemüht sind, können das Gesamtergebnis verzerren und die Antworten von Subpopulationen unzutreffend abbilden. (vgl. Bolt & Johnson, 2009) DIF-Effekte können also nicht nur auf Unterschiede in den Items, sondern auch auf unterschiedlichen Umgang mit den Ratingskalen zurückzuführen sein. In weiterführenden Analysen müssten deshalb auch die Antworttendenzen der befragten Gruppen untersucht werden. Bolt und Johnson (2009) schlagen dafür ein multidimensionales Item- Response Modell vor, mit dem der Einfluss der Antworttendenzen auf DIF von anderen Faktoren getrennt werden kann.

Im Gegensatz dazu korrelierten die Ergebnisse des Tests zur Arbeitsgedächtniskapazität in allen Fällen signifikant ($p < 0.05$) mit den Leistungsdaten. Die Zusammenhänge waren unabhängig von der Länge der Aufgaben und der empirischen Item- bzw. Aufgabenschwierigkeit. Schulformbezogene Analysen ließen stärkere Zusammenhänge im Bereich der Hauptschule erkennen, wobei die Korrelationen aufgabenbezogen waren. Die Arbeitsgedächtniskapazität der Schüler konnte also nicht generell mit den Leistungen bei den Testaufgaben in Verbindung gebracht werden. Höhere Arbeitsgedächtniskapazität ist bei der Beantwortung von Fragen zu Hörstimuli sicherlich hilfreich. Die erhaltenen Ergebnisse legen jedoch nahe, dass die IQB-Testaufgaben zum Hörverstehen andere Kompetenzen erfordern und hohe Arbeitsgedächtniskapazität allein nicht genügt, um die Aufgaben erfolgreich zu bearbeiten.

Auch das Merkmal „Sprachkenntnisse" korrelierte mit den Leistungsdaten der Schüler. Von den Schülern wurde erfragt, wie häufig zu Hause deutsch gesprochen wird bzw. welche Sprache als Muttersprache gelernt wurde. Bessere Sprachkenntnisse, ausgedrückt durch die Angabe, zu Hause immer deutsch zu sprechen und deutsch als Muttersprache gelernt zu haben, hingen erwartungsgemäß mit besseren Testergebnissen zusammen. Dieser Zusammenhang variierte jedoch auf Aufgabenebene und war im Bereich der Hauptschule etwas stärker als im Gymnasium zu beobachten. Dies kann damit zusammenhängen, dass in der Hauptschule mehr Schüler sind, die deutsch u. U. verhältnismäßig schlecht sprechen. Schüler mit Schwierigkeiten in der deutschen Sprache, sind an den Gymnasien hingegen seltener zu finden. Schüler, die ein Gymnasium besuchen und angaben, zu Hause häufig nicht deutsch zu sprechen und deutsch auch nicht als Muttersprache gelernt zu haben, sind i. d. R. bilingual und sprechen sowohl ihre Muttersprache als auch Deutsch fließend.

Fazit und Ausblick

Die Untersuchung von Einzelmerkmalen und Merkmalsgruppen führte zu relativ starken Prädiktoren der Schwierigkeit auf Item- und auf Stimulusebene. In weiterführenden Arbeiten müsste nun der Einfluss dieser Merkmale auf die Schwierigkeit in stärker kontrollierten Experimenten untersucht werden. Dabei gilt es auch, generalisierbare Stufengrenzen für die Erfassung der Merkmale zu definieren.

Der Einfluss dieser Merkmale sollte bei der Erstellung bzw. Überarbeitung von entsprechenden Kompetenzstufenmodellen berücksichtigt werden. Für die Itemgenerierung durch geschulte Entwicklergruppen eignen sich die gefundenen Merkmale jedoch nur bedingt, da sie z. T. sehr aufwändig zu erheben sind. Die Merkmale des Lehrerfragebogens sind für die Auswahl geeigneter Stimuli für die Aufgabenentwicklung und die Unterrichtspraxis hingegen handhabbarer. Allerdings stoßen diese Merkmale aufgrund ihres häufig sehr globalen Charakters und der strengen, vorgegebenen Eingrenzung auf zwei Adjektivpole an ihre Grenzen, wenn mit ihrer Hilfe ausgewählte Materialien in ihrer Schwierigkeit für bestimmte Zielgruppen modifiziert werden sollen. Hier können die Einzelmerkmale bzw. Merkmalsgruppen mehr Hilfestellungen geben.

Von den untersuchten Personenmerkmalen zeigten lediglich die Höhe der Arbeitsgedächtniskapazität und der Grad der Sprachbeherrschung signifikante Zusammenhänge mit den Schülerleistungen. Für die Höhe des Vorwissens, der Motivation und der Grad der Verständlichkeit der Stimuli konnte jedoch kein Einfluss auf die Schülerleistungen im Large-Scale-Assessment nachgewiesen werden. Es ist allerdings davon auszugehen, dass die genannten Variablen im Rahmen des Unterrichts einen größeren Stellenwert einnehmen. In weiteren, u. U. stärker qualitativ angelegten Untersuchungen wäre deshalb nochmals gezielt dem Einfluss der genannten Merkmale nachzugehen.

Die gefundenen Merkmale geben erste Anhaltspunkte zur Dimensionalität des Konstrukts „Zuhören". Beispielsweise scheinen sich die Merkmale auf eine eher inhaltliche und eher sprachliche Dimension zu weisen. Forschungsarbeiten zur Dimensionalität von Sprachkompetenz gibt es dazu bisher beispielsweise von Song (2008), Leucht et al. (2010) und Bremerich-Vos et al. (2009). Die Annahme, Zuhören sei eine einzige, klar separierbare Dimension, Subfacette einer allgemeinen Sprachkompetenz oder in weitere Teilkompetenzen unterscheidbar, hat Auswirkungen auf die zu den Stimuli entwickelten Items. Um die Kompetenz der Schüler auch in ihrer Dimensionalität korrekt abbilden zu können, ist eine weiterführende Analyse der Aufgaben mittels mehrdimensionaler IRT-Modelle notwendig. Ergebnisse stehen für den muttersprachlichen Bereich des Zuhörens in der Sekundarstufe I derzeit noch aus.

Literatur

A

Abram, M. J. & Dowling, W. D. (1979). How Readable are Parenting Books? *The Family Coordinator*, 28(3). 365 – 368.

Abrahamsen, E. & Shelton, K. (1989). Reading comprehension in adolescents with learning disabilities: semantic and syntactic effects. *Journal of Learning Disabilities*, 22, 569 - 572.

Adams, M. J. & Collins, A. (1979). A schema-theoretic view of reading. In R. O. Freedle (Eds.), *New directions in discourse processing*. Norwood, 1 - 22.

Ágel, V. & Hennig, M. (2007). Überlegungen zur Theorie und Praxis des Nähe- und Distanzsprechens. In V. Ágel & M. Hennig (Hrsg.), *Zugänge zur Grammatik der gesprochenen Sprache*. Tübingen, 179 - 214.

Alderson, J. C. (2000). *Assessing reading.* Cambridge.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference. The Experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly,* 3(1), 3 - 30.

Anderson, J. R. (1995). *Cognitive psychology and its implications* (4th ed.). New York.

Anderson, J. R. (2001). *Kognitive Psychologie.* Heidelberg/Berlin.

Artelt, C. & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen? Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, P. Stanat & W. Schneider (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz - vertiefende Analysen im Rahmen von PISA 2000.* Wiesbaden. 169 - 196.

Artelt, C. (2009). Diagnostische Urteile von Lehrkräften im Bereich der Lesekompetenz. In A. Bertschi-Kaufmann & C. Rosebrock (Hrsg.), *Literalität. Bildungsaufgabe und Forschungsfeld.* Weinheim. 125 - 136.

