Zum Zusammenhang von Differentiellen Item Funktionen und Testkultur

Dissertation
zur Erlangung des akademischen Grades

Doktorin der Philosophie (Dr.phil.)

Vorgelegt dem Fachbereich
Psychologie und Sportwissenschaften
der Johann Wolfgang Goethe - Universität
Frankfurt am Main

von
Dipl.Psych. Astrid Jurecka

1. Gutachter: Prof. Dr. Eckhard Klieme
2. Gutachter: Prof. Dr. Marcus Hasselhorn

Frankfurt am Main im Mai 2010
Erklärung

Hiermit erkläre ich,

• dass ich die vorliegende Dissertation selbst verfasst und alle in Anspruch genommenen Hilfsmittel in der Dissertation angegeben habe,

• dass frühere Promotionsverfahren nicht erfolgt und somit nicht erfolglos geblieben sind,

• dass mir die Promotionsordnung zur Erlangung des akademischen Grades eines Doktors der Philosophie an der Johann Wolfgang Goethe-Universität vom 26.06. 2001 bekannt ist, und

• dass ich keine Hilfe einer kommerziellen Promotionsvermittlung in Anspruch genommen habe.

Frankfurt am Main, den 29.04.2010

______________________________
Astrid Jurecka
## Inhalt

**Erklärung** ................................................................. ii

**Zusammenfassung** ...................................................... 1

### 1. Einleitung ............................................................. 11
   1.1. Vergleiche von Leistung und Kompetenz vor dem Hintergrund unterschiedlicher (Bildungs-)Kulturen .......................... 13
   1.2. Beschreibung des Dissertationsvorhabens .......................... 15

### 2. Theoretische Grundlagen ........................................... 18
   2.1. Differentielle Item Funktionen .................................. 19
      2.1.1. DIF und Validität: „Nuisance-Dimension“ oder Ausdruck differentialer Stärken und Schwächen? .................. 28
      2.1.2. Die Erklärung von Differentiellen Item Funktionen ........ 29
   2.2. Fremdsprachenforschung und angewandte Linguistik .......... 34
      2.2.1. Das Konstrukt der fremdsprachlichen Lesekompetenz ...... 35
      2.2.2. Die Messung von Fremdsprachenkompetenzen ............... 44
      2.2.3. Der Gemeinsame Europäische Referenzrahmen für Sprachen .. 50
      2.2.4. Determinanten der Itemschwierigkeit, das Itemkategorisierungssystem Dutch Grid und Beschreibung zugrundeliegender kognitiver Prozesse ........................................ 58
   2.3. Interkulturelle Vergleichbarkeit von Testergebnissen .......... 71
      2.3.1. Validität in interkulturellen Studien: Konzept der Äquivalenz und Methoden zur Überprüfung von Invarianz ........ 75
   2.4. Verknüpfung der Theoriestränge & Rahmenkonzept der Dissertation:
      Messicks Validitätstheorie ........................................... 78
      2.4.1. Samuel Messicks Validitätskonzept ............................. 78
      2.4.2. Quellen der Invalidität ....................................... 81
2.4.3. „Consequencial aspect of validity” und „social values” 82
2.4.4. Zusammenführung der Theoriestränge im Kontext der Validität 85
2.4.5. Zusammenfassung und Relevanz für die Arbeit 87

3. Fragestellung ..................................................... 88
   3.1. Herleitung der Fragestellungen ............................. 88
   3.2. Hauptfragestellung ........................................ 93
   3.3. Fragenkomplex 1: Voraussetzungen und Skalierbarkeit 94
   3.4. Fragenkomplex 2: Erklärung von Itemschwierigkeiten 96
   3.5. Fragenkomplex 3: Erklärung von Differentiellen Item Funktionen 99

4. Methoden .......................................................... 102
   4.1. Datengrundlage .............................................. 102
      4.1.1. Herkunft der Items ................................. 103
      4.1.2. Analysen in der EBAFLS Studie .................. 104
      4.1.3. Design und Stichprobe ............................. 105
      4.1.4. Ergebnisse der EBAFLS-Leseverständnis-Studie 113
      4.1.5. Aufbereitung der Daten ............................. 114
   4.2. Methoden zur Beantwortung von Fragenkomplex 1 ........ 114
      4.2.1. Überprüfung der Rasch-Modellkonformität der Items innerhalb der Länder 115
      4.2.2. Methoden zur Analyse von Differentiellen Item Funktionen 118
      4.2.3. Methoden zur Analyse von Indikatoren nationaler Testkulturen 119
   4.3. Methoden zur Beantwortung der Fragenkomplexe 2 und 3 .... 123
      4.3.1. Methoden zu Fragenkomplex 2: Erklärung der Itemschwierigkeiten innerhalb der Länder 124
      4.3.2. Methoden zu Fragenkomplex 3: Erklärung von Differentiellen Item Funktionen 127
   4.4. Anmerkung zum Umgang mit der Inflation des Alpha-Fehlers 129
5. Ergebnisse ................................................................. 130
  5.1. Ergebnisse Fragenkomplex 1: „Voraussetzungen und Skalierbarkeit“ .... 130
    5.1.1. Zu Frage 1a .................................................. 130
    5.1.2. Zu Frage 1b .................................................. 133
    5.1.3. Zu Frage 1c .................................................. 134
  5.2. Ergebnisse Fragenkomplex 2: Erklärung der Itemschwierigkeiten innerhalb der Länder .................................................. 149
    5.2.1. Zu Frage 2a .................................................. 149
    5.2.2. Zu Frage 2b .................................................. 154
    5.2.3. Zu Fragen 2c und 2d ...................................... 158
  5.3. Ergebnisse zu Fragenkomplex 3: Erklärung von Differentiellen Item Funktionen .................................................. 168
    5.3.1. Zu Frage 3a .................................................. 168
    5.3.2. Zu Fragen 3b und 3c ...................................... 175

6. Interpretation der Ergebnisse, Beantwortung der Hauptfragestellung und Diskussion .................................................. 193
  6.1. Zusammenfassung und Interpretation der Ergebnisse ...................... 193
    6.1.1. Zusammenfassung und Interpretation der Ergebnisse zu Fragenkomplex 1: „Voraussetzungen und Skalierbarkeit“ ............... 193
    6.1.2. Zusammenfassung und Interpretation der Ergebnisse zu Fragenkomplex 2: „Erklärung der Itemschwierigkeiten innerhalb der Länder“ .................................................. 196
    6.1.3. Zusammenfassung und Interpretation der Ergebnisse zu Fragenkomplex 3: „Erklärung von differentiellen Item Funktionen“ ... 201
  6.2. Beantwortung der Hauptfragestellung ..................................... 211
  6.3. Relevanz der Ergebnisse ................................................ 212
    6.3.1. Relevanz der Ergebnisse für den Forschungsbereich „Differentielle Item Funktionen“ .................................................. 212
    6.3.2. Relevanz für Theorie und Forschung im Bereich der fremdsprachlichen Diagnostik .................................................. 216
6.3.3. Relevanz der Ergebnisse für den Bereich der interkulturellen Vergleichbarkeit von Testverfahren .......................... 220

6.3.4. Relevanz für Theorie und Forschung im Bereich der Validität .......................... 223

6.4. Grenzen der Arbeit und zukünftige Forschungsperspektiven .......................... 228

Anhang A. ............................................................................................................. 235

Literaturverzeichnis ............................................................................................ 236
Zusammenfassung


Daten gezeigt werden.

Die Analysen dieser Dissertation werden am Datensatz der europäischen EBAFLS-Studie (European Bank of Anchor Items for Foreign Language Skills; Gille & Sluiter, 2005; Fandel et al., 2007) durchgeführt. Im Rahmen der Studie wurden Daten an ca. 10.500 Schülern in acht europäischen Ländern in den Sprachen Englisch, Deutsch und Französisch erhoben; die Testitems stammten aus den Teilnehmerländern. Im Rahmen dieser Studie zeigte sich, dass viele der Items Differentielle Item Funktionen aufwiesen und daher nicht für einen fairen, kulturübergreifenden Leistungsvergleich geeignet sind (Fandel et al., 2007). Die Erhebung wurde an Schülern der 9.-11. Klasse durchgeführt. Für diese Dissertation werden Items und Datensätze der EBAFLS-Studie zur Messung des fremdsprachlichen Leseverständnisses für Englisch (Länder: Frankreich, Deutschland, Spanien, Ungarn) und Deutsch (Länder: Frankreich, Niederlande, Ungarn, Schweden) verwendet.

In dieser Arbeit wird nun der Frage nachgegangen, ob sich die in der EBAFLS-Studie gefundenen Differentiellen Item Funktionen durch die oben angesprochenen differentiellen Testkulturen erklären lassen. Grundidee dieser Dissertation ist, dass sich DIF, also durch kulturelle Zugehörigkeit verursachte Varianz der Itemschwierigkeit zwischen Gruppen, durch unterschiedliche Stärken und Schwächen der Gruppen im Hinblick auf sprachliche Teilaspekte vorhersagen lassen sollte. Annahme ist, dass diese Stärken und Schwächen durch unterschiedliche Bildungskulturen und unterschiedliche Werte im Bezug darauf, welche Aspekte für das Erlernen einer Sprache als wichtig erachtet werden, verursacht werden. Die so in den unterschiedlichen Bildungskulturen mehr oder weniger häufig unterrichteten sprachlichen Teilaspekte sollten sich ferner auf den aus einem Land stammenden Testitems abbilden. Das heißt, unterschiedliche Schwerpunkte der Länder hinsichtlich dieser sprachlichen Teilaspekte bei den Items der verschiedenen Länder (im Folgenden auch als
Testkulturen bezeichnet) können durch die Analyse von Fremdsprachenitems festgestellt werden und zugleich kann damit die durch Gruppensubjektivität verursachte Varianz erklärt werden. Der Arbeit werden drei Theoriebereiche zugrunde gelegt, nämlich Theorien und Modelle hinsichtlich Differentieller Item Funktionen, der Fremdsprachenforschung und angewandten Linguistik sowie der interkulturellen Vergleichbarkeit von Testergebnissen.


Da es Ziel dieser Arbeit ist, DIF anhand von Testkulturen zu erklären und da es sich bei DIF um Unterschiede hinsichtlich der Itemschwierigkeit bei Gruppen handelt, wird in diesem Theoriebereich darüberhinaus die Frage behandelt, welche Merkmale eines Items zur Messung fremdsprachlichen Leseverständnisses zur Schwierigkeit eines Items beitragen. Modelle aus dem Bereich Fremdsprachenkompetenzen (z.B. Bachman & Palmer, 1996) und verschiedene empirische Studien weisen darauf hin, dass insbesondere Merkmale, die zur Lösung notwendige kognitiv-linguistische Prozesse abbilden, zur Itemschwierigkeit beitragen. Solche sind beispielsweise die Schwierigkeit der grammatischen Strukturen, die Schwierigkeit des Vokabulars, die Abstraktheit des Inhalts oder die Lokalisierung der zur Lösung notwendigen Informationen (wird etwa ein
Zusammenfassung

Detail oder die Hauptidee eines Textes erfragt, liegt die Information explizit oder implizit vor). Für die Kategorisierung der Items hinsichtlich dieser Merkmale wird das Item-Kategorisierungssystem „Dutch-Grid“ (Alderson et al., 2006) zugrundegelegt, welches zum einen auf dem GERS basiert, und zum anderen eine Merkmalkategorie beinhaltet, welche sich auf diese kognitiv-linguistischen Merkmale von Items bezieht.

Der dritte dieser Dissertation zugrunde gelegte Theoriebereich ist die interkulturelle Vergleichbarkeit von Testergebnissen. Dieser Bereich ist ein Teilbereich der Disziplin der interkulturellen Psychologie. Da DIF zwischen Ländern dazu führt, dass Testergebnisse, die mit Items zustande kommen, die DIF aufweisen, interkulturell nicht vergleichbar sind, ist dieser Theoriebereich für die vorliegende Arbeit relevant.


Basierend auf den verschiedenen theoretischen und empirischen Arbeiten der Theoriebereiche lassen sich drei Annahmen ableiten, nämlich dass erstens die Itemschwierigkeit von Item-Anforderungsmerkmalen (mit) bedingt wird und sich Items hinsichtlich dieser Merkmale kategorisieren lassen, dass zweitens Items die Testkultur eines Landes repräsentieren und dass drittens unterschiedliche Testkulturen der Länder unterschiedliche Stärken und Schwächen der Gruppen hinsichtlich der Beantwortung von Items mit bestimmten Item-Anforderungsmerkmalen verursachen.
Von diesen Annahmen ausgehend wird die Hauptfragestellung abgeleitet. Diese lautet:

„Existiert ein Zusammenhang zwischen Differentiellen Item Funktionen und Indikatoren nationaler Testkulturen bei Aufgaben zur Messung des fremdsprachlichen Leseverständnisses in englischer und deutscher Sprache?“

Diese Hauptfragestellung beinhaltet die Frage danach, ob sich durch unterschiedlich häufiges Vorkommen kognitiv-linguistischer Itemmerkmale bei aus unterschiedlichen Ländern stammenden Items differentielle nationale Testkulturen abbilden lassen und ob diese zur Erklärung von Differentiellen Item Funktionen bei fremdsprachlichen Leseverständnis-Items herangezogen werden können. Insgesamt ergeben sich aus den dargelegten Annahmen drei aufeinander aufbauende Komplexe von Fragen, deren sukzessive Bearbeitung zur Beantwortung der Hauptfragestellung notwendig ist.


In einem zweiten Schritt wurden paarweise DIF-Analysen zwischen den Teilnehmerländern durchgeführt, um zu überprüfen, wie groß der Anteil von Items mit signifikanten Differentiellen Item Funktionen ist. Der Anteil bewegt sich, je nach Sprache und Länderpaarung, zwischen 39.6% und 66.4%.

Im zweiten Fragenbereich, „Erklärung von Itemschwierigkeit“, wurde mit Hilfe von korrelationsanalytischen Methoden überprüft, ob es innerhalb der Teilnehmerländer einen Zusammenhang zwischen den verwendeten kognitiv-linguistischen Item-Anforderungsmerkmalen gibt, ob dieser sich zwischen den Ländern unterscheidet und inwieweit sich die Itemschwierigkeits-Varianz anhand der Item-Anforderungsmerkmale in den verschiedenen Ländern aufklären lässt. Dies hatte zum Ziel, die Eignung der gewählten Item-Anforderungsmerkmale und des Item-Kategorisierungssystems „Dutch Grid“ zu überprüfen. Dazu wurden zunächst Korrelationsanalysen zwischen den Item-Anforderungsmerkmalen und den in Fragenbereich 1 erhaltenen Itemschwierigkeits-Parametern in den einzelnen Ländern berechnet. Für die Englisch-Items liegen die signifikanten Korrelationskoeffizienten zwischen $r = -.30 \ (p \leq 0.01)$ und $r = .431 \ (p \leq 0.01)$. Dabei bedeutet ein negativer Koeffizient, dass die Anwesenheit eines Merkmals tendenziell mit einer niedrigen Itemschwierigkeit einhergeht, Umgekehrtes gilt für positive Koeffizienten.

So ließ sich beispielsweise in allen Ländern ein signifikant negativer Korrelationskoeffizient zwischen der Itemschwierigkeit und dem Itemformat „Multiple Choice“, der Ausprägung „ausschließlich konkret“ der Variablen „Abstraktheit des Inhalts“ und der Vokabular-
gung „ausschließlich häufig“ finden, was darauf hinweist, dass diese Itemmerkmale tendenziell mit einer niedrigen Itemschwierigkeit einhergehen. Für die Deutsch-Items lagen die Korrelationen zwischen \( r = -.444 \) (\( p \leq .01 \)) und \( .505 \) (\( p \leq .01 \)). Bis auf wenige Ausnahmen unterschieden sich Größe und Richtung der Korrelationskoeffizienten weder für die Deutsch- noch für die Englisch-Items über die Länder hinweg signifikant.

Wurden die Item-Anforderungsmerkmale in einer multiplen Regression als Prädiktoren der Itemschwierigkeit eingesetzt, ließen sich für die Englisch-Items zwischen 23,5% (\( R^2 = .227 \)) und 36,7% (\( R^2 = .367 \)) der Itemschwierigkeitsvarianz innerhalb der Länder aufklären. Der Anteil der aufgeklärten Varianz lag für die Deutsch-Items zwischen \( R^2 = .208 \) und \( R^2 = .494 \). Die Ergebnisse des zweiten Fragenkomplexes wurden dahingehend interpretiert, dass sowohl die gewählten Item-Anforderungsmerkmale als auch das „Dutch Grid“ Item-Kategorisierungssystem geeignet zu sein scheinen, und dies in den unterschiedlichen Ländern auf ähnliche Art und Weise.


Die Ergebnisse zeigten, dass bei einem großen Anteil der Länderpaarungen die Tatsache, dass ein Item aus dem eigenen Land stammt, signifikant mit einer im Vergleich zur jeweils anderen Gruppe geringeren Itemschwierigkeit (bzw. vorteilhaften DIF) und die Tatsache, dass ein Item aus dem jeweils anderen Land stammt, mit einer signifikant höheren Itemschwierigkeit (bzw. nachteilhaften DIF) einherging. Bei einem weiteren Teil der Länderpaarungen war zumindest einer der beiden Zusammenhänge zu beobachten. Diese Ergebnisse wurden als ein erster deutlicher Hinweis auf den Einfluss von Itemherkunft und Testkultur auf die kulturell bedingte Varianz der Itemschwierigkeit interpretiert.
Zusammenfassung


Auch bei der Betrachtung der Zusammenhänge zwischen DIF-Parametern und schwierigkeitsbestimmenden Itemmerkmalen bei den Deutsch-Items zeigt sich, dass die signifikanten Korrelationen größtenteils in die aufgrund der Testkulturen erwarteten Richtungen deuten. Die Korrelationen bewegen sich auch hier im niedrigen bis moderaten Bereich (zwischen $r = -.470; p \leq .01$ und $r = .416; p \leq .01$). Es waren dort insgesamt 34 Korrelationen mindestens auf einem alpha-Niveau von 5% signifikant, davon entsprechen 25 hinsichtlich ihrer Richtung den Hypothesen. Insgesamt zeigten sich Korrelationen zwischen DIF und schwierigkeitsbestimmenden Merkmalen der Items, die, wenn sie signifikant waren, größtenteils den aufgrund der Testkulturen gemachten Annahmen entsprachen.

Im zweiten Analyseschritt des dritten Fragenbereichs wurde für jedes Länderpaar eine multiple Regression der Differentiellen Item Funktionen auf die Itemmerkmale durchgeführt. Ziel war hier zum einen herauszufinden, wie viel der kulturell bedingten Varianz der Itemschwierigkeiten durch die schwierigkeitsbestimmenden Merkmale insgesamt erklärt werden kann, und zum anderen, ob bzw. welche der Prädiktoren ihrer Richtung nach den aufgrund der Testkulturen aufgestellten Hypothesen hinsichtlich der zu erwartenden Stärken und Schwächen der Länder entsprachen, d.h. welcher Anteil der Varianz auf tesdkulturell bedingte Stärken und Schwächen rückführbar ist.
Bezüglich der Englisch-Items zeigte sich, dass – je nach Länderpaarung – mit Hilfe von Modellen, die ausschließlich signifikante Prädiktoren enthielten, die der Richtung nach den aufgrund der Testkulturen aufgestellten Hypothesen entsprachen, zwischen 22.7 % und 4.1% der Varianz aufgeklärt werden konnten. Bei Deutsch-Items konnten anhand analoger Modelle – je nach Länderpaarung – zwischen 3.1% und 32.7% der Varianz auf testkulturell bedingte Stärken und Schwächen der Gruppen zurückgeführt werden. Je nach Länderpaarung erwiesen sich die Effekte als unterschiedlich stabil. Häufig verringerte sich der Anteil aufgeklärter Varianz bei Ausschluss nicht signifikanter Prädiktoren stark, was insgesamt auf Multikollinearitätsprobleme hinwies.

Neben dem Anteil der anhand der testkulturell konformen Prädiktoren aufgeklärten Varianz stellten auch die Beta-Koeffizienten an sich ein interessantes Ergebnis dar. So ist es anhand ihrer Betrachtung möglich, Informationen bezüglich der Stärken und Schwächen der einzelnen Länder, jeweils relativ zu einer Vergleichsgruppe interpretiert, zu gewinnen. So kann mit Hilfe der standardisierten Beta-Koeffizienten beispielsweise die Aussage getroffen werden, dass die Tatsache, dass ein Item einen authentischen Text (wie beispielsweise einen Zeitungsartikel) verwendet, die Itemsschwierigkeit für die schwedischen Schüler im Vergleich zur französischen Gruppe um 0,41 Logits erhöht. Dies wäre dann so zu interpretieren, dass die Bearbeitung authentischer Texte eine Stärke französischer und eine Schwäche schwedischer Schüler darstellt.


Insgesamt weisen die Ergebnisse dieser Arbeit darauf hin, dass auch bezüglich des fremdsprachlichen Leseverständnisses differentielle Item Funktionen teilweise auf unterschiedliche Testkulturen zurückführbar sind, und dass dieser Umstand bei zukünftigen internationalen Leistungsvergleichen unbedingt beachtet werden sollte, um eine Minderung der Validität von Testaufgaben zu minimieren, und die Fairness und auch Aussagekraft solcher Leistungsvergleich zu erhöhen.


Im Rahmen von Umfragen geben 26% der befragten EU-Bürger an, zusätzlich zu ihrer Muttersprache mindestens zwei weitere Sprachen zu beherrschen (TNS opinion & social, 2005). Nunmehr ist es das erklärte Ziel der EU, jeden europäischen Bürger in die Lage zu versetzen, neben der jeweiligen Muttersprache in noch zwei weiteren in der EU gesprochenen Sprachen kommunizieren zu können: „Die Sprachkenntnisse sind ungleichmäßig auf die Länder und gesellschaftlichen Gruppierungen verteilt. Die Europäer sprechen nur wenige Fremdsprachen: Das Erlernen einer einzigen Lingua Franca reicht nicht aus. Jeder europäische Bürger sollte sich
außer in seiner Muttersprache in mindestens zwei anderen Sprachen gut verständigen können” (Europäische Kommission, 2003).

Gegenwärtig finanziert die Europäische Union verschiedene Programme, die das Erlernen von Fremdsprachen sowie die Forschung in diesem Bereich unterstützen sollen. Zu diesen Programmen gehört unter anderem das SOCRATES-Teilprogramm LINGUA, welches speziell mit der Förderung des Sprachenunterrichts und Sprachenerwerbs befasst ist. Mit neuen Programmen zum lebenslangen Lernen wird diese Förderung fortgesetzt (Europäisches Parlament, 2008).

In einem zeitlich annähernd parallelen Prozess hat sich die Fremdsprachenforschung verstärkt eben diesen oben angesprochenen Fragen zugewandt. In der Erkenntnis, dass die gegebene kulturelle und gesellschaftliche Vielfalt eine Vielzahl von Ansätzen auch im Bereich des Fremdsprachenerwerbs hervorgebracht hat – eine Vielfalt, die, wie noch deutlich werden wird, auch Fragen aufwirft und Probleme mit sich bringt – wurde es und ist es nach wie vor ein zentrales Anliegen der Forschung, länderübergreifende Kriterien bezüglich der Messung von Fremdsprachenkenntnissen zu formulieren und adäquate Messinstrumente zu entwickeln.


1.1. Vergleiche von Leistung und Kompetenz vor dem Hintergrund unterschiedlicher (Bildungs-)Kulturen

Insgesamt hat die Anzahl Länder- und kulturübergreifender Vergleiche von Kompetenzen in den letzten Jahrzehnten zugenommen. Dieser Trend ist vor allem bei durch die Formalbildung angeneigten Kenntnissen wie Mathematik, Naturwissenschaften (TIMSS; Third International Mathematics and Science Study (Schmidt, McKnight, Valverde, Houang & Wiley, 1997) und Leseverständnis (PISA; Programme for International Students Assessment; Prenzel et al., 2007) zu beobachten. Dabei werden üblicherweise zentrale Tests entwickelt, die dann in übersetzter Form in sämtlichen an der Vergleichsstudie teilnehmenden Ländern durchgeführt werden. Solche Vergleiche werden in der Regel in Form von Large Scale Assessments durchgeführt. In den letzten Jahren mehren sich jedoch auch die kritischen Stimmen. Es wird hinterfragt, ob diese Vergleiche vor dem Hintergrund unterschiedlicher Bildungskulturen überhaupt faire und
Ein Gegenvorschlag zu der üblichen Vorgehensweise bei Large Scale Studien kommt beispielsweise von der Gruppe „The European Network of Policy Makers for the Evaluation of Educational Systems“, welche die Ansicht vertritt, jedes Land sollte ausschließlich im eigenen Land und in eigener Sprache konstruierte Items verwenden, um Verzerrungen, die beispielsweise durch Übersetzungen oder unterschiedliche Bildungskulturen zustande kommen, zu verringern oder gar zu vermeiden (Bonnet et al., 2001). Nach Ansicht der Autoren sind Items, die auf der Basis eines gemeinsam definierten Konstrukts wie beispielsweise Leseverständnis in verschiedenen Ländern konstruiert werden, dadurch letztendlich auch miteinander vergleichbar, ohne dass zusätzliche, konstruktirrelevante Varianz, etwa durch fehlerhafte und ungenaue Übersetzungen, eingeführt wird.


Eine mögliche Erklärung für das Vorkommen von Differentiellen Item Funktionen in kulturübergreifenden Leistungsstudien basiert auf der Überlegung, dass verschiedene Kulturen oder Nationalitäten auch unterschiedliche Bildungs- und Testkulturen hervorbringen. Dies könnte sich im Fall von Fremdsprachenkenntnissen beispielsweise dadurch ausdrücken, dass in einem Land
mehr Wert auf kommunikative, in einem anderen jedoch mehr Wert auf grammatische Fähigkeiten gelegt wird. Wird diese Überlegung konsequent weitergedacht, dann könnte dies dazu führen, dass in einem Land bestimmte Teilkompetenzen durch einen diese fördernden Unterrichts- und Teststil stärker ausgebildet werden und somit ausgeprägter sind als in einem anderen Land.


1.2. Beschreibung des Dissertationsvorhabens

Diese Dissertation richtet den Fokus auf die Analyse von Items zur Messung von Fremdsprachenkenntnissen in unterschiedlichen europäischen Ländern sowie auf den Einfluss unterschiedlicher Testkulturen auf die Vergleichbarkeit und Validität dieser Items. Hauptziel ist die Erklärung von Differentiellen Item Funktionen (z.B. Holland & Wainer, 1993) bei Items zur Messung des fremdsprachlichen Leseverständnisses anhand schwierigkeitsbestimmender, kognitiv-linguistischer Itemmerkmale. Mit kognitiv-linguistischen Itemmerkmalen sind hier Charakteristika eines Items gemeint, die zur Lösung eines Items notwendige kognitive und linguistische Prozesse abbilden. Bei den verwendeten Items handelt es sich um Items zur Messung des Leseverständnisses in den Sprachen Englisch und Deutsch. Differentielle Item Funktionen und Fairness sollen ferner bezüglich ihrer Bedeutung für die Konstruktvalidität, und vor dem Hin-
tergrund der in Messicks Validitätskonzept (z.B. 1989) verwendeten Begriffe „Consequential Validity“ (d.h. der Konsequenzen von Testergebnissen und Testwertinterpretationen) und „Social Values“ (d.h. dem Einfluss der sozialen Werte einer Gesellschaft auf den Testinhalt) reflektiert werden.

Die Analysen dieser Dissertation werden am Datensatz der europäischen EBAFLS-Studie (European Bank of Anchor Items for Foreign Language Skills; Gille & Sluiter, 2005; Fandel et al., 2007) durchgeführt. Dabei handelt es sich um eine vom europäischen Rat finanzierte Studie zur Messung fremdsprachlicher Kompetenzen. Im Rahmen der Studie wurden Daten an ca. 10.500 Schülern in acht europäischen Ländern in den Sprachen Englisch, Deutsch und Französisch erhoben; die Items stammten aus den Teilnehmerländern. Im Rahmen dieser Studie zeigte sich, dass viele der Items Differentielle Item Funktionen aufwiesen und daher nicht für einen fairen, kulturübergreifenden Leistungsvergleich geeignet sind (Fandel et al., 2007). Die Studie wird ausführlicher unter 4.1 dargestellt.


signifikanter Unterschiede wird als unterschiedliche Testkulturen und differentielle Lerngelegenheiten gewertet, die wiederum Ausdruck unterschiedlicher zugrundeliegender sozialer Werte sind. Es wird überprüft, ob dies eine Ursache für Differentielle Item Funktionen darstellt.

2. Theoretische Grundlagen

Die Darstellung der theoretischen Grundlagen dieser Arbeit ist in vier Abschnitte unterteilt. Im ersten Theorieabschnitt wird zunächst dargestellt, wie Differentielle Item Funktionen definiert sind, in welchen Forschungstraditionen die Analyse von DIF wurzelt und welches die am häufigsten angewandten Methoden zur Identifikation von DIF sind. Des Weiteren wird ein Überblick über den Stand der Forschung hinsichtlich der Analyse und der Modellierung von DIF gegeben.


Abschließend werden die drei Theoriebereiche mit Messicks (1989) Validitätstheorie verknüpft, die in dieser Dissertation die Rolle einer übergreifenden Rahmentheorie einnimmt. Hierbei spielen insbesondere Aspekte der sozialen Werte, der Konsequenzen von Testwerten, Konstruktvalidität und die verwandten Konzepte von „teaching to the test” und „washback” eine prominente

2.1. Differentielle Item Funktionen


DIF kann wie folgt definiert werden:

„In IRT terms, a scale item displays DIF if examinees with the same latent-trait level have different probabilities of endorsing an item. In other words, in IRT terms, a personality or attitude item is biased if the IRCs (Item Response Curves, Anm. der Autorin) are not the same across two groups of examinees“ (Embretson & Reise, 2000, S. 319).

Das bedeutet, dass die Mitglieder zweier oder mehr Gruppen eine unterschiedliche Wahrscheinlichkeit aufweisen, ein Item korrekt zu lösen, obgleich sie sich hinsichtlich der zu messenden Fähigkeit auf dem gleichen Leistungsniveau befinden. Unterschiedliche Lösungswahrscheinlichkeiten sind in einem solchen Fall ausschließlich durch die Gruppenzugehörigkeit, bzw. durch eine (evtl. zusätzliche, nicht intendierte) Erfassung von Fähigkeiten, die in den Gruppen unterschiedlich ausgeprägt sind, bedingt (Ackerman, 1992; Roussos & Stout, 1996). Damit sind die beobachteten Unterschiede in einem solchen Fall nicht auf real vorhandene Niveauunterschiede hinsichtlich der latenten Fähigkeit zurückzuführen, die mit einem Item eigentlich erfasst werden soll.

Grundlage und unverzichtbare Voraussetzung bei der Anwendung länder- und kulturübergreifender Tests ist jedoch, dass die Items in verschiedenen Ländern oder Kulturen das gleiche Konstrukt messen, also keine DIF aufweisen. Die Kompetenz, die durch einen Test attestiert wird, soll un-
abhängig von Nationalität und kulturellem Hintergrund der Getesteten sein. Sollte sich bei einem Item hingegen zeigen, dass die Wahrscheinlichkeit einer korrekten Antwort für zwei oder mehr Länder trotz gleicher Werte hinsichtlich der latenten Fähigkeit deutlich unterschiedlich ausfällt, kann davon ausgegangen werden, dass dieser Unterschied nicht durch tatsächliche Fähigkeitsunterschiede, sondern durch andere, gruppenpezifische Faktoren, wie beispielsweise bildungskulturelle Unterschiede oder länderspezifische Inhalte und Schwerpunkte, zustande kommt. In einem solchen Fall ist davon auszugehen, dass ein Item in zwei oder mehr verschiedenen Gruppen entweder nicht das gleiche Konstrukt, oder aber unterschiedliche Teile eines Konstrukts erfasst. Ein solches Item weist dann Differentielle Item Funktionen (DIF) auf und kann unter Umständen nicht oder nur eingeschränkt zum fairen Vergleich von Fremdsprachkompetenzen verwendet werden.

Die Überprüfung von Items im Hinblick auf DIF ist ein wichtiger Aspekt von Testfairness und Validität: „The issue of test and selection encompasses many concepts and models. Primary to all of them is DIF. If test items operate in a differential fashion, then the scores for different groups are per se not comparable. This cannot lead to equitable treatments (…)“ (Holland & Wainer, 1993, S. xi). Zur Durchführung von DIF-Analysen existieren verschiedene, aus unterschiedlichen Traditionen stammende Methoden. Diese lassen sich beispielsweise dadurch unterscheiden, ob sie beobachtete oder wahre Werte, oder ob sie eine manifeste oder eine latente Variable zugrunde legen. Hierauf wird weiter unten ausführlicher eingegangen.

Ursprüngliches Ziel von DIF-Analysen war vor allem, die Fairness von Testverfahren zu erhöhen: „Most approaches to detection of differential item functioning have been designed to make tests fairer by focusing on differences between examinee groups defined by characteristics such as gender and ethnicity“ (Li, Cohen & Ibarra, 2004, S. 115). Die ersten Studien im Hinblick auf Item-Bias wurden in den 1960er Jahren vor dem Hintergrund der amerikanischen Bürgerrechtsbewegung entwickelt. Damit sollte die Hypothese überprüft werden, dass beobachtete Leistungsunterschiede zwischen Testpersonen mit unterschiedlichem kulturellen Hintergrund oder unterschiedlicher Hautfarben durch verzerrte Items zustande kommen. Es wurde argumentiert, dass diese aufgrund von kulturellen Vorprägungen und Bedingungen unfair und möglicherweise für Mitglieder anders geprägter kultureller Gruppierungen, häufig Minderheiten, daher nicht korrekt beantwortbar seien, und daher keine Rückschlüsse auf tatsächliche Gruppenunterschiede hinsichtlich bestimmter kognitiver Fähigkeiten zuließen. Das DIF-Konzept wurzelt in dem Versuch, Methoden zur Überprüfung dieser Hypothese und zur Überprüfung der Fairness einzelner Items zu finden (Cole, 1993).

Eine weitere wichtige Begrifflichkeit im Rahmen von DIF-Analysen bezieht sich auf die Unterscheidung von „Item Impact“ und „Item Bias“. Hierbei handelt es sich um den Unterschied zwischen einem tatsächlichen Unterschied der Gruppen in ihrer Leistung (Impact), und einem durch die Gruppenzugehörigkeit (und somit anderen, kulturellen bzw. gruppenspezifischen Faktoren) verursachten Unterschied im Test-Score (Bias): „DIF refers to a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The comparison of matched or comparable groups is critical because it is important to distinguish between differences in item functioning from differences between groups“ (Dorans & Holland, 1993, S. 35). Die Kontrolle von Leistung bzw. der zu messenden latenten Fähigkeit ist somit das Entscheidende bei DIF-Analysen. Als Folge der Unterscheidung von DIF und Impact und der nun eher statistisch-psychometrischen Bedeutung von DIF beschäftigte sich die Forschung in diesem Bereich lange Zeit hauptsächlich mit der Entwicklung neuer Methoden für die Erfassung und Analyse von DIF, um unter anderem folgende Fragen zu klären:

Welche Methoden eignen sich am Besten für welche Gruppen in welchen Kontexten?

Welche Implikationen hat DIF für die Interpretation von Tests und Testwerten?

Ab wann ist DIF als substantiell zu bezeichnen, und wann sollte ein Item aus einem Test entfernt werden?


Lord (1980) beschreibt DIF bzw. Bias im Rahmen von IRT-Modellen wie folgt: „If each test item in a test had exactly the same item response function in every group, then people of the same ability or skill would have exactly the same chance of getting the item right, regardless of their group membership. Such a test would be completely unbiased. If, on the other hand, an item has a different item response function for one group than for another, it is clear that the item is biased” (S. 212). Unterschiedliche ICCs bei ein- und demselben Item in verschiedenen Gruppen können also auf das Vorhandensein von DIF hinweisen. DIF beziehen sich also auf die Differenz von Item-Schwierigkeits-Parametern. Wenn sich die Itemcharakteristischen Funktionen zweier Gruppen unterscheiden, existierten DIF (Thiessen, Steinberg & Wainer, 1993).

Die unterschiedlichen, zur Analyse von Differentiellen Item Funktionen verwendeten eindimensionalen IRT-Modelle unterscheiden sich vor allem dahingehend, welche bzw. wie viele Parameter für die Beschreibung der Antwortfunktion zugrunde gelegt werden. Im einfachsten Modell, dem Rasch-Modell, wird nur der Schwierigkeitsparameter im Hinblick auf Gruppenunterschiede
betrachtet. Hier unterscheiden sich im Falle von DIF die ICCs zweier Gruppen bezüglich eines Items hinsichtlich ihrer Position auf der X-Achse bzw. des Itemschwierigkeitsparameters. Dies wird in Abbildung 2.1 dargestellt.

![Abbildung 2.1. Differentielle Item Funktion in einem 1PL-IRT-Modell (uniform DIF)](image)

Die beiden Item Response Funktionen in Abb. 2.1 stellen die Funktionen zweier unterschiedlicher Gruppen hinsichtlich ein und desselben Items dar. Auf der X-Achse sind die Personenfähigkeit Theta und der Itemschwierigkeitsparameter Sigma gemeinsam abgetragen, als Skaleneinheit wurden Logits gewählt. Auf der Y-Achse ist die Wahrscheinlichkeit abgetragen, ein Item korrekt zu beantworten.

Bei einem Theta-Wert von 0 Logits liegt hier die Wahrscheinlichkeit für eine korrekte Antwort für die eine Gruppe bei ca. 20 %, für die andere Gruppe jedoch bei 40%. Der Unterschied zwischen den beiden Gruppen hinsichtlich der Wahrscheinlichkeit einer richtigen Antwort beträgt somit ca. 20 %, trotz gleichem latenten Fähigkeitsniveaus, und hängt hier somit nicht von der Fähigkeit, sondern ausschließlich von der Gruppenzugehörigkeit ab. Diese Art von DIF wird auch als „uniform DIF“ bezeichnet, da sich die Gruppen hier ausschließlich hinsichtlich der Position auf der X-Achse unterscheiden, die Form der Kurven (Steigung und Schnittpunkt mit der Y-Achse) jedoch dieselbe ist. Man könnte auch sagen, bei „uniform“-DIF handelt es sich im Prinzip um den Haupeffekt der Gruppenzugehörigkeit (Zumbo, 2007). Dieses Modell wird auch als


![Abbildung 2.2. DIF in einem 2PL-IRT-Modell (non-uniform DIF)](image)

Abbildungen 2.2 stellt die unterschiedlichen ICCs verschiedener Gruppen hinsichtlich ein und desselben Items dar. Hier ist zu sehen, dass diese sich nicht nur hinsichtlich der Position auf der X-Achse, sondern auch hinsichtlich der Steigung unterscheiden. Mit diesen Modellen wird „non-uniform“ DIF untersucht.
Ferner existieren auch sogenannte 3PL-Modelle, die einen Rateparameter $c$ mit einbeziehen. Dieser repräsentiert die Wahrscheinlichkeit einer richtig geratenen Antwort für ein Item. Der Parameter bezeichnet den Schnittpunkt der ICC mit der $y$-Achse. Hier wird eine apriori angenommene Rate-Wahrscheinlichkeit zu der Wahrscheinlichkeit einer korrekten Antwort hinzu addiert.


Mittlerweile werden häufig die auf der Item Response Theorie basierenden Methoden zur De- tektion von DIF verwendet. IRT-Methoden besitzen für die Analyse und Erforschung von DIF mehrere Vorteile gegenüber den oben beschriebenen, klassischen Verfahren. So sind beispielsweise Schätzungen der IRT-Parameter weniger konfundiert mit Stichprobencharakteristika (Hambleton, Swaminathan & Rogers, 1991). Auch können die statistischen Eigenschaften eines
Items besser und präziser beschrieben werden und somit auch die Unterschiede zwischen zwei Gruppen im Hinblick auf das Item (Camilli & Shepard, 1994).


Die vorliegende Arbeit folgt dieser Empfehlung, da DIF hier als ein Maß für kulturell bedingte Gruppenunterschiede verwendet wird. Infolgedessen werden alle Items in die Analysen einbezogen. Demzufolge wird jeder Unterschied zwischen zwei Itemcharakteristischen Funktionen bezüglich ein und desselben Items als relevant erachtet.
2.1.1. DIF und Validität: „Nuisance-Dimension” oder Ausdruck differentieller Stärken und Schwächen?

Wie oben ausgeführt, besteht eine Erklärung für DIF in der Annahme, dass es sich dabei um eine nicht modellierte bzw. nicht beachtete Mehrdimensionalität eines Items handelt. Dabei wird deutlich, dass DIF hoch relevant für die Validität und vor allem für die Konstruktvalidität eines Tests ist: DIF kann ein Indikator dafür sein, dass in Abweichung von der ursprünglichen Testintention anhand desselben Items entweder unterschiedliche Konstrukte oder aber zusätzliche Dimensionen desselben Konstrukts mit unterschiedlichen Ausprägungen bei den Gruppen erfasst werden.

Teilweise wird in der Literatur bei der Erörterung und Deutung von DIF der Begriff „Nuisance Dimension“ verwendet (z.B. Ackerman, 1992). Der Terminus beinhaltet, dass durch ein Item mindestens eine zweite, konstruktirrelevante Dimension unbeabsichtigerweise mit erfasst wird. Bezüglich dieser zweiten Dimension existieren in den Gruppen unterschiedliche Leistungsverteilungen, was dieser Argumentation zufolge die Ursache für DIF ist. „The presence of an item that functions differentially on a test indicates that the item is measuring some nuisance dimension”, (…) „If two different groups of examinees have different underlying multidimensional ability distributions and the item tests are capable of discriminating among levels of abilities on these multiple dimensions, then any unidimensional scoring scheme has the potential to produce item bias” (Ackerman, 1992, S. 67). Auch Roussos und Stout (2004) argumentieren in diese Richtung:

„The general purpose of conducting a DIF analysis is to help ensure test equity or fairness. The statistical flagging of items that exhibit evidence of DIF represents an essential contribution toward the achievement of this objective. Because tests are inherently multidimensional and multidimensionality is the basic cause of DIF, increased understanding of test dimensionality and the effects of these dimensions on DIF hold the potential for a more accurate interpretation of the test score, more control over the influence by unintended and irrelevant nuisance dimensions.” (Roussos & Stout, 2004; zitiert aus Geranpayeh & Kunnan, 2007, S. 207).

Demnach sind im Rahmen einer solchen Betrachtungsweise Testwertergebnisse, die auf Items mit signifikanten DIF-Parametern im Rahmen eindimensionaler Modelle basieren, nicht interpretierbar, da die gefundenen Unterschiede nicht eindeutig auf die Fähigkeit zurückgeführt werden können, die zu messen ein Item intendiert. Solche Items werden daher üblicherweise aus
Testverfahren ausgeschlossen, da eine faire Interpretation der Testwerte nicht möglich ist.


„Statistics used to detect differential item functioning can also reflect differential strengths and weaknesses in the performance characteristics of population subgroups. In turn, item features associated with the differential performance patterns are likely to reflect some facet of the item task and hence its difficulty that might previously have been overlooked“ (S. 109).


2.1.2. Die Erklärung von Differentiellen Item Funktionen


Auch für die deutsche Sprachgruppe war ein Vorteil bei deutschen Items zu beobachten. Uneindeutig waren die Ergebnisse jedoch beispielsweise bei schwedischen, spanischen und finnischen Gruppen, hier zeigten sich sowohl Vor- als auch Nachteile durch eigene Items.


Zusammenfassend lässt sich konstatieren, dass empirische Studien Hinweise darauf geben, dass Zusammenhänge zwischen DIF und unterschiedlichen Testkulturen existieren. Diese Zusammenhänge lassen sich anhand von Itemcharakteristika, die aus der Analyse von im Unterricht verwendetem Aufgabenmaterial gewonnen wurden, abbilden, teilweise sogar besser als durch eine videobasierte Unterrichtsanalyse (Klieme & Bos, 2000).

vorliegenden Arbeit für die Analyse des Zusammenhangs zwischen Differentiellen Item Funktionen bei Items des fremdsprachlichen Leseverständnisses und Indikatoren nationaler Testkulturen angewandt werden.