Artelt, C., Drechsel, B., Bos, W. & Stubbe, T. C. (2008). Lesekompetenz in PISA und PIRLS/IGLU - ein Vergleich. *Zeitschrift für Erziehungswissenschaft, Sonderheft 10,* 35 - 52.

Artelt, C., Stanat, P., Schneider, W., Schiefele, U. & Lehmann, R. (2004). Die PISA-Studie zur Lesekompetenz: Überblick und weiterführende Analysen. In U. Schiefele, C. Artelt, P. Stanat & W. Schneider (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz - vertiefende Analysen im Rahmen von PISA 2000.* Wiesbaden. 139 - 168.

Autorengruppe Bildungsberichterstattung (2010). Bildung in Deutschland 2010. Bielefeld.

B

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice. Designing and Developing Useful Language Tests.* Oxford/New York.

Bachman, L. F. (1991). *Fundamental Considerations in Language Testing*. Oxford.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19(4), 453 - 476.

Bachman, L. F., Davidson, F. G. & Foulkes, J. (1990). A comparison of the abilities measured by the Cambridge and Educational Testing Service EFL Test Batteries. *Issues in Applied Linguistics*, 1, 30 - 55.

Bacon, S. M. (1992). Authentic Listening in Spanish: How Learners Adjust Their Strategies to the Difficulty of the Input. *Hispania*, 75(2), 398 - 412.

Baddeley, A. D. & Hitch, G. J. (1974). Working memory. In G. A. Bower (Eds.), *Recent advances in learning and motivation*, 8, 47 - 90.

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

Baddeley, A. D. (1995). Working memory or working attention? In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness & control.* Oxford, 152 - 170.

Baddeley, A. D. (2002). Is Working Memory Still Working? *European Psychologist,* 7(2), 85 - 97.

Bae, J. & Bachman, L. F. (1998). A latent variable approach to listening and reading. Testing factorial invariance across two groups of children in the Korean/English Two-way Immersion program. *Language Testing*, 15(3), 380 - 414.

Banyard, P. & Hayes, N. (1995). Denken und Problemlösen. In J. Gerstenmaier (Hrsg.), *Einführung in die Kognitionspsychologie.* München, 121 - 152.

Bärenfänger, O. (2002). Merkmals- und Prototypensemantik: Einige grundsätzliche Überlegungen. *Linguistik online*, 12/3. Zugriff am 13.11.2009 von http://www.linguistik-online.de/12_02/baerenfaenger.html

Baumert, J. & Schümer, G. (2001). Familiäre Lebensverhältnisse, Bildungsbeteiligung und Kompetenzerwerb. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen, 323 - 407.

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.) (2000a), *Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 1: Mathematisch-naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit*. Opladen.

Baumert, J., Bos, W. & Lehmann, R. (Hrsg.) (2000b), *Dritte Internationale Mathematik- und Naturwissenschaftsstudie: Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 2: Mathematisch-naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit.* Opladen.

Baumert, J., Lehmann, R., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., Köller, O. & Neubrand, J. (1997). *TIMSS: Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich.* Opladen.

Bayerischen Staatsministeriums für Unterricht und Kultus (Hrsg.). *Grund- und Hauptschule: Lehrpläne / Standards Hauptschule*; Zugriff von http://www.isb.bayern.de/isb/index.asp?MNav=3&QNav=4&TNav=0&INav=0&Fach=&LpSta=6&STyp=27am 03.12.10

Bayerischen Staatsministeriums für Unterricht und Kultus (Hrsg.). *Gymnasium: Lehrpläne / Standards Gymnasium G8*; Zugriff von http://www.isb.bayern.de/isb/index.asp?MNav=6&QNav=4&TNav=0&INav=0&Fach=&STyp=14&LpSta=6 am 03.12.10

Bayerischen Staatsministeriums für Unterricht und Kultus (Hrsg.). *Realschule: Lehrpläne / Standards Realschule R6*; Zugriff von http://www.isb.bayern.de/isb/index.asp?MNav=5&QNav=4&TNav=0&INav=0&Fach=&LpSta=6&STyp=5 am 03.12.10

Beck, B. & Klieme, E. (Hrsg.) (2007), *Sprachliche Kompetenzen: Konzepte und Messung - DESI-Studie (Deutsch-Englische-Schülerleistungen-International)*. Weinheim.

Becker-Mrotzek, M. & Meier, C. (2002). Arbeitsweisen und Standardverfahren der Angewandten Diskursforschung. In G. Brünner, R. Fiehler & W. Kindt (Hrsg.), *Angewandte Diskursforschung*, Band 1. Radolfzell, 18 - 45.

Becker-Mrotzek, M. & Vogt, R. (2001). Unterrrichtskommunikation. *Linguistische Analysemethoden und Forschungsergebnisse*. Tübingen.

Becker-Mrotzek, M. (2003). Mündlichkeit – Schriftlichkeit – Neue Medien. In U. Bredel, H. Günther, P. Klotz, J. Ossner & G. Siebert-Ott (Hrsg.), *Didaktik der deutschen Sprache*, Band 1. Paderborn, 69 - 89.

Becker-Mrotzek, M. (2009). Mündliche Kommunikationskompetenz. In M. Becker-Mrotzek (Hrsg.), *Mündliche Kommunikation und Gesprächsdidaktik*. Baltmannsweiler, 66 - 83.

Berkemeyer, N. & Bos, W. (2009). Professionalisierung im Spannungsfeld externer und interner Evaluation. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus & R. Mulder (Hrsg.), *Lehrerprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung.* Weinheim und Basel, 529 - 541.

Bernius, V. & Imhof, M. (2010). Zuhörkompetenz in Unterricht und Schule. *Beiträge aus Wissenschaft und Praxis.* Göttingen.

Best, R., Ozuru, Y., Floyd, R. G. & McNamara, D. S. (2006). Children's text comprehension. Effects of genre, knowledge, and text cohesion. In S. A. Barab, K. E. Hay & D. T. Hickey (Eds.), *Proceedings of the Seventh International Conference of the Learning Sciences*. Mahwah, NJ, 37 - 42.

Best, R., Rowe, M., Ozuru, Y. & McNamara, D. (2005). Deep-Level Comprehension of Science Texts. The Role of the Reader and the Text. *Topics Language Disorders*, 25(1), 65 - 83.

Biemiller, A. & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition Journal of Educational *Psychology*, 93, 498 – 520.

Birenbaum, M. & Tatsuoka, K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11(4), 385 - 395.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & R. M. Novick (Eds.), Statistical Theories of Mental Test Scores. *Reading*, MA, 397 - 472.

Böhme, K. & Robitzsch, A. (2009). Methodische Aspekte der Erfassung der Lesekompetenz. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik*. Weinheim/Basel, 259 - 298.

Böhme, K., Robitzsch, A. & Busé, A.-K. (2010). Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mithilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben. In V. Bernius & M. Imhof (Hrsg.), Zuhörkompetenz in Unterricht und Schule. *Beiträge aus Wissenschaft und Praxis*. Göttingen, 81 - 104.

Bolton, S. (Hrsg.) (2000), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests*. Köln.

Bonsen, M., Kummer, N. & Bos, W. (2008). Schülerinnen und Schüler mit Migrationshintergrund. In W. Bos, M. Bonsen, J. Baumert, M. Prenzel, C. Selter & G. Walther (Hrsg.), *TIMSS 2007. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich.* Münster, 157 - 175.

Borsboom, D., Mellenbergh, G. J. & Van Heerden, J. (2004). The concept of validity. *Psychology Review*, 111, 1061 - 1071.