Aufgrund der vorliegenden Literatur liegt die Hypothese nahe, dass bei der Entstehung von DIF einerseits bestimmte schwierigkeitsbestimmende Eigenschaften von Items, und andererseits Unterschiede zwischen den Gruppen eine Rolle spielen, die nicht nur aufgrund der zusätzlichen, nicht intendierten Messung einer konstruktirelevanten Dimension zustande kommen. Mit der Frage, welche Eigenschaften das im Anwendungskontext dieser Arbeit, also der Fremdsprachenforschung, sein könnten, befassen sich die Ausführungen im nächsten Theorieabschnitt.

2.2. Fremdsprachenforschung und angewandte Linguistik


Zunächst wird in Abschnitt 2.2.1 ein allgemeiner Überblick über das Konstrukt sowie in Abschnitt 2.2.2 über die Messung der fremdsprachlichen Lesekompetenz gegeben. Für diese Arbeit relevante Modelle der Fremdsprachenkompetenz werden dargestellt. Darauf folgend wird im dritten Abschnitt (2.2.3) der Gemeinsame Europäische Referenzrahmen für Sprachen (Europarat, 2001), der für diese Dissertation die zentrale theoretische und praktische Basis darstellt, hinsichtlich der dort zugrunde liegenden theoretischen Modelle und Konstrukte besprochen, sowie näher auf Hintergrund und Funktion des GERS eingegangen. Ferner werden relevante empirische Forschungsergebnisse sowie Kritik am GERS dargestellt. Darauf aufbauend wird im vierten Teil dieses Abschnitts unter 2.2.4 die Eignung der GERS-Skalen für die Testkonstruktion und Messung diskutiert. Das auf dem GERS basierende Item-Kategorisierungs-Instrument „Dutch Grid“ (Alderson et al., 2006), welches in dieser Arbeit für die Kategorisierung von Items hinsichtlich ihrer kognitiv-linguistischen Merkmale verwendet wurde, wird vorgestellt. Weiterhin wird der Stand der Forschung im Hinblick auf schwierigkeitsdeterminierende Itemmerkmale im Bereich des fremdsprachlichen Leseverständnisses dargestellt.
2.2.1. Das Konstrukt der fremdsprachlichen Lesekompetenz


Der „Gemeinsame Europäische Referenzrahmen für Sprachen“, welcher die theoretische Grundlage der Arbeit darstellt (siehe auch 2.2.3), entstammt dieser Tradition, daher wird in dieser Arbeit darauf ein besonderer Fokus gerichtet. Im Folgenden wird nun auf unterschiedliche theoretische Richtungen und Ursprünge von Fremdsprachenmodellierung eingegangen.
**Pragmatische/produktorientierte und psycholinguistische Theorien.** Die unterschiedlichen Theorien und Modelle der Fremdsprachenkompetenz können grob zwei Richtungen zugeordnet werden: zum einen der eher **pragmatischen und produktorientierten**, zum anderen der **kognitiven Forschung** und der Psycholinguistik.


„An alternative approach to examining the process of reading is to inspect the product of reading and, often, to compare that product with the text originally read. It is sometimes said that, although different readers may engage in very different reading processes, the understandings they end up with will be similar. Thus, although there may be many different ways of reaching a given understanding, what matters is not how you reach the understanding, but the fact that you reach it, or, to put it another way, what understanding you reach” (S. 4).

Zweitens bewegt sich diese Arbeit auf einer **systemischen Ebene**, in diesem Fall auf der Ebene von Ländern und Sprachgruppen, und nicht auf der Individualebene. Daher können hier kognitive Prozesse nur auf der Gruppenebene untersucht werden. Da außerdem in den für die vorliegende Arbeit verwendeten Daten keine kognitiven Prozessvariablen erhoben wurden, muss hier der Umweg über die Analyse von Itemmerkmalen und deren Einfluss auf die korrekte Beantwortung von Items gegangen werden, indem davon ausgegangen wird, dass bestimmte Itemmerkmale bestimmte cognitive Prozesse zur korrekten Beantwortung einer Frage erfordern. Diese Merkmale sowie die diesen vermutlich zugrundeliegenden kognitiven Prozesse werden unter 2.2.4 im Rahmen des Abschnitts über Schwierigkeitsdeterminanten von Items beschrieben.
In den folgenden Abschnitten werden verschiedene Aspekte der Geschichte und Entwicklung von Modellierung und Messung der Fremdsprachenkompetenz dargestellt. Dabei spielen einige Aspekte eine besondere Rolle, nämlich die Dimensionalität des Konstrukts, die Frage nach der globalen vs. diskreten Messung, Prozess- vs. Produktmessung, sowie das in den unterschiedlichen Modellen zugrunde gelegte Verhältnis von Kompetenz und Performanz.


Ansätze zur Testung von Performanz stammten vor allem aus dem Bereich des auswärtigen und diplomatischen Dienstes in den USA. Dort wurden Tests zur Erfassung der gesprochenen Sprache im Rahmen von Einstellungstests entwickelt (McNamara, 1996). In den 1960er Jahren ergab sich außerdem durch eine zunehmende Anzahl von ausländischen Studenten an amerikanischen und britischen Universitäten die erhöhte Notwendigkeit zum Testen der Sprachfähigkeiten in...


Brown (1996) beschreibt die Sicht der „Variationisten“, wie sie sie bezeichnet, wie folgt:

„These scholars (Ellis, 1990 „Variable competence model“ ; Anm. d. Autorin) believe that the phenomenon of systematic variability in the utterances produced by second language learners has to be built into any model of second language acquisition and, not only that, but this variability must be represented in the competence of the learner, since the learner does not manifest homogeneous control of structures” (S. 59).

Auch Tarone (1985) postuliert einen Zusammenhang zwischen Performanz und zugrundeliegender Kompetenz: „( … ) the systematic variability which is exhibited in the learner’s performance on a variety of elicitation tasks actually reflects his/her growing capability in IL (Inter Language, Anm. d. Autorin), and is not just a performance phenomenon” (Tarone, 1985, S. 35; zitiert aus Brown, 1996).

Die Gegenposition zu dieser Aussage (z.B. Gregg, 1990) wiederum beinhaltet, dass dies eher eine Beschreibung von Performanz denn der zugrundeliegenden Kompetenz darstellt. Das Problem, das sich im Rahmen dieser Debatte aufzeigt, beschreibt Ellis (1990) wie folgt: „(how are) we supposed to construct a theory of L2 competence when the only data available are performance data?” (Ellis, 1990, S. 388, zitiert aus Brown, 1996).


<table>
<thead>
<tr>
<th>Canale</th>
<th>Van Ek</th>
<th>Bachmann</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical Competence</td>
<td>Linguistic Competence</td>
<td>Language Knowledge</td>
</tr>
<tr>
<td>• Lexical items</td>
<td>• Language functions</td>
<td></td>
</tr>
<tr>
<td>• Rules of word formation</td>
<td>• General notions</td>
<td></td>
</tr>
<tr>
<td>• Sentence formation</td>
<td>• Specific notions</td>
<td></td>
</tr>
<tr>
<td>• Literal meaning</td>
<td>• Grammar &amp; Intonation</td>
<td></td>
</tr>
<tr>
<td>• Pronunciation</td>
<td>• Vocabulary &amp; Idiom</td>
<td></td>
</tr>
<tr>
<td>• Spelling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socio-linguistic Competence</td>
<td>Socio-linguistic Competence</td>
<td>Socio-linguistic Competence</td>
</tr>
<tr>
<td>• Appropriateness of meanings and forms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discourse Competence</td>
<td>Discourse Competence</td>
<td>Textual Competence</td>
</tr>
<tr>
<td>• Cohesion and Coherence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strategic Competence</td>
<td>Compensatory Competence</td>
<td>Strategic Competence</td>
</tr>
<tr>
<td>• Enhances the rhetoric effect of utterances</td>
<td>• Assessment</td>
<td>• Assessment</td>
</tr>
<tr>
<td></td>
<td>• Planning</td>
<td>• Planning</td>
</tr>
<tr>
<td></td>
<td>• Execution</td>
<td>• Execution</td>
</tr>
<tr>
<td>Socio-cultural Competence</td>
<td>Psycho physiological Mechanisms:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Mode: receptive / productive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Channel: oral / aural / visual</td>
<td></td>
</tr>
</tbody>
</table>


fünften Kompetenzbereich. Bei Van Ek handelt es sich dabei um die soziokulturelle Kompetenz und bei Bachman um die psycho-physiologischen Mechanismen, die sich hier auf den Modus, d.h. rezeptiv oder produktiv, sowie den Kanal, d.h. den jeweils angesprochenen Sinn (auditiv/oral vs. visuell), bezieht.


automatischer und routinierter Anwendung des Wissens. Die Kontrolldimension bezieht sich hingegen auf die drei ausführenden Funktionen (Informationsauswahl, Koordination von Informationen, Niveau der Automatisierung und Sprachflüssigkeit). In späteren Veröffentlichungen (Bialystok, 1986) bezeichnet sie die beiden Dimensionen auch als Analyse und Kontrolle (De Jong & Verhoeven, 1992).


**Sprachkompetenz als Entwicklungsprozess: niveau- und kriterienorientiertes Testen.**


“(…) they define achievement testing as a method of indexing stages of competence through indicators such as integration of knowledge, degree of procedural skill, speed of access to memory, and degree of automaticity. Because acquiring language proficiency is a dynamic process, tests may be viewed from a developmental perspective. Given the assumption that in the course of time the learner’s language represents successive interlanguages (Selinker, 1971), a test aims at identifying at what stage of a developmental process a person is located“ (De Jong & Verhoeven, 1992, S. 8).

Als „interlanguage“ gilt ein sich entwickelndes linguistisches System bei einer Person, die eine Sprache zwar lernt, aber noch nicht vollständig und perfekt gemeistert hat. Dem in diesem Zitat ausgedrückten Gedanken liegt zum einen die Idee einer veränderbaren, variablen

**Dimensionalität des Konstrukts.** Bezüglich der Erforschung und Messung von Fremdsprachenkompetenz wird auch die Dimensionalität und notwendige Komplexität des Konstrukts diskutiert. Dabei ist das verwendete Konzept auch immer stark vom jeweiligen Forschungsziel abhängig: sollen Fremdsprachenkenntnisse erfasst werden (Performanz), oder ist es das Ziel, tiefer gehende Kenntnisse hinsichtlich des Konstrukts, d.h. der zugrunde liegenden Kompetenz, zu erlangen? Für Ersteres ist häufig die Annahme der Eindimensionalität des zu messenden Konstrukts vorteilhafter, da unkomplizierter zu operationalisieren; für Letzteres benötigt es eher komplexere, mehrdimensionale Modelle: „From a theoretical point of view, neither the unitary competence hypothesis, nor extremely complex models are beneficiary. Explaining all variation by a single factor, in fact puts an end to all research into a deeper understanding of language, its acquisition and its use. Extremely complex models on the other hand fail to achieve what models are for, i.e., to explain reality by a simplification” (De Jong & Verhoeven, 1992, S. 5). Die Frage nach der Dimensionalität des Konstrukts ist insofern relevant für die vorliegende Arbeit, als von einer Eindimensionalität ausgegangen wird. Dieses basiert auf den theoretischen Annahmen des der Arbeit zugrundeliegenden gemeinsamen europäischen Referenzrahmens (Europarat, 2001). Dieser wird unter 2.2.3 dargestellt.

**2.2.2. Die Messung von Fremdsprachenkompetenzen**

**Internationale Studien zur Messung und zum Vergleich von Fremdsprachenkenntnissen.**


**Ansätze zur Messung von Fremdsprachenkenntnissen.** Unter Berücksichtigung der unter 2.2.1 beschriebenen Debatten hinsichtlich Kompetenz und Performanz, der Dimensionalität des Konstrukts sowie der Frage des holistischen vs. diskreten Testens lassen sich verschiedene Richtungen von Testverfahren unterscheiden.

Zur Erfassung von Sprachkompetenz werden häufig Tests für die vier *Grundfähigkeiten von Sprache*, nämlich Lesen, Sprechen, Hören und Schreiben, getrennt entwickelt und vorgegeben. Neben der Unterteilung in diese vier Fähigkeiten werden verschiedentlich Dichotomisierungen vorgenommen, beispielsweise die Unterscheidung hinsichtlich des Kommunikationskanals (oral vs. geschrieben, bzw. auditiv vs. visuell, wobei jeweils Ersteres sich auf Sprechen und Hören, und Letzteres sich auf Lesen und Schreiben bezieht) und des Kommunikationsmodus, der in produktive (Sprechen, Schreiben) vs. rezeptive (Hören, Lesen) Fähigkeiten unterteilt wird. (siehe Tabelle 2.2).

<table>
<thead>
<tr>
<th>Kommunikationskanal</th>
<th>Kommunikationsmodus</th>
</tr>
</thead>
<tbody>
<tr>
<td>visuell</td>
<td>rezeptiv</td>
</tr>
<tr>
<td></td>
<td>Lesen</td>
</tr>
<tr>
<td></td>
<td>Schreiben</td>
</tr>
<tr>
<td>auditiv</td>
<td>produktiv</td>
</tr>
<tr>
<td></td>
<td>Hören</td>
</tr>
<tr>
<td></td>
<td>Sprechen</td>
</tr>
</tbody>
</table>

**Tabelle 2.2.** Die Grundfähigkeiten von Sprache

Sprachtests können sich ferner dahingehend unterscheiden, ob sie auf eine *direkte* oder eine *indirekte Messung* der Sprachkompetenz abzielen (De Jong & Verhoeven, 1992). Die direkte Messung bezieht sich auf eine Situation, in der die lernende Person möglichst natürlich kommuniziert und so völlig unbewusst auf linguistische und grammatische Regeln zurückgreift. Bei einem indirekten Test formuliert der zu Testende die linguistischen Regeln, die zur Lösung einer Aufgabe notwendig sind, ganz bewusst.

Eine weitere Unterscheidung ergibt sich durch die oben bereits angesprochene Unterscheidung von *diskreten Punktmessungen* und *integrativen, eher globalen Messungen*. Erstere erstellt ein Profil des Getesteten hinsichtlich bestimmter Komponenten linguistischer Kenntnisse und Fähigkeiten, d.h. hinsichtlich diskreter Sprachelemente. Bei Letzteren handelt es sich um globalere Messungen und Interpretationen von Sprachkompetenz.


Bezugssysteme und Referenzrahmen zur Beschreibung von Sprachkompetenz.

Es existieren vielfältige Ansätze bei der Entwicklung von Skalen zur Messung und Beschreibung von Sprachkompetenz: „There have been many attempts to define levels of language proficiency by developing scales, with detailed descriptions of each point, level or band, on the scale.” (Alderson, 2000, S. 278). Grundlage dieser Art von Skalen war der Gedanke, das Konstrukt der Fremdsprachenfähigkeit als Entwicklungsmodell zu betrachten. So wurde, wie oben bereits kurz angesprochen, vor allem in den 80er und 90er Jahren des 20. Jahrhunderts damit begonnen, Skalen oder Referenzrahmen zu entwickeln. Ziel war es dabei zu beschreiben, was beispielsweise Leser auf unterschiedlichen Entwicklungsstufen bzw. Leistungsniveaus voneinander unterscheidet (niveauorientiertes Testen), und was genau ein Leser auf einem jeweiligen Niveau kann bzw. können sollte (kriterienorientiertes Testen). Gemeinsam ist allen diesen Skalen, dass sie Verhalten von Personen, d.h. die Performanz, mit Hilfe sogenannter Deskriptoren beschreiben.


Skalen dieses Typs werden jedoch durchaus auch kritisch betrachtet. Ein Kritikpunkt ist, dass Skalen apriori-definierte Niveaus zugrunde legen, ohne diese empirisch validiert zu haben. Es stellt sich hier also die Frage nach der Validität der Skalen und somit auch des angenommenen zugrundeliegenden Konstrukts. Ferner stellt bei Rating-Skalen die differenzierte Zuordnung von Aufgaben zu Niveaus häufig ein Problem dar: „Since a main purpose of descriptors is to ’anchor’ judgements, as in ’Behaviourally Anchored Rating Scales’, the effect of conventions and clichés not based on any empirical evidence may be to systematise the very judgement error the definitions are intended to help avoid (Landy & Farr, 1983)” (North, 2000, S. 14). Bei einer Zuordnung von Aufgaben zu Niveaus ohne eine explizite empirische Validierung kann es sich daher auch um ein lediglich aus übernommenen Konventionen hervorgehendes und damit wenig aussagefähiges Resultat handeln.


„By the definition of hierarchy, high level skills and text types subsume low ones so that readers demonstrating high levels of reading proficiency should be able to interact with texts and be able to demonstrate the reading skills characteristic of low levels of proficiency. Conversely, readers at low levels of the proficiency scale should neither be able to demonstrate high level skills not interact with high level texts.” (Lee & Musumeci, 1988, S. 173; zitiert nach Alderson, 2000, S. 278).

Die Beschreibung von Lesefähigkeit für die jeweiligen Niveaus der ACTFL-Skala bezieht sich jeweils auf die zu verstehenden Textsorte, die Art der Lesefähigkeit sowie die aufgabenbezogene Performanz (Alderson, 2000). So ist ein spezifisches Entwicklungs niveau meist mit dem Verständnis einer bestimmten Textsorte sowie den zu deren Verständnis notwendigen Lesefähigkeiten assoziiert. Hauptkritikpunkt auch an dieser Skala ist das oben bereits angesprochene Problem, dass sie sich auf die apriori-Definitionen der Niveaus verlässt, ohne diese Annahmen empirisch validiert zu haben.
2.2.3. Der Gemeinsame Europäische Referenzrahmen für Sprachen

Der Gemeinsame Europäische Referenzrahmen für Sprachen (GERS; Europaritat, 2001) stellt die theoretische Basis dieser Arbeit dar. In diesem Abschnitt werden die theoretischen Grundlagen, die Entwicklung, die Validierung, Anwendungsmöglichkeiten sowie die Kritik am GERS dargestellt und diskutiert.


Der GERS unterteilt Sprachkompetenz in sechs unterschiedliche Niveaus, und zwar von A1 bis C2. Dabei entspricht das Niveau A (A1, A2) einer elementaren, Niveau B (B1, B2) einer selbständigen und Niveau C (C1, C2) einer kompetenten Sprachverwendung (siehe auch Tabelle 2.3).

Im Prozess der Entwicklung des GERS wurde versucht, einige der oben genannten Kritikpunkte zu berücksichtigen und eine in dieser Hinsicht verbesserte Skala zu erschaffen. So zieht der GERS beispielsweise als Vergleich zum jeweils erreichten Leistungsstand nicht den Mutter sprachler heran, sondern eine fiktive Person, die eine Sprache als Fremdsprache so gut wie möglich erlernt. Auch sind die verwendeten Skalen — bzw. die dort verwendeten Deskriptoren — empirisch validiert worden (Schneider & North, 2000).

Der GERS beschreibt die Verwendung und das Erlernen von Sprachen als kompetenzbasiert und betrachtet Kompetenzen von einem globalen, plurilingualen, plurikulturellen Standpunkt
aus (Heyworth, 2004): „A given individual does not have a collection of distinct and separate competences to communicate depending on the languages he/she knows, but rather a plurilingual and pluricultural competence encompassing the full range of the languages available to him/her (GERS, S. 168, zitiert aus Heyworth, 2004).”

Dementsprechend ist das „Herz” (Heyworth, 2004) des GERS die sogenannte Globalskala, oder auch „Common Scale of Reference” (GERS, S. 32ff). Durch das Einbeziehen aller vier Sprachgrundfähigkeiten wird Sprachfähigkeit insgesamt als eine global zu betrachtende Fähigkeit angesehen. So liest sich die Beschreibung der Globalskala für das Niveau B1 beispielsweise wie folgt:


Allen im GERS verwendeten Skalen sind folgende Dinge gemein:

1. Alle Statements sind *positiv formuliert*.
2. Im Rahmen des GERS wird Fremdsprachenkompetenz als eine zwar in verschiedene Bereiche *unterteilbare*, insgesamt jedoch *globale Kompetenz* betrachtet.
3. Die Skalen bestehen aus sogenannten Deskriptoren, die sich aus kurzen Aussagen dazu, was ein Sprachenlerner in den verschiedenen Bereichen für ein bestimmtes Niveau mindestens können muss, sogenannten Can-do-statements, zusammensetzen. Diese, auch als *Deskriptorskalen* bezeichneten untergeordneten Skalen wurden von Experten zusammengestellt und überprüft (Schneider & North, 2000; North, 2000).
Die in der Globalskala beschriebenen Niveaus A1 bis C2 entsprechen folgender Unterteilung:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sprachverwendung</td>
<td>Sprachverwendung</td>
<td>Sprachverwendung</td>
</tr>
<tr>
<td>A 1</td>
<td>A 2</td>
<td>B 1</td>
</tr>
<tr>
<td>(Breakthrough)</td>
<td>(Waystage)</td>
<td>(Threshold)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Operational)</td>
</tr>
</tbody>
</table>

Tabelle 2.3. Referenzniveaus GERS (Europarat, 2001, S. 34)

Sprachkompetenz wird im Rahmen des GERS ferner in Grundfähigkeiten unterteilt, was sich wie folgt (Tabelle 2.4) darstellt:

<table>
<thead>
<tr>
<th>C2 bis A1</th>
<th>Verstehen</th>
<th>Sprechen</th>
<th>Schreiben</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hören</td>
<td>Lesen</td>
<td>An Gesprächen teilnehmen</td>
</tr>
</tbody>
</table>

Tabelle 2.4. entnommen aus der Selbsteinschätzungsskala des GERS. (Europarat, 2001, Online-Version)

Wie aus Tabelle 2.4 zu entnehmen ist, existieren zusätzlich zu der Globalskala außerdem Subskalen für die Grundfähigkeiten Verstehen (Hören, Lesen), Sprechen (an Gesprächen teilnehmen; zusammenhängendes Sprechen) und Schreiben.

Die Deskriptorskalen für diese drei verschiedenen Subskalen zielen darauf ab, Sprachen können detaillierter zu beschreiben. Für die Deskriptorskalen von „Hörverständnis“ und „Leseverständnis“ werden im Folgenden einige Beispiele gegeben:


„Skehan (1995 a, S. 16) considers that it is in fact “misconceived to see competence as underlying performance in any straightforward manner” and proposes “ability to use” as something separate from both competence and performance which is itself related to Bachman’s (1990) interpretation of strategic competence.” (S. 2 f.).


„In this study, relating proficiency categories that are meaningful for teachers and assessors to a competence model has been, to say the least, difficult. The approach adopted takes the more behavioural view of proficiency outlined by Parks (1985), broadening the definition of pragmatic competence to include Skehan’s Ability for Use, Spolsky’s Knowing how to use a language, and Fillmore’s Fluency (…). An attempt has also been made to take account of the issue of variability of performance conditions in the design of the rating scale used in conjunction with the descriptors” (North, 2000, S. 53 f).

theory and research is inadequate to provide a basis for it. Whilst relating to theory, it must also be relevant to the contexts of the learning population concerned, and it must remain user-friendly-accessible to practitioners (Council of Europe, 2000: 3.1; S. 25).”


Validierung. Die hier dargestellte Studie von Schneider und North (2000) stellt die empirische Basis des GERS dar und war bereits als Entwicklungsgrundlage für ein gesamteuropäisches Instrument angelegt. Ziel dieser Studie war es, zur Entwicklung der Skalen eines Referenzrahmens für Sprachen in Europa beizutragen. Durchgeführt wurde sie in der Schweiz. Diese wurde aufgrund der dort herrschenden Multilingualität und der durch die Unabhängigkeit der Kantone nicht-zentralistischen Bildungspolitik ausgewählt: „The French and German-Speaking cantons have pedagogic cultures which are quite similar to those of neighboring states speaking the same languages, which gives the Swiss educational System a distinct pluralism” (North, 2000, S. 5). Die Schweiz wurde ferner als komplex genug angesehen, um als Testland für Gesamteuropa zu fungieren: „Switzerland offers the opportunity to develop a common framework scale taking account of different educational sectors, different language regions and pedagogic cultures, and different mother tongues” (North, 2000, S. 5).


**Vergleichbarkeit des GERS in unterschiedlichen Bildungskulturen.** Neben den oben beschriebenen Analysen wurden die Deskriptoren außerdem hinsichtlich Differentieller Item Funktionen, bezogen auf die verschiedenen Sprachregionen und Bildungssysteme der Schweiz, untersucht. Dabei zeigte sich, dass durchaus Differentielle Item Funktionen zu finden waren. Diesbezüglich wurde folgende Schlussfolgerung gezogen: „Many items which showed variability across regions or across sectors appeared nonetheless to be good items, well calibrated, well fitting, sensible, saying something. For example, on Questionnaire ’Independence’ (…), does the fact that the 3 listening comprehension items failed 95% confidence intervals on regions —the French-speaking region considering them much more difficult than the German-speaking region—make them bad items?” (Schneider & North, 2000). Hier ist deutlich herauszulesen, dass mögliche sprachlich-kulturell bedingte Unterschiede bei der Einschätzung der Schwierigkeit bestimmter Testitems nicht als Ausschlusskriterium für Deskriptoren gesehen wurden, obgleich sich schon innerhalb eines einzigen Landes teilweise gravierende Unterschiede hinsichtlich der Einschätzung des Schwierigkeitsniveaus ergaben. Ferner gibt North (2000) weitere Hinweise auf in der
Basisstudie gefundene kulturelle Unterschiede in der didaktischen Gestaltung des Fremdsprachenunterrichts: „A recent analysis of the main course used in the French-speaking cantons concluded that the classroom practice of listening comprehension was minimal. This could be a problem of an inadequate syllabus” (S. 182). Diese Ergebnisse könnten erste Hinweise darauf sein, dass bereits die GERS-Skalen als Grundlage nicht uneingeschränkt in unterschiedlichen Ländern vergleichbar sind.

Andererseits fanden sich auch hinsichtlich der Selbsteinschätzungsskala in einigen Fallstudien auch Hinweise auf eine Vergleichbarkeit der Skalen in unterschiedlichen Ländern. Dabei haben unabhängige Follow-Up-Studien gezeigt, dass die Skalendeskriptoren relativ konsistent bezüglich unterschiedlicher Sprachen und Bildungskontexte verwendet werden. Dies gilt allerdings primär für den Bereich der Selbsteinschätzung (z.B. Kaftandjieva & Takala, 2002; Jones, 2002; North, 2002). Es existieren so gut wie keine Studien zum Bereich der Vergleichbarkeit von Testergebnissen im Hinblick auf Tests, die nicht ausschließlich auf Selbsteinschätzungsskalen basieren.

2.2.4. Determinanten der Itemschwierigkeit, das Itemkategorisierungssystem Dutch Grid und Beschreibung zugrundeliegender kognitiver Prozesse

Die Frage danach, ob bzw. welchen Einfluss Item-Merkmale auf die Schwierigkeit von Testitems und die Leistung von Getesteten haben, wird von Bachman und Palmer (1996) wie folgt formuliert:

„There is also considerable research in language testing that demonstrates the effects of test method on test performance. (…) This research and language teachers’ intuitions both lead to the same conclusion: the characteristics of the tasks used are always likely to affect test scores to some degree, so that there is virtually no test that yields only information about the ability we want to measure. The implication of this conclusion for the design, development, and use of language tests is equally clear: since we cannot totally eliminate the effects of task characteristics, we must learn to understand them and to control them so as to insure that the tests we use will have the qualities we desire and are appropriate for the uses for which they are intended” (Bachman & Palmer, 1996, S. 46).

Überlegungen hinsichtlich des Zusammenhangs zwischen Itemcharakteristika, Itemschwierigkeit und zugrundeliegenden kognitiven Prozessen existieren bereits im Rahmen des Modells der kommunikativen Kompetenz von Bachman und Palmer (1996). Dieses speziell assessment-orientierte Modell befasst sich mit Item- und Texteigenschaften sowie dem Zusammenhang zwischen Itemcharakteristika, Kompetenz und Performanz. Die Sprachfähigkeit wird demnach bedingt durch die Eigenschaften der verwendeten Testaufgaben, sowie durch die Eigenschaften der getesteten Personen: „(…)when we design a language test we need to consider the characteristics of the language use situation and tasks and of the language users and test takers. We need to consider task characteristics in order to insure and demonstrate the ways in which our test tasks correspond to language use tasks” (Bachman & Palmer, 1996, S. 11).

Also sowohl die Merkmale von Items und Testaufgaben als auch die Eigenschaften von Personen sollten demnach Einfluss auf die in diesem Modell angenommene strategische Kompetenz und die tatsächliche Sprachverwendung haben. Auf Seite der Test-Taker werden „Topical Knowledge”, „Language Knowledge” und „Personal Characteristics” genannt, die gemeinsam mit affektiven Komponenten Einfluss auf die strategische Kompetenz haben. Auch die Aufgabencharakteristika beeinflussen die strategische Kompetenz, weshalb die Autoren diese in ihrem Modell explizit mit einbeziehen (siehe 2.2.1).

Auch andere Experten im Bereich des fremdsprachlichen Leseverständnisses beschäftigen sich mit der Analyse von Items und deren Schwierigkeiten sowie dem Zusammenhang von

Abbildung 2.3. Korrespondenz zwischen Sprachgebrauch und Sprachtestleistung (entnommen aus Bachman & Palmer (1996), S. 12.)
Eigenschaften von Items und Lösungswahrscheinlichkeiten. Alderson (2000) beschreibt den Zusammenhang zwischen Analyse des Lesers und Analyse des Texts wie folgt: „(…) Added to this are the inevitable complications when we consider the complexity of analysing texts: since the nature of what we read must have some relation to how we read, then text analysis must be relevant to theories of reading and to research into reading.” (S. 1).

In der der Fremdsprachenforschung zugehörigen Literatur finden sich weitere Taxonomien, die Hinweise auf die schwierigkeitsdeterminierenden Eigenschaften von Items geben. Auf zwei davon wird im Folgenden genauer eingegangen, nämlich auf die bereits erwähnte Taxonomie von Bachman (1990) sowie auf das Instrument „Dutch Grid” (Alderson et al., 2006).


Bachman (1990) beschreibt in seiner Taxonomie fünf für Sprachtestaufgaben wichtige Aspekte: Setting, Teststruktur, Testmaterial (Input), Antwortformat beziehungsweise Antworttyp (Output) sowie die Beziehung zwischen Input und Output. Im Folgenden werden die letzten drei Aspekte behandelt, da diese sich speziell auf die Eigenschaften von Testaufgaben und Testitems beziehen (siehe Tabelle 2.5).


Insgesamt können das Modell an sich, die darauf basierende Taxonomie für Sprachtestaufgaben und die in diesem Umfeld getroffenen theoretischen Annahmen als eine Unterstützung der im Dutch Grid-Kategoriensystem (siehe unten) verwendeten Itemcharakteristika betrachtet werden, insbesondere im Hinblick auf die Kategorien der kognitiv-linguistischen Anforderungsmerkmale der Items. Für eine ausführlichere Darstellung des Modells und der Taxonomie wird auf Bach-
Theoretische Grundlagen 2.2 Fremdsprachenforschung und angewandte Linguistik

<table>
<thead>
<tr>
<th>Aspekte</th>
<th>Charakteristika</th>
<th>Spezifizierung</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input: Charakteristika des Testmaterials</td>
<td>Art des Inputs</td>
<td>Aussage / Aufforderung</td>
</tr>
<tr>
<td></td>
<td>Art der Sprache</td>
<td>Vokabular</td>
</tr>
<tr>
<td></td>
<td>Textlänge</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sprachcharakteristika</td>
<td>organisierend (Grammatik / Textform)</td>
</tr>
<tr>
<td></td>
<td>Inhaltliche Charakteristika</td>
<td>Art der Information (persönlich, technisch, kulturell, etc.)</td>
</tr>
<tr>
<td>Output: Antwort-Charakteristika</td>
<td>Antwortformat</td>
<td>Form / Sprache / Länge</td>
</tr>
<tr>
<td></td>
<td>Antworttyp</td>
<td>Auszahl v. Alternativen / Produktion</td>
</tr>
<tr>
<td>Beziehung zwischen Input und Output</td>
<td>Reaktivität der Aufgabe</td>
<td>reziprok / nicht-reziprok / adaptiv</td>
</tr>
<tr>
<td></td>
<td>Reichweite der Beziehung / Verarbeitungsbreite</td>
<td>breit („Main Idea“) vs. begrenzt (scannen nach spez. Information)</td>
</tr>
</tbody>
</table>

Tabelle 2.5. Auszug aus der Taxonomie für Sprachtestaufgaben nach Bachman (1990)


Der Gemeinsame Europäische Referenzrahmen für Sprachen hat den Anspruch, inhaltliche Spezifizierungen für Fremdsprachentests und -prüfungen zu liefern und Kriterien zur Leistungsbeurteilung bereitzustellen. Darüber hinaus sollen seine Skalen dabei helfen, Kompetenzniveaus bereits bestehender Tests zu beschreiben (Europarat, 2001). In den Skalen des GERS und durch die Formulierung der Deskriptoren sind bisher zwar die für ein gewisses Sprachniveau notwendigen sprachlichen Handlungen (im Sinne von Sprachverwendung) einer Person beschrieben, weni-


Die verwendeten Itemmerkmale beziehen sich beispielsweise auf Oberflächenmerkmale wie Textsorte und Itemtyp, aber auch auf die linguistische und kognitive Komplexität von Texten (z.B. Abstraktionsgrad, Vokabular und Grammatik; siehe auch Tabelle 2.6). Insgesamt können die Itemcharacteristika unterteilt werden in eine Inhaltskategorie und eine Kategorie, die sich auf die kognitive und linguistische Komplexität eines Items, des dazugehörigen Textes bzw. der ganzen Aufgabe bezieht.
Die im „Dutch Grid“ verwendeten Itemcharakteristika sowie deren Ausprägungen sind in Tabelle 2.6 dargestellt.

<table>
<thead>
<tr>
<th>Itemcharakteristik (Inhalt des Textes)</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>Textquelle</td>
<td>(Magazin, Zeitung, etc.)</td>
</tr>
<tr>
<td>Diskurs</td>
<td>Narrativ / beschreibend / Instruktion, etc.</td>
</tr>
<tr>
<td>Domäne</td>
<td>Persönlich, beruflich, öffentlich, etc.</td>
</tr>
<tr>
<td>Thema</td>
<td>Tägliches Leben, Reisen, Gesundheit, etc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemcharakteristik (Linguistische und kognitive Komplexität des Textes)</th>
<th>Beschreibung</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstraktheit des Inhalts</td>
<td>Konkret vs. abstrakt</td>
</tr>
<tr>
<td>(ausschließlich konkret; hauptsächlich konkret; teilweise abstrakt; abstrakt)</td>
<td></td>
</tr>
<tr>
<td>Authentizität des Texts</td>
<td>(angepasst / authentisch)</td>
</tr>
<tr>
<td>Schwierigkeit des Vokabulars</td>
<td>Einfach / häufig vs. schwer / selten</td>
</tr>
<tr>
<td>(ausschließlich häufig / einfach; hauptsächlich häufig / einfach; teilweise erweitert / selten; erweitert / selten)</td>
<td></td>
</tr>
<tr>
<td>Komplexität grammatischer Strukturen</td>
<td>Einfache Strukturen vs. komplexe Strukturen</td>
</tr>
<tr>
<td>(ausschließlich einfach; hauptsächlich einfach; teilweise komplex; komplex)</td>
<td></td>
</tr>
<tr>
<td>Itemtyp</td>
<td>Auswahl vs. Konstruktion einer Antwort</td>
</tr>
<tr>
<td>(z.B. Multiple Choice / offene Antwort)</td>
<td></td>
</tr>
<tr>
<td>Informationsgewinn 1</td>
<td>Information durch Erkennen / Schlussfolgern / Bewerten</td>
</tr>
<tr>
<td>Informationsgewinn 2</td>
<td>Information explizit / implizit</td>
</tr>
<tr>
<td>Informationsgewinn 3</td>
<td>Hauptidee vs. spezifisches Detail</td>
</tr>
</tbody>
</table>

Tabelle 2.6. Inhaltliche und kognitiv-linguistische Itemeigenschaften-Dutch Grid (entnommen und übersetzt aus Alderson et al., 2006).

Die oben angesprochene Kritik am GERS hinsichtlich der Nicht-Verwendbarkeit der Skalen zur Konstruktion von Testverfahren sowie die daraus folgende Entwicklung des „Dutch Grid“ führen zu der Frage, welche Itemeigenschaften tatsächlich die Schwierigkeit eines Items bestimmen. Mittlerweile existieren Anhaltspunkte dafür, dass vor allem solche Item- und Texteigenschaften des Dutch Grid, die bestimmte, zur Lösung eines Items notwendige linguistische und kognitive Prozesse erfordern, für die Schwierigkeit eines Items von Bedeutung sind. Im Rahmen dieser Arbeit liegt daher der Fokus auf diesen Itemmerkmalen. In einer niederländischen Studie


Bezüglich der Charakteristika, die sich schwerpunktmäßig auf das Item beziehen, werden für diese Arbeit die Merkmale „Informationsgewinn 1-3“, und „Itemtyp“ übernommen. Das Merkmal Itemtyp bezieht sich darauf, auf welche Art und Weise das Item formuliert ist. Itemtypen können von geschlossenem oder offenem Format sein. Beispielsweise handelt es sich bei dem Itemtyp „Multiple Choice“ um ein geschlossenes Antwortformat, während die Itemtypen „Offene Antwort“ oder „Kurzantwort“ den offenen Antwortformaten zuzuordnen sind. Das Merkmal „Informationsgewinn“ bezieht sich auf die zur Lösung eines Items notwendigen men-talen Operationen. Diese werden in drei Bereiche eingeteilt: Das Verhalten des Lesers (erkennen, schlussfolgern und bewerten), das „was“, das heißt, der Gegenstand nach dem jeweils gefragt ist (Heraussuchen/Erkennen eines Details oder Verstehen der Hauptaussagen des Textes), und die Informationsquelle (ist die Information im Text explizit oder nur implizit enthalten?).

Bei der Modellierung des Leseprozesses sind dem Autor zufolge unterschiedliche, hierarchische Ebenen zu unterscheiden, nämlich die graphophonische Ebene, die lexikalisch-formale Ebene, die syntaktische Ebene, die lexikalisch-semantische Ebene und die textsemantische Ebene. Darüber hinaus wird von zwei unterschiedlichen Verarbeitungstypen ausgegangen: „datenbasiert“ (engl.: bottom up) und „wissensbasiert“ (engl.: top down). Dabei entsprechen die wissensbasierten Prozesse höherer semantischer Ebenen eher einem erwartungsgeleiteten und hypothetentesenden Lesen, während die datenbasierten Prozesse eher auf der unteren Ebene anzusiedeln sind und zum großen Teil parallel und automatisiert ablaufen.


Demzufolge messen beispielsweise deskriptive Informationsfragen zum Text das Verständnis auf der Textbasis, während Inferenzfragen nur beantwortbar sind, wenn ein mentales Modell konstruiert wurde. Daher ist zu erwarten, dass Letztere zu einer höheren Itemschwierigkeit beitragen als Erstere. Der Autor sieht nach Betrachtung von Modellen kognitiver Prozesse des Leseverstehens hinsichtlich möglicher Determinanten der Itemschwierigkeiten die folgenden Konsequenzen (S. 13):
2.2 Fremdsprachenforschung und angewandte Linguistik

- Es ist der Interaktion zwischen Leser und Aufgabe Rechnung zu tragen.
- Es ist die Abhängigkeit der Lese- und Verstehensprozesse von interindividuell varierenden, L2-unspezifischen Rezipientenmerkmalen zu berücksichtigen.
- Auch die datengeleiteten – beim kompetenten Muttersprachler in der Regel automatisierten – Prozesse auf den unteren Verarbeitungsebenen sind zu erfassen.
- Es ist jeweils die Ebene zu spezifizieren, auf die sich die Verstehensaufgaben beziehen (z.B. Textbasis vs. mentales Modell).


Bezüglich des Itemtyps mag das Nicht-Vorkommen in Grotjahns Taxonomie darauf zurückzuführen sein, dass der Einfluss des Itemtextes an sich (und damit auch des Itemtyps) häufig, so auch von Grotjahn (2000), als konstruktirrelevant angesehen wird. Es existieren diesbezüglich jedoch auch andere Erkenntnisse, dass nämlich mit bestimmten Itemtypen bestimmte, unterschiedliche Konstruktteile des Leseverständnisses erfasst werden können (Rauch & Hartig, in Vorbereitung). Darüber hinaus ist bei der Betrachtung der EBAFLS-Items

67
2.2 Fremdsprachenforschung und angewandte Linguistik

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Zahl der schwierigen / unvertrauten Wörter</td>
<td>Schwierigkeit des Vokabulars</td>
</tr>
<tr>
<td>Abstraktheit des Inhalts</td>
<td>Abstraktheit des Inhalts</td>
</tr>
<tr>
<td>Grammatische Komplexität</td>
<td>Grammatische Komplexität</td>
</tr>
<tr>
<td>Die für eine Antwort benötigte Information ist über den Text verteilt</td>
<td>Erkennen</td>
</tr>
<tr>
<td>Es ist kein unmittelbares Scannen der relevanten Information möglich</td>
<td>Schlussfolgern</td>
</tr>
<tr>
<td>Grad der Implizitheit</td>
<td>Implizit / explizit</td>
</tr>
<tr>
<td>Das Item erfragt eine Hauptinformation („main idea item”)</td>
<td>Hauptidee / detail</td>
</tr>
</tbody>
</table>

Tabelle 2.7. Taxonomie schwierigkeitsdeterminierender Itemeigenschaften, (1. Spalte Auszug aus Grotjahn, 2000)


Bei einem Vergleich der drei oben dargestellten Methoden stellten sich die zweischrittige Methode sowie das LLTM+e den Autoren nach als gleichwertige und, verglichen mit einem „normalen“ LLTM, als die vorteilhafteren Methoden heraus. Es konnten zwischen 39,6% (LLTM+e) und 42,4% (zweischrittige Methode) der Varianz der Itemschwierigkeiten anhand der Item- bzw. Textcharakteristika erklärt werden. Da die Berechnung des LLTM+e sehr aufwändig ist und keinen Vorteil gegenüber der zweischrittigen Methode aufzuweisen scheint, wird in dieser Dissertation die gleichwertige, zweischrittige Methode zur Erklärung der Varianz der Itemschwierigkeiten innerhalb der Länder und zwischen ihnen angewandt (siehe auch 4.2.2).


Insgesamt weisen die vorliegenden empirischen Ergebnisse darauf hin, dass kognitiv-linguistische Anforderungsmerkmale von Items, wie sie im „Dutch Grid“ verwendet werden, als Prädik-
2.3. Interkulturelle Vergleichbarkeit von Testergebnissen


Hinsichtlich der interkulturellen Vergleichbarkeit von Testergebnissen nimmt diese Arbeit eine, wie Van de Vijver und Poortinga (1990) es bezeichnen, moderat universalistische Position der kulturellen Psychologie ein: Es wird von einer generellen Vergleichbarkeit psychologischer Konstrukte ausgegangen und zugleich ein Fokus auf die Erklärung von unterschiedlichem Verhalten gelegt. Es geht also nicht ausschließlich um die Feststellung von kulturellen Unterschieden, sondern auch und maßgeblich um deren Interpretation, Deutung und Erklärung: „The central purpose of cultural psychology is the unambiguous interpretation of cultural differences. The mere observation of differences is not satisfactory; such an observation should be the starting-point for subsequent investigation.” (Van de Vijver & Poortinga, 1990, S. 97). Dieser Empfehlung folgt die vorliegende Arbeit.

Emischer oder etischer Ansatz? Eine Debatte der interkulturellen Psychologie, die auch hinsichtlich der interkulturellen Vergleichbarkeit von Testergebnissen relevant ist, hat die Frage zum Gegenstand, ob theoretische Konstrukte überhaupt über verschiedene Kulturen hinweg vergleichbar sind:

„Whether European, Asian of African, we feel that people from other cultures act differently in many contexts than we do. (…) Are those ’other’ people so different from ’us’ that it is impossible to make comparisons between cultures, or are there (…) the
same deep structures, that is, the same abilities, and the same needs? (…) In general, cross-cultural psychologists use the term culture to mean ‘patterns, explicit and implicit, of and for behaviour acquired and transmitted by symbols, constituting the distinctive achievements of human groups, including their embodiments in artifacts’” (Helfrich, 1999, S. 131f.).