Bortz, J. & Döring, N. (2002). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Heidelberg.

Bos, W. & Pietsch, M. (2005). *KESS 4. Kompetenzen und Einstellungen von Schülerinnen und Schülern Jahrgangstufe 4*. Hamburg.

Bos, W., Bonsen, M. & Gröhlich, C. (Hrsg.) (2010), KESS 7 - Kompetenzen und Einstellungen von Schülerinnen und Schülern an Hamburger Schulen zu Beginn der Jahrgangsstufe 7. *HANSE - Hamburger Schriften zur Qualität im Bildungswesen*, Bd. 5. Münster.

Bos, W., Lankes, E.-M., Prenzel, M, Schwippert, K., Walther, G. & Valtin, R. (Hrsg.) (2003). *Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich.* Münster.

Bos, W., Valtin, R., Voss, A., Hornberg, S. & Lankes, E.-M. (2007). Konzepte der Lesekompetenz in IGLU 2006. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E. M. Lankes, K. Schwippert & R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster.

Bremerich-Vos, A., Böhme, K. & Robitzsch, A. (2009). Sprachliche Kompetenzen im Fach Deutsch - Strukturanalysen und Validierungsbefunde. In D. Granzer, O. Köller, A. Bremerich- Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik.* Weinheim/Basel, 204 - 225.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiplechoice formats. *Journal of Educational Measurement*, 29(3), 253 - 271.

Brindley, G. & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, 19(4), 369 - 394.

Brinker, K. (2005). *Linguistische Textanalyse.* Berlin.

Bolt, D. M. & Johnson, T. R. (2009). Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style. *Applied Psychological Measurement*, 33(5), 335-352.

Buck, B. (1992). Listening Comprehension: construct validity and trait characteristics. *Language Learning*, 42(3), 313 - 357.

Buck, G. & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes of a free response listening test. *Language Testing*, 15 (2), 119 - 157.

Buck, G. (1991). The testing of listening comprehension: an introspective study. *Language Testing*, 8(1), 67 - 91.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11(2), 145 - 170.

Buck, G. (2001). *Assessing Listening.* Cambridge.

Bundesministerium für Bildung und Forschung (2007). *Schavan: „Alphabetisierung gehört zu wichtigsten Aufgaben des Bildungssystems"*. Pressemitteilung. Zugriff am 22.11.2008 von http://www.bmbf.de/press/2119.php

Bundesministerium für Bildung und Forschung (Hrsg.) (2005), *Expertise – Förderung von Lesekompetenz.* Bonn/Berlin.

Buse, A.-K. (2008). *Analyse schwierigkeitsgenerierender Aufgabenmerkmale zum Hörverstehen im Primarbereich*. Norderstedt.

Butler, F. A. & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas. *Language Testing*, 18(4), 409 - 427.

C

Cain, K. (2003). Text comprehension and its relation to coherence and cohesion in children's fictional narratives. *British Journal of Developmental Psychology*, 21, 335 – 351.

Cain, K., Oakhill, J. & Lemmon, K. (2004). Individual Differences in the Influence of Word Meanings From Context: The Influence of Reading Comprehension, Vocabulary Knowledge, and Memory Capacity. *Journal of Educational Psychology*, 96(4), 671 – 681.

Canale, M. & Swain, M. (1981). A theoretical framework for communicative competence. In A.S. Palmer, P. G. Groot & S. A. Trosper (Eds.), *The construct validation of tests of communicative competence. Washington*, D.C, 31 - 36.

Carrell, P. L. (1992). Awareness of Text Structure: Effects on Recall. *Language Learning*, 42(1), 1 - 20.

Carroll, J. B. (1964). Vectors of prose style. In T. A. Sebeok (Ed.), *Style in language*. Cambridge, Mass., 283 - 292.

Cassells, A. & Green, P. (1995). Wahrnehmung. In J. Gerstenmaier (Hrsg.), *Einführung in die Kognitionspsychologie*. München, 41 - 78.

Cassells, A. (1995). Erinnern und Vergessen. In J. Gerstenmaier (Hrsg.), *Einführung in die Kognitionspsychologie*. München. 153 - 194.

Cervantes, R. & Gainer, G. (1992). The effects of syntactic simplification and repetition on listening comprehension. *TESOL Quarterly*, 26, 767 - 771.

Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. In D. Olson, D. Torrance & A. Hildyard, A. (Eds.), *Literacy language and learning*. Cambridge, 105 - 122.

Chalifour, C. & Powers, D. (1989). The Relationship of Content Characteristics of GRE Analytical Reasoning Items to Their Difficulties and Discriminations. *Journal of Educational Measurement*, 26(2), 120 - 132.

Chang, F. R. (1980): Active memory processes in visual sentence comprehension: Clause effects and pronominal reference. *Memory & Cognition*, 8, 58 - 64.

Chase, C. L. (1964). Relative length of options and response set in multiple choice items. *Educational and Psychological Measurement*, 24, 861 - 866.

Chiang, H. H. & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL Quarterly*, 26, 345 - 374.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* New York.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). Applied multiple Regression/ Correlation. *Analysis for the Behavioural Sciences*. Hillsdale, NJ.

Conrad, L. (1989). The effects of time-compressed speech on native and EFL listening comprehension. *Studies in Second Language Acquisition*, 11, 1 - 16.

Cowan, N. (1995). *Attention and memory: An integrated framework*. Oxford: University Press.

Crossley, S. A., McCarthy, P. M. & McNamara, D. S. (2006). Discriminating between Second Language Learning Text-Types. In D. Wilson and G. Sutcliffe (Eds.), *Proceedings of the twentieth International Florida Artificial Intelligence Research Society Conference*. Menlo Park, California, 205 - 210.

Crothers, E. J. (1979). *Paragraph structure inference*. Norwood, NJ.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers in Bilingualism*, 19, 97 - 205.

D

Daneman, M. (1988). Word knowledge and reading skill. In M. Daneman, G., MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 6). San Diego, CA. 145 – 175

Daneman, M. & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450 - 466.

Daneman, M. & Merikle, Ph. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review,* 3(4), 422 - 433.

Daneman, M. & Tardif, T. (1987). Working memory and reading skill reexamined. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. Hove, 491 - 508.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Journal of Experimental Education* 56, 67 – 76.

de Beaugrande, R.-A., Dressler, W. U. (1981). *Einführung in die Textlinguistik*. Tübingen.

de Jonge, P. & de Jong, P. F. (1996). Working memory, intelligence and reading ability in children. *Personality and Individual Differences*, 21(6), 1007 - 1020.

Deppermann, A., Fiehler, R. & Spranz-Fogasy, T. (2006). Zur Einführung: Grammatik und Interaktion. In A. Deppermann, R. Fiehler & T. Spranz-Fogasy (Hrsg.), *Grammatik und Interaktion*. Radolfzell, 5 - 10.

DESI-Konsortium (Hrsg.) *(2007), Sprachliche Kompetenzen. Leistungsverteilungen und Bedingungsfaktoren. DESI-Ergebnisse Band 2*. Weinheim.

Deutsches PISA-Konsortium (Hrsg.) (2001), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen.

Deutsches PISA-Konsortium (Hrsg.) (2002), *PISA 2000. Die Länder der Bundesrepublik Deutschland im Vergleich.* Opladen.

Deutsches PISA-Konsortium (Hrsg.) (2004), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleichs*. Münster.

Deutsches PISA-Konsortium (Hrsg.) (2005), *PISA 2004. Die Länder der Bundesrepublik Deutschland im Vergleich*. Münster.

DiBello, L., Stout, W. & Roussos, L. (2007). Cognitive diagnosis Part I. In C. R. Rao & S. Shinharay (Eds.), Handbook of Statistics: Vol. 26. *Psychometrics*. Amsterdam.

Drum, P. A., Calfee, R. C. & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly,* 16, 486 - 514.

Duden (2005). *Die Grammatik* (7., völlig neu erarb. und erw. Aufl., überarb. Neudr.). Mannheim.

Duden (2006). *Die deutsche Rechtschreibung* (24., völlig neu bearb. und erw. Aufl.). Mannheim.

Dupuy, B. C. (1999). Narrow listening: an alternative way to develop and enhance listening comprehension in students of French as a foreign language. *System*, 27(3), 351 - 361.

E

EFA Global Monitoring Report Team (2008). *EFA Global Monitoring Report 2008 - Summary - Education for All by 2015. Will we make it?* Zugriff am 22.11.2008 von http://unesdoc.unesco.org/images/0015/001548/154820e.pdf

Ehlich, K. (1983). Text und sprachliches Handeln. Die Entstehung von Texten aus dem Bedürfnis nach Überlieferung. In A. Assmann, J. Assmann & C. Hardmeier (Hrsg.), *Schrift und Gedächtnis.* München, 24 – 43.

Embretson, S. E. & Wetzel, C. D. (1987). Component Latent Trait Models for Paragraph Comprehension Tests. *Applied Psychological Measurement*, 11(2), 175 - 193.

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

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

Embretson, S. E. (1999). Generating Items during Testing: Psychometric Issues and Models. *Psychometrika*, 64(4), 407 - 433.

Engel, U. (1988). *Deutsche Grammatik*. Heidelberg.

Engle, R. W., Kane, M. J. & Tuholski, S. W. (1999a). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake/P. Shah (Eds.), *Models of working memory. Mechanisms of active maintenance and executive control*, 102 - 134.

Engle, R. W., Tuholski, S. W., Laughlin, J. E. & Conway, A. R. A. (1999b). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128(3), 309 - 331.

Esser, H. (2006). *Sprache und Integration. Die sozialen Bedingungen und Folgen des Spracherwerbs von Migranten.* Frankfurt/Main: Campus.

Europarat (2001). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen.* Straßburg.

Evetts, J. & Gauthier, M. (2005). *Literacy Task Assessment Guide. National Literacy Secretariat.* Zugriff am 11.11.2008 von http://www.ibd.ab.ca/files/Literacy-task-assessmentguide.pdf

F

Fiehler, R. (2005). Gesprochene Sprache. In *Duden. Die Grammatik* (7., völlig neu erarb. und erw. Aufl., überarb. Neudr.). Mannheim.

Fiehler, R. (2009). Mündliche Kommunikation. In M. Becker-Mrotzek (Hrsg.), *Mündliche Kommunikation und Gesprächsdidaktik*. Baltmannsweiler, 25 - 51.

Fiehler, R., Barden, B., Elstermann, M. & Kraft, B. (2004). *Eigenschaften gesprochener Sprache*. Tübingen.

Field, J. (2003). Promoting perception: lexical segmentation in L2 listening. *ELT Journal*, 57(4), 325 - 334.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York, 131 - 155.

Flowerdew, J. & Miller, L. (1992). Student Perceptions, Problems and Strategies in Second Language Lecture Comprehension. *RELC Journal*, 23, 60 - 80.

Freedle, R. & Fellbaum, C. (1987). An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In R. Freedle & R. Duran (Eds.), *Cognitive and linguistic analyses of test performance*. Norwood, 162 - 192.

Freedle, R. & Kostin, I. (1993a). *The prediction of TOEFL reading item difficulty: implications for construct validity. Language Testing,* 10, 133 - 170.

Freedle, R. & Kostin, I. (1993b). *The Prediction of TOEFL® Reading Comprehension Item Difficulty for Expository Prose Passages for Three Item Types: Main Idea, Inference, and Supporting Idea Items*. ETS Research Report RR 93-13, Princeton, NJ.

Freedle, R. & Kostin, I. (1996). T*he Prediction of TOEFL Listening Comprehension Item Difficulty for Minitalk Passages: Implications for Construct Validity.* ETS Research Report RR 96-29, Princeton, NJ.

Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (Hrsg.) (2003). *Bildungs- und Erziehungsauftrag – Bildungsplan Integrierte Gesamtschule, Sekundarstufe I*. Hamburg.

Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (Hrsg.) (2003). *Bildungs- und Erziehungsauftrag – Bildungsplan neunstufiges Gymnasium, Sekundarstufe I*. Hamburg.

Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (Hrsg.) (2003). *Bildungs- und Erziehungsauftrag, Bildungsplan Hauptschule und Realschule, Sekundarstufe I*. Hamburg.

Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (Hrsg.) (2003). *Rahmenplan Deutsch – Bildungsplan Hauptschule und Realschule, Sekundarstufe I*. Hamburg.

Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (Hrsg.) (2003). *Rahmenplan Deutsch – Bildungsplan Integrierte Gesamtschule, Sekundarstufe I*. Hamburg.

Freie und Hansestadt Hamburg, Behörde für Bildung und Sport (Hrsg.) (2003). *Rahmenplan Deutsch – Bildungsplan neunstufiges Gymnasium, Sekundarstufe I*. Hamburg.

G

Gernsbacher, M. A. (1994) (Ed.). *Handbook of Psycholinguistics*. San Diego, CA.

Gerstenmaier, J. (1995) (Hrsg.). *Einführung in die Kognitionspsychologie*. München.

Gloy, K. (1973). *Die Type-Token-Ratio als Instrument der Quantitativen Linguistik*. Forschungsbericht 4 der Universität Konstanz. Konstanz.

Goh, C.C.M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System,* 28, 55 - 75.

Goldstein, H., Bonnet, G. & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. In *Journal of Educational and Behavioural Statistics*, 32, 252 - 286.

Golub-Smith, M. (1987). A study of the Effects of Item Option Rearrangement on the Listening Comprehension Section of the Test of English as a Foreign Language. *TOEFL Research Reports 24*. Princeton, NJ.

Gorin, J. S. & Embretson, S. E. (2006). Item Difficulty Modeling of Paragraph Comprehension Items. *Applied Psychological Measurement*, 30, 394 - 411.

Graefen, G. & Liedke, M. (2008). *Germanistische Sprachwissenschaft. Deutsch als Erst-, Zweit- oder Fremdsprache*. Tübingen/Basel.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371 – 395.

Graesser, A., McNamara, D.S., Louwerse, M.M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36, 193 - 202.

Griffiths, R. (1990). Facilitating Listening Comprehension Through Rate-Control. *RELC Journal*, 21, 55 - 64.

Griffiths, R. (1991). Language Classroom Speech Rates: A Descriptive Study. *TESOL Quarterly*, 25(1), 189 - 194.

Grotjahn, R. (2000). Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung des TESTDAF. In S. Bolton (Hrsg.), *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. München, 7 - 55.

Grotjahn, R. (2005). Testen und Bewerten des Hörverstehens. In M. Ó Dúill, R. Zahn & K. D. C. Höppner (Hrsg.), *Zusammenarbeiten. Eine Festschrift für Bernd Voss*. Bochum, 115 - 144.

H

Hale, G. A., Rock, D. A. & Jirele, T. J. (1989). *Confirmatory Factor Analysis of the TOEFL® Test*. ETS Research Report RR 96 - 29. Princeton, NJ.

Hambleton, R. K. & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education*, 5(1), 1 - 16.

Hare, V., Rabinowitz, M. & Schieble, K. (1989). Text effects on main idea comprehension. *Reading Research Quarterly*, 24, 72 - 88.