Wie aus dem ersten Teil des obigen Zitats hervorgeht, existiert innerhalb der interkulturellen Psychologie eine kontrovers geführte Debatte über die Art des zu verfolgenden Ansatzes: den emischen oder den etischen. Bei Verwendung des emischen Ansatzes wird davon ausgegangen, dass jede Kultur einzigartig ist, auch und vor allem hinsichtlich der zugrunde liegenden psychologischen Konstrukte. Der etische Ansatz hingegen vertritt die Ansicht, dass trotz oberflächlicher Unterschiede die gleichen zugrunde liegende Variablen und Konstrukte existieren: „The etic approach demands a descriptive system which is equally valid for all cultures and which permits the representation of similarities as well as differences between individual cultures” (Helfrich, 1999, S. 132). Die Unterscheidung von emisch und etisch beinhaltet gleichzeitig auch die Frage nach der Art der Konstruktunterschiede: Während der emische Ansatz diesbezüglich einen Unterschied hinsichtlich der Struktur des Konstrukts postuliert, wird im Rahmen des etischen Ansatzes von Niveauunterschieden ausgegangen.

2.3 Interkulturelle Vergleichbarkeit von Testergebnissen

Unterschiede des Konstrukts zurückgeführt werden. Im Übrigen wäre bei einer ausschließlichen Verwendung des emischen Ansatzes ein Vergleich von Angehörigen verschiedener sprachlicher Kulturen hinsichtlich ihrer Sprachkompetenz nicht möglich.


Nach Van de Vijver und Leung (1997) existieren zwei Dimensionen interkultureller Studien. Die erste Dimension betrifft die Orientierung bzw. den Fokus der Studie (ist sie explorativ oder hypothesenbasiert angelegt): „In exploratory studies, researchers do not have firm ideas about the cross-cultural similarities and differences to be expected. Such occasions are likely to arise when researchers venture into cultures that are unknown to them. Alternatively, there may be insufficient previous research for generating specific hypotheses.” (S. 20). Die dieser Arbeit zugrundeliegende EBAFLS-Studie ist eher als eine explorative Studie anzusehen: Es handelte sich um eine Machbarkeitsstudie zum Feststellen eventuell vorhandener Unterschiede bei der Beantwortung von Items zur Messung fremdsprachlichen Leseverständnisses. Im Kontext dieser Dissertation findet nun eine Verschiebung des Fokus zum anderen Pol dieser Dimension statt, und zwar hin zu einer theoriebasierten, hypothesenbasierten Arbeit: Durch die Analyse von Testitems aus den verschiedenen Ländern wird es möglich, Hypothesen hinsichtlich der Testkultur-Profile und somit auch Hypothesen im Hinblick auf zu erwartende Zusammenhänge differenziert aufzustellen.


„Two closely related concepts play an essential role in cross-cultural comparisons, namely, equivalence and bias (Poortinga, 1989). From a theoretical point of view, the two concepts are the opposite of each other; scores are equivalent when they are unbiased. Nonetheless, the two concepts will be treated separately here because historically, they have become associated with different aspects of cross-cultural comparisons. Equivalence is more often associated with the measurement level at which scores obtained in different cultural groups can be compared, whereas bias indicates the presence of factors that challenge the validity of cross-cultural comparisons. (…) Equivalence is a function of characteristics of an instrument and of the cultural groups involved.” (Van de Vijver & Leung, 1997, 7 ff).


Der Begriff „Bias“ oder „Verzerrung“ spielt mit Blick auf Testfairness, Kultur-faire Vergleiche und die angenommene universelle Gültigkeit psychologischer Konstrukte eine bedeutende Rolle. Es werden mehrere Formen von Bias unterschieden: Konstruktbiased, Methodenbias und Itembias (Van de Vijver & Tanzer, 2004). In dieser Arbeit liegt der Schwerpunkt auf Letzterem. Wie im Zusammenhang mit der Vorstellung des DIF-Konzepts bereits diskutiert wurde, weist ein Item jedoch nicht automatisch einen Bias auf, wenn die Mitglieder zweier Gruppen unterschiedliche Lösungswahrscheinlichkeiten haben, sondern dann, wenn trotz gleicher Fähigkeit die Lösungswahrscheinlichkeiten unterschiedlich sind. Manchmal wird die Existenz von Itembias mit Differential Item Functioning gleichgesetzt (Van de Vijver & Tanzer, 2004), was jedoch nicht immer
 unkritisch gesehen wird (Camilli, 1993). Die hinsichtlich der DIF-Analysen existierenden Methoden wurden bereits unter 2.1 besprochen.


Die höchste Äquivalenzstufe ist die skalare Äquivalenz. Erst wenn diese erreicht ist, können Mittelwertvergleiche zwischen Gruppen durchgeführt werden.


2.4. Verknüpfung der Theoriestränge & Rahmenkonzept der Dissertation: Messicks Validitätstheorie


2.4.1. Samuel Messicks Validitätskonzept


Je nach eingesetztem Test und getesteter Gruppe können die Konsequenzen eines Testergebnisses entweder für einzelne Individuen oder aber für eine ganze Gesellschaft bedeutend sein. Im Fall von Large Scale Assessments sind die Konsequenzen solcher Ergebnisse meist für die ganze Gesellschaft, das heißt für ein gesamtes System relevant, im Falle von Einstufungs- oder Zulassungstests hingegen eher für das einzelne Individuum. In manchem Fall nimmt das Ergebnis eines Tests sogar Einfluss auf den weiteren Lebensweg, die Lebensqualität und die Sicherheit einer Person, beispielsweise im Fall von Einbürgerungstests. Anhand dieser Beispiele lässt sich aufzeigen, dass letztlich demnach vor allem die Konsequenzen dieser Tests (bzw. der Testwertin-
interpretation) eine hoch relevante Rolle spielen. Wie bereits angesprochen wurde, ist darüberhin-
aus die Existenz von DIF ein Hinweis auf die Invalidität eines Testverfahrens für den Vergleich
von Testergebnissen unterschiedlicher Gruppen. Daher bietet sich hier an, bei der Diskussion von
legen.

Er geht davon aus, dass die Konsequenzen der Interpretation von Testwerten ein Teil der Validität
sind. Messick (1995) definiert Validität wie folgt:

„Validity is an overall evaluative judgement of the degree to which empirical evidence
and theoretical rationales support the adequacy and appropriateness of interpretations
and actions on the basis of test scores or other modes of assessment. Validity is not a
property of the test or assessment as such, but rather of the meaning of the test scores”
(S. 741).

Validität ist demnach ein Maß dafür, inwieweit empirische Ergebnisse und theoretische Annah-
men die Adäquatheit und Angemessenheit von Interpretationen und Folgen von Testergebnissen
unterstützen. Validität ist keine Eigenschaft eines Tests an sich, sondern der Bedeutung und In-
terpretation von Testergebnissen.

Dies bedeutet, dass in Messicks Verständnis von Validität vor allem die Bedeutung und Interpre-
tation von Testergebnissen sowie die daraus für eine getestete Person entstehenden Folgen valide
sein müssen. Die Validität von Testverfahren kann also über den Erfolg einer Person und damit
auch über deren weiteren Werdegang mitbestimmen: „Indeed, it is precisely because of such
politically salient potential consequences that the validity of performance assessment needs to be
systematically addressed, as do other basic measurement issues such as reliability, comparability

Fairness beinhaltet hier beispielsweise Fairness hinsichtlich der Testverwendung, die Angemes-
senheit der dem Test und seiner Auswertung zugrunde gelegten Regeln, der verwendeten Kon-
strukte sowie die Regeln hinsichtlich auf dem Test basierender Entscheidungen. Nach Messick
sind Validität, Reliabilität, Vergleichbarkeit und Fairness nicht nur lediglich beim Testen und bei
der Testkonstruktion zu beachtende Prinzipien, sondern repräsentieren soziale Werte, die eine
Bedeutung und Einfluss außerhalb des Tests an sich besitzen, und zwar immer dann, wenn darauf
Bewertungen und Entscheidungen beruhen:
„These issues are critical for performance assessment – as they are for all educational and psychological assessment because validity, reliability, comparability and fairness are not just measurement principles, they are social values that have meaning and force outside of measurement whenever evaluative judgements and decisions are made” (Messick, 1995, S. 741).

Die Interpretation und Folgen von Testwerten sind daher im Rahmen seines Modells von kritischer Relevanz.


1. Inhaltliche Relevanz und Repräsentativität
2. Stichhaltige Theorien und Prozessmodelle
3. Mess- und Score-Modelle als Abbild von Testaufgabe und Struktur der untersuchten Domäne
4. Generalisierbarkeit und Grenzen der Bedeutung von Testergebnissen
5. Konvergente und Divergente Korrelationen mit externen Variablen
6. Konsequenzen als Validitätsbeweis


Für die vorliegende Arbeit spielen vor allem die Konstruktrepräsentation und deren Vergleichbarkeit in verschiedenen Kulturen, die Generalisierbarkeit der Bedeutung von Testwertergeb-
nissen sowie die Konsequenzen als Validitätsbeweis eine Rolle. Nach Messick ist die Konstruktrepräsentation ein fundamentales Charakteristikum von Konstruktvalidität. Aussagen über die Repräsentation eines Konstruks in einem Test können beispielsweise anhand der Untersuchung kognitiver Prozesse, die der Aufgabenbearbeitung zugrunde liegen, getroffen werden. Die Mechanismen, die dem Bearbeiten von Aufgaben zugrunde liegen, können durch das Aufteilen von Aufgaben in theoretisch basierte Komponenten untersucht werden. Darauf basierend sollte dann ein Modell oder eine Prozesstheorie entwickelt werden:

„A fundamental feature of construct validity is construct representation, whereby one attempts to identify through cognitive-process analysis or research on personality and motivation the theoretical mechanisms underlying task performance primarily by decomposing the task into requisite components and assembling them into a functional model or process theory“ (Messick, 1989, S. 742).


2.4.2. Quellen der Invalidität


Hinsichtlich konstruktirrelevanter Varianz unterscheidet Messick zwei Erscheinungsformen, nämlich konstruktirrelevante Schwierigkeit und konstruktirrelevante Leichtigkeit. Im ersten Falle erschweren spezifische, für das Konstrukt möglicherweise nicht relevante Eigenschaften einer Aufgabe die Bewältigung eines Items seitens bestimmter Personengruppen. Letztere Erneu-
nungsform hingegen verringert die Aufgabenschwierigkeit, obgleich die Personenfähigkeit sich nicht erhöht.

Messicks Theorie bezüglich der Ursachen für die Einschränkung der Konstruktvalidität ist für die vorliegende Arbeit zentral von Bedeutung, da die in ihr definierten Quellen der Validitätseinschränkung mit der oben (siehe 2.1.1) bereits diskutierten Interpretation von DIF als „Nuisance Dimension“ oder DIF als Ausdruck differentieller Stärken und Schwächen von Gruppen in Verbindung gebracht werden können. Dabei ist die Einführung konstruktirrelevanter Varianz, also differentieller Leichtigkeit und Schwierigkeit, mit der Betrachtung von DIF als „Nuisance Dimension“, also als einer nicht dem Konstrukt zugehörigen, zusätzlich erfassten, konstruktirrelevanten Dimension gleichzusetzen. Wird davon ausgegangen, dass DIF Ausdruck einer zusätzlich gemessenen Dimension ist, wird im Prinzip DIF auf das Vorhandensein konstruktirrelevanter Varianz zurückgeführt, die im Falle von zwei Gruppen bei der einen als konstruktirrelevante Leichtigkeit und bei der anderen als konstruktirrelevante Schwierigkeit vorliegt.


2.4.3. „Consequential aspect of validity“ und „social values“

Im Zusammenhang mit interkulturellen Kompetenzvergleichen spielt Testvalidität eine sehr wichtige Rolle. Um valide Aussagen über tatsächliche Gruppenunterschiede machen zu können, muss in solchen Fällen besonders auf die Vergleichbarkeit der Testwertergebnisse und auf Fair-
nness geachtet werden. In den letzten Jahren hat sich gezeigt, dass die Ergebnisse internationaler Schulleistungsstudien, zumindest in Deutschland, einen relevanten, systemischen Impact auf die Gesellschaft und bildungspolitische Entscheidungen hatten. So haben beispielsweise PISA-Ergebnisse zu einer tiefgreifenden bildungspolitischen Debatte in Deutschland geführt (siehe auch 1). Dabei wird deutlich, wie angemessen es ist, dass Messick in seinem Validitätskonzept den Hauptfokus auf die Testwertinterpretation und deren Konsequenzen legt.


„The consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning. (…) Rather, because the social values served in the intended and unintended outcomes of test interpretation and use both derive from and contribute to the meaning of the test scores, appraisal of social consequences of the testing is also seen to be subsumed as an aspect of construct validity.” (Messick, 1996, S. 15f).

Besonders deutlich ist der gesellschaftliche Einfluss, d.h. der Einfluss der sozialen Werte, bei sogenannten „high stakes“-Tests wie etwa zentralen Schulleistungsüberprüfungen zu beobachten. Diese Werte werden nach Messicks Theorie letztlich durch die in ihrem Sinne vollzogene Interpretation der Testwertergebnisse und die aus dieser Interpretation erwachsenden Konsequenzen auf eine Testkultur und somit auf den Unterricht übertragen. In unterschiedlichen Bildungskulturen führt dies gegebenenfalls zu differentiellen Lerngelegenheiten.


2.4.4. Zusammenfügung der Theoriestränge im Kontext der Validität


Analysen zur Entdeckung Differentieller Item Funktionen stellen wiederum eine Methode zur Überprüfung der Konstruktvergleichbarkeit dar. DIF können zum einen als konstruktirrelevante


2.4.5. Zusammenfassung und Relevanz für die Arbeit


3. Fragestellung

Ausgehend von den unter Abschnitt 2 diskutierten Theorien und empirischen Befunden zur Erklärung von Itemschwierigkeiten und differentiellen Item Funktionen sowie den weiteren dargestellten Theoriebereichen soll nun im Folgenden die zentrale Fragestellung dieser Arbeit entwickelt werden.

3.1. Herleitung der Fragestellungen

Ziel dieser Arbeit ist die Erklärung von Differentiellen Item Funktionen bei Items zur Messung des fremdsprachlichen Leseverständnisses. Bei Differentiellen Item Funktionen handelt es sich um kulturell bedingte Unterschiede hinsichtlich der Itemschwierigkeit eines Items bei unterschiedlichen Gruppen. Um differentielle Item Funktionen erklären zu können, müssen also zum einen Annahmen darüber getroffen werden, welche Eigenschaften eines Items zur Messung fremdsprachlichen Leseverständnisses dessen Schwierigkeit bedingen, und wie diese determiniert werden können. Zum anderen sind Annahmen darüber notwendig, welche Ursachen für den Unterschied dieser Itemschwierigkeit bei zwei oder mehr Gruppen verantwortlich sein könnten.


In dem vierten, für die vorliegende Arbeit bedeutenden Artikel, nämlich dem von Scheuneman und Gerritz (1990), wurden DIF bei Sprachtestaufgaben in Populations-Subgruppen (Geschlecht, ethnischer Hintergrund) in College-Zulassungstests in den USA untersucht. Per multipler linearer Regression wurden hier DIF mit Hilfe von Itemmerkmalen vorhergesagt. Basie-
rend auf den Ergebnissen schlussfolgerten die Autoren, dass DIF ein Hinweis auf unterschied-
lliche Profile von Stärken und Schwächen der unterschiedlichen Gruppen sind, die sich wieder-
urn durch die Analyse von Items hinsichtlich ihrer Anforderungsmerkmale feststellen lassen.
Ausgehend von diesen empirischen Befunden und den unter Abschnitt 2 diskutierten Theorien
werden in der vorliegenden Arbeit verschiedene Annahmen hinsichtlich der Determination von
Itemschwierigkeiten und DIF getroffen, welche im Folgenden dargelegt werden werden.

1. Die Itemschwierigkeit wird von Anforderungsmerkmalen der Items (mit)bedingt, und
Items lassen sich hinsichtlich ihrer Anforderungsmerkmale kategorisieren

Zunächst wird in dieser Arbeit vorausgesetzt, dass alle Items Anforderungsmerkmale besitzen,
die maßgeblich zur Schwierigkeit eines Items beitragen, da sie zur Lösung des Items notwendige,
kognitiv-linguistische Prozesse abbilden (siehe 2.2.4). Hinsichtlich der Einordnung von Items in
Bezug auf diese Merkmale spielen im Bereich der Messung von Fremdsprachenfähigkeiten vor
allem zwei Item-Kategorisierungs-Systeme in der Literatur eine Rolle: das von Bachman und
Palmer (1996), sowie der auf dem GERS basierende „Dutch Grid“ (Alderson et al., 2006). Ferner
(2000) auch verschiedene empirische Ergebnisse (siehe 2.2.4) für einen Einfluss der in dieser
Arbeit verwendeten Anforderungsmerkmale auf die Itemschwierigkeit.

Das System von Bachman und Palmer (1996) und das des „Dutch Grid“ gehen dabei von ähnli-
chen zugrundeliegenden Item-Anforderungsmerkmalen aus; dies ist nicht verwunderlich, da der
GERS, auf dem wiederum der „Dutch Grid“ aufbaut, unter anderem auch theoretisch durch das
Bachman-Palmer-Modell begründet ist. Im Rahmen des „Dutch Grid“ (Alderson et al., 2006)
 wird außerdem, zumindest implizit, von einer übernationalen Vergleichbarkeit der dort einge-
setzten Itemschwierigkeits-Determinanten und der Kategorien, hinsichtlich derer die Items ein-
zuordnen sind, ausgegangen: Durch seine Anwendung in nationenübergreifenden Projekten wie
EBAFLS, wo die Items unterschiedlicher Länder innerhalb dieses Rahmens kategorisiert wurden,
und auch dadurch, dass die Grundlage des Instruments der Gemeinsame Europäischer Referenz-
rahmen ist, der die angenommene Vergleichbarkeit ja bereits namentlich impliziert.

Im Rahmen dieser Arbeit wird angenommen, dass in allen Ländern zumindest teilweise die glei-
chen kognitiv-linguistischen Anforderungsmerkmale eine Rolle für die Itemschwierigkeit und
somit für das Konstrukt des fremdsprachlichen Leseverständnisses spielen, jedoch hinsichtlich
der Größe und der Richtung des Einflusses aufgrund unterschiedlicher Lerngelegenheiten in den
Ländern möglicherweise differieren können.
2. Items repräsentieren die Testkultur eines Landes


3. Unterschiedliche Testkulturen verursachen unterschiedliche Stärken und Schwächen hinsichtlich der Beantwortung von Items mit bestimmten Anforderungsmerkmalen

wird der Frage nachgehen, inwieweit sich Differentielle Item Funktionen bei fremdsprachlichen Leseverständnis-Items mit Hilfe dieser Annahmen erklären lassen. Zu den Itemmerkmalen, von denen hier angenommen wird, dass sie die kognitive und linguistische Komplexität eines Items verändern und somit einen Einfluss auf Itemschwierigkeiten und DIF haben, gehören zunächst ganz allgemein das Herkunftsland eines Items (Artelt & Baumert, 2004), aber auch die unterschiedlichen Ausprägungen der kognitiv-linguistischen Anforderungsmerkmale (Alderson et al., 2006; Bachman & Palmer, 1996). Diese Merkmale — oder, präziser formuliert, die signifikanten Unterschiede zwischen den Items der jeweiligen Teilnehmerländer hinsichtlich ihrer Merkmalsausprägungen — werden im Folgenden im Zusammenhang mit DIF als „Indikatoren nationaler Testkulturen“ bezeichnet.

**Zusammenfassung der zentralen Annahmen**


3.2. Hauptfragestellung

Diese Dissertation beschäftigt sich primär mit der Frage, ob Differentielle Item Funktionen mit Hilfe von Indikatoren nationaler Testkulturen erklärt werden können. In der vorliegenden Arbeit sollen dazu die Daten der EBAFLS-Studie zum fremdsprachlichen Leseverständnis in den Sprachen Englisch und Deutsch herangezogen werden. Die auf die vorliegenden Daten spezifizierte Hauptfragestellung dieser Dissertation lautet:

Existiert ein Zusammenhang zwischen Differentiellen Item Funktionen und Indikatoren nationaler Testkulturen bei Aufgaben zur Messung des fremdsprachlichen Leseverständnisses in englischer und deutscher Sprache?

Diese Hauptfragestellung beinhaltet die Frage danach, ob sich durch unterschiedliche Ausprägungen kognitiv-linguistischer Itemmerkmale bei aus unterschiedlichen Ländern stammenden Items nationale Testkulturen abbilden lassen, und ob diese zur Erklärung von Differentiellen Item Funktionen bei fremdsprachlichen Leseverständnis-Items herangezogen werden können. Das hier gewählte Vorgehen zur Analyse der durch die unterschiedlichen Merkmalsausprägungen bedingten länderspezifischen Stärken und Schwächen, im Folgenden auch als „Testkultur“ bezeichnet, wird unter 4.2 beschrieben.


Der erste Fragenkomplex beinhaltet dabei das Ziel, festzustellen, ob die Voraussetzungen für die Behandlung der weiteren Fragestellungen gegeben sind. Da für die Analyse der Items hinsichtlich der Itemschwierigkeit und Differentieller Item Funktionen aus theoretischen Gründen das Rasch-Modell gewählt wird (siehe auch 4.2), handelt es sich bei der ersten der insgesamt drei zu überprüfenden Voraussetzung um die Rasch-Modell-Konformität der Items innerhalb der Länder.
Zum Zweiten werden die Items auf das Vorhandensein paarweiser Differentieller Item Funktionen hin überprüft, da im Rahmen der EBAFLS-Studie das OPLM-Modell (Verhelst, Glas & Verstralen, 1995; siehe auch 4.2) und nicht das Rasch-Modell verwendet wurde. Daher muss festgestellt werden, ob Differentielle Item Funktionen auch unter Annahme des strengeren Rasch-Modells existieren.

Die dritte zu überprüfende Voraussetzung bezieht sich auf die Existenz unterschiedlicher Testkulturen in den Teilnehmerländern. Hier wird untersucht, inwieweit anhand der eingereichten Items der Länder differentielle Testkulturen und somit unterschiedliche zu erwartende Stärken und Schwächen der Länder festgestellt werden können.


Der dritte Komplex von Fragen befasst sich mit der Untersuchung von Zusammenhängen zwischen kulturell bedingten Unterschieden der Itemschwierigkeiten zwischen den Ländern, also Differentiellen Item Funktionen, und aufgrund der Testkulturen zu erwartenden differentiellen Stärken und Schwächen der verschiedenen Gruppen. Im nachfolgenden Abschnitt sollen die Einzelfragestellungen dieser aufeinander aufbauenden drei Fragenkomplexe, die im Folgenden „Voraussetzungen und Skalierbarkeit“, „Erklärung von Itemschwierigkeiten“ und „Erklärung von Differentiellen Item Funktionen“ genannt werden, im Detail formuliert werden. Die Teilfragestellungen beziehen sich jeweils auf das fremdsprachliche Leseverständnis in englischer und deutscher Sprache.

3.3. Fragenkomplex 1: Voraussetzungen und Skalierbarkeit

dieser Fragestellungen gewonnenen Schwierigkeitsparameter der Items für die Analysen in Fragenkomplex 2 benötigt. Die zu beantwortende Frage lautet hier:

**Frage 1a: Weisen die Items innerhalb der Länder Rasch-Modellkonformität auf?**

In der ursprünglichen EBAFLS-Studie hatte sich gezeigt, dass innerhalb der Länder Modellkonformität der Items vorhanden war (Fandel et al., 2007). Das bedeutet, dass die Items dort, unabhängig davon aus welchem Land sie stammten, innerhalb der Länder insgesamt dieselbe Dimension erfassten. Obgleich in dieser Arbeit das im Gegensatz zum OPLM-Modell (Verhelst, Glas & Verstralen, 1995) etwas strengere Rasch-Modell angenommen wird, wird erwartet, dass auch hier die Items größtenteils Modellkonformität aufweisen. Daher wird folgende Hypothese aufgestellt:

**Hypothese 1a: Die Items weisen innerhalb der Länder Rasch-Modellkonformität auf.**

Eine weitere Voraussetzung für die Beantwortung der Hauptfragestellung ist das Vorhandensein Differentieller Item Funktionen auch unter Anwendung des Rasch-Modells.

**Frage 1b: Wie groß ist der Anteil von Items mit Differentiellen Item Funktionen?**

Dies traf in der ursprünglichen EBAFLS-Studie auf einen großen Teil der Items zu. Da in der Studie jedoch keine Angaben zur Signifikanz der DIF-Parameter gemacht wurden, ist nicht klar, welcher Anteil der Items tatsächlich signifikante DIF aufwiesen. Außerdem wurde dort das zweiparametrische OPLM-Modell, hier jedoch wird das einparametrische Rasch-Modell angewendet. Aus diesem Grund muss dieser Punkt erneut überprüft werden. Wenn mehr als 35% der Items signifikante DIF-Parameter aufweisen, wird dies hier als ein großer Anteil definiert. Es wird folgende Hypothese aufgestellt:

**Hypothese 1b: Ein großer Anteil, das heißt mehr als 35% der getesteten Items, weist signifikante Differentielle Item Funktionen auf.**

Die dritte zu überprüfende Voraussetzung bezieht sich auf das Vorhandensein unterschiedlicher Testkulturen und somit auch unterschiedlicher zu erwartender Stärken und Schwächen der verschiedenen Ländergruppen:
Frage 1c: Lassen sich unterschiedliche Testkulturen der Länder feststellen?

Um zur Erklärung kulturell bedingter Unterschiede, d.h. DIF, und hinsichtlich der aufgrund der Unterschiedlichkeit der Testkulturen zu erwartenden Stärken und Schwächen der Länder a priori Hypothesen aufstellen zu können, muss zunächst festgestellt werden, ob sich die Testkulturen der Länder hinsichtlich der Häufigkeit des Vorkommens von Item-Anforderungsmerkmalen voneinander unterscheiden und ob eine Analyse von Items überhaupt zur Feststellung solcher Testkulturen als Methode geeignet ist. Aufgrund empirischer Ergebnisse wie beispielsweise von Klieme & Baumert (2001) oder Dogan, Guerrero und Tatsuoka (2005) lässt sich erwarten, dass beides der Fall ist. Die der Fragestellung zugehörige Hypothese lautet daher:

Hypothese 1c: Es lassen sich durch eine Analyse von Items aus unterschiedlichen Ländern unterschiedliche Testkulturen feststellen.

3.4. Fragenkomplex 2: Erklärung von Itemschwierigkeiten

Frage 2a: Weisen die kognitiv-linguistischen Anforderungsmerkmale der Items korrelative Zusammenhänge zu Itemschwierigkeiten innerhalb der Länder auf?


Die Beibehaltung der Hypothese, dass Item-Anforderungsmerkmale zur Itemschwierigkeit beitragen, ist also als Grundlage für die weiteren Analysen wichtig.

Innerhalb der Länder existierende Zusammenhänge können einen Hinweis darauf geben, ob die in der vorliegenden Arbeit verwendeten Itemmerkmale auch tatsächlich schwierigkeitsdeterminierend sind. Auch können von den Ergebnissen Aussagen dahingehend abgeleitet werden, ob
das hier verwendete „Dutch Grid”-Kategoriensystem auch tatsächlich zur Kategorisierung von Testverfahren geeignet ist. Diese Frage wird zunächst korrelationsanalytisch behandelt. Diesbezüglich wird folgende Hypothese aufgestellt:


Ein weiterer Grund für die Behandlung dieser Fragestellung besteht in der Tatsache, dass beim Item-Kategorisierungssystem „Dutch Grid”, wie auch bei seiner theoretischen Basis, dem GERS, zumindest implizit davon ausgegangen wird, dass die verwendeten Itemeigenschaften für die Schwierigkeiten der Items aller Länder gleichermaßen eine Rolle spielen sollten. Daher soll auch folgende Fragestellung hier bearbeitet werden:

Frage 2b: Ist die Höhe der Korrelationen in den Ländern vergleichbar?

Die Verwendung des „Dutch Grid” in internationalen Studien für die Einordnung von Items aus unterschiedlichen Ländern, wie beispielsweise in der EBAFLS-Studie, zeigt, dass zumindest implizit angenommen wird, dass die gleichen Itemeigenschaften auf ähnliche Art und Weise in unterschiedlichen Ländern eine Rolle für die Schwierigkeit von Items spielen. Das Ergebnis dieser Fragestellung sollte demnach nicht nur einen Hinweis darauf geben, ob die Item-Anforderungsmerkmale des „Dutch Grid” überhaupt Zusammenhänge zur Itemschwierigkeit in den unterschiedlichen Ländern aufweisen, sondern auch, ob das in unterschiedlichen Ländern in ähnlichem Maße der Fall ist.

Hypothese 2b: Die Überprüfung dieser Fragestellung erfolgt exploratorisch.

Frage 2c: Weisen die kognitiv-linguistischen Anforderungsmerkmale der Items regressionsanalytische Zusammenhänge zu Itemschwierigkeiten innerhalb der Länder auf? Frage 2d: Wie groß ist der Anteil der durch die Prädiktoren aufgeklärten Varianz (Frage 2d)?

Die Frage danach, ob Zusammenhänge zwischen Itemschwierigkeiten und der kognitiv-linguistischen Kategorie von Item-Anforderungsmerkmalen des „Dutch Grid” bestehen soll ferner

Die beiden Fragen 2c und 2d lassen sich aufgrund der gemeinsamen Methodik nur schwer getrennt betrachten. Daher werden sie im Folgenden gemeinsam abgehandelt. Die beiden Fragen beinhalten zum einen, dass sie den Anteil der aufgrund der verwendeten Prädiktoren aufklärbaren Varianz der Itemschwierigkeit untersuchen, zum anderen betrachten sie Richtung und Größe der Regressionsgewichte.

Diesbezüglich wird folgende Hypothese aufgestellt:


3.5. Fragenkomplex 3: Erklärung von Differentiellen Item Funktionen

Frage 3a: Existieren den Testkulturen entsprechende, signifikante korrelative Zusammenhänge zwischen Testkultur-Indikatoren und DIF?


Frage 3b: Können die Testkultur-Indikatoren als Prädiktoren einen Teil der durch kulturelle Unterschiede verursachten Varianz der Itemschwierigkeiten zwischen den Ländern, d.h. DIF, erklären? Frage 3c: Entspricht die Richtung der Regressionsgewichte den erwarteten Stärken und Schwächen der Länder?

Neben der oben formulierten Fragestellung hinsichtlich der korrelativen Zusammenhänge zwischen differentiellen erwarteten Stärken und Schwächen der Länder und DIF, wird in diesem Schritt als Maß des Gesamtzusammenhangs die multiple Regression als Analysemethode heran-


Hypothese 3b: Differentielle Item Funktionen, d.h. die kulturell bedingte Unterschiede der Itemschwierigkeiten zwischen den Ländern, können mit Hilfe von Indikatoren nationaler Testkulturen teilweise erklärt werden.

Des Weiteren sollte die Richtung der Regressionsgewichte jeweils den durch die im Rahmen von Frage 1c durchgeführten Analysen der Testitems festgestellten testkulturellen Schwerpunkten der Länder entsprechen, so dass damit Aussagen bezüglich der relativen Stärken und Schwächen der Ländergruppen, nämlich jeweils im Vergleich zu einer anderen Gruppe, hinsichtlich der Beantwortung von Items gemacht werden können. Bezüglich der oben formulierten dritten Teilfragestellung wird folgende Hypothese aufgestellt:
Hypothese 3c: Die Richtung der Regressionsgewichte der verwendeten Prädiktoren sollte den aufgrund der Analyse der Items erwarteten Stärken und Schwächen in den nationalen Testkulturen entsprechen.

Nachdem nun die Herleitung der Hauptfragestellung aus den im Theorie teil dargestellten theoretischen und empirischen Arbeiten erfolgt ist sowie näher auf die zur Beantwortung der Hauptfragestellung notwendigen Einzelfragestellungen eingegangen wurde, werden im folgenden Teil der Arbeit die zur Beantwortung der Fragestellungen verwendeten Methoden beschrieben.
4. Methoden

In diesem Teil der Arbeit werden zunächst die Datengrundlage, die Stichprobe, die Herkunft der Items, die verwendeten Testinstrumente sowie das Design der EBAFLS-Studie beschrieben.

Nachdem unter Abschnitt 3 die Fragestellungen und die dazugehörigen Hypothesen formuliert wurden, werden in diesem Teil die dazugehörigen Analysemethoden dargelegt. Zunächst kommen die bei der Bearbeitung von Fragenkomplex 1 „Voraussetzungen und Skalierbarkeit“ zur Anwendung kommenden Methoden zur Sprache, und zwar hinsichtlich der Überprüfung der Rasch-Modellkonformität der Items, der Analysen zur Feststellung von DIF, der Einordnung der Items hinsichtlich ihrer kognitiv-linguistischen Anforderungsmerkmale sowie der Feststellung länderspezifischer testkultureller Schwerpunkte.


4.1. Datengrundlage

Diese Dissertation ist in den unter Abschnitt 1 beschriebenen europäischen Gesamtkontext und dort im Speziellen in das im Folgenden beschriebene EBAFLS-Projekt (Fandel et al., 2007) eingebettet.


Die in diesem Projekt gewonnenen internationalen Daten bilden die Grundlage für diese Dissertation.

Im EBAFLS-Projekt wurde angestrebt, einen Kultur-fairen, länderübergreifenden Kompetenzvergleich sowie das Testen von Fremdsprachen vor dem Hintergrund des Gemeinsamen

4.1.1. Herkunft der Items


103
4.1 Datengrundlage

Abbildung 4.1. Herkunft der Items zur Messung fremdsprachlichen Leseverständnisses

4.1.2. Analysen in der EBAFLS Studie


4.1.3. Design und Stichprobe


Design
Methoden 4.1 Datengrundlage

Abbildung 4.2. Getestete Sprachen in den Teilnehmerländern

wurde innerhalb jedes Landes eingesetzt, d.h. 15 Testhefte pro Land und getesteter Sprache. Die jeweiligen Stichprobengrößen waren von den finanziellen und personellen Möglichkeiten der einzelnen Länder abhängig (CITO, 2008).

Ziel war eine Zahl von insgesamt 1500 Beobachtungen pro Item; Minimale Anzahl der Antworten pro Item pro Sprache pro Land war 200, was einer minimalen Stichprobengröße von 600 Personen pro Sprache und Land entspricht. Die Anzahl der Beobachtungen pro Item pro Land werden in Tabelle 4.2 dargestellt. Wie aus dieser Tabelle ersichtlich wird, wurde nicht in allen Ländern für alle Sprachen die Minimal-Stichprobengröße von 600 Personen erreicht. Das Ziel von insgesamt 1500 Antworten pro Item konnte hier nur für Englisch erreicht werden, Deutsch liegt mit 1360 Antworten pro Item noch relativ nahe daran, und Französisch mit 947 Antworten pro Item weit unter dem Ziel. Nicht zuletzt aus diesem Grund wird der Datensatz für das französischsprachige Leseverständnis nicht in die Analysen dieser Dissertation mit einbezogen. Die Anzahl der Antworten pro Item pro Sprache pro Land entspricht jeweils in etwa einem Drittel der Gesamtzahl der Testhefte pro Sprache pro Land, da jedes Item in fünf, das heißt einem Drittel der Testhefte, enthalten war. Die durchschnittliche Antwortquote betrug 83% für Englisch, 88% für Französisch und 87% für Deutsch.


### Positionen

<table>
<thead>
<tr>
<th>Testheft</th>
<th>Position 1</th>
<th>Position 2</th>
<th>Position 3</th>
<th>Position 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1</td>
<td>A2</td>
<td>B1</td>
<td>B2</td>
</tr>
<tr>
<td>2</td>
<td>A2</td>
<td>A1</td>
<td>C1</td>
<td>C2</td>
</tr>
<tr>
<td>3</td>
<td>D1</td>
<td>D2</td>
<td>A1</td>
<td>A2</td>
</tr>
<tr>
<td>4</td>
<td>E1</td>
<td>E2</td>
<td>A2</td>
<td>A1</td>
</tr>
<tr>
<td>5</td>
<td>A1</td>
<td>A2</td>
<td>F1</td>
<td>F2</td>
</tr>
<tr>
<td>6</td>
<td>B1</td>
<td>B2</td>
<td>C1</td>
<td>C2</td>
</tr>
<tr>
<td>7</td>
<td>B2</td>
<td>B1</td>
<td>D1</td>
<td>D2</td>
</tr>
<tr>
<td>8</td>
<td>E2</td>
<td>E1</td>
<td>B1</td>
<td>B2</td>
</tr>
<tr>
<td>9</td>
<td>F1</td>
<td>F2</td>
<td>B2</td>
<td>B1</td>
</tr>
<tr>
<td>10</td>
<td>D2</td>
<td>D1</td>
<td>C2</td>
<td>C1</td>
</tr>
<tr>
<td>11</td>
<td>C1</td>
<td>C2</td>
<td>E1</td>
<td>E2</td>
</tr>
<tr>
<td>12</td>
<td>C2</td>
<td>C1</td>
<td>F2</td>
<td>F1</td>
</tr>
<tr>
<td>13</td>
<td>D1</td>
<td>D2</td>
<td>E2</td>
<td>E1</td>
</tr>
<tr>
<td>14</td>
<td>F2</td>
<td>F1</td>
<td>D2</td>
<td>D1</td>
</tr>
<tr>
<td>15</td>
<td>E1</td>
<td>E2</td>
<td>F1</td>
<td>F2</td>
</tr>
</tbody>
</table>

**Tabelle 4.1.** Das Design der EBALFS Leseverständnis-Studie: complete balanced block design (EBALFS-Bericht; CITO, 2008).

<table>
<thead>
<tr>
<th></th>
<th>Englisch</th>
<th>Französisch</th>
<th>Deutsch</th>
<th>Gesamt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>1238</td>
<td>—</td>
<td>1412</td>
<td>2650</td>
</tr>
<tr>
<td>Deutschland</td>
<td>716</td>
<td>529</td>
<td>—</td>
<td>1245</td>
</tr>
<tr>
<td>Ungarn</td>
<td>750</td>
<td>—</td>
<td>683</td>
<td>1433</td>
</tr>
<tr>
<td>Luxemburg</td>
<td>—</td>
<td>738</td>
<td>722</td>
<td>1460</td>
</tr>
<tr>
<td>Niederlande</td>
<td>526</td>
<td>—</td>
<td>434</td>
<td>960</td>
</tr>
<tr>
<td>Schottland</td>
<td>—</td>
<td>321</td>
<td>249</td>
<td>570</td>
</tr>
<tr>
<td>Spanien</td>
<td>1500</td>
<td>580</td>
<td>—</td>
<td>2080</td>
</tr>
<tr>
<td>Schweden</td>
<td>—</td>
<td>673</td>
<td>641</td>
<td>1314</td>
</tr>
<tr>
<td>Booklets Gesamt</td>
<td>4730</td>
<td>2841</td>
<td>4141</td>
<td>11712</td>
</tr>
<tr>
<td>Antworten pro Item</td>
<td>1576</td>
<td>947</td>
<td>1360</td>
<td>3883</td>
</tr>
</tbody>
</table>

**Tabelle 4.2.** Anzahl der Testhefte pro Sprache pro Land; Anzahl der Beobachtungen pro Item insgesamt (Internationaler EBALFS-Bericht; CITO, 2008).

Außer den Testheften an sich wurde ein Schülerfragebogen zur Erfassung von Hintergrundvariablen erstellt. Der Fragebogen enthielt Fragen zu Geschlecht, Alter, Herkunft der Schüler und
Eltern, sozioökonomischem Status, Muttersprache der Schüler und Eltern sowie eine Selbsteinschätzung in der jeweils getesteten Sprache und eine Skala zur Testmotivation. Aus unterschiedlichen nationalen Gründen wurde nicht der komplette Fragebogen in allen Teilnehmerländern ausgefüllt bzw. ausgewertet. Teilweise wurden auch in unterschiedlichen Ländern unterschiedliche Fragen beantwortet bzw. ausgewertet, so dass kein einheitlicher Fragebogen-Datensatz zusammengestellt werden kann. In Spanien wurde der Fragebogen überhaupt nicht bearbeitet. Die Gründe dafür waren verschieden und reichten von innerpolitischen bis hin zu finanziellen und personellen Gründen. Aus diesem Grund kann der Fragebogen nicht für Analysen herangezogen werden, was auch die Verwendung von Personenvariablen ausschließt.

Die Hauptauswertung der EBAFLS-Daten erfolgte am nationalen Testinstitut der Niederlande CITO. Erste bereinigte Daten lagen im Frühjahr 2007 vor. Für diese Dissertation werden die Daten der EBAFLS Leseverständnis-Studie für die Sprachen Deutsch und Englisch verwendet.

**Stichprobenbeschreibung**

Im Folgenden werden die in dieser Arbeit verwendeten nationalen Stichproben im Hinblick auf Geschlecht, Alter und Herkunftsland beschrieben. Bei den in dieser Studie erhobenen Stichproben handelt es sich größtenteils um sogenannte convenience-Samples, die nicht repräsentativ für die einzelnen Länder sind. Aus diesem Grund können hier auch keine länderübergreifenden Vergleiche der Personenzfähigkeiten durchgeführt werden, da dies zu verzerrten Ergebnissen führen kann. Obgleich in dieser Dissertation nicht die Schüler und Schülerleistungen der einzelnen Länder, sondern die in den unterschiedlichen Ländern konstruierten Items im Fokus der Analysen stehen, sollte dieser Sachverhalt bei der Interpretation der Ergebnisse beachtet werden. Ferner werden die Daten Rasch-skaliert. Dabei wird davon ausgegangen, dass die so gewonnenen Itemschwierigkeitsparameter, unter der Voraussetzung dass das Rasch-Modell gültig ist, stichprobennunabhängig sind (Spezifische Objektivität; Rost, 2004). Da hier ausschließlich Itemparameter betrachtet und vorhergesagt werden sollen, können die Fragestellungen demnach anhand der vorhandenen Stichproben beantwortet werden.

Für die Erhebung des englischsprachigen Leseverständnisses werden im Folgenden die Stichproben der deutschen, französischen und ungarischen Schülerinnen und Schüler beschrieben. Da in Spanien aus innerpolitischen Gründen keine Fragebögen ausgefüllt wurden, sind über diese Stichproben, abgesehen von der Anzahl der Schüler, die insgesamt am Test teilgenommen haben, leider keine weiteren Angaben möglich. Da in dieser Dissertation jedoch nicht die Schüler und Schülerleistungen der einzelnen Länder, sondern die dort konstruierten Items im Fokus stehen, ist die Tatsache der unausgefüllten Fragebögen in Spanien zwar sehr bedauerlich, schließt
aber die Verwendung der spanischen Stichprobe nicht aus. Die Daten der niederländischen Schülerinnen und Schüler werden für das englischsprachige Leseverständnis nicht verwendet. Grund dafür ist, dass keine niederländischen Testitems Teil des EBAFLS Itempools waren, somit keine Kenntnis über die niederländische Testkultur vorhanden ist und daher die oben aufgestellten Hypothesen nicht anhand der niederländischen Stichprobe überprüft werden können.