Hartig, J. (2008). Psychometric Models for the Assessment of Competencies. In J. Hartig, E. Klieme & D. Leutner (Eds.), *Assessment of Competencies in Educational Contexts*. Toronto, 69 - 90.

Hartland, J. (1995). Sprache und Denken. In J. Gerstenmaier (Hrsg.), *Einführung in die Kognitionspsychologie.* München, 195 - 244.

Hartmann, C. (2008). *Schwierigkeitserklärende Merkmale von Englisch- Leseverstehens-aufgaben.* Unveröffentlichte Diplomarbeit.

Hayes, N. (1995). Kognitive Prozesse. Eine Einführung. In J. Gerstenmaier (Hrsg.), *Einführung in die Kognitionspsychologie.* München, 11 - 40.

Heckhausen, H. (1989). *Motivation und Handeln.* Berlin.

Helbig, G. & Buscha, J. (2001). *Deutsche Grammatik. Ein Handbuch für den Ausländer-unterricht.* München.

Helmke, A. (2006). *Unterrichtsqualität erfassen, bewerten, verbessern*. Seelze.

Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In: R. Arnold & C. Griese (Hrsg.), *Schulmanagement und Schulentwicklung*. Hohengehren.

Hennig, M. (2006). *Grammatik der gesprochenen Sprache in Theorie und Praxis*. Kassel.

Henning, G. (1991). *A study of the effects of variation of short-term memory load, reading response length, and processing hierarchy on TOEFL®. Listening Comprehension Item Performance.* ETS Research Report 90-18. Princeton, NJ.

Hessisches Kultusministerium (Hrsg.), *Lehrplan Gymnasialer Bildungsgang, Jahrgangsstufen 5 bis 13*; Zugriff von http://www.kultusministerium.hessen.de/irj/HKM_Internet?cid=9e0b5517dfc688683c15ce252202d4b9 am 03.12.2010.

Hessisches Kultusministerium (Hrsg.), *Handreichung zur Arbeit mit den Lehrplänen der Bildungsgänge Hauptschule, Realschule und Gymnasium, Deutsch, an schulformübergreifenden (integrierten) Gesamtschulen und Förderstufen.* Zugriff von http://www.kultusministerium.hessen.de/irj/HKM_Internet?cid=e9791809d4e92c28a030b4f9340b0d5e am 03.12.2010.

Hessisches Kultusministerium (Hrsg.), *Lehrplan Deutsch, Bildungsgang Hauptschule, Jahrgangsstufen 5 bis 9/10*; Zugriff von http://www.kultusministerium.hessen.de/irj/HKM_Internet?cid=770244b3f3f61faf79f08f0f0db32a30 am 03.12.2010.

Hessisches Kultusministerium (Hrsg.), *Lehrplan Deutsch, Bildungsgang Realschule, Jahrgangsstufen 5 bis 10*; Zugriff von http://www.kultusministerium.hessen.de/irj/HKM_Internet?cid=f1e079cc428af80d07f4fe2db20fe301 am 03.12.2010.

Hessisches Kultusministerium, Pressemitteilung vom 02. Dezember 2010, Zugriff am 09.12.2010 von http://www.kultusministerium.hessen.de/irj7HKM_Internet?rid=HKM_15/HKM_Internet/nav/8e0/8e0703e0cf26-2901-be59-2697ccf4e69f,66338d7e-cc2a-c21f-012f-31e2389e4818,,,11111111-2222-3333-4444-100000005004%26_ic_uCon_zentral=66338d7e-cc2a-c21f-012f-31e2389e4818%26overview=true.htm&uid=8e0703e0-cf26-2901-be59-2697ccf4e69f

Hirschfeld, U. (1992). Wer nicht hören will… *Fremdsprache Deutsch* 7, 17 - 20.

Hollingworth, L., Beard, J. J. & Proctor, T. (2007). An Investigation of Item Type in a Standards-Based Assessment. *Practical Assessment Research & Evaluation*, 12(18). Zugriff von http://pareonline.net/pdf/v12n18.pdf am 27.01.10

I

Imhof, M. (2003). *Zuhören. Psychologische Aspekte auditiver Informationsverarbeitung*. Göttingen.

In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System,* 34(3), 317 - 340.

J

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, 10, 409 - 416.

Jensen, C., Hansen, C., Green, S. B. & Akey, T. (1997). An investigation of item difficulty incorporating the structure of listening tests: A hierarchical linear modeling analysis. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96*. Jyväskylä: University of Jyväskylä, 151 - 164.

Jude, N. & Klieme, E. (2007). Definition sprachlicher Kompetenz – Ein Differenzierungsansatz. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen: Konzepte und Messung - DESI-Studie (Deutsch-Englische-Schülerleistungen-International)*. Weinheim, 9 - 22.

Just, M. A. & Carpenter, P. A. (1992). A Capacity Theory of Comprehension: Individual Differences in Working Memory. *Psychological Review*, 99(1), 122 - 149.

K

Kane, M. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, 38(4). 319 - 342.

Katz, S, Lautenschlager, G, Blackburn, A. & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Science*, 1, 122 - 127.

Keshavarz, M. H., Atai, M. R. & Ahmadi, H. (2007). Content schemata, linguistic simplification, and EFL readers' comprehension and recall. *Reading in a Foreign Language*, 19, 19 - 33.

Kieras, D. E. (1985). Thematic processing in the comprehension of technical prose. In D. Britton & J. Black (Eds.), *Understanding expository text.* Hillsdale, 89 - 107.

Kinder-/Jugendliteratur und Medien, in Forschung, Schule und Bibliothek/Kjl&m, 08.3., 60(3), 2008

Kintsch, W. & Keenan, J. (1973). Reading Rate and Retention as a Function of the Number of Propositions in the Base Structure of Sentences. *Cognitive Psychology*, 5, 257 - 274.

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

Kintsch, W. (1998). *Comprehension. A paradigm for cognition.* Cambridge.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G. & Keenan, J.M. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behaviour*, 14, 196 - 214.

Klare, G. R. (1963). *The measurement of readability*. Ames, Iowa.

Klare, G. R. (1974-1975). Assessing readability. *Reading Research Quarterly*, 10, 62 - 102.

Klein, W. (1985). Gesprochene Sprache - geschriebene Sprache. *Zeitschrift für Literaturwissenschaft und Linguistik*, 15(59), 9 - 35.

Klein-Braley, C. (1994). *Language Testing with the C-Test*. Unveröffentlichte Habilitationsschrift, Universität Duisburg.

Kletzien, S. B. (1992). Proficient and less proficient comprehenders' strategy use for different top-level structures. *Journal of Reading Behavior*, 24(2), 191 - 232.

Klieme, E., Eichler, W., Helmke, A., Lehmann, R. H., Nold, G., Rolff, H.-G., Schröder, K., Günther, T. & Willenberg, H. (Hrsg.) (2003), *DESI: Bericht über die Entwicklung und Erprobung der Erhebungsinstrumente*. Frankfurt am Main.

Knoche, N. & Lind, D. (2005). *Strukturanalysen zur Mathematikleistung und eine differentielle Itemanalyse der PISA-2000-Items zu den Faktoren Bildungsgang und Geschlecht.* Zugriff am 11.01.10 von http://www.mathematik.unidortmund.de/didaktik/BzMU/BzMU2005/Beitraege/knoche-lind-gdm

Koch, P. & Oesterreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36, 15 - 43.

Kohler, B. (2005). *Rezeption internationaler Schulleistungsstudien*. Münster/New York/München/Berlin.

Köller et al. (Hrsg.) (2010), *Sprachliche Kompetenzen im Ländervergleich*. Münster.