**Alter der Schüler/innen**


<table>
<thead>
<tr>
<th>Geburtsjahr</th>
<th>Schüler F</th>
<th>Schüler D</th>
<th>Schüler U</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 1984</td>
<td>1 (1)</td>
<td>2 (2)</td>
<td>7 (13)</td>
</tr>
<tr>
<td>1985</td>
<td>2 (2)</td>
<td>—</td>
<td>18 (3.4)</td>
</tr>
<tr>
<td>1986</td>
<td>8 (0.6)</td>
<td>—</td>
<td>12 (2.3)</td>
</tr>
<tr>
<td>1987</td>
<td>41 (3.3)</td>
<td>—</td>
<td>59 (11.1)</td>
</tr>
<tr>
<td>1988</td>
<td>291 (23.6)</td>
<td>1 (1)</td>
<td>135 (25.5)</td>
</tr>
<tr>
<td>1989</td>
<td>808 (65.5)</td>
<td>25 (3.6)</td>
<td>133 (25.1)</td>
</tr>
<tr>
<td>1990</td>
<td>47 (3.8)</td>
<td>336 (47.7)</td>
<td>120 (22.6)</td>
</tr>
<tr>
<td>1991</td>
<td>2 (2)</td>
<td>287 (40.8)</td>
<td>29 (5.5)</td>
</tr>
<tr>
<td>1992</td>
<td>—</td>
<td>4 (6)</td>
<td>—</td>
</tr>
<tr>
<td>1993</td>
<td>—</td>
<td>1 (1)</td>
<td>—</td>
</tr>
<tr>
<td>1994</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1995</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>≥ 1996</td>
<td>3 (2)</td>
<td>10 (1.4)</td>
<td>2 (0.4)</td>
</tr>
<tr>
<td>Keine Angabe</td>
<td>30 (2.4)</td>
<td>38 (5.4)</td>
<td>—</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fragebögen gesamt</th>
<th>Schüler F</th>
<th>Schüler D</th>
<th>Schüler U</th>
</tr>
</thead>
<tbody>
<tr>
<td>1233</td>
<td>704</td>
<td>530</td>
<td></td>
</tr>
</tbody>
</table>

*Tabelle 4.3.* Englische Items: Altersverteilung nach Geburtsjahrgängen der deutschen (D), französischen (F) und ungarischen (U) Schüler
4.1 Datengrundlage

<table>
<thead>
<tr>
<th>Geburtsjahr</th>
<th>Schüler F</th>
<th>Schüler U</th>
<th>Schüler NL</th>
<th>Schüler SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤ 1984</td>
<td>—</td>
<td>8</td>
<td>1.3</td>
<td>—</td>
</tr>
<tr>
<td>1985</td>
<td>1</td>
<td>.1</td>
<td>8</td>
<td>1.2</td>
</tr>
<tr>
<td>1986</td>
<td>2</td>
<td>.1</td>
<td>26</td>
<td>3.9</td>
</tr>
<tr>
<td>1987</td>
<td>11</td>
<td>.8</td>
<td>99</td>
<td>15.0</td>
</tr>
<tr>
<td>1988</td>
<td>217</td>
<td>15.4</td>
<td>221</td>
<td>33.4</td>
</tr>
<tr>
<td>1989</td>
<td>1026</td>
<td>72.9</td>
<td>167</td>
<td>25.3</td>
</tr>
<tr>
<td>1990</td>
<td>107</td>
<td>7.6</td>
<td>75</td>
<td>11.3</td>
</tr>
<tr>
<td>1991</td>
<td>1</td>
<td>.1</td>
<td>29</td>
<td>4.4</td>
</tr>
<tr>
<td>1992</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1993</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1994</td>
<td>1</td>
<td>.1</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1995</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>≥ 1996</td>
<td>7</td>
<td>.5</td>
<td>3</td>
<td>.5</td>
</tr>
<tr>
<td>Keine Angabe</td>
<td>35</td>
<td>2.5</td>
<td>25</td>
<td>3.8</td>
</tr>
<tr>
<td>Fragebögen gesamt</td>
<td>1408</td>
<td>661</td>
<td>421</td>
<td>592</td>
</tr>
</tbody>
</table>

Tabelle 4.4. Deutsche Items: Altersverteilung nach Geburtsjahrgängen der französischen (F), ungarischen (U), niederländischen (NL) und schwedischen (SW) Schüler

**Geschlecht**

Wie sich in Tabelle 4.5 zeigt, haben bezüglich der Englisch-Tests sowohl in Deutschland als auch in Frankreich deutlich mehr Schülerinnen als Schüler an der Studie teilgenommen. In Ungarn hingegen ist der Anteil beider Geschlechter in etwa gleich groß, dort ist der Anteil der männlichen Schüler sogar etwas größer. Bei den Deutsch-Stichproben (Tabelle 4.6) zeigt sich, dass auch hier insgesamt in allen Ländern mehr Schülerinnen als Schüler an den Tests teilgenommen haben.

**Herkunftsland**

Da davon ausgegangen wird, dass die durch die Testkulturen transportierten sozialen Werte der Gesellschaft einen Einfluss auf die Testleistung und Itemschwierigkeit der Ländergruppen haben sollten, wird an dieser Stelle zusätzlich zu Alter und Geschlecht in den Ländern der Anteil der Schülerinnen und Schüler analysiert, die nicht ursprünglich aus dem jeweiligen Testland stammen. Möglicherweise könnte ein sehr großer oder sehr unterschiedlicher Anteil von aus ande-
Tabelle 4.5. Geschlecht der französischen (F), deutschen (D) und ungarischen (U) Schülerinnen und Schüler in den nationalen Englisch-Stichproben

<table>
<thead>
<tr>
<th>Geschlecht</th>
<th>Schüler F</th>
<th>Schüler D</th>
<th>Schüler U</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anzahl (%)</td>
<td>Anzahl (%)</td>
<td>Anzahl (%)</td>
</tr>
<tr>
<td>Männlich</td>
<td>493 40</td>
<td>274 38.9</td>
<td>267 50.4</td>
</tr>
<tr>
<td>Weiblich</td>
<td>723 58.6</td>
<td>412 58.5</td>
<td>245 46.2</td>
</tr>
<tr>
<td>Keine Angabe</td>
<td>17 1.4</td>
<td>18 2.6</td>
<td>18 3.4</td>
</tr>
<tr>
<td>Fragebögen gesamt</td>
<td>1233 —</td>
<td>704 —</td>
<td>530 —</td>
</tr>
</tbody>
</table>

Tabelle 4.6. Geschlecht der französischen (F), ungarischen (U), niederländischen (NL) und schwedischen (SW) Schülerinnen und Schüler in den nationalen Deutsch-Stichproben

<table>
<thead>
<tr>
<th>Geschlecht</th>
<th>Schüler F</th>
<th>Schüler U</th>
<th>Schüler NL</th>
<th>Schüler SW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Anzahl (%)</td>
<td>Anzahl (%)</td>
<td>Anzahl (%)</td>
<td>Anzahl (%)</td>
</tr>
<tr>
<td>Männlich</td>
<td>588 41.8</td>
<td>290 43.9</td>
<td>179 42.5</td>
<td>249 42.1</td>
</tr>
<tr>
<td>Weiblich</td>
<td>796 56.5</td>
<td>355 53.7</td>
<td>230 54.6</td>
<td>301 50.8</td>
</tr>
<tr>
<td>Keine Angabe</td>
<td>24 1.7</td>
<td>16 2.4</td>
<td>12 2.9</td>
<td>42 7.1</td>
</tr>
<tr>
<td>Fragebögen gesamt</td>
<td>1408 661</td>
<td>421 592</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ren Ländern stammenden Schüler/innen den Einfluss der gesellschaftsspezifischen Testkultur in den jeweiligen Ländern verringern, da diese nicht so lange der in dem jeweiligen Land gültigen Testkultur exponiert waren. Dadurch könnten die Ergebnisse möglicherweise verzerrt werden. Es wird hier ausschließlich der Anteil der Schülerinnen und Schüler überprüft, die nicht im Testland geboren sind, da davon ausgegangen wird, dass alle übrigen Teilnehmer/innen, auch wenn sie einen Migrationshintergrund besitzen, jedoch im Testland geboren sind, das Schulsystem der jeweiligen Länder von Anfang an durchlaufen haben und somit alle für die gleiche Dauer der jeweils vorherrschenden Testkultur exponiert waren. Bezüglich der Englisch-Stichprobe zeigt sich, dass die Anteile der in anderen Ländern geborenen Schülerinnen und Schüler sich zwar geringfügig unterscheiden, vor allem zwischen Frankreich und Ungarn, insgesamt jedoch nicht sehr weit auseinander liegen mit 5.7%, 4.9% und 4.2%. Der Prozentsatz von Personen, die möglicherweise vorher der Testkultur eines anderen Landes ausgesetzt waren, unterscheidet sich hier zwischen den Ländern also nicht deutlich. Obgleich die Unterschiede hier nicht auf Signifikanz überprüft wurden, wird doch davon ausgegangen, dass der Anteil von aus einem anderen Land bzw. Testkultur stammenden Schüler keinen unterschiedlich starken, die Ergebnisse mög-
4.1 Datengrundlage

<table>
<thead>
<tr>
<th>Herkunftsland</th>
<th>Schüler F</th>
<th>Schüler D</th>
<th>Schüler U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testland</td>
<td>1140 92.5</td>
<td>657 93.3</td>
<td>493 93</td>
</tr>
<tr>
<td>Anderes Land</td>
<td>71 5.7</td>
<td>34 4.9</td>
<td>22 4.2</td>
</tr>
<tr>
<td>Keine Angabe</td>
<td>22 1.8</td>
<td>13 1.8</td>
<td>15 2.8</td>
</tr>
<tr>
<td>Fragebögen gesamt</td>
<td>1233</td>
<td>704</td>
<td>530</td>
</tr>
</tbody>
</table>

Tabelle 4.7. Herkunft der französischen (F), deutschen (D) und ungarischen (U) Schülerinnen und Schüler in den nationalen Englisch-Stichproben

...licherweise verzerrenden Einfluss auf die Wirkung der hier betrachteten Testkulturen in den verschiedenen Ländern haben sollte. Bezüglich des Anteils der in einem anderen Land geborenen Schüler/innen zeigt sich für die Deutsch-Stichprobe, dass die Anteile sich insgesamt deutlicher unterscheiden als das bei der Englisch-Stichprobe der Fall ist. Mit 3.1% und 3.2% ist der Anteil in Frankreich und Ungarn relativ gering, in den Niederlanden und Schweden mit jeweils über 7% deutlich höher.

4.1.4. Ergebnisse der EBAFLS-Leseverständnis-Studie

Modell-Konformität auf.
Sowohl hinsichtlich eines globalen Modelltests, als auch bei lokalen Modelltests, d.h. der Analyse von einzelnen Items, zeigten sich von der Leistung unabhängige beobachtete Leistungsunterschiede, d.h. DIF, bei den Leseverständnis-Items, und zwar sowohl in allen Sprachen, als auch über alle Länder hinweg.
Im Rahmen der vorliegenden Arbeit sollen nun, wie unter Abschnitt 3 dargelegt, mögliche Ursachen für diese Differentiellen Item Funktionen aufgespürt und analysiert werden. Dieses Ergebnis hinsichtlich der Existenz von Differentiellen Item Funktionen sowie die Tatsachen, dass die in der Studie verwendeten Items zum einen von den Ländern eingereicht Item wurden und zum anderen innerhalb der Länder in nationalen Testverfahren bereits verwendet wurden und somit repräsentativ sein sollten, tragen dazu bei, dass die EBAFLS-Daten hervorragend für die Untersuchung von Zusammenhängen zwischen nationalen Testkulturen und DIF geeignet sind.

4.1.5. Aufbereitung der Daten

Vor der erneuten Skalierung der Daten für diese Dissertation mussten die ursprünglichen EBAFLS-Rohdaten mit Hilfe einer MySQL Datenbank-Anwendung in eine von ConQuest und SPSS verwendbare Struktur gebracht werden. Die Datenbank ist im Anhang einsehbar.

4.2. Methoden zur Beantwortung von Fragenkomplex 1

Im Folgenden werden die Methoden dargestellt, die zur Bearbeitung der im Rahmen von Fragenkomplex 1 formulierten Fragestellungen herangezogen werden. Diese Methoden dienen dazu, die Voraussetzungen zur Beantwortung der weiteren Fragestellungen zu überprüfen. Es wird dabei zunächst auf die Methode zur Überprüfung der Modellkonformität der Items eingegangen, danach auf die Methoden zur Analyse von Differentiellen Item Funktionen, und auf die Methoden für die Analyse der Items hinsichtlich ihrer Anforderungsmerkmale und der Existenz nationaler Testkulturen.
4.2.1. Überprüfung der Rasch-Modellkonformität der Items innerhalb der Länder

Frage 1a: „Weisen die Items innerhalb der Länder Rasch-Modellkonformität auf?“


Diese Formel (Abbildung 4.3) sagt aus, dass die Wahrscheinlichkeit dass eine Person \( v \), ein
4 Methoden zur Beantwortung von Fragenkomplex 1

\[
p(X_{vi} = 1) = \frac{\exp(\theta_v - \sigma_i)}{1 + \exp(\theta_v - \sigma_i)}
\]

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Bedeutung</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p(X_{vi} = 1))</td>
<td>die Wahrscheinlichkeit, dass Person v Item i korrekt beantwortet</td>
</tr>
<tr>
<td>(\theta_v)</td>
<td>Personenparameter der Person v, d.h. die Position der Person hinsichtlich der latenten Fähigkeit</td>
</tr>
<tr>
<td>(\sigma_i)</td>
<td>Itemschwierigkeit von Item i</td>
</tr>
<tr>
<td>(\exp(\theta_v - \sigma_i))</td>
<td>die Exponentialfunktion der Differenz des Personenparameters und der Itemschwierigkeit</td>
</tr>
</tbody>
</table>

**Abbildung 4.3.** Modellgleichung des eindimensionalen Rasch-Modells (entnommen aus Rost, 2004)

bestimmtes Item i korrekt beantwortet, abhängig ist von einerseits der Ausprägung der Person auf der gemessenen latenten Variable, und zum anderen von der Schwierigkeit des betrachteten Items. Ist die Differenz der beiden Parameter \(\leq 0\), dann beträgt die Lösungswahrscheinlichkeit einer Person mindestens 50%. Für eine ausführlichere Darstellung des Rasch-Modells wird auf Rost (2004) verwiesen.

Die Rasch-Analysen im Rahmen dieser Arbeit werden mit dem Programm ConQuest (Wu, Adams & Wilson, 1998) durchgeführt. Wie es im Rahmen von IRT-Modellen üblich ist, stellt auch ConQuest die Itemschwierigkeiten in Einheiten auf einer Logit-Skala zur Verfügung. Durch den „constraint cases“ Befehl in der Syntax wird der Mittelwert der latenten Verteilung auf Null gesetzt. Die daraus errechneten Werte sind somit zentriert und untereinander vergleichbar. Zur Testung der Rasch-Modellkonformität werden in ConQuest gewichtete MNSQ-Fit-Statistiken (weighted mean square; gewichtete quadrierte Abweichungswerte; Erwartungswert=1) zur Verfügung gestellt. Um die Modellkonformität eines Items zu überprüfen, wird um den unter der H0 (das Antwortverhalten beim fokussierten Item erfolgt Rasch-konform) erwar-

title Rasch EBAFLS German
Schweden;
set warnings=no,
constraints=cases;
DATAFILE Germany.dat;
FORMAT studID 1-5 itemID 6-15 responses 16 Country 17;
CODES 0,1;
MODEL Item;
estimate;
show !estimates=latent >> Rasch.Germany.shw;
quit;

Abbildung 4.4. Beispielsyntax ConQuest

oben bereits dargelegt. Neben dem Datensatz wird ConQuest die Struktur des Datensatzes mitgeteilt. „StudID 1-5“ bedeutet hier beispielsweise, dass sich die Schüler-Vpn in den Spalten 1-5 des Datensatzes befindet.

4.2.2. Methoden zur Analyse von Differentiellen Item Funktionen

Frage 1b: Wie groß ist der Anteil von Items mit Differentiellen Item Funktionen?

Verschiedene existierende Methoden zur DIF-Analyse wurden bereits unter 2.1 beschrieben. Im Folgenden wird auf die in der vorliegenden Dissertation verwendeten Methoden eingegangen. Ausgangspunkt für alle unter 4.3 beschriebenen Methoden ist die eindimensionale Rasch-Skalierung der Daten zur Berechnung der Schwierigkeits- und DIF-Parameter.

Die Skalierung und die Berechnung der DIF-Parameter erfolgten ebenfalls in Conquest (Wu, Adams & Wilson, 1998). Auch die unter 4.2.1 bereits dargestellte Syntax findet sich zur Modellierung der DIF wieder, allerdings unterscheidet sich der Model-Befehl. Differentielle Item Funktionen werden in ConQuest wie in der folgenden Beispielsyntax modelliert:

MODEL Item + Gruppe + Item * Gruppe

Dieser Ausdruck beinhaltet zwei Facetten, nämlich das Item und die Gruppenzugehörigkeit. Zur Berechnung von DIF identifiziert ConQuest nun alle möglichen Kombinationen von Item und

Auch für die DIF-Parameter wird ein Standardfehler errechnet und zur Verfügung gestellt, anhand dessen festgestellt werden kann, ob die Differentielle Item Funktion signifikant ist. Wenn der DIF-Parameter eines Items größer als zweimal der dazugehörige Standardfehler ist, dann unterscheiden sich die beiden Gruppen signifikant hinsichtlich ihrer Itemschwierigkeit (Wu, Adams & Wilson, 1998).


### 4.2.3. Methoden zur Analyse von Indikatoren nationaler Testkulturen

#### Frage 1c: Lassen sich unterschiedliche Testkulturen der Länder feststellen?

Die Einordnung der Items hinsichtlich der Itemmerkmale ist ein für diese Arbeit kritischer Punkt, da daraus die Indikatoren nationaler Testkulturen sowie später die spezifischen Hypothesen in Fragenkomplex 3 abgeleitet werden. Dies geschieht in mehreren Schritten, die im Folgenden beschrieben werden:
Schritt 1: Einordnung der Items im Hinblick auf Item- und Anforderungsmerkmale


Schritt 2: Auswahl der Ratings und Überprüfung der Inter-Rater-Übereinstimmung


Methoden

4.2 Methoden zur Beantwortung von Fragenkomplex 1


Schritt 3: Auswahl der schwierigkeitsdeterminierenden Eigenschaften und Prädiktoren für Itemschwierigkeit und Differenzielle Item Funktionen

Dieser Schritt dient dazu, aus den Itemmerkmals-Kategorien des „Dutch Grid“ diejenigen Item-Anforderungsmerkmale auszuwählen, die für die Feststellung der Testkulturen und auch die weiteren Analysen hinsichtlich der untersuchten Zusammenhänge mit Itemschwierigkeit und Differenziellen Item Funktionen verwendet werden sollen. Dabei handelt es sich um die bereits unter 2.2.4 beschriebenen Itemcharakteristika. Für die Auswahl wird auf die dort dargelegten theoretischen Ansätze und empirischen Ergebnisse zurückgegriffen.

Schritt 4: Häufigkeiten von Itemeigenschaften: Die Bildung nationaler Testprofile

Die in dieser Arbeit verwendeten Item-Anforderungsmerkmale werden für jedes der Teilnehmerländer hinsichtlich der Häufigkeit ihres Vorkommens bzw. des Vorkommens einer bestimmten Ausprägung der Eigenschaft analysiert.


Zur Feststellung der Testkulturen werden alle eingereichten Items untersucht, nicht nur die tatsächlich getesteten. Dies sollte aufgrund einer größeren Itemanzahl die Repräsentativität der Items für die Testkulturen verbessern.
Die Items jedes Landes werden nun zunächst hinsichtlich der Häufigkeit (in Prozent) des Vorkommens der verschiedenen Item-Anforderungsmerkmale analysiert. Danach wird überprüft, ob sich die Items unterschiedlicher Länder hinsichtlich der Häufigkeit des Vorkommens der Anforderungsmerkmale signifikant unterscheiden. Es wird dabei angenommen, dass signifikante Unterschiede auch auf unterschiedliche Testkulturen und somit auf zu erwartende Stärken und Schwächen der einzelnen Gruppen, jeweils relativ zur Vergleichsgruppe gesehen, hinweisen.

Um für die spätere Aufstellung der Hypothesen hinsichtlich der Zusammenhänge von Testkultur-Indikatoren und DIF und zu erwartende Stärken und Schwächen der Gruppen sicherzustellen zu können, dass deskriptiv gefundenen Häufigkeitsunterschiede auch relevant sind, wird eine Effektgröße zur Überprüfung von Unterschieden bei Prozentwerten, Cohens $h$ (Cohen, 1988), verwendet. Zunächst werden dazu die Prozentanteile der verschiedenen Itemeigenschaften bei den Items eines Landes Arcussinus-transformiert:

$$
\Phi = 2\text{acsin}\sqrt{P}
$$

(wobei $P = $ Prozentanteil)

Danach erfolgt die Berechnung der Effektstärke $h$:

$$
h = |\Phi_1 - \Phi_2|
$$

(zweiseitig)

Bei der Effektstärke $h$ handelt es sich um die Differenz der Arcussinus-transformierten Prozentanteile der Itemeigenschaften bei den Items von jeweils zwei Ländern. Die Arcussinus-Transformation hat den Vorteil, dass damit berücksichtigt wird, ob sich der gefundene Unterschied eher an den Rändern der Verteilung oder in der Mitte befindet. Da die Wahrscheinlichkeit, kleine Unterschiede an den Verteilungs­rändern zu entdecken, hier als geringer angesehen wird als das in der Mitte einer Verteilung der Fall ist, werden dort auch kleinere Unterschiede schneller signifikant. Bei der Effektstärke kann bei einem Wert von 0.2 von einem kleinen, ab 0.5 von einem mittleren und ab 0.8 von einem großen Effekt ausgegangen werden. Zusätzlich zu den Effektstärken werden darauf basierende Signifikanztests durchgeführt (siehe Cohen, 1988).
4.3. Methoden zur Beantwortung der Fragenkomplexe 2 und 3


4.3.1. Methoden zu Fragenkomplex 2: Erklärung der Itemschwierigkeiten innerhalb der Länder

Frage 2a: Weisen die kognitiv-linguistischen Anforderungsmerkmale der Items korrelative Zusammenhänge zu den Itemschwierigkeiten innerhalb der Länder auf?


alle in die Analysen mit einbezogenen Länder. Für die Analysen wird SPSS verwendet.

Frage 2b: Sind die Korrelationsmuster in den Ländern vergleichbar?

Um die in den jeweiligen Ländern vorhandenen Korrelationsmuster vergleichbar zu machen, werden die signifikanten Korrelationen jeweils Fisher-z-transformiert. Dies führt zu einer annähernden Normalverteilung und ermöglicht einen Vergleich der Korrelationskoeffizienten zweier Grundgesamtheiten. Die Transformation erfolgt nach folgender Formel (Bortz, 2005):

\[ z = \frac{1}{2} \ln \left( \frac{1 + r}{1 - r} \right) \]

(wobei \( r \) = Korrelationskoeffizient)

Es soll hier überprüft werden, ob sich zwei Korrelationen, die für zwei voneinander unabhängige Stichproben ermittelt wurden, signifikant voneinander unterscheiden. Dazu wird im nächsten Schritt \( z \) ermittelt (nach Bortz, 2005):

\[ z = \frac{Z_1 - Z_2}{\sigma(Z_1-Z_2)} \]

wobei

\[ \sigma(Z_1-Z_2) = \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}} \]


Frage 2c: Weisen die kognitiv-linguistischen Anforderungsmerkmale der Items regressionsanalytische Zusammenhänge zu Itemschwierigkeiten innerhalb der Länder auf? Frage 2d: Wie groß ist der Anteil der durch die Prädiktoren aufgeklärten Varianz?

Da die Berechnung des LLTM+e sehr aufwändig ist und keinen Vorteil gegenüber der zweiseitigen Methode aufzuweisen scheint, wird in dieser Dissertation die gleichwertige, zweiseitige Methode zur Erklärung der Varianz der Itemschwierigkeiten innerhalb und zwischen den Ländern angewandt.

Ferner wird untersucht, welche der Prädiktoren signifikant zur Erklärung der Itemschwierigkeit beitragen. Die Vorhersage der Itemschwierigkeit erfolgt für alle in die Analyse einbezogenen Länder. Die Analysen erfolgen in SPSS.

4.3.2. Methoden zu Fragenkomplex 3: Erklärung von Differentiellen Item Funktionen

| Frage 3a: Existieren den Testkulturen entsprechende, signifikante korrelative Zusammenhänge zwischen Testkultur-Indikatoren und DIF? |

Wenn die Testkultur und somit die Itemherkunft einen Einfluss auf den kulturell bedingten Unterschied der Itemschwierigkeit zwischen zwei Gruppen (in diesem Fall Länder) hat, dann sollte sich das zunächst darin zeigen, dass „eigene“ Items für Schüler leichter korrekt zu beantworten sind als für Schüler, die aus einer anderen Testkultur stammen.

Zur Untersuchung dieser Hypothese werden unter Anwendung der oben unter 2.1.2 beschriebenen Methode von Klieme & Baumert (2001) jeweils paarweise DIF Parameter zwischen jeweils 2 Ländern berechnet (siehe auch Frage 1b) und mit der Variablen „Herkunft der Items“ korreliert. Hypothese ist hier, dass die Korrelation zwischen DIF und der Tatsache, dass ein Item aus dem eigenen Land stammt, negativ ausfallen sollte. Dies stellt in diesem Fall einen Vorteil für die Fokusgruppe dar, da DIF dann zu deren Vorteil ausfällt und die Items für diese Gruppe somit leichter sind. Die Korrelation zwischen DIF und Items aus dem anderen Land, d.h. der jeweiligen Referenzgruppe, sollte hingegen positiv sein, d.h. in diesem Fall zu Ungunsten der Fokusgruppe, und die Itemschwierigkeit im Vergleich zur anderen Gruppe deutlich niedriger sein.

Hintergrund und Methode

Frage 3b: Können die Testkultur-Indikatoren als Prädiktoren einen Teil der durch kulturelle Unterschiede verursachten Varianz der Itemschwierigkeiten zwischen den Ländern, d.h. DIF, erklären? Frage 3c: Entspricht die Richtung der Regressionsgewichte den erwarteten Stärken und Schwächen der Länder?


Daher würden so die Zusammenhänge zwischen den differentiellen Testkulturen, d.h. den erwarteten Stärken und Schwächen der Länder, und der kulturell verursachten Varianz, d.h. DIF, verdeckt. Da aber diese differentiellen Zusammenhänge besonders interessieren, wird die Itemherkunft hier als Prädiktor nicht mit aufgenommen. Auch hier wird im Rahmen der multiplen Regression das Einschluss-Verfahren angewandt. Es werden jeweils mehrere Regressionsmodelle gerechnet: In das erste Modell werden alle Prädiktoren mit aufgenommen, in das Endmodell hingegen ausschließlich die Prädiktoren, die einerseits signifikant sind und andererseits den aufgrund der Testkulturen erwarteten Stärken und Schwächen der Gruppen entsprechen. Der so aufgeklärte Anteil der Varianz der Differentiellen Item Funktionen entspricht somit dem Anteil, der tatsächlich durch Stärken und Schwächen der Gruppen verursacht wird und nicht durch eine mit erfasste, konstrukturrelevante Dimension begründet ist. Die unterschiedlichen Modelle sind auch im Rahmen der Ergebnisdarstellung nochmals beschrieben.
4.4. Anmerkung zum Umgang mit der Inflation des Alpha-Fehlers

Im Rahmen der vorliegenden Arbeit werden zur Analyse der Daten hauptsächlich Zusammenhangsanalysen, d.h. Korrelationen und Regressionen, durchgeführt. Bei einer großen Anzahl von Signifikanztests wird üblicherweise eine Korrektur des Alpha-Fehlers notwendig. Diese wird aus folgenden Gründen in dieser Arbeit nicht durchgeführt:

Zum Ersten werden die Korrelationen an unterschiedlichen Stichproben durchgeführt. So wurden im Rahmen von Fragenkomplex 2 beispielsweise bezüglich der Englisch-Items insgesamt 56 Korrelationen (4 Länder * 14 Itemmerkmale) durchgeführt. Da es sich jedoch um 4 Stichproben handelt, reduziert sich hier die Anzahl der Korrelationen pro Stichprobe auf 14.

Zweitens wurden in Bezug auf den dritten Fragenkomplex in diesem Fall die Stichproben aus jeweils 2 Ländern zusammengesetzt, nämlich jeweils den beiden Ländern, für die DIF-Parameter berechnet wurden. Aufgrund der unterschiedlichen Länderkombinationen handelt es sich hierbei aus diesem Grund ebenfalls um unterschiedliche Stichproben. So wurden bezüglich der Englisch-Items im dritten Fragenkomplex 108 Korrelationen durchgeführt (18 Itemmerkmale * 6 Länderpaarungen), jedoch verringert sich ob der oben genannten Tatsache die Anzahl der Korrelationen pro Stichprobe auf 18.


Zum Dritten handelt es sich bei den letztlich berechneten Regressionsmodellen um die eigentlichen „Endmodelle“. Hier wurden pro DIF-Länder-Paarung, d.h. pro Stichprobe, 2-3 sukzessive Regressionsmodelle gerechnet. Aus diesem Grund sollte hier eine Korrektur des Alpha-Fehlers nicht notwendig sein.
5. Ergebnisse

In diesem Teil der Arbeit werden die Ergebnisse der in den drei Bereichen „Voraussetzungen und Skalierbarkeit“ (5.1), „Erklärung von Itemsschwierigkeiten“ (5.2) und „Erklärung von DIF“ (5.3) formulierten Fragestellungen nacheinander dargestellt.

5.1. Ergebnisse Fragenkomplex 1: „Voraussetzungen und Skalierbarkeit“


5.1.1. Zu Frage 1a

Frage 1a: Entsprechen die Items innerhalb der Länder dem Rasch-Modell?


Im Folgenden werden die Ergebnisse bezüglich der Rasch-Modellkonformität der Items inner-
halb der einzelnen Länder tabellarisch berichtet. Dargestellt wird die Anzahl modellkonformer Items pro Land, darüberhinaus wird differenziert auf diejenigen Items eingegangen, die möglicherweise nicht perfekt dem Rasch-Modell entsprechen. Die Schwierigkeitsparameter der Items, die dazugehörigen Standardfehler sowie die WMNSQ Fit-Statistiken aller Items sind im Anhang einzusehen. Die Überprüfung der Modellkonformität erfolgt ausschließlich auf Itemebene, da zum einen in ConQuest kein Gesamt-Modell-Fit vorgesehen ist, und zum anderen, wie auch im Methodenteil (4.2) bereits dargelegt wurde, das Rasch-Modell aus theoretischen Überlegungen zugrunde gelegt wird, da auch im GERS von einer eindimensionalen Skalierung der fremdsprachlichen Lesefähigkeit ausgegangen wird. Im Folgenden werden zunächst die Ergebnisse für die englischen, danach für die deutschen Items dargestellt. Wie aus der Tabelle ersichtlich

<table>
<thead>
<tr>
<th>Land</th>
<th>Anzahl modellkonformer Items</th>
<th>Nicht modellkonforme Items</th>
<th>MNSQ (weighted) KI</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>119/122</td>
<td>REFR_4_2</td>
<td>0.90 (0.91, 1.09)</td>
<td>-2.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REHU_24_1</td>
<td>0.83 (0.90, 1.10)</td>
<td>-3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REHU_25_6</td>
<td>1.09 (0.92, 1.08)</td>
<td>2.0</td>
</tr>
<tr>
<td>Deutsch</td>
<td>121/122</td>
<td>REFR_6_6</td>
<td>1.12 (0.89, 1.11)</td>
<td>2.1</td>
</tr>
<tr>
<td>Ungarn</td>
<td>121/122</td>
<td>REGE_9_10</td>
<td>1.17 (0.86, 1.14)</td>
<td>2.3</td>
</tr>
<tr>
<td>Spanien</td>
<td>119/122</td>
<td>REFR_4_7</td>
<td>0.90 (0.91, 1.09)</td>
<td>-2.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>REFR_6_3</td>
<td>0.90 (0.91, 1.09)</td>
<td>-2.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RESP_18_4</td>
<td>1.11 (0.93, 1.07)</td>
<td>2.7</td>
</tr>
</tbody>
</table>

**Tabelle 5.1.** Anzahl der Rasch-modellkonformen Items pro Land (Englisch-Items). N = 122

wird, sind in allen Teilnehmerländer jeweils nur sehr wenige der 122 getesteten Englisch-Items nicht perfekt modellkonform. Um die Modellkonformität eines Items zu überprüfen, werden um den unter der H0 (der geschätzte Parameter weicht nicht vom angenommenen Rasch-Modell ab) erwarteten Parameter (mit dem Wert 1) 95%-Konfidenzintervalle (KI) gelegt. Wenn die gewichtete MNSQ-Fit-Statistik außerhalb des KI liegt und der dazugehörige T-Wert den Wert 2.0 (bzw. 1.96) überschreitet, dann ist das Item nicht perfekt Rasch-modellkonform, da der aus den vorliegenden Daten geschätzte Parameter von dem unter Gültigkeit des Modells erwarteten Wert signifikant abweicht (Wu, Adams & Wilson, 1998; Wilson, 2005). Ein positiver T-Wert, der den Wert 2 überschreitet, weist hier auf eine niedrige Trennschärfe des Items hin.

Nach Adams und Khoo (1996) sollte der WMNSQ-Wert nicht die Werte 0.75 unter- bzw.
Ergebnisse 5.1 Ergebnisse Fragenkomplex 1


<table>
<thead>
<tr>
<th>Land</th>
<th>Anzahl modellkonformer Items</th>
<th>Nicht modellkonforme Items</th>
<th>MNSQ (weighted)</th>
<th>KI</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>99/101</td>
<td>RGGE_27_3</td>
<td>1.08 (0.93, 1.07)</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGGE_28_3</td>
<td>1.09 (0.92, 1.08)</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>Niederlande</td>
<td>101/101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schweden</td>
<td>100/101</td>
<td>RGNE_11_22</td>
<td>0.86 (0.87, 1.13)</td>
<td>-2.3</td>
<td></td>
</tr>
<tr>
<td>Ungarn</td>
<td>98/101</td>
<td>RGGE_27_4</td>
<td>0.88 (0.88, 1.12)</td>
<td>-2.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGNE_10_3</td>
<td>1.16 (0.88, 1.12)</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGNE_10_6</td>
<td>0.84 (0.85, 1.15)</td>
<td>-2.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGSW_25_5</td>
<td>1.16 (0.85, 1.15)</td>
<td>1.9</td>
<td></td>
</tr>
</tbody>
</table>

Tabelle 5.2. Anzahl der Rasch-modellkonformen Items pro Land (Deutsch-Items). N = 101


Hypothese 1a: Die Items weisen innerhalb der Länder Rasch-Modellkonformität auf.

Die im Rahmen dieser Analysen geschätzten Item-Schwierigkeitsparameter werden für die Analysen in Fragenkomplex 2 als abhängige Variable verwendet.

5.1.2. Zu Frage 1b

Diese Fragestellung bezieht sich auf die zweite Voraussetzung, die für weitere Analysen erfüllt sein sollte: Es sollte ein nicht geringer Anteil von Items signifikante Differentielle Item Funktionen aufweisen. Es sollten etwa 35% der Items signifikante DIF aufweisen. Wenn der geschätzte DIF-Parameter größer ist als zweimal der dazugehörige Standardfehler, dann weist das Item signifikant Differentielle Item Funktionen auf (Wu, Adams & Wilson, 1998). Das bedeutet, die ICCs der jeweiligen beiden Gruppen unterscheiden sich signifikant. Für die Analysen in Fragenkomplex 3 werden zwar alle DIF Parameter, unabhängig von der Signifikanz, einbezogen; nur wenige signifikante DIF Parameter würden jedoch auf insgesamt nur sehr geringe Unterschiede hinsichtlich der Itemfunktionen hinweisen, was möglicherweise zu einer Varianzeinschränkung führen könnte. Im Folgenden werden hinsichtlich der paarweise durchgeführten Analysen der Anteil der Items mit signifikanten Differentiellen Item Funktionen berichtet. Aufgrund der großen Anzahl von Items wird bezüglich der DIF Parameter und der dazugehörigen Standardfehler auf den Anhang verwiesen.

Wie aus Tabelle 5.3 ersichtlich wird, weisen bei paarweisen DIF-Analysen zwischen jeweils zwei Ländern bei den Englisch-Items zwischen 44.2% und 66.4% der Items signifikante Differentielle Item Funktionen auf.

In Tabelle 5.4 wird gezeigt, dass auch bezüglich der Deutsch-Items eine große Anzahl signifikante DIF aufweisen, nämlich zwischen 39.6% und 59.4%. Auch hier werden im Anhang die DIF-Parameter sowie die dazugehörigen Standardfehler berichtet. Diese sind analog zu den Englisch-Items zu interpretieren.
Ergebnisse

5.1 Ergebnisse Fragenkomplex 1

<table>
<thead>
<tr>
<th>Paarweise DIF-Analyse zwischen</th>
<th>Anzahl/Anteil Items mit DIF (sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deutschland-Frankreich</td>
<td>56/122 (44.2%)</td>
</tr>
<tr>
<td>Frankreich-Ungarn</td>
<td>66/122 (54.1%)</td>
</tr>
<tr>
<td>Frankreich-Spanien</td>
<td>81/122 (66.4%)</td>
</tr>
<tr>
<td>Deutschland-Ungarn</td>
<td>52/122 (42.6%)</td>
</tr>
<tr>
<td>Deutschland-Spanien</td>
<td>79/122 (64.8%)</td>
</tr>
<tr>
<td>Ungarn-Spanien</td>
<td>54/122 (44.3%)</td>
</tr>
</tbody>
</table>

Tabelle 5.3. Anzahl der Englisch-Items mit signifikanten differentiellen Item Funktionen

<table>
<thead>
<tr>
<th>Paarweise DIF-Analyse zwischen</th>
<th>Anzahl Items mit DIF (sig.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich-Ungarn</td>
<td>50/101 (49.5%)</td>
</tr>
<tr>
<td>Frankreich-Niederlande</td>
<td>58/101 (57.4%)</td>
</tr>
<tr>
<td>Frankreich-Schweden</td>
<td>60/101 (59.4%)</td>
</tr>
<tr>
<td>Niederlande-Schweden</td>
<td>48/101 (47.5%)</td>
</tr>
<tr>
<td>Niederlande-Ungarn</td>
<td>48/101 (47.5%)</td>
</tr>
<tr>
<td>Schweden-Ungarn</td>
<td>40/101 (39.6%)</td>
</tr>
</tbody>
</table>

Tabelle 5.4. Anzahl der Deutsch-Items mit signifikanten differentiellen Item Funktionen

Hypothese 1b: Ein großer Anteil der getesteten Items, d.h. mehr als 35% weist signifikante differentielle Item Funktionen auf.

Die in Hypothese 1b getroffenen Annahmen können beibehalten werden. Sowohl bezüglich der Deutsch- als auch der Englisch-Items weist bei paarweisen DIF-Analysen ein Anteil von deutlich über den oben genannten 35% der Items signifikant Differentielle Item Funktionen auf. Diese Voraussetzung kann damit als erfüllt angesehen werden. Die hier geschätzten DIF-Parameter werden für die weiteren Analysen unter 5.3 verwendet.

5.1.3. Zu Frage 1c

Frage 1c: Lasse sich unterschiedliche Testkulturen der Länder feststellen?

Im Folgenden wird nun die dritte der für weitere Analysen notwendigen Voraussetzungen bearbeitet. Dies erfolgte in mehreren Schritten, und zwar waren dies zunächst die Einordnung der
Items durch Experten mit Hilfe des „Dutch Grid“ Kategoriensystems, sodann –zweitens– die Auswahl der Ratings und die Überprüfung der Inter-Rater-Reliabilität, drittens die Auswahl der für die Analyse verwendeten Item-Anforderungsmerkmale, und schließlich die Bildung nationaler Testprofile. Die Ergebnisse dieser vier Schritte werden im Folgenden berichtet. Wie im Rahmen des Methodenteils bereits angeführt, waren in einem ersten Schritt die Items im Rahmen des EBAFLS-Projekts bereits von den Fremdsprachen-Experten innerhalb der Länder eingeordnet worden. Über diese Prozedur, die bereits im Rahmen der ursprünglichen EBAFLS-Studie stattfand, wurde bereits im Abschnitt 4 ausführlicher berichtet, weshalb diese hier nicht mehr genauer dargestellt wird. Im Folgenden wird direkt auf den zweiten Schritt zur Analyse der Testkulturen eingegangen:

**Schritt 2: Auswahl der Ratings und Überprüfung der Inter-Rater-Übereinstimmung.**


Insbesondere bei den Itemeigenschaften, die die Art der für die Beantwortung eines Items notwendige Informationsaufnahme (die Variablen Informationsgewinn 1-3; siehe 2.2.4 sowie die Einschätzung der Authentizität einer Aufgabe zum Gegenstand haben scheint es bei den Ratern unterschiedliche Einschätzungen zu geben (Englisch: $r = .11$ bis $r = .38$; Deutsch: höchster Wert bei $r = .23$). Teilweise existieren zwischen verschiedenen Rater-Paarungen dabei sogar negative Korrelationen (Komplexität der Grammatik in Englisch; Authentizität des Inhalts in Deutsch). Eine mögliche Erklärung für die ungenügenden Inter-Rater-Übereinstimmungen ist, dass diese Variablen und die Unterschiede ihrer verschiedenen Abstufungen im Rahmen von
Schulungen nur schwer vermittelbar sind; ferner sind sie möglicherweise in der Theorie, und damit auch im hier verwendeten „Dutch Grid”, nicht ausreichend genau beschrieben, um die Unterschiede den Ratern näher zu bringen.

Bei den Antwort- bzw. Itemtypen sind die Zusammenhänge insgesamt etwas höher (zwischen \( r = .58 \) und \( r = 1.0 \)). Die Inter-Rater-Korrelation zwischen den Original-Ratings der Länder und dem Konsens-Urteil der neuen Rater ist nur hinsichtlich des Itemtyps „Multiple Matching” nicht signifikant. Die übrigen Korrelationen bei den Ausprägungen der Variablen „Itemtyp” bei den Englisch-Items bewegen sich mit \( r = .58 \) als niedrigstem (Multiple Choice) und \( r = 1.0 \) (Ordnen) als höchstem Wert im Bereich eines mittleren bis sehr hohen Zusammenhangs. Bei den Deutsch-Items zeigen sich hohe Übereinstimmungen bezüglich des Itemtyps (\( r = .59 \) bis \( r = .99 \)) zwischen den beiden neu rekrutierte Beurteilern; die Zusammenhänge zu den Original-Ratings sind hier allerdings etwas niedriger und bewegen sich zwischen nicht signifikant bezüglich der Itemtypen „Multiple Matching” und „Lückentext” und \( r = 1.0 \) bezüglich des Itemtyps „Richtig - Falsch”. Die Einschätzung der Item-bzw. Antworttypen scheint den Ratern insgesamt deutlich leichter gefallen zu sein, da die Inter-Rater-Übereinstimmungen dort größer sind.

Auch die Rater-Übereinstimmung bezüglich der Schwierigkeit von Grammatik und Vokabular ist insgesamt eher niedrig bis moderat und liegt für Englisch zwischen \( r = .17 \) (n.s.) und \( r = .54 \), für Deutsch zwischen nicht-signifikant und \( r = .31 \). Bei der Komplexität der Grammatik findet sich für die Deutsch-Items sogar ein negativer Zusammenhang zwischen den Original-Ratings und den neuen Ratern. Hier ist jedoch deutlich zu erkennen, dass dies auf einen der beiden neuen Rater zurückzuführen ist, der sich hier im Konsens-Urteil durchgesetzt hat (siehe Anhang). Die Übereinstimmung der neuen Rater der Englisch-Items ist insgesamt niedriger als die Übereinstimmung dieser Rater mit den Original-Ratings der Länder, bei den Deutsch-Ratern ist sie dagegen deutlich höher.

Das Problem der insgesamt relativ niedrigen Rater-Übereinstimmung bei der Beurteilung von Items hinsichtlich ihrer Itemeigenschaften lässt sich in der Literatur häufig finden (z.B. Alderson, 2000; Alderson et al., 2006)

So beschreiben die „Dutch Grid”-Autoren den Einigungsprozess, der zwischen den Ratern stattfand, als sehr schwierig. Mögliche Ursachen werden in der Ergebnisdiskussion (siehe Abschnitt 6) dargelegt. Basierend auf den Ergebnissen kann nun für die vorliegende Arbeit zwar festgestellt werden, dass die Inter-Rater-Übereinstimmung der neuen Beurteiler zu den Original-Urteilen insgesamt nicht besonders hoch ist; welche der Beurteilungen allerdings „richtiger” oder besser geeignet für die entsprechenden Items ist, kann daraus nicht geschlussfolgert werden.


Schritt 4: Häufigkeiten von Itemeigenschaften: Die Bildung nationaler Testprofile. In der vorliegenden Arbeit wurden für die Operationalisierung der Testkulturen nicht nur die in der EBAFLS Studie letztlich verwendeten Items, sondern sämtliche eingereichte Items analysiert. Dies erhöht die Anzahl der analysierbaren Items beträchtlich (Englisch insgesamt: 122 Items getestet, 204 Items eingereicht und analysiert; Deutsch insgesamt: 101 Items getestet, 234 Items eingereicht und analysiert) und sollte daher die Repräsentativität der Items für die jeweiligen
Länder so gut wie möglich gewährleisten. Dadurch wird die Gesamtheit der Testkulturen valider abgebildet. Es wird davon ausgegangen, dass die Items und die ihnen zugeordneten Itemeigenschaften für das jeweilige Land möglichst repräsentativ sind und daher auch die Testkultur eines Landes abbilden.


**Analyse der Items zur Erfassung englischsprachigen Leseverständnisses.** Im Folgenden werden die prozentualen Häufigkeiten des Vorkommens der unter Schritt 3 ausgewählten Itemmerkmale bei den in den verschiedenen Ländern konstruierten Items tabellarisch dargestellt. Nicht alle theoretisch möglichen Merkmalsausprägungen kommen auch tatsächlich bei den Items vor. Dies gilt beispielsweise für die Variable Itemtyp, und für die höchsten Ausprägungen der Variablen „Grammatische Strukturen“ und „Vokabular“ bei den Englisch-Items.

<table>
<thead>
<tr>
<th></th>
<th>Multiple Choice</th>
<th>Multiple Matching</th>
<th>Richtig-Falsch</th>
<th>Zitieren</th>
<th>Ordnen</th>
<th>Lückentext</th>
<th>Kurzantwort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>62.3</td>
<td>15.1</td>
<td>0</td>
<td>9.4</td>
<td>11.3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Deutschland</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Spanien</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ungarn</td>
<td>0</td>
<td>43.33</td>
<td>23.3</td>
<td>0</td>
<td>0</td>
<td>33.33</td>
<td>0</td>
</tr>
</tbody>
</table>

*Tabelle 5.5. Prozentuale Anteile der Ausprägungen der Variablen Itemtyp in den Teilnehmerländern (Englisch-Items)*

Bei der Betrachtung der Verteilung der verschiedenen Itemtypen (Tabelle 5.5) zeigt sich, dass fast alle Länder, bis auf Ungarn, Multiple-Choice-Items eingereicht haben. Für Deutschland und Spanien betrifft dies sogar 100% der Items. Auch hat Ungarn als einzige Nation eine große Anzahl an Multiple Matching Aufgaben eingebracht; daher sollten diese für die ungarischen Schüler einfacher sein. Insgesamt ist über die verschiedenen Itemeigenschaften hinweg ein gewisses Maß an Unterschiedlichkeit hinsichtlich der Ausprägungen der Variable „Itemtyp“ zwischen den Ländern beobachtbar. In Tabelle 5.6 sind die Häufigkeiten für die zwei Variablen, die unter den Bereich der für die Lösung notwendigen kognitiven Operationen und Informationsgewinn fallen,
Ergebnisse 5.1 Ergebnisse Fragenkomplex 1

<table>
<thead>
<tr>
<th>Informationsgewinn 1</th>
<th>Informationsgewinn 2</th>
<th>Authentizität Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erkennen Schlussfolgern</td>
<td>Explizit Implizit</td>
<td>angepasst/ vereinfacht authentisch</td>
</tr>
<tr>
<td>Frankreich 88.68</td>
<td>11.32</td>
<td>100</td>
</tr>
<tr>
<td>Deutschland 63.0</td>
<td>37.0</td>
<td>50</td>
</tr>
<tr>
<td>Spanien 62.5</td>
<td>37.5</td>
<td>62.5</td>
</tr>
<tr>
<td>Ungarn 66.7</td>
<td>33.3</td>
<td>100</td>
</tr>
</tbody>
</table>

Tabelle 5.6. Prozentuale Anteile der Ausprägungen der Variablen Informationsgewinn 1, Informationsgewinn 2 und Authentizität in den Teilnehmerländern (Englisch-Items)

sowie die Variable Authentizität des Texts, gemeinsam dargestellt.

Bezüglich des für die Lösung eines Items notwendigen Informationsgewinns „Erkennen“ vs. „Schlussfolgern“ heben sich hinsichtlich der Häufigkeit der beiden Ausprägungen der Variablen nur die französischen Items von den der drei anderen Länder ab. Das weist hier darauf hin, dass bei der Bearbeitung französischer Items häufiger als in den drei anderen Ländern eine Information im Text nur erkannt werden muss, während bei den Items der drei anderen Ländern häufiger als bei französischen Items Schlussfolgerungen vonnöten sind.

Hinsichtlich der Variablen „Informationsgewinn 2“, die sich darauf bezieht, ob die für die Lösung notwendige Information im Text explizit oder implizit enthalten ist, unterscheiden sich die Items der vier Länder teilweise. Bei den französischen und ungarischen Items sind die notwendigen Informationen in sämtlichen Items explizit gegeben, bei spanischen Items in 62.5% der Fälle, und bei deutschen Items in 50% der Fälle.

Im Hinblick auf die Authentizität des Textes unterscheiden sich deutsche und spanische Items nur minimal. Alle deutschen Items sind eher vereinfacht/pädagogisch, bei den spanischen Items sind dies 96.6%, hingegen bei französischen Items nur 20.8% und keines der ungarischen. Hinsichtlich des Grades der Abstraktheit des Inhalts zeigt sich in Tabelle 5.7, dass alle deutschen und französischen Items der mittleren Kategorie der Variable, nämlich „hauptsächlich konkret“, zugeordnet wurden, während dies bei den ungarischen Items nur auf 66.7% bei den spanischen Items lediglich auf 9.4% zutrifft. Als „teilweise abstrakt“ wurden ausschließlich ungarische Items eingestuft.

Bezüglich der Schwierigkeit des Vokabulars (Tabelle 5.8) zeigt sich, dass in der einfachsten Kategorie „ausschließlich häufig“ mit 45.3% am häufigsten französische Items eingestuft wurden. Danach folgen spanische respektive ungarische Items mit 15.6% bzw. 16.7%. In der Kategorie „teilweise häufig“ werden am meisten deutsche Items eingeordnet, gefolgt von spanischen, französischen und ungarischen. In die Kategorie „teilweise erweitert / selten“ werden mit 60% am
Ergebnisse 5.1 Ergebnisse Fragenkomplex 1

Tabelle 5.7. Prozentuale Anteile der Ausprägungen der Variable Abstraktheit des Inhalts in den Teilnehmerländern (Englisch-Items)

<table>
<thead>
<tr>
<th></th>
<th>ausschließlich konkret</th>
<th>hauptsächlich konkret</th>
<th>teilweise abstrakt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Deutschland</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Spanien</td>
<td>90.6</td>
<td>9.4</td>
<td>0</td>
</tr>
<tr>
<td>Ungarn</td>
<td>16.7</td>
<td>66.7</td>
<td>26.7</td>
</tr>
</tbody>
</table>

Tabelle 5.8. Prozentuale Anteile der Ausprägungen der Variable Vokabular in den Teilnehmerländern (Englisch-Items)

<table>
<thead>
<tr>
<th></th>
<th>ausschließlich häufig/einfach</th>
<th>hauptsächlich häufig/einfach</th>
<th>teilweise erweitert/selten</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>45.3</td>
<td>34.0</td>
<td>20.8</td>
</tr>
<tr>
<td>Deutschland</td>
<td>0</td>
<td>78.3</td>
<td>21.7</td>
</tr>
<tr>
<td>Spanien</td>
<td>15.6</td>
<td>46.9</td>
<td>46.6</td>
</tr>
<tr>
<td>Ungarn</td>
<td>16.7</td>
<td>23.3</td>
<td>60</td>
</tr>
</tbody>
</table>

häufigsten ungarischen Items eingeordnet, am zweit häufigsten spanische Items mit 46.6%, darauf folgen deutsche und französischen Items.

Tabelle 5.9. Prozentuale Anteile der Ausprägungen der Variable Grammatische Strukturen in den Teilnehmerländern (Englisch-Items)

<table>
<thead>
<tr>
<th></th>
<th>ausschließlich einfach</th>
<th>hauptsächlich einfach</th>
<th>teilweise komplex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>13.2</td>
<td>28.3</td>
<td>58.5</td>
</tr>
<tr>
<td>Deutschland</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Spanien</td>
<td>0</td>
<td>75.0</td>
<td>25.0</td>
</tr>
<tr>
<td>Ungarn</td>
<td>16.7</td>
<td>26.7</td>
<td>56.7</td>
</tr>
</tbody>
</table>

Bezüglich der Komplexität der grammatischen Strukturen (Tabelle 5.9) werden in die einfachste Kategorie „ausschließlich einfache grammatische Strukturen“ nur einige wenige Items aus Ungarn und Frankreich eingeordnet. In der zweiten Kategorie „hauptsächlich einfache grammatische Strukturen“ fallten die meisten der spanischen Items (75%), gefolgt von französischen und ungarischen. In der dritten Kategorie „teilweise komplexe grammatische Strukturen“ werden alle der deutschen Items eingeordnet, gefolgt von französischen, ungarischen und spanischen Items.

Insgesamt zeigt sich, dass sich die Items der Länder hinsichtlich ihrer Einordnung in die unterschiedlichen Ausprägungen der schwierigkeitsdeterminierenden Merkmale der Items unterscheiden. Dies lässt auf unterschiedliche Testkulturen und somit auf unterschiedliche Stärken und
Ergebnisse 5.1 Ergebnisse Fragenkomplex 1

Schwächen der Schülerinnen und Schüler im Umgang mit Testitems aus jeweils anderen Ländern schließen.


Da später zur Beantwortung der Fragestellungen im Fragenkomplex 3 die Länder jeweils paarweise gegenübergestellt werden und so für jedes Länderpaar eine Hypothese bezüglich der relativen Stärken und Schwächen aufgestellt werden soll, wird in einem nächsten Schritt eine Rangreihe der Häufigkeiten der Itemeigenschaften bei den aus den unterschiedlichen Ländern stammenden Items gebildet. Unterschiede, die hier mit einem „größer“-Zeichen dargestellt werden, sind mindestens auf einem 5%-Niveau signifikant; Sind die Unterschiede zwischen zwei Ländern nicht signifikant, wird dies mit einem Gleichheitszeichen dargestellt. Daraus ergeben sich über die Länder hinweg für die Englisch-Items Häufigkeits-Rangreihen der Itemmerkmale (Tabelle 5.10)

Die Aussage dieser Rangreihen ist wie folgt zu verstehen: Beispielsweise ist bezüglich der Itemeigenschaft „Grammatische Strukturen“ der prozentuale Anteil an Items, welche die Ausprägung „Grammatik komplex“ besitzen, in Deutschland am größten, in Spanien hingegen am niedrigsten. Der Unterschied zwischen Deutschland und Frankreich sowie zwischen Ungarn und Spanien ist ferner mindestens auf einem 5%-Niveau signifikant, auf den Unterschied zwischen Frankreich und Ungarn hingegen trifft dies nicht zu.

Demzufolge sollten Items mit hauptsächlich komplexer Grammatik für deutsche Schüler leichter sein als für spanische. Es zeigt sich hier, dass unterschiedliche Profile in den Häufigkeiten von Itemeigenschaften bzw. deren Ausprägungen bei den verschiedenen Ländern zu beobachten sind. Es kann also bei der Messung englischsprachiger Lesekompetenz von unterschiedlichen Testkulturen und somit auch von unterschiedlichen zu erwartenden Stärken und Schwächen der Länder ausgegangen werden. Aufgrund dieser Ergebnisse lassen sich für den Fragenkomplex 3
<table>
<thead>
<tr>
<th>Itemeigenschaft</th>
<th>Ausprägung</th>
<th>Rangreihe der Häufigkeiten</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itemtyp</td>
<td>Multiple Choice</td>
<td>F&gt;U=SP&gt;D</td>
</tr>
<tr>
<td></td>
<td>Multiple Matching</td>
<td>D&gt;SP=F&gt;U</td>
</tr>
<tr>
<td></td>
<td>Richtig-Falsch</td>
<td>U=SP&gt;F=D</td>
</tr>
<tr>
<td></td>
<td>Zitieren</td>
<td>F&gt;U=U=D=SP</td>
</tr>
<tr>
<td></td>
<td>Ordnen</td>
<td>F&gt;U=U=D=SP</td>
</tr>
<tr>
<td></td>
<td>Lückentext</td>
<td>U&gt;F=D=SP</td>
</tr>
<tr>
<td></td>
<td>Kurzantwort</td>
<td></td>
</tr>
<tr>
<td>Informationsgewinn 1</td>
<td>Schlussfolgern/Erkennen</td>
<td>SP=D=U=F</td>
</tr>
<tr>
<td>Informationsgewinn 2</td>
<td>Implizit/Explizit</td>
<td>D=SP&gt;F=U</td>
</tr>
<tr>
<td>Authentizität des Texts</td>
<td>Text authentisch/angepasst= vereinfacht</td>
<td>U&gt;F=D=SP</td>
</tr>
<tr>
<td>Abstraktheit des Inhalts</td>
<td>Ausschließlich konkret</td>
<td>SP=U&gt;D=F</td>
</tr>
<tr>
<td></td>
<td>Hauptsächlich konkret</td>
<td>F=D&gt;U=Sp</td>
</tr>
<tr>
<td></td>
<td>Ziemlich abstrakt</td>
<td>U=D=F=SP</td>
</tr>
<tr>
<td>Vokabular</td>
<td>Ausschließlich häufig/einfach</td>
<td>F&gt;U=Sp&gt;D</td>
</tr>
<tr>
<td></td>
<td>Hauptsächlich häufig/einfach</td>
<td>D=Sp=F&gt;U</td>
</tr>
<tr>
<td></td>
<td>Teilweise erweitert/selten</td>
<td>U=Sp=F=D</td>
</tr>
<tr>
<td>Grammatik</td>
<td>Ausschließlich einfache Strukturen</td>
<td>U=F=D=Sp</td>
</tr>
<tr>
<td></td>
<td>Hauptsächlich einfache Strukturen</td>
<td>Sp=F=U=D</td>
</tr>
<tr>
<td></td>
<td>Teilweise komplexe Strukturen</td>
<td>D=F=U=Sp</td>
</tr>
</tbody>
</table>

Tabelle 5.10. Prozentuale Anteile der Ausprägungen der Variable Grammatische Strukturen in den Teilnehmerländern (Englisch-Items)

Einzelhypothesen über die zu erwartenden relativen Stärken und Schwächen der Länder aufstellen (Tabelle 5.11).

Die Tabelle ist wie folgt zu lesen: Für den Itemtyp „Multiple Choice“ weisen Items zur Messung fremdsprachlichen Leseverständnisses aus Frankreich signifikant häufiger diese Merkmalsausprägung auf, als dies bei den aus Deutschland stammenden Items der Fall ist. Daher ist zu erwarten, dass der Umstand, dass ein bestimmtes Item ein Multiple-Choice-Format aufweist, bewirken sollte, dass dieses für die französischen Schüler, verglichen mit den deutschen Schülern, leichter zu beantworten ist. Obgleich aufgrund der vorher aufgestellten Hypothesen auch Annahmen darüber gemacht werden könnten, dass sich zwei Länder hinsichtlich eines Merkmals nicht unterscheiden und daher ein Zusammenhang nicht signifikant sein sollte, (d.h. die Hypothese aufgestellt werden könnte, dass die H0 nicht zurückgewiesen wird), besteht dabei doch das Problem, dass hier nicht klar abzugrenzen ist, ob eine Korrelation tatsächlich aufgrund des fehlenden Unterschieds hinsichtlich der Stärken und Schwächen (ein bestimmtes testkulturelles Merkmal betreffend) nicht signifikant wird, oder ob dies auf andere Faktoren wie beispielsweise Varianzeinschränkung zurückzuführen ist. Hier ist eine Konfundierung von Ursachen möglich.
Ergebnisse 5.1 Ergebnisse Fragenkomplex 1

Tabelle 5.11. Hypothesen der zu erwartenden Stärken und Schwächen der Länder bei Englisch-Items

Daher lassen sich ausschließlich signifikante Korrelationen eindeutig interpretieren.

Analyse der Items zur Messung deutschsprachigen Leseverständnisses. Auch für die Deutsch-Items wurden die oben bereits beschriebenen Analysen durchgeführt. Diese werden im Folgenden berichtet. Wie schon bei den Englisch-Items werden in einem letzten Schritt dann die Rangfolgen der Häufigkeiten und somit die Testkulturen bzw. die zu erwartenden Stärken und Schwächen dargestellt.

Tabelle 5.12. Prozentuale Anteile der Ausprägungen der Variablen Itemtyp in den Teilnehmerländern (Deutsch-Items)

Bezüglich der Ausprägungen des Merkmals „Itemtyp“ (Tabelle 5.12) zeigt sich, dass sich auch die in unterschiedlichen Ländern konstruierten Deutsch-Items unterscheiden. So weisen die französischen Items beispielsweise den größten Anteil an Multiple-Choice-Items auf, hingegen
sind die niederländischen und schwedischen Items die einzigen, die das Format „Short Answer“ aufweisen.

<table>
<thead>
<tr>
<th>Informationsgewinn 1</th>
<th>Informationsgewinn 2</th>
<th>Informationsgewinn 3</th>
<th>Authentizität Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erkennen</td>
<td>Schlussfolgern</td>
<td>Explizit</td>
<td>Implizit</td>
</tr>
<tr>
<td>Frankreich</td>
<td>76.3</td>
<td>23.7</td>
<td>100</td>
</tr>
<tr>
<td>Ungarn</td>
<td>41.2</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Niederlande</td>
<td>54.8</td>
<td>45.2</td>
<td>100</td>
</tr>
<tr>
<td>Schweden</td>
<td>22.0</td>
<td>78.0</td>
<td>22</td>
</tr>
</tbody>
</table>

Tabelle 5.13. Prozentuale Anteile der Ausprägungen der Variablen Informationsgewinn 1, Informationsgewinn 2 und Authentizität in den Teilnehmerländern (Deutsch-Items)

Wie auch schon bei den Englisch-Items werden die drei Variablen, die sich auf den zur Lösung des Items notwendigen Informationsgewinn beziehen, sowie die Variable „Authentizität“ hier gemeinsam dargestellt (Tabelle 5.13). In Bezug auf die Variable „Informationsgewinn 1“, die sich darauf bezieht, ob die zur Lösung notwendige Information im Text lediglich erkannt bzw. gefunden werden muss oder ob zur Lösung des Items Schlussfolgerungen gezogen werden müssen, ist die Ausprägung „Schlussfolgerungen“ bei schwedischen Items am häufigsten, gefolgt von den niederländischen und französischen Items. Sowohl bei französischen als auch ungarischen und niederländischen Items sind die zur Lösung notwendigen Informationen explizit gegeben. Bei schwedischen Items hingegen liegt diese Information in 78% der Fälle implizit vor.

Hinsichtlich der Authentizität des Textes wurden in Frankreich und Ungarn sämtliche Items als authentisch eingestuft, in den Niederlanden nur 85.7%. In Schweden hingegen wurden dagegen 70.7% der Items als vereinfacht bzw. pädagogisch eingeordnet.

<table>
<thead>
<tr>
<th>ausschließlich konkret</th>
<th>hauptsächlich konkret</th>
<th>teilweise abstrakt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frankreich</td>
<td>73.7</td>
<td>26.3</td>
</tr>
<tr>
<td>Ungarn</td>
<td>14.7</td>
<td>58.8</td>
</tr>
<tr>
<td>Niederlande</td>
<td>35.7</td>
<td>64.3</td>
</tr>
<tr>
<td>Schweden</td>
<td>0</td>
<td>82.9</td>
</tr>
</tbody>
</table>

Tabelle 5.14. Prozentuale Anteile der Ausprägungen der Variable Abstraktheit des Inhalts in den Teilnehmerländern (Deutsch-Items)

Bezüglich der Abstraktheit des Inhalts der Items (Tabelle 5.14) zeigt sich, dass die meisten französischen Items der Kategorie „ausschließlich konkret“ zugehörig sind, gefolgt von den niederländischen (35.7%) und ungarischen (14.7%) Items. Bezüglich der Kategorie „hauptsächlich konkret“ wurden 82.9% der schwedischen Items so bewertet, 64.3% der niederländischen, 58.8%
der ungarischen und 26.3% der französischen. In die Kategorie „hauptsächlich abstrakt“ fallen nur Items aus Ungarn (26.5%) und Schweden (17.1%). Tabelle 5.15 zeigt die Variable Vokabular.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frankreich</th>
<th>Ungarn</th>
<th>Niederlande</th>
<th>Schweden</th>
</tr>
</thead>
<tbody>
<tr>
<td>ausschließlich häufig/einfach</td>
<td>0</td>
<td>68.4</td>
<td>4.8</td>
<td>95.2</td>
</tr>
<tr>
<td>hauptsächlich häufig/einfach</td>
<td>0</td>
<td>58.8</td>
<td>9.8</td>
<td>90.2</td>
</tr>
<tr>
<td>teilweise erweitert/selten</td>
<td>31.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>erweitert/selten</td>
<td>0</td>
<td>26.5</td>
<td>26.5</td>
<td>0</td>
</tr>
</tbody>
</table>

Tabelle 5.15. Prozentuale Anteile der Ausprägungen der Variable Vokabular in den Teilnehmerländern (Deutsch-Items)

In keinem der Länder fallen Items in die Kategorie „ausschließlich häufig“. Als „hauptsächlich häufig“ werden 68.8% der französischen Items bewertet, gefolgt von ungarischen (14.7%) und niederländischen (4.8%). Der Kategorie „hauptsächlich erweitert / selten“ wurden mit 95.2% am meisten niederländische Items zugeordnet, gefolgt von ungarischen (58.8%), französischen (31.6%) und schwedischen (9.8%). In die Kategorie „erweitert / selten“ fallen die meisten der schwedischen (90.2%) und 26.5% der ungarischen Items.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Frankreich</th>
<th>Ungarn</th>
<th>Niederlande</th>
<th>Schweden</th>
</tr>
</thead>
<tbody>
<tr>
<td>ausschließlich einfach</td>
<td>97.4</td>
<td>0</td>
<td>58.8</td>
<td>73.2</td>
</tr>
<tr>
<td>hauptsächlich einfach</td>
<td>0</td>
<td>28.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>teilweise komplexe Strukturen</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>komplexe Strukturen</td>
<td>2.6</td>
<td>26.5</td>
<td>26.8</td>
<td>0</td>
</tr>
</tbody>
</table>

Tabelle 5.16. Prozentuale Anteile der Ausprägungen der Variable Grammatische Strukturen in den Teilnehmerländern (Deutsch-Items)

Hinsichtlich der Komplexität grammatischer Strukturen (Tabelle 5.16) sind 97.4% der französischen Items der Kategorie „ausschließlich einfach“ zugeordnet, gefolgt von den niederländischen (71.4%) und ungarischen (14.7%) Items. In die Kategorie „hauptsächlich einfach“ wurden 28.6% der niederländischen Items eingeordnet, keines von den jeweiligen anderen Ländern. Bezüglich der Kategorie „teilweise komplexe Strukturen“ wurden 73.2% der schwedischen Items und 58.8% der ungarischen Items dort eingeordnet. 26.8% der schwedischen, 26.5% der ungarischen und 2.6% der französischen Items fallen unter die Kategorie „komplexe Strukturen“.

Insgesamt ist auch für die Deutsch-Items Varianz hinsichtlich des Vorkommens der verschiede-
nen Itemmerkmale beobachtbar. Daraus ergeben sich über die Länder hinweg für die Deutsch-Items Häufigkeits-Rangreihen der Itemmerkmale (Tabelle 5.17).

<table>
<thead>
<tr>
<th>Itemeigenschaft</th>
<th>Ausprägung</th>
<th>Rangreihe der Häufigkeiten</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itemtyp</td>
<td>Multiple Choice</td>
<td>F&gt;NL&gt;U=SW</td>
</tr>
<tr>
<td></td>
<td>Banked Multiple Choice</td>
<td>SW&gt;F=U=NL</td>
</tr>
<tr>
<td></td>
<td>Multiple Matching</td>
<td>SW&gt;U=F=NL</td>
</tr>
<tr>
<td></td>
<td>Richtig-Falsch</td>
<td>U&gt;F=NL=SW</td>
</tr>
<tr>
<td></td>
<td>Zitieren</td>
<td>F=U=NL=SW</td>
</tr>
<tr>
<td></td>
<td>Lückentext</td>
<td>NL&gt;SW&gt;F=U</td>
</tr>
<tr>
<td></td>
<td>Kurzantwort</td>
<td>NL&gt;SW&gt;F=U</td>
</tr>
<tr>
<td>Informationsgewinn 1</td>
<td>Schlussfolgern vs.Erkennen/Evaluieren</td>
<td>SW&gt;NL&gt;F&gt;U</td>
</tr>
<tr>
<td>Informationsgewinn 2</td>
<td>Implizit/Explizit</td>
<td>U=F&gt;NL&gt;SW</td>
</tr>
<tr>
<td>Informationsgewinn 3</td>
<td>Hauptidee/Detail</td>
<td>U=SW&gt;NL&gt;F</td>
</tr>
<tr>
<td>Authentizität des Texts</td>
<td>Text authentisch vs. angepasst/vereinfacht</td>
<td>F=U=NL&gt;SW</td>
</tr>
<tr>
<td>Abstraktheit des Inhalts</td>
<td>Ausschließlich konkret</td>
<td>F&gt;NL&gt;U&gt;SW</td>
</tr>
<tr>
<td></td>
<td>Hauptsächlich konkret</td>
<td>SW&gt;NL&gt;U&gt;F</td>
</tr>
<tr>
<td></td>
<td>Ziemlich abstrakt</td>
<td>U=NL=F=SW</td>
</tr>
<tr>
<td>Vokabular</td>
<td>Ausschließlich häufig/einfach</td>
<td>F=U=NL=SW</td>
</tr>
<tr>
<td></td>
<td>Hauptsächlich häufig/einfach</td>
<td>F&gt;U&gt;NL&gt;SW</td>
</tr>
<tr>
<td></td>
<td>Teilweise erweitert/selten</td>
<td>NL&gt;U=F&gt;SW</td>
</tr>
<tr>
<td></td>
<td>Erweitert/selten</td>
<td>SW&gt;U=F&gt;NL</td>
</tr>
<tr>
<td>Grammatik</td>
<td>Ausschließlich einfache Strukturen</td>
<td>F&gt;NL&gt;U&gt;SW</td>
</tr>
<tr>
<td></td>
<td>Hauptsächlich einfache Strukturen</td>
<td>NL&gt;F=U=SW</td>
</tr>
<tr>
<td></td>
<td>Teilweise komplexe Strukturen</td>
<td>SW&gt;U=F=NL</td>
</tr>
<tr>
<td></td>
<td>Komplexe Strukturen</td>
<td>SW&gt;U=F=NL</td>
</tr>
</tbody>
</table>

Tabelle 5.17. Prozentuale Anteile der Ausprägungen der Variable Grammatische Strukturen in den Teilnehmerländern (Deutsch-Items)

Auch für die Deutsch-Items zeigt sich, dass sich unterschiedliche Profile hinsichtlich der Häufigkeiten der Item-Anforderungsmerkmale bei den aus unterschiedlichen Ländern stammenden Items feststellen lassen. Wie schon bei den Englisch-Items sind die dargestellten Unterschiede mit Cohen’s h (Cohen, 1988) sowie dem darauf basierenden Signifikanztest überprüft worden. Die hier dargestellten Unterschiede sind daher gleichfalls mindestens auf einem 5%-Niveau signifikant. Die dazugehörigen Tests und Ergebnisse sind im Anhang einzusehen.

Basierend auf diesen Ergebnissen der Analysen zur Existenz differentieller Testkulturen lassen sich Einzelhypothesen für die Deutsch-Items aufstellen (Tabelle 5.18).
Tabelle 5.18. Hypothesen der zu erwartenden Stärken und Schwächen der Länder bei Deutsch-Items

Die hier aufgestellten Hypothesen sind wie folgt zu lesen: Beispielsweise ist bei den Ländern Frankreich und den Niederlanden zu erwarten, dass die Tatsache, dass ein Item ein Multiple-Choice-Format besitzt, dieses für die französischen Schüler, verglichen mit den niederländischen Schülern, vereinfachen sollte. Auch hier werden später ausschließlich signifikante Zusammenhänge bezüglich der Hypothesen, die eine Unterschiedlichkeit von Testkulturen annehmen, interpretiert.
Die hinsichtlich der Unterschiedlichkeit der Testkulturen eingangs aufgestellte Hypothese lautete:

**Hypothese 1c: Es lassen sich durch eine Analyse von Items aus unterschiedlichen Ländern unterschiedliche Testkulturen feststellen.**

Die in Hypothese 1c getroffene Annahme kann insgesamt beibehalten werden: Es lassen sich sowohl bezüglich der Deutsch- als auch der Englisch-Items signifikante Unterschiede hinsichtlich des Vorkommens von Item-Anforderungsmerkmalen bei den Items unterschiedlicher Länder feststellen. Dadurch lassen sich unterschiedliche Schwerpunkte von Item-Anforderungsmerkmalen bei den Items unterschiedlicher Länder ableiten. Diese Voraussetzung, das heißt die Existenz unterschiedlicher Testkulturen in den unterschiedlichen Ländern, wird somit sowohl für die englische als auch die deutsche Sprache als erfüllt angesehen.

Mit Hilfe dieser Rangreihen können, wie erwartet, einzelne Hypothesen darüber abgeleitet werden, welche Itemeigenschaft für die Testkultur welchen Landes in welcher Ausprägung eine besondere Rolle spielt, das heißt positiven oder negativen Einfluss auf die Itemschwierigkeit haben sollte. Das bedeutet, es können **konkrete Annahmen bezüglich der länderspezifischen Stärken und Schwächen getroffen werden.** Auch hier sind die dargestellten Unterschiede allesamt signifikant.

**Zusammenfassung und Fazit.**

Unter 5.1 wurden die Voraussetzungen für die weiteren Analysen unter 5.2 und 5.3 überprüft. Es wurde zunächst festgestellt, dass die Items innerhalb der Länder dem Rasch-Modell entsprechen. Dann wurde dargestellt, dass ein großer Teil der Items bei paarweisen Analysen Differentielle Item Funktionen aufweisen. Zum dritten wurde aufgezeigt, dass sich die Länder hinsichtlich ihrer Testkulturen, dargestellt durch die Häufigkeit des Vorkommens von Anforderungsmerkmalen bei den Items eines Landes, teilweise signifikant voneinander unterscheiden und sich auf diese Weise a-priori-Hypothesen hinsichtlich der Zusammenhänge von Testkulturen und Differentiellen Item Funktionen aufstellen lassen. Ferner wurden die für die in 5.2 und 5.3 durchzuführenden Analysen notwendigen Parameter, nämlich Item-Schwierigkeitsparameter und DIF-Parameter, berechnet. Es konnten alle Hypothesen weitestgehend beibehalten werden, und die Items können für die Analysen in den Fragenkomplexen 2 und 3 verwendet werden.
5.2. Ergebnisse Fragenkomplex 2: Erklärung der Itemschwierigkeiten innerhalb der Länder

Im Folgenden werden die Zusammenhänge zwischen Itemschwierigkeitsparametern und Itemcharakteristika innerhalb der Länder dargestellt, zunächst anhand von Einzelkorrelationen und danach anhand multipler Regressionsanalysen.

5.2.1. Zu Frage 2a

Frage 2a: Weisen die kognitiv-linguistischen Anforderungsmerkmale der Items korrelative Zusammenhänge zu den Itemschwierigkeiten innerhalb der Länder auf?


Insgesamt wird hier davon ausgegangen, dass signifikante Korrelationen zwischen den Itemschwierigkeiten innerhalb der Länder und den kognitiv-linguistischen Item-Angabeigenschaften existieren. Im Folgenden werden die korrelativen Zusammenhänge zwischen den Itemschwierigkeiten und den Item-Anforderungsmerkmalen berichtet. Dies geschieht zunächst für die Englisch- und danach für die Deutsch-Items.

Die Ergebnisse in Tabelle 5.19 sind wie folgt zu lesen: In der linken Spalte sind jeweils die Item-Anforderungsmerkmale inklusive ihrer verschiedenen Abstufungen abgetragen. In den restlichen Spalten die Korrelationen (Pearson Produkt-Moment-Korrelation r) mit den Itemschwierigkeitsparametern. Die hier dargestellten, signifikanten Ergebnisse weisen eine auf einem 5- bzw. 1-Prozent-Niveau signifikante Korrelation zwischen der Itemschwierigkeit innerhalb des jeweili-
<table>
<thead>
<tr>
<th>Itemeigenschaft</th>
<th>Frankreich</th>
<th>Deutschland</th>
<th>Ungarn</th>
<th>Spanien</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itemtyp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>-.241**</td>
<td>-.300**</td>
<td>-.229*</td>
<td>-.213*</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>.02</td>
<td>.004</td>
<td>.02</td>
<td>.043</td>
</tr>
<tr>
<td>Ordnen</td>
<td>-.077</td>
<td>-.073</td>
<td>-.021</td>
<td>-.057</td>
</tr>
<tr>
<td>Zitieren</td>
<td>.023</td>
<td>.229*</td>
<td>.238**</td>
<td>-.052</td>
</tr>
<tr>
<td>Lückentext</td>
<td>.431**</td>
<td>.406**</td>
<td>.213*</td>
<td>.392**</td>
</tr>
</tbody>
</table>

| Informationsgewinn 1        |            |             |        |        |
| Schlussfolgern/Erkennen     | .212*      | .165        | .105   | .141   |

| Informationsgewinn 2        |            |             |        |        |
| Implizit (vs. Explizit)     | -.082      | -.137       | -.042  | -.080  |

| Informationsgewinn 3        |            |             |        |        |
| Detail (vs. Hauptidee)      | Durch Mehrfachennungen in Englisch nicht auswertbar ('Detail' & 'main': obgleich gegensätzlich, konnte beides angewählt werden) |

| Authentizität               |            |             |        |        |
| Authentisch (vs. angepasst/ vereinfacht) | .106       | .248**      | .132   | .112   |

| Abstraktheit des Inhalts    |            |             |        |        |
| aussließlich konkret        | -.136      | -.224*      | -.192* | -.188* |
| hauptsächlich konkret       | -.134      | -.038       | .063   | .063   |
| teilweise abstrakt          | .431**     | .406**      | .213*  | .392** |

| Vokabular                   |            |             |        |        |
| aussließlich häufig/einfach | -.264**    | -.236**     | -.300**| -.221* |
| hauptsächlich häufig/einfach| .046       | -.025       | .138   | .028   |
| teilweise erweitert/selten  | .199*      | .254**      | .128   | .179*  |

| Grammatik                   |            |             |        |        |
| aussließlich einfache Strukturen | .188*      | .184*       | .137   | .184*  |
| hauptsächlich einfache Strukturen | -.168     | -.149       | -.127  | -.260**|
| teilweise komplexere Strukturen | .039       | .025        | .032   | .125   |

N=122; * p ≤ 0.05, α = 5%; ** p ≤ 0.01, α = 1%

Tabelle 5.19. Korrelationen zwischen kognitiv-linguistischen Anforderungsmerkmalen und Itemschwierigkeiten innerhalb der Länder bei Englisch-Items

gegen Landes und der entsprechenden Itemeigenschaft auf. Signifikante Korrelationen weisen auf die Richtigkeit der Annahme hin, dass die in dieser Arbeit gewählten Item-Anforderungsmerkmale innerhalb der Länder einen Zusammenhang zur Itemschwierigkeit aufweisen. Die signifikanten Korrelationen bewegen sich zwischen r = -.300 und r = .431. Eine negative Korrelation bedeutet hier, dass ein Zusammenhang zwischen der Anwesenheit einer Merkmalsausprägung der Item-Anforderungsmerkmale und der Tatsache, dass ein Item leichter ist, besteht. Umgekehrt
bedeutet dann ein positiver Korrelationskoeffizient, dass ein Zusammenhang zwischen einem Merkmal und der Tatsache, dass ein Item für die Schüler schwieriger ist, besteht. Dies ist durch die für die Itemschwierigkeit verwendete Logit-Skala bedingt: Je kleiner dort der Wert ist, desto geringer ist die Itemschwierigkeit. Insgesamt zeigt sich, dass die Korrelationen innerhalb der Länder in dieselben Richtungen weisen, wenn auch teilweise in unterschiedlicher Höhe. So weist beispielsweise der in allen Ländern zu beobachtende positive Zusammenhang zwischen der Itemschwierigkeit und dem Merkmal „hauptsächlich abstrakt“ darauf hin, dass diese Items in allen Ländern mit einer eher größeren Itemschwierigkeit einhergehen. Ferner weisen die Richtungen der Zusammenhänge innerhalb der Länder größtenteils in die theoretisch (wie beispielsweise durch den „Dutch Grid“) erwartete Richtung.

Das heißt, der Umstand, dass ein Item nach Experteneinschätzung eher einfaches Vokabular enthält, hängt auch eher damit zusammen, dass ein Item eine niedrigere Itemschwierigkeit besitzt, wohingegen Items mit tendenziell seltenem, schwierigem Vokabular zugleich auch eine eher größere Itemschwierigkeit aufweisen. Ausnahmen finden sich hinsichtlich der grammatischen Strukturen: Hier zeigt sich beispielsweise in Spanien, dass dort ein signifikanter Zusammenhang zwischen einfachen grammatischen Strukturen und der Tatsache, dass ein Item eher eine höhere Itemschwierigkeit aufweist, besteht Darüber wird in Abschnitt 6 nochmals eingegangen. Einen Zusammenhang zur Itemschwierigkeit innerhalb der Länder weisen vor allem die Merkmale Itemtyp, Abstraktheit des Inhalts, Schwierigkeit des Vokabulars sowie die Komplexität der grammatischen Strukturen auf. Kaum Zusammenhänge zeigen sich unerwarteter Weise zwischen der geschätzten Itemschwierigkeit und denjenigen Item-Anforderungsmerkmalen, welche die Art des zur korrekten Beantwortung des Items notwendigen Informationsgewinns abbilden (Informationsgewinn 1-3). Dies ist möglicherweise darauf zurückzuführen, dass die Experten im Hinblick auf die Einschätzung von Items bezüglich dieser Itemmerkmale insgesamt größere Schwierigkeiten zu haben scheinen (siehe auch 5.1.3) und daher der Zusammenhang hier geringer ausfällt.

Bezüglich der Interpretation der Korrelationen zwischen den Anforderungsmerkmalen der Deutsch-Items und der geschätzten Itemschwierigkeit (Tabelle 5.20) innerhalb der Länder gilt Ähnliches wie für die oben beschriebenen Ergebnisse bezüglich der Englisch-Items. Die Höhe der signifikanten Korrelationen bewegt sich zwischen $r = -.44$ und $r = .51$. Ähnlich wie bei den Englisch-Items zeigt sich hier, dass der Zusammenhang innerhalb der Länder größtenteils der im „Dutch Grid“ erfolgten Abstufung der Itemmerkmale folgt: So weist ein schwieriges Vokabular einen Zusammenhang zu schwereren Items, leichteres Vokabular einen Zusammenhang zu einer
### Tabelle 5.20. Korrelationen zwischen kognitiv-linguistischen Anforderungsmerkmalen und Itemschwierigkeiten innerhalb der Länder bei Deutsch-Items

<table>
<thead>
<tr>
<th>Itemeigenschaft</th>
<th>Frankreich</th>
<th>Niederlande</th>
<th>Ungarn</th>
<th>Schweden</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Itemtyp</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>-.110</td>
<td>-.152</td>
<td>.078</td>
<td>.023</td>
</tr>
<tr>
<td>Banked Multiple Choice</td>
<td>.036</td>
<td>.060</td>
<td>-.014</td>
<td>.077</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>-.243**</td>
<td>-.167</td>
<td>-.272***</td>
<td>-.267***</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>.162</td>
<td>.168</td>
<td>-.077</td>
<td>.012</td>
</tr>
<tr>
<td>Kurzantwort</td>
<td>.247**</td>
<td>.187</td>
<td>.298***</td>
<td>.223**</td>
</tr>
<tr>
<td>Lückentext</td>
<td>.020</td>
<td>.060</td>
<td>-.032</td>
<td>-.024</td>
</tr>
<tr>
<td><strong>Informationsgewinn 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schlussfolgern/Erkennen</td>
<td>.343***</td>
<td>.383***</td>
<td>.363***</td>
<td>.264**</td>
</tr>
<tr>
<td><strong>Informationsgewinn 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implizit (vs. Explizit)</td>
<td>.441***</td>
<td>.434***</td>
<td>.332***</td>
<td>.276**</td>
</tr>
<tr>
<td><strong>Informationsgewinn 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detail (vs. Hauptidee)</td>
<td>-.087</td>
<td>-.252**</td>
<td>.022</td>
<td>-.002</td>
</tr>
<tr>
<td><strong>Authentizität</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentisch (vs. angepasst/ vereinfacht)</td>
<td>-.239**</td>
<td>-.150</td>
<td>-.209*</td>
<td>-.138</td>
</tr>
<tr>
<td><strong>Abstraktheit des Inhalts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ausschließlich konkret</td>
<td>-.362***</td>
<td>-.393***</td>
<td>-.444***</td>
<td>-.284***</td>
</tr>
<tr>
<td>hauptsächlich konkret</td>
<td>.272**</td>
<td>.270**</td>
<td>.334***</td>
<td>.204*</td>
</tr>
<tr>
<td>teilweise abstrakt</td>
<td>.151</td>
<td>.210*</td>
<td>.184</td>
<td>.136</td>
</tr>
<tr>
<td><strong>Vokabular</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ausschließlich häufig / einfach</td>
<td>-.10</td>
<td>-.131</td>
<td>-.199**</td>
<td>-.165 *</td>
</tr>
<tr>
<td>hauptsächlich häufig / einfach</td>
<td>-.355***</td>
<td>-.363***</td>
<td>-.294***</td>
<td>-.259**</td>
</tr>
<tr>
<td>teilweise erweitert / selten</td>
<td>-.067</td>
<td>-.021</td>
<td>-.009</td>
<td>.020</td>
</tr>
<tr>
<td>erweitert / selten</td>
<td>.505***</td>
<td>.478***</td>
<td>.430***</td>
<td>.339***</td>
</tr>
<tr>
<td><strong>Grammatik</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ausschließlich einfache Strukturen</td>
<td>-.399***</td>
<td>-.359***</td>
<td>-.256**</td>
<td>-.194*</td>
</tr>
<tr>
<td>hauptsächlich einfache Strukturen</td>
<td>.139</td>
<td>.048</td>
<td>.062</td>
<td>.079</td>
</tr>
<tr>
<td>teilweise komplexe Strukturen</td>
<td>.120</td>
<td>.170</td>
<td>-.030</td>
<td>-.005</td>
</tr>
<tr>
<td>komplexe Strukturen</td>
<td>.250**</td>
<td>.251**</td>
<td>.326***</td>
<td>.183*</td>
</tr>
</tbody>
</table>

N=104; * p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%

Im Gegensatz zu den Englisch-Items weisen hier auch die Anforderungsmerkmale, welche die Art der zur Beantwortung notwendigen Information abbilden (Variablen: Informationsgewinn 1,2,3), signifikante Zusammenhänge zur Itemschwierigkeit in den Ländern auf. Die Ergebnisse für die Deutsch-Items weisen hier in eine ähnliche Richtung, wie es bei den Englisch-Items bereits zu beobachten war; hier wird der Zusammenhang noch deutlicher. Auch zeigen sich hier keine Ausnahmen hinsichtlich der grammatischen Strukturen.

Fazit:
Die eingangs aufgestellte Hypothese bezüglich korrelativer Zusammenhänge zwischen Item-Anforderungsmerkmalen und Itemschwierigkeit innerhalb der einzelnen Länder lautete:

**Hypothese 2a:** Die kognitiv-linguistischen Item-Anforderungsmerkmale des „Dutch Grid“-Kategoriensystems weisen einen korrelativen Zusammenhang mit den Itemschwierigkeiten innerhalb der Länder auf.

5.2.2. Zu Frage 2b

Frage 2b: Ist die Höhe der Korrelationen in den Ländern vergleichbar?

Englisch-Items


Die Ergebnisse dieser Tabelle sind wie folgt zu lesen (Beispiel): Der Unterschied der Korrelation zwischen der Itemschwierigkeit und dem Itemtyp „Multiple Choice“ ist in Deutschland und Frankreich nicht signifikant (Prüfgröße z (0.49) < zkrit (1.96)). Die Nullhypothese „die Korrelationen unterscheiden sich nicht“ wird beibehalten. Die übrigen Ergebnisse sind analog zu diesem Beispiel zu interpretieren. Insgesamt zeigen sich nur sehr wenige signifikante Unterschiede, in 64 von 68 Fällen unterscheiden sich die Korrelationen nicht.