Köller, O., Leucht, M., Harsch, C. (2009, September). *Schwierigkeitsgenerierende Merkmale von Items zum Lese- und Hörverstehen im Fach Englisch*. Vortrag auf der 12. Fachtagung für Pädagogische Psychologie (PAEPS) der Deutschen Gesellschaft für Psychologie, Saarbrücken.

Kürschner, C. & Schnotz, W. (2008). Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen. *Psychologische Rundschau*, 59(3), 139 - 149.

Kultusministerium des Landes Sachsen-Anhalt (Hrsg.) (1999), *Rahmenrichtlinien Sekundarschule, Schuljahrgänge 7 - 10*, Deutsch. Magdeburg.

Kultusministerium des Landes Sachsen-Anhalt (Hrsg.) (2003), *Rahmenrichtlinien Gymnasium, Schuljahrgänge 5 - 12*, Deutsch. Magdeburg.

Kultusministerkonferenz, Universität Duisburg/Essen, IQB. *Kompetenzstufenmodell zu den Bildungsstandards im Kompetenzbereich Sprechen und Zuhören – hier Zuhören – für den Mittleren Schulabschluss*, Zugriff am 07.09.09 von http://www.iqb.huberlin.de/arbbereiche/testentw/projekte?pg=p_36&spg=r_6

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

L

Langer, I. F., Schulz von Thun, F. & Tausch, R. (1974). *Verständlichkeit in Schule, Verwaltung, Politik und Wissenschaft: mit einem Selbsttrainingsprogramm zur verständlichen Darstellung von Lehr- und Informationstexten.* München.

Lehmann, R. H., Peek, R., Gänsfuß, R. & Husfeldt, V. (2001). *LAU 9. Aspekte der Lernausgangslage und der Lernentwicklung - Klassenstufe 9.* Hamburg.

Lehrndorfer, A. (1996). *Kontrolliertes Deutsch. Linguistische und sprachpsychologische Leitlinien für eine (maschinell) kontrollierte Sprache in der Technischen Dokumentation.* Tübingen.

Leonhart, R. (2008). Psychologische Methodenlehre/Statistik. München.

Leucht, M., Harsch, C. & Köller, O. (in Revision). *Schwierigkeitsgenerierende Merkmale von Items zum Lese- und Hörverstehen im Fach Englisch.*

Leucht, M., Retelsdorf, J., Möller, J. & Köller, O. (2010). Zur Dimensionalität rezeptiver englischsprachiger Kompetenzen. *Zeitschrift für Pädagogische Psychologie*, 24, 123 - 138.

Levine, A. & Revers, T. (1988). The FL receptive skills: Same or different? *System*, 16, 326 – 336.

Lewkowicz, J. A. (1996). Authenticity for whom? Does authenticity really matter? In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96.* Jyväskylä: University of Jyväskylä, 165 - 184.

Lienert, G. A. & Raatz, U. (1998). Testaufbau und Testanalyse. Weinheim.

Long, D. L., & Chong, J. L. (2001). Comprehension skill and global coherence: A paradoxical picture of poor comprehenders' abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1424 – 1429.

Lüer, G. (1988). Kognitive Prozesse und Augenbewegungen. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie*. München, 386 - 399.

Lund, R. J. (1991). A Comparison of Second Language Listening and Reading Comprehension. *Modern Language Journal*, 75, 196 - 204.

M

Mandl, H., Tergan, S. & Ballstaedt, S. P. (1982). Textverständlichkeit – Textverstehen. In B. Treiber & F. T. Weinert (Hrsg.), *Lehr-Lern-Forschung.* München, 66 - 88.

Martinez, M. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement*, 28(2), 131 - 145.

Mattys, S.L., Brooks, J. & Cooke, M. (2009). Recognizing speech under a processing load: Dissociating energetic from informational factors. *Cognitive Psychology,* 59, 203 - 243.

McNamara, D. S. & Kintsch, W. (1996). Learning From Texts: Effects of Prior Knowledge and Text Coherence. *Discourse Processes,* 22, 247 - 288.

Meyer, B. J. F. & Freedle, R. O. (1984). Effects of discourse type on recall. *American Educational Research Journal*, 21, 121 - 143.

Ministère de l'Education nationale et de la Formation professionelle (2008). *Bildungsstandards Sprachen. Leitfaden für den kompetenzorientierten Sprachenunterricht an Luxemburger Schulen.* Zugriff am 15.07.2009 von http://www.men.public.lu/publications/sy

Ministerium für Bildung, Jugend und Sport des Landes Brandenburg (Hrsg.) (2002). *Rahmenlehrplan, Deutsch, Sekundarstufe I.* Berlin.

Ministerium für Bildung, Kultur und Wissenschaft Saarland (Hrsg.) (2005). *Achtjähriges Gymnasium - Lehrplan für das Fach Deutsch, Klassenstufe 9.* Saarbrücken.

Ministerium für Bildung, Wissenschaft und Kultur Mecklenburg-Vorpommern (Hrsg.) (2002). *Rahmenplan Deutsch - Gymnasium, Integrierte Gesamtschule, Jahrgangsstufen 7 -10.* Schwerin.

Ministerium für Bildung, Wissenschaft und Kultur Mecklenburg-Vorpommern (Hrsg.) (2002). *Rahmenplan Deutsch - Regionale Schule, Verbundene Haupt- und Realschule, Hauptschule, Realschule, Integrierte Gesamtschule, Jahrgangsstufen 7 -10.* Schwerin.

Ministerium für Bildung, Wissenschaft und Weiterbildung, Mainz (Hrsg.) (1998). *Lehrplan Deutsch (Klassen 5 - 9/10), Hauptschulen, Realschulen, Gymnasien, Regionale Schulen, Gesamtschulen.* Grünstadt.

Ministerium für Bildung, Wissenschaft, Forschung und Kultur des Landes Schleswig-Holstein (Hrsg.) (1997). *Lehrplan für die Sekundarstufe I der weiterführenden allgemeinbildenden Schulen Hauptschule, Realschule, Gymnasium, Gesamtschule - Deutsch.* Zugriff am 17.12.2010 von http://www.schulrecht-sh.de/texte/l/lehrplaene_97.htm.

Ministerium für Kultus und Sport Baden-Württemberg (Hrsg.) (1994). *Kultus und Unterricht - Amtsblatt des Ministeriums für Kultus und Sport Baden-Württemberg, Ausgabe C, Lehrplanhefte, Bildungsplan für die Hauptschule.* Villingen-Schwenningen.

Ministerium für Kultus und Sport Baden-Württemberg (Hrsg.) (1994). *Kultus und Unterricht - Amtsblatt des Ministeriums für Kultus und Sport Baden-Württemberg, Ausgabe C, Lehrplanhefte, Bildungsplan für die Realschule.* Villingen-Schwenningen.

Ministerium für Kultus und Sport Baden-Württemberg (Hrsg.) (2001). *Kultus und Unterricht - Amtsblatt des Ministeriums für Kultus und Sport Baden-Württemberg, Ausgabe C, Lehrplanhefte, Bildungsplan für das allgemein bildende Gymnasium mit achtjährigem Bildungsgang.* Villingen-Schwenningen.

Ministerium für Schule, Jugend und Kinder des Landes Nordrhein-Westfalen (Hrsg.) (2004). *Kernlehrplan für die Hauptschule in Nordrhein-Westfalen*, Deutsch. Frechen.

Moosbrugger, H. & Schermelleh-Engel, K. (2007). Exploratorische (EFA) und Konfirmatorische Faktorenanalyse (CFA). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion*. Heidelberg, 307 - 324.

Moosbrugger, H. (2007). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion*. Heidelberg, 215 - 260.

Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology*, 88, 314 - 332.