Es zeigt sich, dass bei der Variablen „Zitieren“ signifikante Unterschiede zwischen den Korrelationen in Deutschland und Spanien (p = 0.028) bzw. zwischen Spanien und Ungarn (p = 0.023) zu finden sind. Des Weiteren zeigt sich bei der Variablen „hauptsächlich abstrakt“ ein auf einem 10%-Niveau signifikanter Unterschied bei den Korrelationen von Frankreich und Ungarn (p = 0.059) und Deutschland und Ungarn (p = 0.098). Es wurde zusätzlich zu dem üblichen 5%-Signifikanz-Niveau ein 10%-Signifikanz-Niveau festgelegt. Dies hatte zum Ziel, auch geringe Unterschiede aufzudecken und diese als relevant zu betrachten. Alle übrigen Korrelationsvergleiche weisen darauf hin, dass sich die Zusammenhänge zwischen Itemschwierigkeiten und Itemmerkmalen hinsichtlich der Größe und Richtung kaum unterscheiden.
<table>
<thead>
<tr>
<th>Itemeigenschaft</th>
<th>Frankreich-Deutschland</th>
<th>Frankreich-Ungarn</th>
<th>Frankreich-Spanien</th>
<th>Deutschland-Spanien</th>
<th>Deutschland-Ungarn</th>
<th>Spanien-Ungarn</th>
<th>( z )</th>
<th>( p(z) )</th>
<th>( z )</th>
<th>( p(z) )</th>
<th>( z )</th>
<th>( p(z) )</th>
<th>( z )</th>
<th>( p(z) )</th>
<th>( z )</th>
<th>( p(z) )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Itemtyp</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>.49</td>
<td>.623</td>
<td>.1</td>
<td>.922</td>
<td>.68</td>
<td>.82</td>
<td>.57</td>
<td>.566</td>
<td>.59</td>
<td>.556</td>
<td>.13</td>
<td>.897</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>.12</td>
<td>.902</td>
<td>.0</td>
<td>.1</td>
<td>.18</td>
<td>.859</td>
<td>.3</td>
<td>.763</td>
<td>.12</td>
<td>.902</td>
<td>.18</td>
<td>.859</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ordnen</td>
<td>.03</td>
<td>.975</td>
<td>.43</td>
<td>.665</td>
<td>.15</td>
<td>.877</td>
<td>.12</td>
<td>.901</td>
<td>.4</td>
<td>.866</td>
<td>.28</td>
<td>.819</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zitieren</td>
<td>.162</td>
<td>.105</td>
<td>1.069</td>
<td>.09</td>
<td>.58</td>
<td>.563</td>
<td>2.2**</td>
<td>.028</td>
<td>.07</td>
<td>.941</td>
<td>2.27**</td>
<td>.023</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lückentext</td>
<td>.23</td>
<td>.815</td>
<td>1.89</td>
<td>.059</td>
<td>.36</td>
<td>.717</td>
<td>.13</td>
<td>.898</td>
<td>1.65</td>
<td>.098</td>
<td>.153</td>
<td>.127</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Information gewinn 1**

| Schlussfolger / Erkennen | .38 | .707 | .85 | .397 | .57 | .566 | .19 | .850 | .47 | .637 | .28 | .778 |

**Information gewinn 2**

| Implizit (vs. Explizit) | .44 | .657 | .31 | .757 | .0 | .988 | .45 | .656 | .74 | .46 | .29 | .656 |

| **Authentizität**       | 113 | .257 | .2 | .839 | .05 | .963 | 1.09 | .277 | .93 | .35 | .16 | .876 |

| ausschließlich konkret   | .7  | .483 | .44 | .657 | .41 | .68  | .29 | .772 | .26 | .79 | .03 | .974 |
| hauptsächlich konkret   | .75 | .455 | 1.53 | .127 | .75 | .58  | .19 | .847 | .78 | .435 | .97 | .33 |
| teilweise abstrakt      | .23 | .815 | 1.89* | .06 | .36 | .717 | .13 | .898 | 1.65* | .1 | .153 | .127 |

| **Vokabular**           | .23 | .818 | .3 | .763 | .35 | .724 | .12 | .903 | .53 | .595 | .65 | .513 |
| ausschließlich häufig / einfach | .55 | .584 | .72 | .474 | .14 | .889 | .41 | .683 | 1.26 | .206 | .86 | .392 |
| hauptsächlich häufig / einfach | .45 | .655 | .56 | .573 | .16 | .873 | .61 | .544 | 1.01 | .312 | .40 | .687 |
| teilweise erweitert / selten | .11 | .904 | .05 | .957 | .67 | .504 | .78 | .438 | .05 | .957 | .72 | .47 |

**Grammatik**

| ausschließlich einfache Strukturen | .03 | .975 | .4 | .696 | .03 | .975 | 0 | 10 | .37 | .71 | .37 | .710 |
| hauptsächlich einfache Strukturen | .15 | .88  | .32 | .746 | .74 | .457 | .89 | .371 | .17 | .863 | 1.07 | .286 |
| teilweise komplexe Strukturen     | .11 | .904 | .05 | .957 | .67 | .504 | .78 | .438 | .05 | .957 | .72 | .47 |

\( N=104; * p \leq 0.1, \alpha = 10\%; ** p \leq 0.05, \alpha = 5\%; *** p \leq 0.01, \alpha = 1\% \)
**Deutsch-Items**

Im Folgenden werden die Ergebnisse der Korrelationsvergleiche bei den Deutsch-Items dargestellt (Tabelle 5.22). Der Tabelleninhalt ist analog zu den oben dargestellten Ergebnissen der Englisch-Items zu interpretieren.

Für die Deutsch-Items zeigt sich ein ähnliches Muster wie auch schon bei den Englisch-Items. In 79 von 84 Fällen gleichen sich die Korrelationen zwischen Itemschwierigkeiten und den verschiedenen schwierigkeitsbestimmenden Itemmerkmalen hinsichtlich Größe und Richtung. Lediglich bei der Variable „Informationsgewinn 3“ lassen sich auf einem 5%-Niveau signifikante Unterschiede zwischen den Niederlanden und Ungarn (p=0.031) und den Niederlanden und Schweden (p=0.049) finden. Des Weiteren gibt es auf einem 10%-Niveau signifikante Unterschiede bei den Variablen „Multiple Matching“ (Frankreich-Ungarn, p=0.063; Niederlande-Ungarn; p=0.057) und „Multiple choice“ (Niederlande-Ungarn; p=0.074).

<table>
<thead>
<tr>
<th>Itemtyp</th>
<th>Frankreich-Niederlande</th>
<th>Frankreich-Ungarn</th>
<th>Frankreich-Schweden</th>
<th>Niederlande-Ungarn</th>
<th>Niederlande-Schweden</th>
<th>Ungarn-Schweden</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>z</td>
<td>p(z)</td>
<td>z</td>
<td>p(z)</td>
<td>z</td>
<td>p(z)</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>.33 .742</td>
<td>1.45 .146</td>
<td>1.03 .303</td>
<td>1.78 .074 *</td>
<td>.174 .136</td>
<td>.43 .671</td>
</tr>
<tr>
<td>Banked Multiple Choice</td>
<td>.19 .853</td>
<td>.39 .70</td>
<td>.32 .751</td>
<td>1.55 .121</td>
<td>.895 .13</td>
<td>.7 .482</td>
</tr>
<tr>
<td>Multiple Choice</td>
<td>.61 .54</td>
<td>.24 .811</td>
<td>.2 .843</td>
<td>.85 .394</td>
<td>.418 .81</td>
<td>.04 .967</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>.05 .962</td>
<td>1.86 .063 *</td>
<td>1.17 .243</td>
<td>1.9 .057 *</td>
<td>.224 .122</td>
<td>.69 .492</td>
</tr>
<tr>
<td>Kurzantwort</td>
<td>.49 .627</td>
<td>.43 .671</td>
<td>.2 .854</td>
<td>.91 .362</td>
<td>.772 .29</td>
<td>.62 .535</td>
</tr>
<tr>
<td>Lückentext</td>
<td>.31 .757</td>
<td>.4 .688</td>
<td>.34 .734</td>
<td>.71 .478</td>
<td>.517 .65</td>
<td>.06 .951</td>
</tr>
</tbody>
</table>

| Informationsgewinn 1         | Schlussfolgern/Erkennen| .36 .722          | .18 .860            | .67 .502          | .18 .858            | .304 1.03       | .85 .396        |
| Informationsgewinn 2         | Implizit(vs. Explizit) | .07 .947          | .99 .322            | 1.47 .142         | .92 .356            | .162 1.14       | .48 .634        |
| Informationsgewinn 3         | Detail (vs. Hauptidee) | .31 .189          | .84 .40             | .66 .511          | 2.16 ** .031        | .049 ** 1.97    | .19 .853        |
| Authentizität                 | Authentisch (vs. angepasst/vereinfacht) | .71 .475          | .24 .807            | .81 .419          | .47 .638            | .925 .09        | .56 .572        |
| Abstraktheit des Inhalts     | ausschließlich konkret | .28 .78           | .76 .45             | .67 .501          | .48 .633            | .341 .95        | 1.14 .153       |
|                              | hauptsächlich konkret  | .02 .987          | .53 .598            | .56 .578          | .54 .587            | .584 .55        | 1.09 .275       |
|                              | teilweise abstrakt     | .47 .638          | .26 .793            | .12 .906          | .21 .835            | .556 .59        | .38 .704        |
| Vokabular                    | ausschließlich häufig/einfach | .24 .808          | .78 .434            | .51 .61           | .55 .585            | .789 .27        | .27 .786        |
|                              | hauptsächlich häufig/einfach | .07 .944          | .53 .599            | .82 .413          | .6 .55              | .374 .89        | .29 .77         |
|                              | teilweise erweitert/sehrten | .68 .497          | .59 .557            | .36 .716          | .23 .817            | .994 .01        | .22 .823        |
|                              | erweitert/sehrten       | .27 .784          | .74 .459            | 1.57 .117         | .47 .641            | .197 1.29       | .82 .409        |
| Grammatik                    | ausschließlich einfache Strukturen | .34 .719          | 1.24 .215           | 1.74 .081         | .88 .380            | .167 1.38       | .5 .614         |
|                              | hauptsächlich einfache Strukturen | .71 .479          | .60 .548            | .47 .63           | .11 .914            | .810 .24        | .13 .895        |
|                              | teilweise komplexe Strukturen | .39 .694          | 1.16 .245           | .97 .333          | .156 .12            | .173 .36        | .19 .847        |
|                              | komplexe Strukturen     | .01 .993          | .64 .522            | .54 .587          | .63 .528            | .582 .55        | .118 .237       |

N=104; * p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%
5.2.3. Zu Fragen 2c und 2d

| Frage 2c: Weisen die kognitiv-linguistischen Anforderungsmerkmale der Items regressionsanalytische Zusammenhänge mit den Itemsschwierigkeiten innerhalb der Länder auf? | Frage 2d: Wie groß ist der Anteil der durch die Prädiktoren aufgeklärten Varianz? |

Die Ergebnisse der beiden Teilfragestellungen 2c und 2d werden im Folgenden gemeinsam berichtet.

Zur Beantwortung der Fragen wurden die unter 2.2.4 genannten Anforderungsmerkmale der Items dazu verwendet, innerhalb der einzelnen Länder die bei Frage 1a geschätzten Itemsschwierigkeiten vorherzusagen. Dies diente dem Zweck zu überprüfen, inwieweit die hier verwendeten Itemeigenschaften des Dutch-Grid-Kategorisierungssystems dazu geeignet sind, die Varianz der Itemsschwierigkeiten innerhalb der verschiedenen Länder aufzuklären. Bezüglich der Richtung und Größe der standardisierten Regressionsgewichte (Beta-Gewichte) wurden keine Hypothesen aufgestellt.

Als Prädiktoren in einer multiplen Regression wurden die kognitiv-linguistischen Anforderungsmerkmale verwendet, die Itemsschwierigkeiten innerhalb der Länder stellten jeweils die abhängige Variable dar.


Das zweite Modell zeigt dann letztendlich jeweils den Anteil der Varianz der Itemsschwierigkeit innerhalb eines Landes, der durch signifikante Item-Anforderungsmerkmale aufgeklärt wird. Der Anteil der anhand der Itemmerkmale aufgeklärten Varianz bezieht sich auf die Beantwortung von Frage 2c. Es wird im Folgenden die Einschluss-Methode verwendet, da es keine theoretischen Überlegungen zu Reihenfolgeeffekten der Prädiktoren gibt. Die Ergebnisse der multiplen Regressionen der Itemsschwierigkeit auf die Item-Anforderungsmerkmale in den einzelnen Länder-
Ergebnisse für die Englisch-Items

<table>
<thead>
<tr>
<th></th>
<th>Modell 1</th>
<th></th>
<th>Modell 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \beta ) (sig)</td>
<td>( \beta ) (sig)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schlussfolgern/Erkennen (Info 1)</td>
<td>0.180 (.203)</td>
<td></td>
<td>0.245 (.010)**</td>
<td></td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>-0.116 (.281)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>0.102 (.487)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt hauptsächlich konkret</td>
<td>0.230 (.040)**</td>
<td>0.245 (.010)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>0.401 (.024)**</td>
<td>0.508 (.000)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>0.254 (.060) *</td>
<td>0.200 (.066) *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>0.381 (.003)***</td>
<td>0.275 (.011)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haupts. einfache grammatische Strukturen</td>
<td>-0.477 (.002)***</td>
<td>-0.408 (.003)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>-0.602 (.001)***</td>
<td>-0.564 (.000)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itemtyp Multiple Choice</td>
<td>0.196 (.142)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itemtyp Ordnen</td>
<td>-0.076 (.516)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itemtyp Zitieren</td>
<td>0.295 (.002)***</td>
<td>0.256 (.001)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.387</td>
<td>0.367</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abhängige Variable: Itemanforderungs-Merkmale; Methode: Einschluß; Zelleninhalt: Standardisierte \( \beta \)-Gewichte; 
* \( p \leq 0.1, \alpha = 10\% \); ** \( p \leq 0.05, \alpha = 5\% \); *** \( p \leq 0.01, \alpha = 1\% \)

Tabelle 5.23: Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Deutschland (Englisch-Items)

Die Ergebnisse in Tabelle 5.23 sind wie folgt zu interpretieren: Wenn beispielsweise komplexe grammatische Strukturen vorliegen (zu interpretieren im Vergleich zur Variablenausprägung „ausschließlich einfache grammatische Strukturen“, da dies in diesem Fall die Kontrastvariable ist), dann verringert sich die Itemschwierigkeit für die Schüler in Deutschland um 0.6 Logits. Wenn hingegen das Vokabular eines Items von den Experten als eher selten oder schwierig eingeschätzt wurde, dann erschwert dies das Item um mehr als ein Drittel Logit im Gegensatz zu Items, die ausschließlich sehr häufiges bzw. sehr einfaches Vokabular beinhalten (Kontrastvariable). Die weiteren Beta-Gewichte sind auf die gleiche Art und Weise zu interpretieren.

Das berichtete Ergebnis hinsichtlich der komplexen grammatischen Strukturen entspricht nicht den im „Dutch Grid“ angenommenen Schwierigkeitsabstufungen der Itemmerkmale. Bei den Englisch-Items ist dieses Phänomen in allen Ländern zu beobachten (siehe unten). Auf mögliche
Gründe wird in der Ergebnisdiskussion unter Abschnitt 6 eingegangen. Insgesamt können mit Hilfe aller kognitiv-linguistischen Itemeigenschaften 38.7% der Varianz der Itemschwierigkeiten in Deutschland aufgeklärt werden. Unter Verwendung der signifikanten Prädiktoren aus Modell 1 wird in Modell 2 noch immer 36.7% der Varianz erklärbar. Der Anteil aufgeklärter Varianz verringert sich nur wenig von Modell 1 zu Modell 2, was für die hier vorgenommene Reduktion der Prädiktorenanzahl spricht. Wie schon bei den Einzelkorrelationen zeigen sich auch im Rahmen einer multiplen Regression keine signifikanten Zusammenhänge zwischen der Itemschwierigkeit und den Variablen, welche die zur Lösung notwendige Art der Information abbilden, d.h. „Erkennen/Schlussfolgern“ und „implizit/explicit“.

<table>
<thead>
<tr>
<th></th>
<th>Modell 1</th>
<th>Modell 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schlussfolgern/Erkennen (Info 1)</td>
<td>-0.046 (.763)</td>
<td></td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>-0.018 (.876)</td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>0.016 (.919)</td>
<td></td>
</tr>
<tr>
<td>Inhalt hauptsächlich konkret</td>
<td>0.127 (.291)</td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>0.446 (0.021)**</td>
<td>0.383 (0.000)**</td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>0.312 (.115)</td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>0.206 (.140)</td>
<td></td>
</tr>
<tr>
<td>Haupts. einfache grammatische Strukturen</td>
<td>-0.449 (0.006)**</td>
<td>-0.448 (0.001)**</td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>-0.397 (0.034)**</td>
<td>-0.316 (0.023)**</td>
</tr>
<tr>
<td>Itemtyp Multiple Choice</td>
<td>0.028 (.849)</td>
<td></td>
</tr>
<tr>
<td>Itemtyp Ordnen</td>
<td>-0.045 (.727)</td>
<td></td>
</tr>
<tr>
<td>Itemtyp Zitieren</td>
<td>-0.063 (.533)</td>
<td></td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.273</td>
<td>0.227</td>
</tr>
</tbody>
</table>

Tabelle 5.24. Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Spanien (Englisch-Items)

Bei den spanischen Schülern (Tabelle 5.24) können insgesamt 27.3% der Varianz der Itemschwierigkeit der Englisch-Items auf die hier als Prädiktoren verwendeten Item-Anforderungsmerkmale zurückgeführt werden. Im zweiten Modell wurden erneut nur die signifikanten Prädiktoren aufgenommen. Es zeigt sich, dass hier insgesamt nur drei der Prädiktoren signifikant zur Itemschwierigkeit beitragen, und zwar die Variablen „abstrakter Inhalt“, und „hauptsächlich einfache“ sowie „teilweise komplexe...
grammatische Strukturen”. Mit Hilfe dieser drei Prädiktoren können insgesamt noch 22.7% der Varianz der geschätzten Itemschwierigkeiten in Spanien aufgeklärt werden.

<table>
<thead>
<tr>
<th></th>
<th>Modell 1</th>
<th>Modell 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β (sig)</td>
<td>β (sig)</td>
</tr>
<tr>
<td>Schlussfolgern/Erkennen (Info 1)</td>
<td>.057 (.702)</td>
<td></td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>-.048 (.673)</td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>-.032 (.837)</td>
<td></td>
</tr>
<tr>
<td>Inhalt hauptsächlich konkret</td>
<td>.133 (.253)</td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>.523 (.005)***</td>
<td>.474 (.000)***</td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>.334 (.019)**</td>
<td>.348 (.002)***</td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>.223 (.099)*</td>
<td>.204 (.069)*</td>
</tr>
<tr>
<td>Haupts./einfache grammaticale Strukturen</td>
<td>-.318 (.016)**</td>
<td>-.347 (.013)**</td>
</tr>
<tr>
<td>Teilweise komplexe grammaticale Strukturen</td>
<td>-.542 (.003)***</td>
<td>-.458 (.001)***</td>
</tr>
<tr>
<td>Itemtyp Multiple Choice</td>
<td>.044 (.754)</td>
<td></td>
</tr>
<tr>
<td>Itemtyp Ordnen</td>
<td>-.069 (.577)</td>
<td></td>
</tr>
<tr>
<td>Itemtyp Zitieren</td>
<td>.051 (.599)</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.319</td>
<td>.301</td>
</tr>
</tbody>
</table>

Abhängige Variable: Itemschwierigkeit Frankreich; Methode: Einschluss; Zelleninhalt: Standardisierte β-Gewichte; * $p \leq 0.1, \alpha = 10\%$; ** $p \leq 0.05, \alpha = 5\%$; *** $p \leq 0.01, \alpha = 1\%

Tabelle 5.25. Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Frankreich (Englisch-Items)

In der französischen Schülergruppe (Tabelle 5.25) lassen sich mit Hilfe aller Prädiktoren 31.9% der Varianz der Itemschwierigkeiten aufklären. In einem zweiten Modell werden auch hier lediglich die Prädiktoren aufgenommen, die in Modell 1 ein signifikantes signifikantes Beta-Gewicht aufwiesen. Hier zeigt sich, dass insgesamt 5 Prädiktoren das 30.1% der Varianz aufklären können. Auch hier spricht der geringe Verlust an aufgeklärter Varianz durch die Herausnahme nicht signifikanter Prädiktoren für die Reduktion der Prädiktorenanzahl.

Bezüglich der Gruppe ungarischer Schüler (Tabelle 5.26) zeigt sich hier, dass 25.4% der Varianz der Itemschwierigkeiten auf Item-Anforderungsmerkmale zurückgeführt werden können. In Modell 2, welches ausschließlich Prädiktoren mit einem signifikanten bzw. tendenziell signifikanten Beta-Gewicht aufnimmt, werden noch immer 25% der Varianz aufgeklärt. Dies spricht auch hier für die Reduktion der Anzahl der Prädiktoren.
Zusammenfassend lässt sich bezüglich der Englisch-Items sagen, dass (im jeweils zweiten Modell) zwischen 36.7% und 22.7% der Varianz der Itemschwierigkeiten innerhalb der Länder auf Item-Anforderungsmerkmale zurückführbar sind. Meist handelt es sich dabei um Ausprägungen der Merkmale „Abstraktheit des Inhalts“, „Schwierigkeit des Vokabulars“ und „Komplexität grammatischer Strukturen“. In Ungarn und Deutschland trägt außerdem der Itemtyp „Zitieren“ zur Itemschwierigkeit bei. Wie auch schon bei den Einzelkorrelationen zu beobachten war, tragen die Merkmale „Informationsgewinn 1+2“ nicht signifikant zur Itemschwierigkeit der Englisch-Items bei.

**Ergebnisse für die Deutsch-Items**

Im Folgenden werden nun die Ergebnisse der Regressionen der Itemschwierigkeiten der Deutsch-Items innerhalb der Länder auf die Item-Anforderungsmerkmale berichtet. Hier sind als Prädiktoren teilweise mehr Stufen der Itemmerkmale mit einbezogen, als dies bei den Englisch-Items der Fall war. So existiert hier beispielsweise die Ausprägung „komplexe grammatische Strukturen“ der Variable „Komplexität grammatischer Strukturen“. Auch bezüglich des Vokabulars wird die Abstufung „Vokabular selten/schwierig“ ergänzt. Das ist darauf zurückzuführen,
dass keines der Englisch-Items in die dem „Dutch Grid“ zufolge schwierigsten Merkmalsstufen (wie „Grammatik komplex“) eingeordnet wurde. Bei den Deutsch-Items hingegen war das der Fall. Da immer die niedrigste Stufe als Kontrastvariable gewählt wurde und daher als Prädiktor nicht aufgeführt ist, führt dies dazu, dass beispielsweise das Itemmerkmal „Grammatische Strukturen“ bei den Englisch-Items nur zwei, bei den Deutsch-Items jedoch drei Abstufungen aufweist. Ferner konnte für die Deutsch-Items auch die Variable „Informationsgewinn 3“ einbezogen werden, die sich auch auf die zur Beantwortung notwendige Information bezieht, nämlich darauf, ob zur Beantwortung des Items im Text ein bestimmtes Detail gefunden werden muss, beispielsweise ein Name, oder ob es notwendig ist, den Gesamtzusammenhang des Texts verstanden zu haben. Auch bezüglich des Itemtyps der Aufgaben ist bei den Deutsch-Aufgaben die Variabilität größer. So gibt es hier auch Aufgaben des Typs „Kurzantwort“. Im Folgenden werden die Regressionen der Itemschwierigkeiten auf die Item-Anforderungsmerkmale innerhalb der Länder Frankreich, Ungarn, Niederlande und Schweden berichtet.

In der ungarischen Schülergruppe (Tabelle 5.28) können 56.9% der Varianz der Itemschwierigkeiten anhand kognitiv-linguistischer Anforderungsmerkmale der Items aufgeklärt werden. Auch hier werden in einem zweiten Modell ausschließlich die Prädiktoren aufgenommen, die signifikante Beta-Gewichte aufweisen. Es können 48.8% der Varianz auf Ausprägungen der Variablen „Itemtyp“, „Abstraktheit des Inhalts“, „Schwierigkeit des Vokabulars“ und „Komplexität grammatischer Strukturen“ zurückgeführt werden. Darüber hinaus deutet auch hier die Anwesenheit von Beta-Koeffizienten > 1 auf Suppressionseffekte hin.

In der niederländischen Gruppe (Tabelle 5.29) lassen sich insgesamt 44.7% der Varianz der Itemschwierigkeiten der Deutsch-Items auf die Item-Anforderungsmerkmale zurückführen. Hier hat im Endmodell ausschließlich die Variable „Schwierigkeit des Vokabulars“ ein signifikant das Beta-Gewicht und klärt 22.8% der Varianz der Itemschwierigkeiten auf.

### Tabelle 5.27. Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Frankreich (Deutsch-Items)

<table>
<thead>
<tr>
<th>Itemtyp</th>
<th>Modell 1 $\beta$ (sig)</th>
<th>Modell 2 $\beta$ (sig)</th>
<th>Modell 3 $\beta$ (sig)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itemtyp Multiple Choice</td>
<td>-.016 (.967)</td>
<td>- .220 (.019)**</td>
<td>-.201 (.030)**</td>
</tr>
<tr>
<td>Itemtyp Banked Multiple Choice</td>
<td>-.140 (.456)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richtig/Falsch</td>
<td>-.589 (.087)</td>
<td>- .202 (.019)**</td>
<td>-.201 (.030)**</td>
</tr>
<tr>
<td>Itemtyp Multiple Matching</td>
<td>-.047 (.894)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itemtyp Kurzantwort</td>
<td>.106 (.622)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schlussfolger/Erkennen (Info 1)</td>
<td>-.102 (.531)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implizit/Explitzt (Info 2)</td>
<td>-.041 (.828)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detail/Hauptidee</td>
<td>-.104 (.584)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentizität</td>
<td>-.441 (.005)**</td>
<td>- .133 (.186)</td>
<td></td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>.113 (.439)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>-.037 (.816)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise häufig / einfach</td>
<td>.847 (.005)**</td>
<td>.410 (.044)**</td>
<td>.319 (.095)*</td>
</tr>
<tr>
<td>Vokabular hauptsächlich ausgeweitet/ selten</td>
<td>1.105 (.001)**</td>
<td>.685 (.003)**</td>
<td>.567 (.006)**</td>
</tr>
<tr>
<td>Vokabular ausgeweitet/ selten</td>
<td>1.359 (.005)**</td>
<td>1.023 (.000)**</td>
<td>.985 (.000)**</td>
</tr>
<tr>
<td>Haupts. einfache grammatische Strukturen</td>
<td>.626 (.001)**</td>
<td>.432 (.000)**</td>
<td>.450 (.000)**</td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>.316 (.329)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komplexe grammatische Strukturen</td>
<td>-.326 (.417)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.584</td>
<td>.463</td>
<td>.451</td>
</tr>
</tbody>
</table>

Abhängige Variable: Itemschwierigkeit Frankreich; Methode: Einschluß; Zelleninhalt: Standardisierte $\beta$-Gewichte:

* $p \leq 0.1, \alpha = 10\%$; ** $p \leq 0.05, \alpha = 5\%$; *** $p \leq 0.01, \alpha = 1\%$
### Ergebnisse

#### 5.2 Ergebnisse Fragenkomplex 2

<table>
<thead>
<tr>
<th>Modell 1</th>
<th>Modell 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>β (sig)</td>
<td>β (sig)</td>
</tr>
<tr>
<td>Banked Multiple Choice</td>
<td>-0.128</td>
</tr>
<tr>
<td>Richtig-Falsch</td>
<td>-0.527 (.011)**</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>-0.559 (.008)*****</td>
</tr>
<tr>
<td>Kurzantwort</td>
<td>0.132 (.473)</td>
</tr>
<tr>
<td>Lückentext</td>
<td>-0.073 (.553)</td>
</tr>
<tr>
<td>Schlussfolgerung/Erkennen (Info 1)</td>
<td>-0.199 (.233)</td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>-0.024 (.901)</td>
</tr>
<tr>
<td>Detail/Hauptidee</td>
<td>-0.310 (.111)</td>
</tr>
<tr>
<td>Authentizität</td>
<td>-0.199 (.204)</td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>0.379 (.013)**</td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>0.042 (.795)</td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>0.759 (.013)**</td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>0.858 (.011)**</td>
</tr>
<tr>
<td>Vokabular ausgeweitet/selten</td>
<td>1.654 (.001)*****</td>
</tr>
<tr>
<td>Haupts.einfache grammatische Strukturen</td>
<td>0.423 (.002)**</td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>-0.176 (.592)</td>
</tr>
<tr>
<td>Komplexe grammatische Strukturen</td>
<td>-0.660 (.108)</td>
</tr>
</tbody>
</table>

**Abhängige Variable: Itemschwierigkeit Ungarn; Methode: Einschluss; Zelleninhalt: Standardisierte β-Gewichte; * p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R²</td>
<td>0.569</td>
</tr>
<tr>
<td></td>
<td>0.494</td>
</tr>
</tbody>
</table>

**Tabelle 5.28.** Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Ungarn (Deutsch-Items)


Die Überprüfung von Frage 2c erfolgte exploratorisch. Der Anteil der durch die Itemmerkmale aufklärbaren Varianz liegt bei den Englisch-Items nach Ausschluss aller nicht-signifikanten Prädiktoren zwischen $R^2 = .227$ und $R^2 = .367$. Der Anteil der aufgeklärten Varianz liegt für die
Tabelle 5.29. Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in den Niederlanden (Deutsch-Items)

Deutsch-Items zwischen $R^2 = .208$ und $R^2 = .494$.

**Hypothese 2d:** Die kognitiv-linguistischen Item-Anforderungsmerkmale des Dutch-Grid Kategoriensystems weisen einen regressionsanalytischen Zusammenhang mit den Itemschwierigkeiten innerhalb der Länder auf.

Insgesamt kann diese Hypothese beibehalten werden. Es zeigt sich sowohl für die Englisch- als auch für die Deutsch-Items, dass die in dieser Arbeit ausgewählten Item-Anforderungsmerkmale innerhalb der Länder zur Varianzaufklärung beitragen. Dabei fällt auf, dass dies vor allem für die Merkmale „Komplexität der Grammatik” und Schwierigkeit des Vokabulars” zutrifft. Jedoch weisen auch die anderen Merkmale Zusammenhänge auf, wenn auch nicht gleichermaßen in
Tabelle 5.30. Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Schweden (Deutsch-Items)

<table>
<thead>
<tr>
<th>Merkmal</th>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$ (sig)</td>
<td>$\beta$ (sig)</td>
<td>$\beta$ (sig)</td>
</tr>
<tr>
<td>Banked Multiple Choice</td>
<td>.011 (.931)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richtig-Falsch</td>
<td>-.365 (.120)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>-.383 (.111)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kurzantwort</td>
<td>.354 (.097)*</td>
<td>.413 (.007)***</td>
<td>.449 (.004)***</td>
</tr>
<tr>
<td>Lückentext</td>
<td>.063 (.656)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schlussfolgern/Erkennen (Info 1)</td>
<td>-.270 (.162)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>-.180 (.416)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detail/Hauptidee</td>
<td>-.305 (.173)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentizität</td>
<td>-.166 (.357)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>.296 (.087)*</td>
<td>.563 (.575)</td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>.170 (.368)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>.514 (.141)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>.680 (.078)*</td>
<td>.161 (.220)</td>
<td></td>
</tr>
<tr>
<td>Vokabular ausgeweitet/selten</td>
<td>1.750 (.002)***</td>
<td>.482 (.006)***</td>
<td>.435 (.002)***</td>
</tr>
<tr>
<td>Haupts. einfache grammatische Strukturen</td>
<td>.272 (.196)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukuren</td>
<td>-.342 (.367)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komplexe grammatische Strukturen</td>
<td>-.1240 (.010)***</td>
<td>-.399 (.035)**</td>
<td>-.456 (.014)**</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.426</td>
<td>.240</td>
<td>.208</td>
</tr>
</tbody>
</table>

Abhängige Variable: Itemschwierigkeit Schweden; Methode: Einschluß; Zelleninhalt: Standardisierte $\beta$-Gewichte; 
* $p \leq 0.1, \alpha = 10\%$; ** $p \leq 0.05, \alpha = 5\%$; *** $p \leq 0.01, \alpha = 1\%$.

5.3. Ergebnisse zu Fragenkomplex 3: Erklärung von Differentiellen Item Funktionen

Im Folgenden wird auf die Beantwortung der Fragestellungen eingegangen, die sich mit der Erklärung von Differentiellen Item Funktionen, das heißt der kulturell bedingten Varianz der Itemschwierigkeiten zwischen den Ländern, beschäftigen. Dabei stellt sich vor allem die Frage, ob die unter 5.1 herausgearbeiteten Testkulturen der Länder zur Erklärung dieser Unterschiede beitragen können, und ob und inwieweit sich Testkulturen zur Analyse erwarteter Stärken und Schwächen von Gruppen heranziehen lassen. Dazu werden zunächst Korrelationen zwischen DIF und der Itemherkunft berichtet. Danach wird auf die Ergebnisse der Korrelationsanalysen zwischen DIF und den itemschwierigkeitsbestimmenden Merkmalen eingegangen sowie darauf, ob die Ergebnisse den durch die Testkulturen prognostizierten Stärken und Schwächen entsprechen. Im dritten Schritt werden die Ergebnisse multipler Regressionsanalysen berichtet, mit den DIF-Parametern der unter 5.1 durchgeführten paarweisen DIF-Analysen als abhängigen Variablen und den Item-Anforderungsmerkmalen als unabhängigen Variablen. Auch wird darauf eingegangen, ob die Beta-Gewichte den durch die Testkulturen prognostizierten Stärken und Schwächen der Gruppen entsprechen.

5.3.1. Zu Frage 3a

| Frage 3a: Existieren den Testkulturen entsprechende, signifikante korrelative Zusammenhänge zwischen Testkultur-Indikatoren und DIF? |

darstellen, differentielle Item Funktionen sollten also zum Vorteil der Gruppe ausfallen (d.h. die Itemschwierigkeit ist niedriger als für die Referenzgruppe). Die Korrelation zwischen DIF und Items aus dem anderen Land, also der jeweiligen Referenzgruppe, sollte hingegen positiv sein: In diesem Fall sind die Differentiellen Item Funktionen von Nachteil für die Fokusgruppe, und die Itemschwierigkeit vergrößert sich im Vergleich zur Referenzgruppe, aus deren Land das Item stammt. Zunächst werden die Ergebnisse für die Englisch-Items berichtet.

<table>
<thead>
<tr>
<th>Itemherkunft Land A (für A-B)</th>
<th>F-D</th>
<th>F-U</th>
<th>F-Sp</th>
<th>D-U</th>
<th>D-Sp</th>
<th>U-Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>.407**</td>
<td>-.201*</td>
<td>-.004</td>
<td>-.30**</td>
<td>-.312*</td>
<td>-.209*</td>
<td></td>
</tr>
<tr>
<td>Itemherkunft Land B (für A-B)</td>
<td>.281*</td>
<td>.251*</td>
<td>.054</td>
<td>.249**</td>
<td>-.058</td>
<td>.11</td>
</tr>
</tbody>
</table>

Tabelle 5.31. Korrelation zwischen DIF und Itemherkunft (Englisch-Items)

Tabelle 5.31 ist wie folgt zu lesen: In den Spalten sind jeweils die beiden Länder dargestellt, zwischen denen DIF-Analysen durchgeführt wurden. Bezüglich der ersten Spalte bedeutet dies beispielsweise, dass hier paarweise DIF-Analysen zwischen Frankreich und Deutschland berechnet wurden. Frankreich ist als erstes Land aufgeführt, was bedeutet, dass Frankreich als Fokusgruppe gewählt wurde und die Ergebnisse daher aus Sicht dieser Gruppe interpretiert werden müssen. In der ersten Zeile (Herkunft aus Land A) ist jeweils die Korrelation der Differentiellen Item Funktionen zweier Länder mit der Tatsache, dass das Item aus dem Land der Fokusgruppe stammt, dargestellt, in der zweiten Zeile hingegen die Korrelation der Differentiellen Item Funktionen mit dem Sachverhalt, dass das Item aus dem Land der Referenzgruppe stammt. In der ersten Spalte zeigt die Korrelation von $r = -.407$ also, dass der Umstand, dass ein Item aus Frankreich (der Fokusgruppe) stammt, signifikant ($p \leq .01$) mit einer niedrigeren Itemschwierigkeit für die französische Stichprobe einhergeht. Das Gegenteil ist der Fall, wenn das Item aus Deutschland (der Referenzgruppe) stammt: Die Korrelation von $r = .281$ ($p \leq .05$) bedeutet, dass in diesem Fall die Tatsache, dass ein Item aus Deutschland stammt, signifikant mit einer höheren Itemschwierigkeit für die französische Stichprobe einhergeht. Hier zeigt sich, dass für die DIF-Analysen zwischen Frankreich und Deutschland, Frankreich und Ungarn sowie Deutschland und Ungarn gilt, dass die Tatsache, dass ein Item aus dem eigenen Land stammt, die Itemschwierigkeit im Vergleich zur Referenzgruppe verringert, wohingegen die Tatsache, dass Items aus dem Land der Referenzgruppe stammten, deren Schwierigkeit für die Fokusgruppe im Vergleich zur Referenzgruppe erhöht.
Ferner gilt für die DIF-Analysen zwischen Deutschland und Spanien bzw. Ungarn und Spanien, dass Items aus dem eigenen Land die Itemschwierigkeit für die Fokusgruppe zwar verringern, Items aus dem Land der Referenzgruppe die Itemschwierigkeit jedoch nicht erhöhen. Keine signifikanten Korrelationen wurden zwischen den DIF-Parametern von Frankreich-Spanien und der Itemherkunft gefunden.

<table>
<thead>
<tr>
<th>Herkunft Land A (für A-B)</th>
<th>FR-NL</th>
<th>FR-SW</th>
<th>FR-U</th>
<th>NL-SW</th>
<th>U-NL</th>
<th>SW-U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herkunft Land B (für A-B)</td>
<td>.302*</td>
<td>.395**</td>
<td>.135</td>
<td>.242*</td>
<td>.205*</td>
<td>-.177</td>
</tr>
</tbody>
</table>

* p ≤ 0.05, α = 5%; ** p ≤ 0.01, α = 1%;

Tabelle 5.32. Korrelation zwischen DIF und Itemherkunft (Deutsch-Items)


Ein nicht hervorgehobener Eintrag, der mit einem „ungleich“-Zeichen versehen ist, weist auf einen den Hypothesen entgegen gesetzten Zusammenhang hin.

<table>
<thead>
<tr>
<th>Itemeigenschaft</th>
<th>DIF</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Itemtyp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itemtyp Multiple Choice</td>
<td>.194**</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Banked Multiple Choice</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Richtig-Falsch</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Zitieren</td>
<td>-.396***</td>
<td>-.283***</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.419***</td>
</tr>
<tr>
<td>Lückentext</td>
<td></td>
<td>.400***</td>
<td>.207**</td>
<td>.467***</td>
<td>—</td>
<td>-.276***</td>
</tr>
<tr>
<td>Informationsgewinn 1</td>
<td></td>
<td>—</td>
<td>.198**</td>
<td>.209**</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Informationsgewinn 2</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-.194**</td>
<td>—</td>
</tr>
<tr>
<td>Informationsgewinn 3</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Authentizität</td>
<td></td>
<td>—</td>
<td>.304***</td>
<td>—</td>
<td>—</td>
<td>.283***</td>
</tr>
<tr>
<td>Abstraktheit des Inhalts</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ausschließlich konkret</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>hauptsächlich konkret</td>
<td></td>
<td>—</td>
<td>-.296**</td>
<td>-.183**</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>teilweise abstrakt</td>
<td></td>
<td>-.217***</td>
<td>.40***</td>
<td>.207**</td>
<td>.467***</td>
<td>.198**</td>
</tr>
<tr>
<td>Vokabular</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ausschließlich häufig / einfach</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>hauptsächlich häufig / einfach</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>-.216**</td>
<td>—</td>
</tr>
<tr>
<td>teilweise erweitert / selten</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>.297**</td>
<td>.198**</td>
</tr>
<tr>
<td>erweitert/selten</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Grammatik</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ausschließlich einfache Strukturen</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>hauptsächlich einfache Strukturen</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>teilweise komplexe grammatische Strukturen</td>
<td></td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

*D = Deutschland; F = Frankreich; SP = Spanien; U = Ungarn
* \( p \leq 0.1; \alpha = 10%); ** \( p \leq 0.05; \alpha = 5%); *** \( p \leq 0.01; \alpha = 1%)
≠ = entgegen der Hypothese; kursiv = neutral

| Tabelle 5.33. Korrelation zwischen DIF und kognitiv-linguistischen Itemmerkmalen (Englisch-Items) |

Dort sollte DIF zugunsten von Ungarn geringer sein, das Gegenteil ist aber der Fall. Das lässt sich möglicherweise dadurch erklären, dass die Einschätzung der Experten hinsichtlich der Itemeigenschaft nicht ganz korrekt war.

Insgesamt zeigt sich, dass Differentielle Item Funktionen einen Zusammenhang zu kognitiv-linguistischen schwierigkeitsbestimmenden Merkmalen aufweisen, und zwar größtenteils entsprechend der Richtung der testkulturellen Ausprägungen der Gruppen.

Die aufgrund der Testkultur für die Deutsch-Items erwarteten Stärken und Schwächen der Gruppen wurden unter 5.1.3 tabellarisch dargestellt. Aus dem Sachverhalt, dass die Items eines Landes signifikant häufiger das jeweilige Itemmerkmal bzw. die jeweilige Ausprägung aufweisen als die Items eines Vergleichslandes resultierte die Hypothese, dass Items mit diesem Merkmal für die Schüler der ersten Gruppe einfacher sein sollten als für die Schüler der Vergleichsgruppe, da deren Items seltener die entsprechende Merkmalsausprägung beinhalten. Auch hier werden nur die Unterschiedshypothesen bearbeitet und interpretiert. Im Folgenden werden nun die Ergebnisse der Korrelationen zwischen den paarweise berechneten DIF-Parametern und den Item-Anforderungsmerkmalen dargestellt. Auch hier werden die nicht-signifikanten Korrelationen aus Gründen der Übersichtlichkeit in der Tabelle nicht dargestellt. Die komplett Korrelationsmatrix ist im Anhang einsehbar.

Tabelle 5.34 ist analog zu der Korrelationstabelle der Englisch-Items zu lesen und zu interpretieren. Der größte Teil der signifikanten Korrelationen, nämlich 25 von 34, entspricht den Einzelhypothesen, was darauf hindeutet, dass DIF einen Zusammenhang zu den Testkulturen der Länder aufweist. Auch hier finden sich lediglich zwei signifikante Korrelationen, die sich den oben aufgestellten Hypothesen entgegengesetzt darstellen, nämlich bezüglich des Zusammenhangs zwischen DIF/Ungarn-Schweden und der Tatsache, dass ein Item einen konkreten bzw. hauptsächlich konkreten Inhalt aufweist. Die Korrelationen, die signifikant werden, obwohl sich keine Unterschiede bezüglich eines Merkmals bei den Testkulturen zweier Länder zeigen („neutral“), nämlich fünf der 34 signifikanten Korrelationen, könnten möglicherweise darauf hinweisen, dass die in dieser Studie verwendeten Items eines Landes bezüglich dieser Merkmalsausprägung nicht repräsentativ sind.
### Tabelle 5.34. Korrelation zwischen DIF und kognitiv-linguistischen Itemmerkmalen (Deutsch-Items)

<table>
<thead>
<tr>
<th>Itemtyp</th>
<th>FR-NL</th>
<th>FR-SW</th>
<th>FR-HU</th>
<th>NL-SW</th>
<th>HU-NL</th>
<th>SW-HU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Choice</td>
<td>—</td>
<td>-0.264**</td>
<td>-0.369***</td>
<td>-0.279***</td>
<td>0.332***</td>
<td>—</td>
</tr>
<tr>
<td>Banked Multiple Choice</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Richtig-Falsch</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>—</td>
<td>0.304***</td>
<td>0.470***</td>
<td>0.256**</td>
<td>-0.354***</td>
<td>—</td>
</tr>
<tr>
<td>Kurzantwort</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Lückentext</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

| Informationsgewinn 1 | Schlussfolgern/Erkennen | — | 0.227** | — | 0.256** | — | — |
| Informationsgewinn 2 | Implizit(vs. Explizit) | — | 0.396*** | 0.275** | 0.321*** | — | — |
| Informationsgewinn 3 | Detail (vs. Hauptidee) | 0.242** | — | — | -0.400*** | 0.397*** | — |
| Authentizität | Authentisch (vs. angepasst / vereinfacht) | — | -0.235** | — | — | — | — |
| Abstraktheit des Inhalts | ausschließlich konsequent | — | -0.230*** | — | -0.245*** | — | -0.328*** ≠ |
| | hauptsächlich konsequent | — | — | — | — | — | — |
| | teilweise abstrakt | — | — | — | — | — | — |
| Vokabular | ausschließlich häufig/einfach | — | — | — | — | — | — |
| | hauptsächlich häufig/einfach | — | -0.258** | — | -0.231** | — | — |
| | teilweise erweitert/selten | — | — | — | — | — | — |
| | erweitert/selten | — | 0.416*** | 0.228** | 0.306*** | — | — |
| Grammatik | ausschließlich einfache Strukturen | — | -0.452*** | — | -0.311*** | — | — |
| | hauptsächlich einfache Strukturen | — | — | — | — | — | — |
| | teilweise komplexe Strukturen | — | 0.241* | 0.287** | 0.278** | -0.289** | — |
| | komplexe Strukturen | — | — | — | -0.330 *** | — | 0.286** |

* p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%

≠ = entgegen der Hypothese; kursiv = neutral

F = Frankreich; NL = Niederlande; SW = Schweden; U = Ungarn
Ergebnisse 5.3 Ergebnisse Fragenkomplex 3

Hypothese 3a: Es existieren signifikante korrelative Zusammenhänge zwischen Indikatoren der Testkulturen (Herkunft des Items/Itemmerkmale) und Differentiellen Item Funktionen. Die Richtung des Zusammenhang sollte jeweils den aufgrund der Testkultur-Profile erwarteten Stärken und Schwächen der Gruppen entsprechen.


5.3.2. Zu Fragen 3b und 3c

Frage 3b: Können die Testkultur-Indikatoren als Prädiktoren einen Teil der durch kulturelle Unterschiede verursachten Varianz der Itemschwierigkeiten zwischen den Ländern, d.h. DIF, erklären? Frage 3c: Entspricht die Richtung der Regressionsgewichte den erwarteten Stärken und Schwächen der Länder?

Hier soll, im Gegensatz zu Frage 2b, nicht die Varianz innerhalb, sondern DIF, also die kulturell bedingte Varianz zwischen den Ländern, in einer multiplen Regression anhand der Testkultur-Indikatoren erklärt werden. Auch hier werden die im Rahmen von Frage 1b paarweise berechneten DIF-Parameter verwendet, die dann jeweils als abhängige Variable anhand der schwierigkeitsbestimmenden Itemmerkmale als Prädiktoren vorhergesagt werden. Es wird die Hypothese aufgestellt, dass mit Hilfe der Itemeigenschaften bzw. deren Ausprägungen ein Teil der Varianz aufgeklärt werden kann (Frage 3b). Ferner sollte die Richtung der Beta-Gewichte den nationalen Testprofilen entsprechen (Frage 3c). In das jeweilige Endmodell (Modell 3) werden ausschließlich Prädiktoren aufgenommen, die a) in Modell 1 signifikant sind und b) der durch die Testkultur erwarteten Richtung entsprechen. Auf diese Weise kann der Anteil der aufgeklärten Varianz $R^2$ des jeweiligen Endmodells als der Anteil der Varianz interpretiert werden, der ausschließlich auf die durch die Testkultur und differentielle Lerngelegenheiten erwarteten Stärken und Schwächen der Gruppen zurückzuführen ist.
Auch in diesen Modellen wurde ein drittes Signifikanz-Niveau (\(\alpha = 10\%\)) eingeführt. Die Modelle beinhalten also jeweils folgende Prädiktoren:

**Modell 1:** Alle Itemmerkmale werden als Prädiktoren für DIF in das Modell einbezogen

**Modell 2:** Basierend auf Modell 1 werden Prädiktoren ausgeschlossen, die hinsichtlich der Richtung nicht den erwarteten Stärken und Schwächen der Gruppen entsprechen. Dieses Modell beinhaltet noch die Prädiktoren, die im Folgenden als „neutral“ bezeichnet werden. Diese weisen Zusammenhänge zu DIF auf, obgleich aufgrund der Testkultur-Analysen eigentlich keine Zusammenhänge zu erwarten sind. Dieses Modell wird nur gerechnet, wenn solche Prädiktoren in Modell 1 auftreten. Ansonsten wird jeweils direkt Modell 3 berechnet:

**Modell 3:** Basierend auf Modell 1 werden nur die Prädiktoren eingeschlossen, die aufgrund ihrer Richtung den im Anschluss an die Testkultur-Analysen aufgestellten Hypothesen entsprechen.

**Ergebnisse der Englisch-Items**

5.3 Ergebnisse Fragenkomplex 3

<table>
<thead>
<tr>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta(sig)</td>
<td>Beta(sig)</td>
<td>Beta(sig)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Itemtyp</th>
<th>Modell 1 Beta(sig)</th>
<th>Modell 2 Beta(sig)</th>
<th>Modell 3 Beta(sig)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Matching</td>
<td>-.240 (.032)**</td>
<td>-.090 (.312)</td>
<td>.016 (.856)</td>
</tr>
<tr>
<td>Typ Ordnung</td>
<td>-.175 (.186)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zitieren</td>
<td>.351 (.001)***</td>
<td>.414 (.000)***</td>
<td>.396 (.000)***</td>
</tr>
<tr>
<td>Schlussfolgerung/Erkennen (Info 1)</td>
<td>.251 (.090)*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implizit/ explizit (Info 2)</td>
<td>-.145 (.200)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>.241 (.118)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>.227 (.052)*</td>
<td>.232 (.013)**</td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>-.228 (.258)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular häufig/einfach</td>
<td>-.037 (.790)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular ausgeweit/selt</td>
<td>.374 (.006)***</td>
<td>.304 (.001)***</td>
<td></td>
</tr>
<tr>
<td>Haupts. einfache grammatische Strukturen</td>
<td>-.311 (.047)** ≠</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>-.298 (.097)*</td>
<td>-.138 (.133)</td>
<td>-.080 (.381)</td>
</tr>
<tr>
<td>R²</td>
<td>.327</td>
<td>.251</td>
<td>.163</td>
</tr>
</tbody>
</table>

Abhängige Variable: DIF Deutschland - Frankreich; Methode: Einschluß; Zelleninhalt: β-Gewichte
* p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%
≠ entgegen der Hypothese; kursiv = neutral

Tabelle 5.35. Regression von DIF Deutschland - Frankreich auf Anforderungsmerkmale (Englisch-Items)

Merkmal betreffend) erklärt werden, oder aber mit einer nicht ganz korrekten Einordnung durch
die Rater.
Daher könnte es sich hier um verdeckte testkulturelle Einflüsse handeln. Mit diesem Modell
werden noch 25.1% der DIF-Varianz aufgeklärt. Im dritten und letzten Modell werden ausschließlich
die signifikanten Prädiktoren aus Modell 1 aufgenommen, die den Testkulturen eindeutig entspre-
chen. Das Endmodell erklärt 16.3% der Varianz. Dabei handelt es sich nun um die aufgeklärte
Varianz, die den durch die differentiellen Lerngelegenheiten erwarteten Stärken und Schwächen
der Gruppen entspricht. Die starke Veränderung einiger der Beta-Gewichte, beispielsweise der
Variablen „Multiple Choice“ bei Herausnahme von nicht-signifikanten Prädiktoren, weist in die-
sem Modell auf starke Kollinearitätseffekte hin.

Der Zelleninhalt enthält Informationen über die Natur der Stärken und Schwächen der Gruppen
und ist beispielsweise wie folgt zu lesen: Die Tatsache, dass ein Item dem Itemtyp „zitieren“,
entspricht, erschwert ein Item für die deutschen Schüler im Vergleich zu den französischen Schü-
lern um ca. 0.4 Logits. Gleiches gilt für den Prädiktor „Vokabular erweitert / selt“ (Modell 2):
die Tatsache, dass ein Item schwieriges Vokabular beinhaltet, erschwert dies für die deutschen Schüler im Vergleich zu den französischen um 0.3 Logits. Bei diesem Prädiktor handelt es sich jedoch um eine Variable, bei der ein signifikanter Zusammenhang besteht, obgleich aufgrund der Testkultur keiner zu erwarten gewesen wäre. Das zweite Modell beinhaltet diese Prädiktoren, die folgendermaßen interpretiert werden:

Da in Modell 2 aufgrund oben genannter Gründe nicht auszuschließen ist, dass die dort gefundenen, nicht erwarteten Zusammenhänge durch testkulturelle Einflüsse bedingt sind, werden die $R^2$ dieser beiden Modelle als die Obergrenze (Modell 2) bzw. die Untergrenze (Modell 3) der durch die theoretischen Annahmen und die hier verwendeten Testkultur-Indikatoren aufklärbaren Varianz interpretiert. Demnach werden zwischen 16.3% und 25.1% der DIF-Varianz dieser Länderpaarung durch testkulturelle Variablen erklärt. Es zeigt sich also, dass zumindest ein Teil der durch kulturelle Unterschiede verursachten Varianz zwischen den deutschen und den französischen Schülern auf Stärken und Schwächen zurückgeführt werden kann, die vermutlich wiederum auf differentielle Lerngelegenheiten durch unterschiedliche Testkulturen der Länder rückführbar sind.

<table>
<thead>
<tr>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta(sig)</td>
<td>Beta(sig)</td>
<td>Beta(sig)</td>
</tr>
<tr>
<td><strong>Itemtyp Multiple Matching</strong></td>
<td>.183 (.094)*</td>
<td>.099 (.638)</td>
</tr>
<tr>
<td><strong>Itemtyp Ordnen</strong></td>
<td>.217 (.095)*</td>
<td>.123 (.157)</td>
</tr>
<tr>
<td><strong>Itemtyp Zitieren</strong></td>
<td>-108 (.265)</td>
<td></td>
</tr>
<tr>
<td><strong>Schlussfolgern/Erkennen (Info 1)</strong></td>
<td>-227 (.118)</td>
<td></td>
</tr>
<tr>
<td><strong>Implizit/explicit (Info 2)</strong></td>
<td>.189 (.089)*</td>
<td>.239 (.007)**</td>
</tr>
<tr>
<td><strong>Authentischer Text</strong></td>
<td>-.146 (.334)</td>
<td></td>
</tr>
<tr>
<td><strong>Inhalt haupts. konkret</strong></td>
<td>-128 (.263)</td>
<td></td>
</tr>
<tr>
<td><strong>Inhalt teilweise abstrakt</strong></td>
<td>-192 (.330)</td>
<td></td>
</tr>
<tr>
<td><strong>Vokabular hauptsächlich häufig/einfach</strong></td>
<td>-.003 (.982)</td>
<td></td>
</tr>
<tr>
<td><strong>Vokabular teilweise ausgeweitet/selten</strong></td>
<td>-.264 (.046)**</td>
<td>-.344 (.000)***</td>
</tr>
<tr>
<td><strong>Haupts.einfache grammatikalische Strukturen</strong></td>
<td>.449 (.004)***</td>
<td>.201 (.023)**</td>
</tr>
<tr>
<td><strong>Teilweise komplexe grammatische Strukturen</strong></td>
<td>.440 (.013)***</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.352</td>
<td>.186</td>
</tr>
</tbody>
</table>

Abhängige Variable: DIF Ungarn - Deutschland; Methode: Einschluss; Zelleninhalt: β-Gewichte
* $p \leq 0.1$, ** $p \leq 0.05$, *** $p \leq 0.01$, α = 10%;
≠ = entgegen der Hypothese; kursiv = neutral

Tabelle 5.36. Regression von DIF Ungarn - Deutschland auf Anforderungsmerkmale (Englisch-Items)
In Tabelle 5.36 zeigt sich, dass mit Hilfe aller verwendeten Prädiktoren 35.2% der durch kulturelle Faktoren verursachten Varianz der Itemschwierigkeiten zwischen Ungarn und Deutschland erklärt werden können. In einem zweiten Modell werden nun diejenigen Variablen als Prädiktoren aufgenommen, die sowohl signifikant sind, als auch hinsichtlich der Richtung ihrer Gewichte den eingangs aufgestellten Hypothesen entsprechen. Ebenso werden die oben beschriebenen „neutralen“ Prädiktoren, die sich in Modell 1 als signifikant erweisen, verwendet. Mit Hilfe dieser Variablen lassen sich noch 18.6% der Varianz erklären. Es wird dann das dritte Modell mit den vier in Modell 1 signifikanten, der Testkultur entsprechenden Prädiktoren gerechnet. Hier können noch 17.1% der Varianz aufgeklärt werden. Dies entspricht dem Anteil der durch die erwarteten Stärken und Schwächen der Gruppen aufgeklärten Varianz. Dabei zeigt sich, dass vor allem die Variablen „Informationsgewinn 2“, „Vokabular teilweise erweitert / selten“ und „hauptsächliche einfache grammatische Strukturen“ eine testkulturelle Rolle zu spielen scheinen. Die Effekte erweisen sich hier als etwas stabiler als bei den Ländergruppen Deutschland-Frankreich.

<table>
<thead>
<tr>
<th>Modell 1</th>
<th>Modell 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta(sig)</td>
<td>Beta(sig)</td>
</tr>
<tr>
<td><em>Itemtyp Multiple Matching</em></td>
<td>-.050 (.212)</td>
</tr>
<tr>
<td><em>Itemtyp Ordnen</em></td>
<td>.053 (.695)</td>
</tr>
<tr>
<td><em>Itemtyp Zitieren</em></td>
<td>.246 (.017)**</td>
</tr>
<tr>
<td>Schlussfolgern/Erkennen (Info 1)</td>
<td>.015 (.924)</td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>.091 (.566)</td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>.091 (.566)</td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>.091 (.566)</td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>-.442 (.035)**</td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>-.042 (.773)</td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>.101 (.468)</td>
</tr>
<tr>
<td>Haupts.einfache grammatische Strukturen</td>
<td>.101 (.468)</td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>.167 (.367)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.277</td>
</tr>
</tbody>
</table>

*Tabelle 5.37.* Regression von DIF Ungarn - Frankreich auf Anforderungsmerkmale (Englisch-Items)
In Modell 1 von Tabelle 5.37 werden mit Hilfe der verwendeten Prädiktoren 27.7% der Varianz der Differentiellen Item Funktionen zwischen Ungarn und Frankreich aufgeklärt. Zwei der Prädiktoren sind signifikant, nämlich der Itemtyp „zitieren“ und die Variable „Inhalt teilweise abstrakt“. Diese werden in einem zweiten Modell nochmals überprüft. Bezüglich dieser beiden Ländergruppen existieren keine „neutralen“ Prädiktoren, daher ist ein Modell, das diese berücksichtigt, nicht notwendig. Im Endmodell (hier Modell 2) zeigt sich, dass dort noch 22.7% der kulturellen Varianz zwischen Ungarn und Frankreich hypothesenkonform aufgeklärt werden können, d.h. durch die aufgrund der Testkulturen erwarteten Stärken und Schwächen der Gruppen. In diesem Fall ist das Ergebnis so zu interpretieren, dass die Tatsache, dass ein Item vom Typ „Zitieren“ ist, das Item um 0.26 Logits für die ungarischen Schüler im Vergleich zu den französischen Schülern erschwert, Items mit teilweise abstraktem Inhalt diese hingegen um 0.384 Logits erleichtern (d.h. die Itemschwierigkeit um 0.384 Logits verringern).

<table>
<thead>
<tr>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta(sig)</td>
<td>Beta(sig)</td>
<td>Beta(sig)</td>
</tr>
<tr>
<td><strong>Itemtyp Multiple Matching</strong></td>
<td>-.220 (.041)**</td>
<td>-.021 (.812)</td>
</tr>
<tr>
<td>Itemtyp Ordnien</td>
<td>-.196 (.122)</td>
<td></td>
</tr>
<tr>
<td><strong>Itemtyp Zitieren</strong></td>
<td>.416 (.000)***</td>
<td>.492 (.000)***</td>
</tr>
<tr>
<td>Schlussfolger/Erkennen (Info 1)</td>
<td>.333 (.020)**</td>
<td>.216 (.014)**</td>
</tr>
<tr>
<td>Implizit/explizit (Info 2)</td>
<td>-.161 (.137)</td>
<td>.142 (.337)</td>
</tr>
<tr>
<td>Authentischer Text</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inhalt haupts. konkret</strong></td>
<td>.218 (.052)* ≠</td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>-.029 (.879)</td>
<td></td>
</tr>
<tr>
<td><strong>Vokabular hauptsächlich häufig/einfach</strong></td>
<td>.140 (.298)</td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>.367 (.005)***</td>
<td>-.109 (.224)</td>
</tr>
<tr>
<td><strong>Haupts.einfache grammatikalische Strukturen</strong></td>
<td>-.246 (.100)</td>
<td></td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>-.501 (.004)***</td>
<td>-.195 (.043)**</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.381</td>
<td>.256</td>
</tr>
</tbody>
</table>

*Tabelle 5.38. Regression von DIF Deutschland - Spanien auf Anforderungsmerkmale (Englisch-Items)


Betrachtet man im Vergleich dazu die Einzelkorrelationen zwischen den Itemeigenschaften und DIF, dann fällt auf, dass dort einige der Variablen einen signifikanten Zusammenhang zu den DIF aufweisen. In diesem Falle sind die Einzelkorrelationen hinsichtlich der Stärken und Schwächen der beiden Gruppen vermutlich als Informationsquellen vorzuziehen.

Bezüglich der beiden Gruppen Spanien und Ungarn (Tabelle 5.40) können in Modell 1 mit Hilfe aller verwendeter Prädiktoren 34.5% der DIF-Varianz aufgeklärt werden. Allerdings fällt hier auf, dass nur ein einziger Prädiktor ein signifikantes Beta-Gewicht aufweist. Mit diesem Prä-
Tabelle 5.39. Regression von DIF Frankreich - Spanien auf Anforderungsmerkmale (Englisch-Items)

diktor können 17.6% der Varianz aufgeklärt werden. Allerdings ist dieser Prädiktor „neutral”,
das heißt, es zeigt sich ein signifikanter Einfluss des Prädiktors auf die abhängige Variable,
obgleich sich die Gruppen hinsichtlich dieser Variablen in den Testkulturen nicht signifikant un-
terscheiden. Dieser Prädiktor widerspricht zwar somit nicht den oben aufgestellten Hypothesen,
allerdings ist er auch nicht eindeutig zu interpretieren. Möglicherweise hängt dieses Ergebnis
mit einer fehlerhaften Einordnung der Items hinsichtlich ihrer Anforderungsmerkmale durch die
Experten zusammen, weshalb kein Unterschied in der Testkultur festgestellt werden konnte.

Fazit:
Inwieweit Differentielle Item Funktionen mit Hilfe der testkulturellen Merkmale vorhergesagt
werden können, scheint bezüglich der Englisch-Items teilweise von der jeweiligen Länder-
paarung mit beeinflusst zu sein: Paarungen betreffend, die sich aus den Ländern Deutschland,
Frankreich und Ungarn konstituieren (R² = .163 bis R² = .227), scheinen besser geeignet zu sein
als Paarungen, die die spanische Stichprobe mit einschließen (R² = .0 bis R² = .041). Ob dies
lediglich auf die spezifische, im Rahmen der EBAFLS Studie verwendete spanische Stichprobe

182
Tabelle 5.40. Regression von DIF Spanien - Ungarn auf Anforderungsmerkmale (Englisch-Items)

zurückführbar ist, oder ob sich diese Beobachtung auf spanische Schüler generalisieren lässt, muss in weiterführender Forschung genauer betrachtet werden. Auch sind in einigen Fällen die Effekte über unterschiedliche Modelle hinweg stabiler als in anderen.

**Ergebnisse der Deutsch-Items**

Im Folgenden werden die Ergebnisse der Regression von paarweisen DIF bei Deutsch-Items auf kognitiv-linguistische Anforderungsmerkmale der Items berichtet. Die Ergebnisse jedes Prädiktors sind immer im Vergleich zu einer Kontrastvariable zu interpretieren. Das ist bei der Variablen „Itemtyp“ die Ausprägung „Multiple Choice“, bei den anderen Variablen handelt es sich dabei immer um die leichteste Ausprägung. So müssen beispielsweise die Ergebnisse bei „Vokabular“ immer im Vergleich zu der Ausprägung „ausschließlich einfaches/häufiges Vokabular“ interpretiert werden. Es werden auch hier, wie bereits für die Englisch-Items, für jede Länderpaarung verschiedene Modelle dargestellt: zunächst jeweils ein Gesamtmodell, welches alle Prädiktoren beinhaltet; danach ein zweites Mo-
5.3 Ergebnisse Fragenkomplex 3

dell, das Prädiktoren beinhaltet, die der Richtung nach aufgrund der Testkulturen aufgestellten Hypothesen entsprechen, bzw. die Prädiktoren beinhaltet, die hier als „neutral“ interpretiert werden. Neutral bedeutet auch hier, dass ein Prädiktor ein signifikantes Beta-Gewicht aufweist, sich die Testkulturen zweier Länder jedoch hinsichtlich dieser Variablen nicht signifikant unterscheiden.


Die Tabelle 5.41 beinhaltet drei multiple Regressionen von DIF zwischen Frankreich und Ungarn auf Item-Anforderungsmerkmale. Dabei erklärt das erste Modell, welches alle Prädiktoren enthält, 56.7% der Varianz der Differentiellen Item Funktionen. Allerdings entspricht die Richtung des Beta-Gewichts der Variablen „Inhalt hauptsächlich konkret“ nicht der Hypothese; daher wird diese im zweiten Modell nicht mehr einbezogen. Dieses beinhaltet nur noch signifikante Prädiktoren, die nicht der Richtung nach aufgrund der Testkulturen gemachten Annahmen widersprechen. Da das zweite Modell mit der Variablen „Authentischer Text“ noch einen signifikanten Prädiktor beinhaltet, der nicht eindeutig auf die Testkultur zurückführbar ist, wird der aufgrund dieses Modells erklärbare Anteil der Varianz als die Obergrenze der aufgrund der Testkultur-Indikatoren erklärbaren Varianz interpretiert. Das letzte Modell beinhaltet ausschließlich signifikante Prädiktoren, die auch aufgrund der Testkultur aufgestellten Hypothesen entsprechen. Es zeigt sich, dass sich aufgrund der Testkultur bei der Länderpaarung Frankreich-Ungarn zwischen 45.7 % und 34.7% der DIF-Varianz erklären lassen. Dabei sind die Beta-Koeffizienten wie folgt zu interpretieren: Die Tatsache, dass ein Item dem Itemtyp „Multiple Matching“ entspricht, erhöht die Itemschwierigkeit für die französische Schülergruppe im Vergleich zu den ungarischen
Tabelle 5.41. Regression von DIF Frankreich - Ungarn auf Anforderungsmerkmale (Deutsch-Items)

Schülern um 0.52 Logits, und die Tatsache, dass ein Item hauptsächlich einfache grammatische Strukturen besitzt (im Vergleich zu ausschließlich einfachen Strukturen, da dies die Kontrastvariable ist) erhöht die Itemschwierigkeit um 0.337 Logits.

Bezüglich der Varianz Differentieller Item Funktionen zwischen Frankreich und den Niederlanden (Tabelle 5.42 lassen sich in Modell 1 unter Verwendung aller Prädiktoren 25.6% der Varianz aufklären. Werden allerdings die nicht-signifikanten bzw. den Hypothesen nicht entsprechenden Variablen nicht mehr in das Modell aufgenommen, lassen sich nur noch 8.3% der Varianz erklären und ohne die „neutralen“ Prädiktoren nur noch 3.1%. Die Effekte sind über die Modelle hinweg insgesamt nicht stabil. Der relativ starke Verlust bezüglich des Anteils der aufgeklärten
### Ergebnisse Fragenkomplex 3

<table>
<thead>
<tr>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itemtyp Banked Multiple Choice</td>
<td>-.375 (.015)**</td>
<td>-.162 (.202)</td>
</tr>
<tr>
<td>Itemtyp Richtig-Falsch</td>
<td>-.850 (.002)**</td>
<td>-.275 (.053)*</td>
</tr>
<tr>
<td>Itemtyp Multiple Matching</td>
<td>-.122 (.653)</td>
<td></td>
</tr>
<tr>
<td>Itemtyp Kurzantwort</td>
<td>-.263 (.277)</td>
<td></td>
</tr>
<tr>
<td>Itemtyp Lückentext</td>
<td>-.157 (.334)</td>
<td></td>
</tr>
<tr>
<td>Erkennen/Schlussfolgern (Info 1)</td>
<td>-.153 (.483)</td>
<td></td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>-.058 (.817)</td>
<td></td>
</tr>
<tr>
<td>Hauptidee/Detail (Info 3)</td>
<td>.221 (.384)</td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>-.484 (.201)**</td>
<td>-.109 (.393)</td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>-.177 (.366)</td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>-.271 (.209)</td>
<td></td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>1.086 (.007)*** ≠</td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>1.222 (.006)***</td>
<td>.034 (.784)</td>
</tr>
<tr>
<td>Vokabular ausgeweitet/selten</td>
<td>.596 (.348)</td>
<td></td>
</tr>
<tr>
<td>Haupts.einfache grammatikalische Strukturen</td>
<td>.607 (.013)**</td>
<td>.269 (.054)*</td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>.926 (.034)**</td>
<td>.133 (.349)</td>
</tr>
<tr>
<td>Komplexe grammatische Strukturen</td>
<td>.538 (.317)</td>
<td></td>
</tr>
</tbody>
</table>

| R² | .256 | .083 | .031 |

Abhängige Variable: DIF Frankreich - Niederlande; Methode: Einschluß; Zelleninhalt: β-Gewichte

* p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%
≠ entgegen der Hypothese; kursiv = neutral

__Tabelle 5.42. Regression von DIF Frankreich - Niederlande auf Anforderungsmerkmale (Deutsch-Items)__

Varianz durch die Herausnahme nicht-signifikanter Prädiktoren spricht dafür, dass möglicherweise auch hier ein Multikollinearitätsproblem vorhanden ist. Auch existieren kaum Einzelkorrelationen (siehe 5.3.1) zwischen den DIF der beiden Länder und den hier verwendeten Prädiktoren. Das spricht dafür, dass bezüglich dieser beiden Länder möglicherweise noch andere, hier nicht mit erhobene kulturelle Variablen für die Entstehung von DIF eine Rolle spielen, was jedoch im Rahmen dieser Arbeit leider nicht überprüft werden kann. Ferner existieren Beta-Gewichte größer 1, was auf die Anwesenheit auf Suppressionseffekte hinweist.
### Ergebnisse 5.3 Ergebnisse Fragenkomplex 3

<table>
<thead>
<tr>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta(Sig.)</td>
<td>Beta(Sig.)</td>
<td>Beta(Sig.)</td>
</tr>
<tr>
<td><strong>Itemtyp Banked Multiple Choice</strong></td>
<td>-2.77 (.030)** ≠</td>
<td></td>
</tr>
<tr>
<td><strong>Itemtyp Richtig-Falsch</strong></td>
<td>-0.517 (.023)**</td>
<td>-0.147 (.240)</td>
</tr>
<tr>
<td><strong>Itemtyp Multiple Matching</strong></td>
<td>0.569 (.014)**</td>
<td>0.394 (.006)**</td>
</tr>
<tr>
<td><strong>Itemtyp Kurzantwort</strong></td>
<td>-0.361 (.074)* ≠</td>
<td></td>
</tr>
<tr>
<td><strong>Itemtyp Lückentext</strong></td>
<td>-0.095 (.477)</td>
<td></td>
</tr>
<tr>
<td><strong>Schlussfolgern/Erkennen (Info 1)</strong></td>
<td>0.249 (.171)</td>
<td></td>
</tr>
<tr>
<td><strong>Implizit/explizit (Info 2)</strong></td>
<td>0.219 (.294)</td>
<td></td>
</tr>
<tr>
<td><strong>Hauptidee/Detail (Info 3)</strong></td>
<td>0.306 (.149)</td>
<td></td>
</tr>
<tr>
<td><strong>Authentischer Text</strong></td>
<td>-0.579 (.001)**</td>
<td>-0.313 (.018)**</td>
</tr>
<tr>
<td><strong>Inhalt haupts. konkret</strong></td>
<td>-0.272 (.096)* ≠</td>
<td></td>
</tr>
<tr>
<td><strong>Inhalt teilweise abstrakt</strong></td>
<td>-0.355 (.049)** ≠</td>
<td></td>
</tr>
<tr>
<td><strong>Vokabular hauptsächlich häufig/einfach</strong></td>
<td>0.789 (.018)** ≠</td>
<td></td>
</tr>
<tr>
<td><strong>Vokabular teilweise ausgeweitet/selten</strong></td>
<td>0.104 (.006)**</td>
<td>0.130 (.283)</td>
</tr>
<tr>
<td><strong>Vokabular ausgeweitet/selten</strong></td>
<td>-0.270 (.608)</td>
<td></td>
</tr>
<tr>
<td><strong>Haupts. einfache grammatische Strukturen</strong></td>
<td>0.762 (.000)**</td>
<td>0.381 (.001)**</td>
</tr>
<tr>
<td><strong>Teilweise komplexe grammatische Strukturen</strong></td>
<td>1.181 (.001)**</td>
<td>0.382 (.009)**</td>
</tr>
<tr>
<td><strong>Komplexe grammatische Strukturen</strong></td>
<td>1.429 (.002)**</td>
<td>0.249 (.054)*</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.487</td>
<td>0.369</td>
</tr>
</tbody>
</table>

*Abhängige Variable: DIF Frankreich - Schweden; Methode: Einschluß; Zelleninhalt: β-Gewichte

* $p < 0.1, \alpha = 10\%$; ** $p < 0.05, \alpha = 5\%$; *** $p < 0.01, \alpha = 1\%$

≠ = entgegen der Hypothese; kursiv = neutral

#### Tabelle 5.43. Regression von DIF Frankreich - Schweden auf Anforderungsmerkmale (Deutsch-Items)

Hinsichtlich der Varianz Differentieller Item Funktionen zwischen Frankreich und Schweden (Tabelle 5.43) werden mit Hilfe aller Prädiktoren 48.7% aufgeklärt. Wie ersichtlich wird, entsprechen einige der signifikanten Prädiktoren hinsichtlich ihrer Richtung nicht den Testkulturen. In diesen Fällen überlagern möglicherweise andere, nicht mit erfasste Einflüsse die der Testkultur, oder aber die Items wurden von den Experten nicht korrekt eingeordnet. Ferner zeigt sich im Vergleich zu den Einzelkorrelationen, dass die sich hier als den Hypothesen entgegengesetzt darstellenden Prädiktoren „Inhalt hauptsächlich konkret“, „Inhalt teilweise abstrakt“, und „Short Answer“ keine signifikanten Einzelkorrelationen mit DIF aufweisen. Die Variable „Vokabular häufig einfach“ weist eine signifikant negative Einzelkorrelation auf, was der Hypothese ent-

Anhand von Modell 1 aus Tabelle 5.44 lassen sich mit Hilfe aller Prädiktoren 36.1% der Varianz der Differentiellen Item Funktionen zwischen den Niederlanden und Schweden aufklären. Werden die Prädiktoren herausgenommen, die nicht signifikant sind, werden noch 20.4% der Varianz aufgeklärt. Da dieses Modell wieder „neutrale“ Prädiktoren beinhaltet, ist dies als die Obergrenze der anhand dieser Prädiktoren erklärbaren Varianz zu interpretieren. Die Untergrenze bildet hier Modell 3, in dem noch 12.7% der Varianz erklärt werden können. Hinsichtlich der auf die Testkulturen zurückführbaren relativen Stärken und Schwächen der beiden Gruppen zeigt sich, dass die beiden Variablen „Multiple Matching“ und „Schlussfolgern“ die Itemschwierigkeit für die niederländischen Schüler im Vergleich zu der schwedischen Schülergruppe erhöht.

Tabelle 5.45 zeigt, dass sich mit Hilfe aller Prädiktoren 46.9% der Varianz der differentiellen Item Funktionen zwischen Ungarn und den Niederlanden erklären lassen. Es zeigen sich zwei den Hypothesen entgegengesetzte Regressionskoeffizienten, „Lückentext“ und „Vokabular häufig/einfach“. Auch hier zeigt ein Abgleich mit den Einzelkorrelationen, dass dort keine signifi-
Ergebnisse

5.3 Ergebnisse Fragenkomplex 3

<table>
<thead>
<tr>
<th>Itemtyp</th>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beta (Sig.)</td>
<td>Beta (Sig.)</td>
<td>Beta (Sig.)</td>
</tr>
<tr>
<td>Banked Multiple Choice</td>
<td>.185 (.190)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richtig-Falsch</td>
<td>.514 (.040)**</td>
<td>.298 (.007)**</td>
<td></td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>.596 (.020)**</td>
<td>.299 (.006)**</td>
<td>.260 (.018)**</td>
</tr>
<tr>
<td>Kurzantwort</td>
<td>-.007 (.977)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lückentext</td>
<td>.094 (.531)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schlussfolger/Fn Erkennen (Info 1)</td>
<td>.370 (.070)*</td>
<td>.272 (.015)**</td>
<td>.179 (.10)*</td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>.114 (.624)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haupt/Detail (Info 3)</td>
<td>.006 (.978)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>-.059 (.759)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>-.028 (.876)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>.008 (.968)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>-.546 (.138)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>-.513 (.206)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular ausgeweitet/selten</td>
<td>-.870 (.141)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haupts./einfache grammatischen Strukturen</td>
<td>-.044 (.841)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>-.053 (.894)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komplexe grammatische Strukturen</td>
<td>.573 (.251)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>.361</td>
<td>.204</td>
<td>.127</td>
</tr>
</tbody>
</table>

Abhängige Variable: DIF Niederlande - Schweden; Methode: Einschluß; Zelleninhalt: β-Gewichte
* p < 0.1; α = 10%; ** p < 0.05; α = 5%; *** p < 0.01; α = 1%
≠ = entgegen der Hypothese; kursiv = neutral

Tabelle 5.44. Regression von DIF Niederlande - Schweden auf Anforderungsmerkmale (Deutsch-Items)

kanten Zusammenhänge existieren. Es handelt sich hierbei also möglicherweise auch um Suppressionseffekte und methodische Artefakte. Dies trifft allerdings auch auf die beiden hypothesenkonformen Prädiktoren „Vokabular erweitert / selten” und „Richtig-Falsch” zu. Ersterer ist in Modell zwei jedoch nicht mehr signifikant, und Letzterer könnte dadurch erklärt werden, dass die Variable „Multiple Choice”, die eine signifikante Einzelkorrelation aufweist, hier als Kontrastvariable dient. In dem zweiten Modell, welches wieder die hypotheseskonformen und die „neutralen” Prädiktoren beinhaltet, können mit Hilfe dieser beiden Prädiktoren noch 26.8% der Varianz erklärt werden. In einem dritten Modell sind ausschließlich die hypotheseskonformen Prädiktoren mit aufgenommen, hier können noch 23.6% der Varianz erklärt werden. Die Ergebnis-
Tabelle 5.45. Regression von DIF Ungarn - Niederlande auf Anforderungsmerkmale (Deutsch-Items)

<table>
<thead>
<tr>
<th>Itemtyp</th>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banked Multiple Choice</td>
<td>-.351(.007)**</td>
<td>-.196(.061)*</td>
<td></td>
</tr>
<tr>
<td>Richtig-Falsch</td>
<td>-.707(.002)**</td>
<td>-.255(.015)**</td>
<td>-.217 (.036)**</td>
</tr>
<tr>
<td>Multiple Matching</td>
<td>-.802(.001)**</td>
<td>-.511(.000)**</td>
<td>-.506 (.000)**</td>
</tr>
<tr>
<td>Kurzantwort</td>
<td>-.261 (.060)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lückentext</td>
<td>-.260 (.060) * ≠</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schlussfolgern/Erinnern</td>
<td>-.265 (.154)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implizit/explicit (Info 2)</td>
<td>.086 (.684)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hauptidee/Detail (Info 3)</td>
<td>-.033 (.879)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>-.108(.535)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>.154(.352)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>-.163 (.371)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>.855(.012)** ≠</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>.749(.045)**</td>
<td>-.062(.573)</td>
<td>-.119 (.273)</td>
</tr>
<tr>
<td>Vokabular ausgeweitet/selten</td>
<td>.791(.143)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haupts.einfache grammatisch Strukturen</td>
<td>.256(.206)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>.243(.505)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komplexe grammatische Strukturen</td>
<td>.148(.743)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.469</td>
<td>.268</td>
<td>.236</td>
</tr>
</tbody>
</table>

Abhängige Variable: DIF Ungarn - Niederlande; Methode: Einschluß; Zeileninhalt: $\beta$-Gewichte
* $p < 0.1, \alpha = 10\%$; ** $p < 0.05, \alpha = 5\%$; *** $p < 0.01, \alpha = 1\%$
≠ entgegen der Hypothese; kursiv = neutral

Die Ergebnisse werden so interpretiert, dass zwischen 26.8% und 23.6% der Varianz auf durch die Testkultur verursachte differentielle Lerngelegenheiten zurückzuführen sind. Für die ungarischen Schüler bedeutet hier die Tatsache, dass ein Item ein Richtig-Falsch-Item ist, dass das Item um etwa 0.2 Logit leichter wird als es für die niederländischen Schüler der Fall ist; ist das Item vom Typ „Multiple Matching“, sind es 0.506 Logits.

Anhand der Prädiktoren können 32.9% der Varianz der differentiellen Item Funktionen zwischen Schweden und Ungarn (Tabelle 5.46) erklärt werden. Entfernt man die nicht-signifikanten Prädiktoren, sind es nur noch 10.8% der Varianz, bzw. 5.5% ohne den „neutralen“ Prädiktor „Komplexe grammatische Strukturen“. Schon unter 5.3.1, bei den korrelativen Zusammenhän-
<table>
<thead>
<tr>
<th>Itemtyp Banked Multiple Choice</th>
<th>Modell 1</th>
<th>Modell 2</th>
<th>Modell 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta(Sig.)</td>
<td>-.261(.072)*</td>
<td>-.160(.134)</td>
<td>-.166(.095)*</td>
</tr>
<tr>
<td>Itemtyp Richtig-Falsch</td>
<td>-.339(181)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itemtyp Multiple Matching</td>
<td>-.368(156) ≠</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Itemtyp Kurzantwort</td>
<td>-.383(098)*</td>
<td>-.065(.683)</td>
<td>.154(.121)</td>
</tr>
<tr>
<td>Itemtyp Lückentext</td>
<td>-.249(108)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schlussfolger/Erkennen(Info 1)</td>
<td>.107(606)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implizit/explicit(Info 2)</td>
<td>.275(.250)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hauptidee/Detail (Info 3)</td>
<td>-.038(873)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentischer Text</td>
<td>-.078(690)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt haupts. konkret</td>
<td>.184(322)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inhalt teilweise abstrakt</td>
<td>-.223(276)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular hauptsächlich häufig/einfach</td>
<td>.509(177)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular teilweise ausgeweitet/selten</td>
<td>.401(333)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vokabular ausgeweitet/selten</td>
<td>-.010(987)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haupts.einfache grammaticielle Strukturen</td>
<td>.309(175)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teilweise komplexe grammatische Strukturen</td>
<td>.279(.495)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Komplexe grammatische Strukturen</td>
<td>.976(.061)*</td>
<td>.321(.046)**</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>.329</td>
<td>.108</td>
<td>.055</td>
</tr>
</tbody>
</table>

* p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%
≠ entgegen der Hypothese; kursiv = neutral

Tabelle 5.46. Regression von DIF Schweden - Ungarn Anforderungsmerkmale (Deutsch-Items)

gen, zeigten sich kaum signifikante Zusammenhänge zwischen den DIF der beiden Länder und den Testkultur-Indikatoren. Dies ist möglicherweise durch eine Einschränkung der Varianzen der DIF zu erklären, oder aber durch nicht mit erfasste kulturelle Indikatoren, die bei diesen beiden Ländern eine Rolle bei der Entstehung Differentieller Item Funktionen spielen.

Fazit:
Auch bei den Deutsch-Items scheint es teilweise von der jeweiligen Länderpaarung abzuhängen, ob sich die kulturell bedingte Varianz mit Hilfe von Testkultur-Indikatoren erklären lässt. Hier lässt sich jedoch nicht, wie es bezüglich der Englisch-Items der Fall war, eine Stichprobe herauskristallisieren, bei der möglicherweise deren Zusammensetzung oder andere gruppenspezifische
Faktoren eine Rolle spielen. Es muss hier davon ausgegangen werden, dass noch weitere, kulturspezifische Variablen beim Entstehen von DIF eine Rolle spielen, die jedoch im Rahmen der vorliegenden Arbeit nicht überprüft werden können.

**Hypothese 3b:** Differentielle Item Funktionen, d.h. die kulturell bedingte Varianz zwischen den Ländern, können mit Hilfe von Indikatoren nationaler Testkulturen teilweise erklärt werden. Die Richtung der Regressionsgewichte sollte den aufgrund der Analyse der Items erwarteten Stärken und Schwächen in den nationalen Testkulturen entsprechen.

Diese Hypothese kann, abhängig von den analysierten Länderpaarungen, insgesamt teilweise beibehalten werden. Ein Teil der Varianz der Differentiellen Item Funktionen zwischen Ländern kann immer mit den Testkultur-Indikatoren aufgeklärt werden, die Höhe des $R^2$ unterscheidet sich jedoch je nach Länderpaarung deutlich, und zwar in den jeweiligen Endmodellen zwischen $R^2 = 0.031$ und $R^2 = 0.44$. Die spricht zum einen dafür, dass noch weitere kulturelle Indikatoren vermutlich eine Rolle bei der Entstehung differentieller Item Funktionen spielen. Zum anderen weisen die Ergebnisse darauf hin, dass auch die Testkulturen durchaus eine Rolle spielen. Die Bedeutung dieser Ergebnisse für die Beantwortung der Hauptfragestellung wird unter Punkt 6 diskutiert.

**Hypothese 3c:** Die Regressionsgewichte entsprechen hinsichtlich ihrer Richtung den aufgrund der Testkulturen erwarteten Stärken und Schwächen der Gruppen.


Im folgenden Teil der vorliegenden Arbeit werden die Ergebnisse kurz zusammengefasst und interpretiert. Auch von den Hypothesen abweichende Ergebnisse sowie mögliche Gründe dafür werden diskutiert.
6. Interpretation der Ergebnisse, Beantwortung der Hauptfragestellung und Diskussion


6.1. Zusammenfassung und Interpretation der Ergebnisse


6.1.1. Zusammenfassung und Interpretation der Ergebnisse zu Fragenkomplex 1: „Voraussetzungen und Skalierbarkeit“

Zusammenfassung und Interpretation der Ergebnisse


Die aus diesen Analysen stammenden Item-Schwierigkeitsparameter wurden für die Analysen in Fragenkomplex 2 verwendet.

Die zweite notwendige Voraussetzung für die Beantwortung der Hauptfragestellung war das Vorhandensein von Differentiellen Item Funktionen zwischen den Ländern. Dies wurde im Rahmen von Frage 1b überprüft. Zweck war hier zum einen, zu überprüfen, ob auch unter Zugrundelegung des Rasch-Modells signifikante DIF-Parameter existieren, zum anderen sollte eine genügend große Anzahl (in diesem Fall 35%) von signifikanten DIF-Parametern sicherstellen, dass die Einschränkung der Varianz der Itemschwierigkeiten zwischen den Ländern begrenzt ist. Bezüglich der Englisch-Items zeigte sich, dass bei jeweils paarweisen DIF-Analysen zwischen 44.2% und 66.4% der Items signifikante DIF-Parameter aufwiesen. In Bezug auf die Deutsch-Items traf dies, abhängig von der jeweils betrachteten Länder-Paarung, auf 39.6% bis 59.4% der Items zu. Auch hier kann also die Annahme, dass ein großer Anteil der Items signifikante DIF-Parameter aufweist, als bestätigt angesehen werden. Die aus diesen Analysen stammenden DIF-Parameter wurden daher für die Analysen in Fragenkomplex 3 verwendet. Dies schließt aufgrund

Das Vorhandensein von DIF ist dahingehend zu interpretieren, dass dieselben Items in unterschiedlichen Ländern vermutlich teilweise unterschiedliche oder zusätzliche Dimensionen messen. DIF ist also ein Hinweis darauf, dass diese Items nicht oder nur teilweise dasselbe messen. Ob und inwieweit dies zum Teil auf unterschiedliche Testkulturen zurückführbar ist, wird im Rahmen der Hauptfragestellung beantwortet.