N

Neumann, A. (2007). *Briefe schreiben in Klasse 9 und 11. Beurteilungskriterien, Messungen, Textstrukturen und Schülerleistungen.* Münster.

Newsome, R. S. & Gaite, J. H. (1971). Prose learning: Effects of pretesting and reduction of passage length. *Psychology Reports*, 28, 123 - 129.

Niedersächsisches Kultusministerium (Hrsg.) (2006), *Kerncurriculum für das Gymnasium, Schuljahrgänge 5 -10, Deutsch*. Hannover.

Niedersächsisches Kultusministerium (Hrsg.) (2006), *Kerncurriculum für die Hauptschule, Schuljahrgänge 5 -10, Deutsch.* Hannover.

Niedersächsisches Kultusministerium (Hrsg.) (2006), *Kerncurriculum für die Integrierte Gesamtschule, Schuljahrgänge 5 -10, Deutsch*. Hannover.

Niedersächsisches Kultusministerium (Hrsg.) (2006), *Kerncurriculum für die Realschule, Schuljahrgänge 5 -10, Deutsch.* Hannover.

Nissan, S., DeVincenzi, F. & Tang, K. L. (1996). *An Analysis of Factors Affecting the Difficulty of Dialogue Items in TOEFL Listening Comprehension*. TOEFL Research Reports 51, Princeton, NJ.

Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). Upper Saddle River, NJ.

Nold, G. & Rossa, H. (2007). Hörverstehen. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen: Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. Weinheim, 174 - 192.

O

O'Malley, M., Chamot, A. U. & Küpper, L. (1989). Listening Strategies Comprehension in Second Language Acquisition. *Applied Linguistics*, 10(4), 418 - 437.

Oberauer, K. (2005). The measurement of working memory capacity. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring Intelligence*. Thousand Oaks, CA, 393 - 407.

Oberauer, K., Bristol, U., Mayr, E. & Kluwe, R. (2006). Gedächtnis und Wissen. In H. Spada (Hrsg.), *Lehrbuch Allgemeine Psychologie*. Bern, 115 - 197.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O. & Wittmann, W. W. (2000). Working memory capacity - facets of a cognitive ability construct. *Personality and Individual Differences*, 29, 1017 - 1045.

Organisation for Economic Co-operation and Development (OECD). (2002). *PISA 2000 Technical Report*. Zugriff am 24.06.2009 von http://www.oecd.org/dataoecd/53/19/33688233.pdf

Ozuru, Y., Rowe, M., O'Reilly, T. & McNamara, D. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behaviour Research Methods*, 40(4), 1001 - 1015.

P

Pallier, C., Christophe, A. & Mehler, J. (1997). Language-specific listening. *Trends in Cognitive Sciences* 1(4), 129 - 132.

Parker, K. & Chaudron, C. (1987). The effects of linguistic simplification and elaborative modification on L2 comprehension. *University of Hawaii, Working Papers in EST*, 6, 107 - 133.

Perfetti, C. (1985). *Reading ability*. New York.

Perfetti, C. A. (1994). Psycholinguistics and reading ability. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics*. San Diego, CA, 849 – 894.

Perfetti, C. A. & Goldman, S. R. (1976). Discourse memory and reading comprehension skill. *Journal of Verbal Learning and Verbal Behavior*, 14, 33 - 42.

Perkins, K. & Brutten, S. R. (1993). A model of ESL reading comprehension difficulty. In: A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96*. Jyväskylä: University of Jyväskylä, 205 - 218.

Pfitzinger, H. (2001). Phonetische Analyse der Sprechgeschwindigkeit. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München* (FIPKM), 38, 117 - 264.

Pica, T., Young, R. & Doughty, C. (1987). The Impact of Interaction on Comprehension. *TESOL Quarterly*, 21(4), 737 - 758.

PISA-Konsortium Deutschland. (2008). *PISA ´06. PISA 2006 in Deutschland. Die Kompetenzen der Jugendlichen im dritten Ländervergleich.* Münster.

R

Ramm, G., Prenzel, M., Heidemeier, H. & Walter, O. (2004). Soziokulturelle Herkunft: Migration. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, R. Pekrun, H.-G. Rolff, J. Rost & U. Schiefele (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland – Ergebnisse des zweiten internationalen Vergleichs.* Münster, 254 – 272.

Rauch, D. & Hartig, J. (2007). Interpretation von Testwerten in der IRT. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion*. Heidelberg, 240 - 250.

Raudenbusch, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models, second edition.* Thousand Oaks, CA.

Reddy, P. (1995). Aufmerksamkeit und das Lernen von Fertigkeiten. In J. Gerstenmaier (Hrsg.), *Einführung in die Kognitionspsychologie.* München, 91 - 120.

Richards, J. C. (1983). Listening Comprehension: Approach, Design, Procedure. *TESOL Quarterly*, 17(2), 219 - 240.

Richgels, D. J., McGee, L. M., Lomax, R. G. & Sheard, C. (1987). Awareness of Four Text Structures: Effects on Recall of Expository Text. *Reading Research Quarterly*, 22(2), 177 - 196.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests. In D. Ganzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik.* Weinheim/ Basel, 44 - 111.

Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104(3), 192 - 233.

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing,* 19, 305 – 335.

Rost, M. (2002). *Teaching and Researching Listening.* Harlow.

Rowe, M. & McNamara, D.S. (2008). Inhibition needs no negativity: Negativity links in the construction-integration model. In V. Sloutsky, B. Love & K. McRae (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society.* Washington, DC, 1777 - 1782.

Royer, J. (1990). The sentence verification technique: A new direction in the assessment of reading comprehension. In S. Legg & J. Algina (Eds.), *Cognitive assessment of language and math outcomes.* Norwood, NJ, 144 - 191.

Rubin, J. (1994). A Review of Second Language Listening Comprehension Research. *The Modern Language Journal,* 78(2), 199 - 221.

S

Sächsisches Staatsministerium für Kultus (Hrsg.) (2001). *Lehrplan Gymnasium - Gewichtete Fassung, Deutsch, Klassen- und Jahrgangsstufen 5 – 12.* Dresden.

Schank, R. C. & Abelson, R. P. (1977). Scripts, plans, and knowledge. In P. N. Johnson-Laird & P. C. Wason (Eds.), *Thinking: readings in cognitive science.* Cambridge, 421 - 432.

Schiefele, U. & Krapp, A.(1996). Topic interest and free recall of expository text. *Learning and Individual Differences*, 8(2), 141 - 160.

Schiefele, U. (1996). Topic Interest, Text Representation, and Quality of Experience. *Contemporary Educational Psychology*, 21, 3 – 18.

Schmalt, H.-D. & Sokolowski, K. (2006). Motivation. In H. Spada (Hrsg.), *Lehrbuch Allgemeine Psychologie.* Bern, 500 - 551.

Schneider, W. & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review,* 84(1), 1 - 66.

Schweitzer, K. (2007). *Der Schwierigkeitsgrad von Textverstehensaufgaben. Ein Beitrag zur Differenzierung und Präzisierung von Aufgabenbeschreibungen*. Frankfurt.

Schwippert, K., Hornberg, S., Freiberg, M. & Stubbe, T. C. (2007). Lesekompetenzen von Kindern mit Migrationshintergrund im internationalen Vergleich. In W. Bos, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, E.-M. Lankes, K. Schwippert & R. Valtin (Hrsg.), *IGLU 2006. Lesekompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Münster, 249 – 269.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2004), *Bildungsstandards im Fach Deutsch für den Hauptschulabschluss*. Beschluss vom 15.10.2004, München.

Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (2004), *Bildungsstandards im Fach Deutsch für den Mittleren Schulabschluss*. Beschluss vom 4.12.2003, München.

Senator für Bildung und Wissenschaft (Hrsg.) (2006), Deutsch, *Bildungsplan für die Gesamtschule, Jahrgangsstufe 7 und 8*. Bremen.

Senator für Bildung und Wissenschaft (Hrsg.) (2006), Deutsch, *Bildungsplan für die Sekundarschule, Jahrgangsstufe 7 und 8*. Bremen.

Senatsverwaltung für Bildung, Jugend und Sport Berlin (Hrsg.) (2006), *Rahmenlehrplan für die Sekundarstufe I, Jahrgangsstufe 7-10, Hauptschule, Realschule, Gesamtschule, Gymnasium, Deutsch*. Berlin.

Sheehan, K. M. & Ginther, A. (2001, April). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section.* Paper presented at the 2001 Annual Meeting of the National Council of Measurement in Education.

Sherman, J. (1997). The effect of question preview in listening comprehension tests. *Language Testing*, 14(2), 185 - 213.

Shohamy, E. & Inbar, O. (1991). Validation of listening comprehension tests: the effect of text and question type. *Language Testing*, 8, 23 - 40.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1, 147 - 170.

Song, M.-Y. (2008). Do divisible subskills exist in second language (L2) comprehension? A structural equation modeling approach. *Language Testing*, 25(4), 435 - 464.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L. & Voss, J. F. (1979). Text processing of domainrelated information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, 18, 275 – 290.

SPSS Inc. (2007). SPSS for Windows release 15 standard version. Chicago, IL: SPSS Inc.

Statistisches Bundesamt (2010). *Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund. Ergebnisse des Mikrozensus 2009*. Wiesbaden.

Stiggins. R. J. (2002). *Assessment crisis: The absence of assessment for learning*. Phi Delta Kappan, 83, 758 - 765.

Strube, G., Becker, B., Freksa, C., Hahn U., Opwis, K. & Palm, G. (Hrsg.) (1996). *Wörterbuch der Kognitionswissenschaft*. Stuttgart.

T

Tannen, D. (1982). The oral literate continuum of discourse. In D. Tannen (Ed.), *Spoken and written language*. Norwood, NJ, 1 - 16.

Tannen, D. (1985). Relative focus of involvement in oral and written discourse. In D. Olson, D. Torrance & A. Hildyard (Eds.), *Literary language and learning*. Cambridge, 124 - 147.

Thaler, E. (2007). Schulung des Hör-Seh-Verstehens. *Praxis Fremdsprachenunterricht*, 4, 12 - 17.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. In *Cognitive Psychology*. New York, NY.121 - 152.

Thüringer Kultusministerium (Hrsg.) (1999), *Lehrplan für das Gymnasium – Deutsch*. Erfurt.

Thüringer Kultusministerium (Hrsg.) (1999), *Lehrplan für die Regelschule und für die Förderschule mit dem Bildungsgang der Regelschule – Deutsch*. Erfurt.

Tuholski, S. W., Engle, R. W. & Baylis, G. C. (2001). Individual differences in working memory capacity and enumeration. *Memory & Cognition*, 29(3), 484 - 492.

Tuinman, J. (1973-1974). Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 9, 206 - 223.

Turner, M. L. & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, 28, 127 - 154.

V

Valencia, S. W. & Pearson, P. D. (1988). Principles for Classroom Comprehension. *RASE, 9(1*), 26 - 35.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. N.Y.

Van Dijk, T. A. (1980). *Macrostructures: an interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, N.J.

Vogt, R. (2002). *Im Deutschunterricht diskutieren. Zur Linguistik und Didaktik einer kommunikativen Praktik*. Tübingen.

Von Davier, M. (2009). Some Notes on the Reinvention of Latent Structure Models as Diagnostic Classification Models. *Measurement: Interdisciplinary Research & Perspective*, *7(1)*, 67 - 74.

W

Walter, O. & Taskinen, P. (2007). Kompetenzen und Bildungsrelevante Einstellungen von Jugendlichen mit Migrationshintergrund in Deutschland: Ein Vergleich mit ausgewählten OECD-Staaten. In M. Prenzel, C. Artelt, J. Baumert, W. Blum, M. Hammann, E. Klieme & R. Pekrun (Hrsg.), *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie*. Münster, 337 – 366.

Weinert, F. E. (1999). *Concepts of competence. Definition and selection of competencies*. München.

Willenberg, H. (1995). Die Strategien des Lesens und Lernens sind individuell gemischt. *Empirische Pädagogik, 9,* 263 - 283.

Willenberg, H. (2007). Lesen. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. Weinheim, 107 - 117.

Winkelmann, H. & Robitzsch, A. (2009). Modelle mathematischer Kompetenzen: Empirische Befunde zur Dimensionalität. In D. Ganzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik*. Weinheim/Basel, 175 - 203.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen.

Wittrock, M. C., Marks, C. & Doctorow, M. (1975). Reading as a generative process. *Journal of Educational Psychology*, *67*, 484 – 489.

Wolff, D. (1999). Hörverstehen in einer Fremdsprache: Ein psycholinguistisches Ratespiel? In D. Eggers (Hrsg.), *Sprachandragogik Jahrbuch 1998: Hörverstehen aus andragogischer Sicht. Sprachlern- und Spracherwerbsstrategien im Fremdsprachunterricht mit Erwachsenen*. 17 - 35.

Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S. A. (2007). *ACER ConQuest. Version 2.0. Camberwell*, Victoria: ACER Press (Australian Council for Educational Research).

Y

Yanagawa, K. & Green, A. (2008). To show or not to show: The effects of item stems and answer options on performance on a multiple-choice listening comprehension test. *System, 36(1)*, 107 - 122.

Yi'an, W. (1998). What do tests of listening comprehension test? - A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*, 21 - 44.

Yin-Kum, L. (1995). Effects of Text Structures on Recall. *Education Journal, 23(1)*, 115 - 133.

Yuill, N. M., Oakhill, J. V., & Parkin, A. J. (1989). Working memory, comprehension skill and the resolution of text anomaly. *British Journal of Psychology, 80,* 351 – 361.

Yousif, A. A. (2006). Listening Comprehension Difficulties as Perceived by. *Journal of King Saud University Language & Translation, 19*, 35 - 47.

Weblinks

http://www.kmk.org/schul/pisa/konstanz.htm, Zugriff am 29.12.08

http://wortschatz.uni-leipzig.de/, Zugriff am 09.12.09

http://www.bmbf.de/de/6626.php, Zugriff am 20.04.09

http://www.hrsdc.gc.ca/eng/hip/lld/nls/Surveys/ialsintro.shtml, Zugriff am 26.06.09

http://www.iea.nl/reading_literacy.html, Zugriff am 26.06.09

http://www.iea.nl/readingcomprehension.html, Zugriff am 26.06.09

http://www.iqb.hu-berlin.de/aktuell?pg=a_7, Zugriff am 24.06.09

http://www.iqb.hu-berlin.de/vera2, Zugriff am 02.05.2010

http://www.standardsicherung.schulministerium.nrw.de/lernstand8/, Zugriff am 11.05.09

http://www.uni-landau.de/vera/, Zugriff am 20.04.09

Verzeichnis der Anhänge

Anhang A: Synopsen der Lehrpläne für den Hauptschulabschluss und den Mittleren Schulabschluss im Bereich „Sprechen und Zuhören“

Anhang B: Aufmerksamkeitstest

Anhang C: Korrelationsübersichten

Anhang D: Tabellen – Faktorenanalyse

Anhang E: Ergebnisse der Regressionsanalysen

Anhang F: Synopse Bildungsstandards „Sprechen und Zuhören“

Anhang G: Tabellen Zusammenhangsanalysen