Signifikante Unterschiede wurden hier als ein Hinweis auf die Unterschiedlichkeit zweier Testkulturen hinsichtlich des jeweiligen Merkmals gewertet. Es zeigte sich, dass die Länder unterschiedliche Testkultur-Profile aufweisen, was für die Richtigkeit der Annahme spricht, dass unterschiedliche Testkulturen in den Ländern existieren. Dies deutet auf unterschiedliche Schwerpunkte hinsichtlich der Verwendung schwierigkeitsbestimmender Itemmerkmale bei der Konstruktion von fremdsprachlichen Testitems hin. Basierend auf diesen unterschiedlichen Testkultur-Profilen wurden Hypothesen hinsichtlich der zu erwartenden Stärken und Schwächen der Gruppen bei der Beantwortung von Items mit bestimmten schwierigkeitsdeterminierenden Merkmalen aufgestellt. Diese wurden zur Erklärung Differentieller Item Funktionen in Fragenkomplex 3 benötigt.

Insgesamt lässt sich hinsichtlich dieses Bereichs von Fragen konstatieren, dass die Voraussetzungen für eine Bearbeitung der Hauptfragestellung durch die Rasch-Skalierbarkeit der Items innerhalb der Länder, das Vorhandensein kulturell bedingter Varianz zwischen den Ländern sowie das Vorhandensein unterschiedlicher Testkultur-Profiles und somit unterschiedlicher zu erwartender Stärken und Schwächen der Gruppen bei der Beantwortung von Items mit bestimmten Merkmalen gegeben sind.

Es muss allerdings darauf hingewiesen werden, dass die Rasch-Schwierigkeitsparameter zunächst nur für eine vergleichbare Stichprobe innerhalb der jeweiligen Länder gültig sind, das heißt, jeweils für Schülerinnen und Schüler, die sich nach Stand des schulischen Curriculums etwa auf dem GERS-Niveau B1 befinden sollten. Da die Stichproben in den meisten Ländern außerdem nicht repräsentativ erhoben wurden, sondern es sich zumeist um Convenience-Stichproben handelte, sollten die Parameter nochmals anhand einer repräsentativen Stichprobe überprüft werden.


6.1.2. Zusammenfassung und Interpretation der Ergebnisse zu Fragenkomplex 2: „Erklärung der Itemschwierigkeiten innerhalb der Länder”

Im Folgenden werden Ergebnisse der einzelnen Fragestellungen zu diesem Bereich zusammengefasst und interpretiert. Dabei handelt es sich zum einen um die Ergebnisse der Korrelationsanalysen zwischen Itemschwierigkeit und den schwierigkeitsbestimmenden Merkmalen innerhalb der unterschiedlichen Länder (Frage 2a), sowie um deren Vergleichbarkeit über die Länder hinweg (Frage 2b). Zum anderen handelt es sich um die Ergebnisse der multiplen Regressionsanalysen, in denen die Itemschwierigkeit innerhalb der Länder jeweils mit Hilfe der schwierigkeitsbestimmenden Itemmerkmale vorhergesagt wurde (Frage 2c / 2d). Die Verwendung die-

In einem ersten Schritt wurden für jedes Land Pearson Produkt-Moment-Korrelationskoeffizienten für den Zusammenhang zwischen den in Frage 1a berechneten Itemschwieigkeitsparametern und den schwierigkeitsdeterminierenden Itemmerkmalen berechnet. Dies hatte zum Zweck, zunächst einen Überblick über einzelne Zusammenhänge zwischen den Itemmerkmalen und den Itemschwierigkeiten innerhalb der Länder zu erhalten, frei von Multikollinearitätsproblemen, wie sie bei Regressionsanalysen häufig zu finden sind. Für die Englisch-Items liegen die signifikanten Korrelationskoeffizienten zwischen r = -.30 (p ≤ 0.01) und r = .431 (p ≤ 0.01) und sind damit als niedrig bis moderat einzustufen. Dabei bedeutet ein negativer Koeffizient, dass die Anwesenheit eines Merkmals tendenziell mit einer niedrigen Itemschwierigkeit einhergeht. Ein positiver Koeffizient hingegen bedeutet das Einhergehen eines Merkmals mit einer hohen Itemschwierigkeit. Die einzelnen Ergebnisse sind unter 5.2 einzusehen.


Entgegen der Erwartung zeigte sich auch eine positive Korrelation zwischen Itemschwierigkeit und „ausschließlich einfachen grammatischen Strukturen“ bei den Englisch-Items. Auf eine mögliche Interpretation dieses Ergebnisses wird weiter unten eingegangen. Ferner gibt es in sämtlichen Ländern keinerlei signifikanten korrelativen Zusammenhang zwischen der Itemschwierig-
Diskussion

6.1 Zusammenfassung und Interpretation der Ergebnisse

keit und den Variablenausprägungen „Multiple Matching“, „Ordnen“, „Informationsgewinn 2“, „Inhalt hauptsächlich konkret“, „Vokabular einfach / häufig“ und „teilweise komplexe grammatische Strukturen“.


Die Ergebnisse für die Deutsch-Items ähneln denen der Englisch-Items: Die Korrelationen liegen hier zwischen r = -0.444 (p ≤ .01) und r = 0.505 (p ≤ .01) und somit gleichfalls im niedrigen bis moderaten Bereich. Auch hier ähneln sich die Korrelationen der Länder. So existiert in allen Ländern ein positiver Zusammenhang zwischen der Itemschwierigkeit und den Variablen „Richtig - Falsch“, „Kurzantwort“, „Informationsgewinn 1“ (Schlussfolgern), „Informationsgewinn 2“ (Implizit), „erweitertes / seltenes Vokabular“ und „komplexe grammatische Strukturen“. Mit einer niedrigen Itemschwierigkeit gehen hingegen die Variablen „ausschließlich konkreter Inhalt“, „hauptsächlich häufiges Vokabular“ und „ausschließlich einfache grammatische Strukturen“ einher. Im Gegensatz zu den Englisch-Items weisen bei den Deutsch-Items ferner teilweise auch die Variablen, die sich auf die Informationsgewinnung beziehen, signifikante Zusammenhänge mit der Itemschwierigkeit auf.

Der Vergleich der Korrelationen der unterschiedlichen Länder weist auch bei den Deutsch-Items darauf hin, dass diese sich nicht signifikant voneinander unterscheiden. Insgesamt sprechen die Ergebnisse bezüglich der Korrelationen zwischen Itemschwierigkeiten und Merkmalen innerhalb der Länder dafür, dass sowohl das Instrument „Dutch Grid“ als auch die dort verwendeten Merkmalsausprägungen mit deren zunehmendem Schwierigkeitsgrad zum großen Teil für die Analyse der Itemschwierigkeit geeignet zu sein scheinen.
Im Rahmen der Fragen 2c und 2d wurde dann per multipler linearer Regression innerhalb der verschiedenen Länder die Itemschwierigkeit (AV) mit Hilfe der schwierigkeitsbestimmenden Itemmerkmale (UV) bzw. deren Abstufungen vorhergesagt. Hier zeigte sich, dass der Anteil der erklärten Varianz $R^2$ bei den Englisch-Items nach Ausschluss aller nicht-signifikanten Prädiktoren zwischen $R^2 = .227$ und $R^2 = .367$ lag. Der Anteil der aufgeklärten Varianz liegt für die Deutsch-Items zwischen $R^2 = .208$ und $R^2 = .494$.

Wie auch bereits bei den Korrelationskoeffizienten der Englisch-Items zeigte sich, dass die Merkmale, welche die Art der zu erfassenden Information (Variablen Informationsgewinnung 1-3) abbilden, kaum zur Varianzerklärung beitragen, und keine signifikanten beta-Gewichte aufwiesen. Eine mögliche Erklärung wäre hier, dass die Rater Probleme damit haben, die Items bezüglich dieser Kategorien korrekt zuzuordnen. Dafür spricht auch, dass im Rahmen der erneuten Einordnung der Items die Rater teilweise äußerten, Probleme bezüglich der Einordnung von Items in diese Kategorien zu haben, da der Unterschied zwischen den Ausprägungen der Merkmale nicht eindeutig sei. Auch die Authentizität des Textes spielte hier keine Rolle für die Varianzaufklärung. Möglicherweise fehlt hier im Rahmen des „Dutch Grid“ eine ausreichende Beschreibung und Definition dieser Merkmale.


Wie im Rahmen der Ergebnisdarstellung bereits angemerkt wurde, finden sich in einigen der Modelle standardisierte Regressionskoeffizienten größer als eins, was in der Regel auf Suppressionseffekte hindeutet (Smith, Ager & Williams, 1992; Kline, 2005). Das bedeutet, Variablen, die nicht unbedingt einen hohen Zusammenhang zum Kriterium aufweisen, unterdrücken irrelevante Varianz, was zu höheren Beta-Gewichten führt (Bortz, 2005). So weist beispielsweise die Variable „Vokabular teilweise selten/erweitert“ in Frankreich ein Beta-Gewicht von 1.105, jedoch nur einen nicht-signifikanten Korrelationskoeffizienten von \( r = -.07 \). Die Variable „Vokabular selten/erweitert“ weist einen Koeffizienten von Beta= 1.359, und einen Korrelationskoeffizienten von \( r = .505 \) auf. Beide Male sind die Regressionskoeffizienten deutlich höher als die Einzelkorrelationen. Dies trifft im Übrigen auf alle Beta-Koeffizienten größer 1 zu. Ferner lassen sich Beta-Koeffizienten dieser Größe primär bei den beiden Variablen „Vokabular“ und „Grammatik“ finden. Insgesamt deuten dies auf die Anwesenheit von Suppressionsvariablen bezüglich dieser beiden Merkmale hin. Ferner weisen auch diese beiden Merkmale eine relativ hohe Interkorrelation zwischen \( r=-.442 \) (\( p \leq .01 \)) und \( r = .695 \) (\( p \leq .01 \)) auf.


6.1.3. Zusammenfassung und Interpretation der Ergebnisse zu Fragenkomplex 3: „Erklärung von differentiellen Item Funktionen“

Im Folgenden werden die Ergebnisse des dritten Fragenkomplexes, der sich mit der Erklärung der Differentiellen Item Funktionen befasst, zusammengefasst und interpretiert. In einem ersten Schritt (Frage 3a) wurden dabei Korrelationen zwischen den aus Frage 1b stammenden, paarweise berechneten Differentiellen Item Funktionen (bzw. DIF-Parametern) und den schwierigkeitsbestimmenden Itemmerkmalen berechnet. In einem zweiten Schritt (Fragen 3b und 3c) wurden in multiplen linearen Regressionen die DIF-Parameter als abhängige Variable mit Hilfe der Itemmerkmale vorhergesagt. Auch die für diese Fragestellungen verwendeten Methoden lehnen sich an die Arbeiten von Scheuneman und Gerritz (1990) sowie von Klieme und Baumert (2001) an.

Für die Beantwortung von Frage 3a wurden die im Rahmen von Frage 1b erhaltenen DIF-Parameter verwendet. Es existierten also DIF-Parameter für jede Paarung von Ländern innerhalb einer getesteten Sprache, was auf insgesamt jeweils sechs Paarungen von DIF-Parametern pro Sprache hinauslief. Diese Parameter wurden jeweils mit den schwierigkeitsbestimmenden Itemmerkmalen korreliert.
Im Rahmen der Korrelationen wurde als zusätzliche Variable außerdem die Itemherkunft eingeführt. Diese hatte die Funktion einer Screening-Variable, mit deren Hilfe zunächst ein Überblick darüber gewonnen werden sollte, ob die Herkunft von Items (und somit auch die auf den Items abgebildete Testkultur) insgesamt überhaupt einen Zusammenhang zu DIF, also der kulturell bedingten Varianz der Itemschwierigkeit zwischen jeweils zwei Ländern, aufweist. Die Annahme war hier, dass die Tatsache, dass ein Item aus dem Land der Fokusgruppe stammt, dieses für diese Schüler im Vergleich zu den Schülern der Referenzgruppe erleichtern sollte. Dies sollte sich in einer aus Sicht der Fokusgruppe signifikant negativen Korrelation zwischen DIF und Itemherkunft für Items aus dem eigenen Land widerspiegeln. Umgekehrt sollten die Items aus dem Land der Vergleichsgruppe mit einer höheren Itemschwierigkeit für die Fokusgruppe und einer signifikant positiven Korrelation einhergehen.

Für die Englisch-Items zeigte sich, dass beides für die Paarungen Frankreich-Deutschland, Frankreich-Ungarn und Deutschland-Ungarn der Fall war. Die Größe der Korrelationen bewegte sich zwischen $r=-.407$ ($p \leq .01$) und $r=.281$ ($p \leq .05$) und damit im niedrigen bis moderaten Bereich.

Diese Ergebnisse lassen sich insgesamt so interpretieren, dass die Tatsache, dass ein Item aus dem eigenen Land stammt, tendenziell mit für die Gruppe vorteilhaften DIF einhergeht und diese Items für diese Gruppe somit leichter sind als für die jeweilige Vergleichsgruppe. Ein Item aus dem Land der Referenzgruppe hingegen geht tendenziell mit für die Fokusgruppe nachteilhaften DIF einher. Das bedeutet, die Itemschwierigkeit erhöht sich für diese Gruppe im Vergleich zur Referenzgruppe.

Für die Paarungen Deutschland-Spanien und Ungarn-Spanien zeigte sich jeweils, dass eine signifikant negative Korrelation zwischen DIF und der Herkunft der Items aus dem Land der Fokusgruppe besteht. Die Tatsache, dass das Item aus dem Land der Referenzgruppe stammt, ging jedoch nicht mit einer Vergrößerung von nachteilhaften DIF einher, das heißt solche Items zeigten sich nicht als schwieriger für die Fokusgruppe im Vergleich zur Referenzgruppe.

Kein Zusammenhang mit der Itemherkunft zeigte sich bei der Paarung Spanien-Frankreich. Wie aus den Ergebnissen zu Frage 1b jedoch ersichtlich wird, weisen 66.4% aller Items signifikante Differentielle Item Funktionen zwischen Spanien und Frankreich auf. Daher ist zu vermuten, dass hier entweder weitere schwierigkeitsbestimmende Itemmerkmale, die nicht im „Dutch Grid” enthalten sind, oder aber andere, stichproben spezifische Merkmale verantwortlich sind, die nicht auf die Testkultur zurückführbar und somit im Rahmen dieser Arbeit nicht überprüfbar sind. Eine weitere Rolle mag hier auch die Nicht-Repräsentativität der Stichproben spielen.
Ferner zeigt sich bei Betrachtung der deskriptiven Statistiken der DIF-Parameter (siehe Anhang), dass die DIF-Parameter dieser Paarung nur eine geringe Varianz aufweisen. Aus diesem Grund wäre eine Einschränkung der Varianz eine weitere mögliche Ursache für fehlende oder unter- schätzte Zusammenhänge.

Hinsichtlich der Deutsch-Items zeigt sich ein ähnliches Bild wie bei den Englisch-Items: Auch hier finden sich für drei Paarungen, nämlich Frankreich-Schweden, Niederlande-Schweden und Ungarn-Niederlande, die aufgrund der Testkultur erwarteten positiven und negativen Korrelationen. Das bedeutet, bei diesen Paarungen geht die Beantwortung von Items, die aus dem eigenen Land stammen, tendenziell mit vorteilhaften DIF einher, während die Beantwortung von Items aus dem jeweils anderen Land tendenziell mit nachteilhaften DIF einhergeht.

Bei der Paarung Frankreich-Niederlande zeigt sich hingegen lediglich eine positive Korrelation zwischen DIF und Itemherkunft. Das ist so zu interpretieren, dass Items aus den Niederlanden die Itemschwierigkeit für die französische Stichprobe im Vergleich zur niederländischen Schülergruppe tendenziell erhöhen, französische Items für die niederländische Gruppe jedoch nicht schwieriger sind als für die französische Gruppe.

Bei der Paarung Frankreich-Ungarn existiert hingegen nur eine negative Korrelation mit dem Sachverhalt, dass ein Item aus Frankreich stammt, was dafür spricht, dass hier französische Items für die französische Gruppe einfacher als für die ungarische Gruppe zu beantworten sind, ungarische Items jedoch nicht schwieriger.


In einem nächsten Schritt wurden die DIF-Parameter der jeweiligen Länderpaarungen mit den schwierigkeitsbestimmenden Itemmerkmalen korreliert. Dabei waren, basierend auf den in Fragenkomplex 1 dargestellten Testkultur-Profilen, Hypothesen bezüglich der zu erwartenden relativen Stärken und Schwächen der Gruppen und somit auch bezüglich der Richtung der Korrelationen aufgestellt worden.

Betrachtet werden hier ausschließlich die signifikanten Korrelationen. Obgleich aufgrund der vorher aufgestellten Hypothesen auch angenommen werden könnte, dass sich zwei Länder hinsichtlich eines Merkmals nicht unterscheiden und daher die Korrelation nicht signifikant sein sollte (also letztlich angenommen wird, dass die H0 nicht zurückgewiesen wird), besteht dabei doch das Problem, dass hier nicht klar abzugrenzen ist, ob eine Korrelation tatsächlich aufgrund des fehlenden Unterschieds hinsichtlich der Stärken und Schwächen (ein bestimmtes testkulturelles Merkmal betreffend) nicht signifikant wird, oder ob dies auf andere Faktoren wie beispielsweise Varianzeinschränkung zurückzuführen ist. Hier ist eine Konfundierung von Ursachen möglich. Daher lassen sich ausschließlich signifikante Korrelationen eindeutig interpretieren.

Insgesamt erwiesen sich bei den Englisch-Items 29 der Korrelationen zwischen DIF und Itemmerkmalen mindestens auf einem alpha-Niveau von 5% als signifikant. Die Größe der Korrelationskoeffizienten ist als niedrig bis moderat einzuschätzen (zwischen $r = -.396; p \leq .01$ und $r = .467; p \leq .01$). Von diesen 29 Korrelationen entsprechen 23 der aufgrund der erwarteten Stärken und Schwächen der Gruppen angenommenen Richtung. Das heißt, kommt beispielsweise das Merkmal „komplexe grammatische Stukturen” bei den Items einer Gruppe signifikant seltener vor als bei den Items der jeweiligen Vergleichsgruppe, geht dies mit einer aus Sicht dieser Gruppe signifikant positiven Korrelation zwischen dem Merkmal und den DIF-Parametern der beiden Gruppen einher, und die Items sind für die Gruppe schwerer zu lösen als für die Vergleichsgrup-
Diskussion 6.1 Zusammenfassung und Interpretation der Ergebnisse

pe. Dies entspricht der aufgrund der Testkultur erwarteten Schwäche der Gruppe bei der Beantwortung von Items mit komplexer grammatischer Struktur.


Auch bei der Betrachtung der Zusammenhänge zwischen DIF-Parametern und schwierigkeitsbestimmenden Itemmerkmalen bei den Deutsch-Items zeigt sich, dass die signifikanten Korrelationen größtenteils in die aufgrund der Testkulturen erwarteten Richtungen deuten. Die Korrelationen bewegen sich auch hier im niedrigen bis moderaten Bereich (zwischen \( r = -.470; p \leq .01 \) und \( r = .416; p \leq .01 \)). Es sind dort insgesamt 34 Korrelationen mindestens auf einem alpha-Niveau von 5% signifikant, davon entsprechen 25 hinsichtlich ihrer Richtung den Hypothesen. Ferner sind 5 der Korrelationen signifikant, obgleich aufgrund der Testkulturen eigentlich keine signifikante Korrelation zu erwarten war. Wie bereits bei den Englisch-Items ist dies möglicherweise mit einer nicht-Repräsentativität der Items der entsprechenden Länder bezüglich des jeweiligen Merkmals erklärbar, was jedoch in weiteren Studien überprüft werden müsste.


Hier ist zu beobachten, dass die gleichen Variablen bzw. Prädiktoren in unterschiedlichem Kontext und auf unterschiedlichen Ebenen eine unterschiedliche Bedeutung erlangen können: Innerhalb der Länder stehen die Itemmerkmale ausschließlich für Determinanten von Itemschwierigkeit und eine Abbildung zugrunde liegender kognitiver Prozesse; zwischen den Ländern wird hingegen betrachtet, ob eine Differenz hinsichtlich dieser schwierigkeitsbestimmenden Merkmale existiert. Diese Differenzen werden dann an dieser Stelle zu einem Indikator für Testkultur.

Ein Beispiel für die Englisch-Items stellt das Ergebnis der Variablen „Authentizität des Texts“ bei der Länder-Paarung Frankreich und Deutschland dar. Innerhalb der beiden Länder geht die Tatsache, dass ein Item authentisch (d.h. nicht vereinfacht oder adaptiert) ist, eher mit einer größeren Itemschwierigkeit einher, so wie es theoretisch auch zu erwarten wäre. Ferner wurden si-
signifikant mehr der französischen als der deutschen Items als „authentisch“ eingeordnet. Bei der Betrachtung des Zusammenhangs zwischen DIF-Parametern und der Authentizität des Textes zeigt sich nun eine aus französischer Sicht signifikant negative Korrelation (r = -.304; p ≤ .01), das heißt Items dieser Art gehen für diese Gruppe mit einer im Vergleich zur deutschen Gruppe signifikant niedrigeren Itemschwierigkeit einher. Dies deutet darauf hin, dass bei kulturell bedingten Unterschieden hinsichtlich der Itemschwierigkeit testkulturelle Variablen eine Rolle zu spielen scheinen. Ein weiteres Beispiel im Rahmen der Englisch-Items stellen die Ergebnisse der Variablen „Inhalt hauptsächlich abstrakt“ bei der Länderpaarung Ungarn-Spanien dar. Innerhalb beider Länder existiert eine signifikant positive Korrelation mit der Itemschwierigkeit (r = .213; p ≤ .05 bzw. r = .392; p ≤ .01), was darauf hindeutet, dass abstrakter Inhalt in beiden Gruppen die Itemschwierigkeit erhöht. Die Korrelation mit den DIF-Parametern der Länderpaarung hingegen ist aus ungarischer Sicht signifikant negativ (r = -.276; p ≤ .01). Dieses Ergebnis entspricht den aufgrund der Testkulturen gemachten Annahmen: Die ungarischen Items weisen signifikant häufiger die Ausprägung „Inhalt ziemlich abstrakt“ auf als dies bei den spanischen Items der Fall ist, und diese Items sind für die ungarische Gruppe leichter als für die spanische. Beispiele im Rahmen der Deutsch-Items finden sich, um nur einige zu nennen, bei der Länderpaarung Frankreich-Ungarn bezüglich der Variablen „komplexe grammatische Strukturen“ und bei der Paarung Ungarn-Niederlande die Variable „teilweise komplexe grammatische Strukturen“ betreffend.

Zusammenfassend läßt sich sagen, dass sich Korrelationen zwischen DIF und schwierigkeitsbestimmenden Merkmalen der Items zeigen, die, wenn sie signifikant sind, größtenteils den aufgrund der Testkulturen aufgestellten Hypothesen entsprechen. Dies trifft sowohl für die Englisch- als auch für die Deutsch-Items zu. Ähnliche Ergebnisse berichteten Klieme und Baumert (2001) für Mathematik-Items.

Im zweiten Analyseschritt des dritten Fragenbereichs wurde eine multiple Regression der Differenziellen Item Funktionen auf die Itemmerkmale durchgeführt. Ziel war hier, zum einen herauszufinden, inwieweit die kulturell bedingte Varianz der Itemschwierigkeiten durch die schwierigkeitsbestimmenden Merkmale insgesamt erklärt werden kann, und zum anderen, ob bzw. welche der Prädiktoren ihrer Richtung nach den aufgrund der Testkulturen aufgestellten Hypothesen hinsichtlich der zu erwartenden Stärken und Schwächen der Länder entsprechen. Dazu wurden pro Länderpaarung mehrere Modelle gerechnet: zunächst ein Modell mit allen Itemmerkmalen als Prädiktoren, in einem zweiten Schritt ein Modell, welches ausschließlich signifikante, nicht den Hypothesen widersprechende Prädiktoren enthält. Das heißt einschließlich der „neutralen“


Auch bezüglich der Deutsch-Items wurden analog zu den Englisch-Items Regressionsanalysen berechnet. Wie bereits bei den Analysen der Englisch-Items war es auch hier von der jeweiligen Länder-Paarung abhängig, wie gut DIF anhand testkulturell-konformer Merkmale vorhergesagt werden konnte. Hier konnten anhand der Endmodelle, die gleichfalls ausschließlich noch signifikante, den Testkulturen entsprechende Prädiktoren beinhalteten, je nach Paarung zwischen 3.1% und 32.7% der Varianz auf testkulturell bedingte Stärken und Schwächen der Gruppen zurückgeführt werden. Auch hier erhöht sich der Anteil der aufgeklärten Varianz unter Einschluss der „neutralen“ Prädiktoren teils deutlich (beispielsweise von 5.5 % auf 10.8% bei der Paarung Schweden-Ungarn). Darüber hinaus zeigten sich auch hier zwar einige der im ersten Modell signifikanten Prädiktoren den Hypothesen entgegengesetzt, jedoch handelte es sich gleichfalls um einen deutlich kleineren Anteil als den der hypothesenkonformen Prädiktoren.


Neben dem Anteil der anhand der testkulturell konformen Prädiktoren aufgeklärten Varianz stellen auch die beta-Koeffizienten an sich ein interessantes Ergebnis dar. So ist es anhand ihrer Betrachtung möglich, Informationen bezüglich der Stärken und Schwächen der einzelnen Länder, jeweils in Relation zu der Vergleichsgruppe interpretiert, zu gewinnen. Aufgrund der nicht eindeutigen Interpretierbarkeit der „neutralen“ Prädiktoren werden hier dazu ausschließlich die hypothesenkonformen Prädiktoren der Endmodelle herangezogen. Es kann mit Hilfe der standardisierten beta-Koeffizienten beispielsweise die Aussage getroffen werden, dass die Tatsache, dass ein Item einen authentischen Text (wie beispielsweise einen Zeitungsartikel) verwendet, die Itemschwierigkeit für die schwedischen Schüler im Vergleich zur französischen Gruppe um 0.41 Logits erhöht. Dies wäre dann so zu interpretieren, dass die Bearbeitung authentischer Texte eine Stärke französischer und eine Schwäche schwedischer Schüler darstellt. Die übrigen beta-Koeffizienten sind analog dazu zu interpretieren. Die Anwendungsbereiche solcher Informationen werden unter 6.3 diskutiert.
Besonders häufig stellen in den Endmodellen Ausprägungen der Variablen „Itemtyp“ einen signifikanten Prädiktor dar, außerdem Ausprägungen der Variablen „Grammatik“, und etwas seltener von „Abstraktheit“ und „Vokabular“. Wie auch schon bei den Korrelationsanalysen spielen die Merkmale, die sich mit der Art der für eine korrekte Beantwortung notwendigen Information befassen, kaum eine relevante Rolle.

Auch hier finden sich in einigen der Regressionsanalysen (bezüglich der Deutsch-Items) standardisierte Regressionskoeffizienten größer als eins, was auch hier auf die Anwesenheit von Suppressorvariablen hindeutet. Diesbezüglich gilt das Gleiche wie oben bereits angemerkt wurde: Suppressionseffekte lassen sich üblicherweise dadurch mindern, dass entweder Variablen zusammengefasst oder aber aus der Analyse ausgeschlossen werden. In dieser Arbeit war es jedoch zunächst das Ziel, die einzelnen Merkmalsausprägungen und deren Zusammenhang zu DIF zu untersuchen und darzustellen. In weiterführenden Studien sollte ausführlich analysiert werden, bei welchen Variablen es sich um Suppressionavariablen handelt, ob es sich bei allen Analysen um gleichen sind handelt, und ob sich diese Effekte durch das Zusammenfassen von Variablen verringern lassen.

6.2. Beantwortung der Hauptfragestellung

Die Beantwortung der Hauptfragestellung lässt sich aus den oben dargestellten Ergebnissen herleiten. Die Fragestellung lautete:

Existiert ein Zusammenhang zwischen Differentiellen Item Funktionen und Indikatoren nationaler Testkulturen bei Aufgaben zur Messung des fremdsprachlichen Leseverständnisses in englischer und deutscher Sprache?

Insgesamt kann die Frage mit einem vorsichtigen „Ja“ beantwortet werden. In den Ergebnissen der diesbezüglichen Einzelfragestellungen finden sich deutliche Hinweise auf einen Zusammenhang zwischen Testkultur und kulturell bedingter Varianz der Itemschwierigkeit. Insgesamt sprechen fünf Argumente für die Existenz eines solchen Zusammenhangs:

Erstens zeigt sich im Rahmen von Frage 1c, dass die hier analysierten Länder sich teilweise signifikant hinsichtlich ihrer Testkultur unterscheiden, das heißt hinsichtlich der Häufigkeit des Vorkommens von itemschwierigkeitsbestimmenden Merkmalen bei den in der EBAFLS-Studie eingereichten Items. Damit konnte gezeigt werden, dass unterschiedliche nationale Testkulturprofile existieren.

Zweitens existieren hypothesenkonforme Korrelationen zwischen DIF und der Itemherkunft, sowie hypothesenkonforme Korrelationen zwischen DIF und den schwierigkeitsbestimmenden Merkmalen. Zusätzlich existieren nur in sehr wenigen Ausnahmen Korrelationen, die hier den aufgrund der Testkulturen aufgestellten Hypothesen widersprechen.

Drittens lässt sich bei der überwiegenden Zahl der Regressionsanalysen zumindest ein Teil der kulturell bedingten Varianz auf Unterschiede der Testkulturen und Unterschiede hinsichtlich der Stärken und Schwächen der Gruppen zurückführen, wenn auch in einigen Analysen Multikollinearität eine Rolle spielt und diese somit nicht völlig stabil interpretierbar sind.


Das fünfte Argument für einen Zusammenhang zwischen Testkultur und DIF ist die Tatsache, dass sich die Ergebnisse replizieren lassen: Obgleich es sich mit Englisch und Deutsch um zwei
unterschiedliche Sprachen handelt und die Analysen teilweise anhand unterschiedlicher Länder durchgeführt wurden, weisen die Ergebnisse in sämtlichen Analysen in die gleiche Richtung. Gerade diese Replizierbarkeit weist darauf hin, dass möglicherweise trotz der Nicht-Repräsentativität der Stichproben und der anderen oben bereits genannten möglichen Probleme mit den Daten die Ergebnisse nicht zufällig zustande kommen, sondern dass tatsächlich ein Zusammenhang zwischen Testkulturen und DIF besteht.


6.3. Relevanz der Ergebnisse

In diesem Abschnitt der Ergebnisdiskussion wird auf die Relevanz der Ergebnisse für die unter Punkt 2 dargestellten, dieser Arbeit zugrundeliegenden Theoriebereiche eingegangen. Zunächst wird diesbezüglich der Bereich „Differentielle Item Funktionen“ thematisiert, darauf folgen die Bereiche „Fremdsprachenforschung und angewandte Linguistik“ und „Interkulturelle Vergleichbarkeit von Testverfahren“. Ferner soll die Bedeutung der Ergebnisse für Validität und Fairness sowie für die Konstruktion zukünftiger Testverfahren erörtert werden. Für jeden dieser Bereiche werden darüber hinaus die jeweils zu nennenden Kritikpunkte angesprochen.

6.3.1. Relevanz der Ergebnisse für den Forschungsbereich „Differentielle Item Funktionen”

Wie bereits oben ausgeführt, beziehen sich Differentielle Item Funktionen auf das Phänomen, dass die Mitglieder zweier oder mehr Gruppen eine unterschiedliche Wahrscheinlichkeit aufweisen, ein Item korrekt zu lösen, obgleich sie sich hinsichtlich der zu messenden Fähigkeit auf dem gleichen Leistungsniveau befinden (Holland & Wainer, 1993). Der vorliegenden Arbeit
liegt nun die Frage zugrunde, worin mögliche Ursachen für DIF liegen könnten, das heißt, worin genau sich die Mitglieder zweier Gruppen im Bereich der Fremdsprachenkompetenzen bei gleichen Fähigkeiten unterscheiden.

Unter 2.1 wurden zwei grundsätzlich mögliche Ansätze für das Entstehen von DIF diskutiert: Im Rahmen des ersten sind DIF das Ergebnis eines nicht intendierten Messens einer oder mehrerer zusätzlicher, konstruktirrelevanter Dimensionen. In diesem Fall weist eine der betrachteten Gruppen bezüglich dieser Dimension bzw. Dimensionen eine höhere Fähigkeit auf. Differentielle Item Funktionen sind demnach also auf konstruktirrelevante Fähigkeitsunterschieden zurückzuführen. DIF werden im Rahmen dieses Ansatzes auch als „Nuisance Dimension“ (Ackerman, 1992; Roussos & Stout, 2004) bezeichnet.

Der zweite Ansatz hinsichtlich des Zustandekommens von DIF verfolgt einen im diagnostischen Sinne konstruktiveren Ansatz. Er geht davon aus, dass DIF zumindest zum Teil durch unterschiedliche Stärken und Schwächen von Gruppen bedingt ist (Scheuneman & Gerritz, 1990; Klieme & Baumert, 2001; Dogan, Guerrera & Tatsuoka, 2005). In dieser Arbeit wurde dieser Ansatz zugrunde gelegt, es wurde also davon ausgegangen, dass aufgrund testkultureller Unterschiede die analysierten Ländergruppen unterschiedlichen Lerngelegenheiten ausgesetzt waren.


Es können somit zumindest hinsichtlich einiger Länderpaarungen und Itemmerkmale diagnostische Aussagen dahingehend gemacht werden, welches Itemmerkmal (und damit welcher diesem zugrunde liegende kognitive Prozess) beim Lösen eines Fremdsprachen-Items bei einer Gruppe eine Stärke oder Schwäche im Vergleich zu einer jeweils anderen Gruppe darstellt. Dabei handelt es sich in diesem Fall nicht um Unterschiede auf einer konstruktirrelevanten, sondern hinsichtlich einer konstruktirelevanten Dimension, wie beispielsweise grammatische Strukturen oder Häufigkeit des Vokabulars.
Dennoch kann DIF nur teilweise durch solche unterschiedlichen testkulturellen Prädigungen der Gruppen erklärt werden. Dies mag durchaus auch darin begründet sein, dass auch konstruktirelevante, durch das Item zusätzlich erfasste Fähigkeiten für die Itemschwierigkeiten, also mit anderen Worten eine „Nuisance Dimension“ (Ackerman, 1992), bei der Entstehung von DIF eine Rolle spielen. Dies müsste allerdings noch in weiteren Studien überprüft werden, beispielsweise auch durch das Hereinnehmen weiterer konstruktirelevanter Itemmerkmale für die Modellierung von DIF. Es kann gemutmaßt werden, dass letztendlich ein Teil von DIF durch konstruktirelevante, differentielle Stärken und Schwächen der Gruppen zustande kommt, ein weiterer Teil jedoch durch die oben genannte „Nuisance Dimension“ und konstruktirelevante Multidimensionalität.


„When DIF is conceptualized this way, items displaying DIF cannot be regarded as unwelcome ‘biased’ items any more. These items become indicators of micro-level performance differences among countries after controlling for their macrolevel, or overall, performance“(Dogan, Guerrera & Tatsuoka, 2005, S. 24).

nen untermauert werden. Unter diesem Aspekt sind die Ergebnisse der vorliegenden Dissertation daher als ein wichtiger Schritt bei der Erforschung und Modellierung von Differentiellen Item Funktionen einzustufen.

Allerdings sind gegenüber dem in dieser Arbeit gewählten Ansatz hinsichtlich einiger Punkte Vorbehalte zu formulieren. Eine Einschränkung betrifft den Umstand, dass für die Erklärung von DIF mögliche weitere erklärende, itembasierte Variablen nicht mit einbezogen wurden. Zu diesen gehören beispielsweise die inhaltliche Kategorie des Dutch Grid (Alderson et al., 2006), welche Variablen wie beispielsweise das Thema (z.B. Reisen, Alltag, Musik, etc.) oder die Quelle des Items (z.B. Zeitungsartikel, Magazin, eMail, etc.) berücksichtigt.


Neben solchen rein itemseitigen Variablen werden in anderen Forschungsarbeiten außer Itemmerkmalen auch Eigenschaften bestimmter Itemgruppen oder auch Personeneigenschaften zur Erklärung von DIF herangezogen, wie es etwa von den Noortgate und deBoeck (2005) im Rahmen der Anwendung sogenannter „logistic mixed models“ vorschlagen: „Note that there are other kinds of interaction effects that could be regarded as DIF. For example, it is possible that there is a differential functioning of groups of items, instead of, or in addition to, a differential item functioning of individual items. Moreover, it is not uncommon that items function differently over persons belonging to the same group” (S. 456). Auch diese Faktoren konnten bedauerlicherweise nicht berücksichtigt werden: Wegen der Unvollständigkeit der Fragebögen in der EBAFLS-Studie standen Personenmerkmale nicht zur Verfügung. Zwar bedeuten diese hier formulierten Vorbehalte gewisse Einschränkungen und zeigen auf, welche weiteren Aspekte zukünftige Studien berücksichtigen könnten bzw. sollten, gleichwohl können die Ergebnisse dieser Arbeit als fundiert angenommen werden: DIF kann nicht ausschließlich als eine konstruktirrelevante „Nuisance Dimension“ betrachtet werden, und DIF kann relevante diagnostische Informationen beinhalten und Ursachen für die Invalidität von Testverfahren aufzeigen.
6.3.2. Relevanz für Theorie und Forschung im Bereich der fremdsprachlichen Diagnostik

Abgesehen von Ihrem Nutzen für die Forschung im Bereich der Differentiellen Item Funktionen und somit auch für die Validität von Testverfahren sind die Ergebnisse der vorliegenden Arbeit auch für den Bereich der Fremdsprachenforschung relevant. Wie unter 2.2.3 dargelegt wurde, ist der Gemeinsame Europäische Referenzrahmen für Sprachen die theoretische Grundlage dieser Arbeit. Im Rahmen von dessen Entwicklung wurde die Eindimensionalität von fremdsprachlichem Leseverständnis postuliert und dies auch anhand von Rasch-Analysen der Skalendescriptoren bestätigt (North, 2000).

Die Ergebnisse dieser Arbeit weisen darauf hin, dass dies zumindest innerhalb der einzelnen Länder bestätigt werden kann: Es zeigte sich, dass innerhalb der Länder das Rasch-Modell auf die Daten anwendbar ist, was für eine Eindimensionalität des Konstrukts „fremdsprachliches Leseverständnis“ innerhalb der verschiedenen Länder spricht.

Hinsichtlich der kulturell bedingten Unterschiede zwischen den Ländern zeigt sich hingegen, dass auch unter Zugrundelegung des Rasch-Modells (im Vergleich zum OPLM-Modell der EBAFLS-Studie) eine große Anzahl Items signifikante Differentielle Item Funktionen aufweisen. Insofern kann die Annahme der Eindimensionalität des Konstrukts fremdsprachlichen Leseverständnisses nicht über unterschiedliche Länder hinweg beibehalten werden. Die Ergebnisse dieser Arbeit zeigen, dass die Ergebnisse der EBAFLS-Studie (Fandel et al., 2007), deren Daten dieser Dissertation zugrunde liegen, diesbezüglich auch unter Anwendung des strengeren Rasch-Modells bestehen bleiben.


Ein weiteres diesbezüglich relevantes Ergebnis ist die Beobachtung, dass sich die Korrelationen zwischen Itemschwierigkeit und Itemmerkmalen innerhalb der unterschiedlichen Länder nicht signifikant voneinander unterscheiden. Dies könnte ein Hinweis darauf sein, dass das Instrument in unterschiedlichen Ländern gleichermaßen für die Kategorisierung von Items hinsichtlich ihrer Schwierigkeit geeignet zu sein scheint. Auch gehen die theoretisch als schwieriger eingestuften Abstufungen der Itemmerkmale (beispielsweise „sehr komplexe grammatische Strukturen“ im Gegensatz zu „ausschließlich einfache grammatische Strukturen“) fast durchgehend mit einer höheren Itemschwierigkeit einher, was den theoretischen Annahmen des Instruments und auch des GERS entspricht. Die vorliegende Arbeit trägt also auch diesbezüglich zur Validierung des „Dutch Grid“-Kategoriensystems bei.


Diskussion 6.3 Relevanz der Ergebnisse

ren, Freedle & Kostin (1993) 30%-52% und Bachman, Davidson & Milanovic, 1996 44.5-68%.
Der im Rahmen dieser Arbeit aufgeklärte Anteil der Varianz der Itemschwierigkeit liegt mit
20.8%-48.3% zwar teilweise etwas niedriger, was wohl auch mit einer geringeren Anzahl von
Prädiktoren erklärt werden könnte, weisen jedoch insgesamt in die gleiche Richtung und bestä-
tigen damit bisherige empirische Forschung auf diesem Gebiet.

Darüber hinaus lässt sich anhand der aufgezeigten Zusammenhänge zwischen Differentiellen
Item Funktionen und den unterschiedlichen Profilen von Stärken und Schwächen der Gruppen
etwas über die Unterschiedlichkeit des Konstrukts „fremdsprachliches Leseverständnis“ in den
hier untersuchten Ländern aussagen. So deutet die Existenz der systematischen Stärken und
Schwächen der Ländergruppen darauf hin, dass vermutlich zumindest Teile des Konstrukts sich
unterscheiden und einen unterschiedlich gewichteten Stellenwert in den fremdsprachlichen Cur-
icula und Testkulturen der Länder einnehmen. Ob es sich dabei um Struktur- oder Niveauunter-
schiede handelt, ist dabei, wie oben bereits angesprochen wurde, nicht klarbar.

Interessant ist dieses Ergebnis hinsichtlich der unter 1.1 dargestellten Debatte um die Frage
„Ist mein B1 auch dein B1“? (Fandel et al., 2007). Diese Debatte im Rahmen des GERS bezieht
sich genau auf die Frage nach der Vergleichbarkeit des Konstrukts und könnte auch umbenannt
werden in die Frage „Ist mein Konstrukt auch dein Konstrukt“? Aufgrund der vorliegenden
Ergebnisse könnte man diese Frage vermutlich mit „ja, teilweise“ beantworten:

Die Ergebnisse zu Frage 1c zeigen, dass zumindest innerhalb der Länder die gleichen Itemmerk-
male einen ähnlichen Zusammenhang mit den Itemschwierigkeiten aufweisen, und zwar sowohl
hinsichtlich der Richtung als auch der Größe der Zusammenhänge. Es zeigten sich kaum signi-
fikante Unterschiede. Das spricht dafür, dass zumindest Teilkomponenten des Konstrukts über
die Länder hinweg dieselben sein sollten. Die Existenz von DIF und von Testkulturen spricht
wiederum dafür, dass sich das Konstrukt über die verschiedenen Länder hinweg jedoch auch
unterscheidet, sei es nun aufgrund von Niveau- oder aufgrund von Strukturunterschieden. Die
Ergebnisse weisen somit darauf hin, dass die Annahme der Eindimensionalität des Konstrukts
zwischen den Ländern vermutlich nicht haltbar ist. Die Existenz von DIF weist darauf hin, dass
noch mindestens eine zusätzliche Dimension anhand der Testaufgaben erfasst wird. Die Dimen-
sionalität des Konstrukts lässt sich im Rahmen dieser Arbeit jedoch nicht endgültig klären. Auch
ob es sich bei den Konstruktunterschieden zwischen den Ländern um Struktur- oder Niveauunter-
schiede handelt, ist in dieser Arbeit aufgrund des Multi-Matrix-Designs der Studie und der damit
einhergehenden Nicht-Anwendbarkeit von z.B. Mehrgruppenanalysen nicht endgültig klärbar.


6.3.3. Relevanz der Ergebnisse für den Bereich der interkulturellen Vergleichbarkeit von Testverfahren


Im Hinblick auf die Frage nach dem emischen (jede Kultur ist einzigartig; Unterschiede = strukturelle Unterschiede) oder etischen Ansatz (es existieren die gleichen zugrundeliegenden Variablen und Konstrukte in unterschiedlichen Kulturen; Unterschiede = Niveaunterschiede; z.B. Helfrich, 1999) der interkulturellen Psychologie scheint mit Blick auf die Ergebnisse die in dieser Arbeit eingenommene Zwischenposition sinnvoll:

Die Ähnlichkeit der Korrelationen innerhalb der Länder bezüglich des Zusammenhangs zwischen Itemschwierigkeit und Itemeigenschaften sprechen dafür, dass sich das Konstrukt fremdsprachlichen Leseverständnisses über die Länder hinweg nicht vollständig unterscheidet: die gleichen Merkmale spielen in allen untersuchten Ländern in vergleichbarem Ausmaß eine Rolle für die Itemschwierigkeit. Möglicherweise existieren jedoch auch Unterschiede hinsichtlich des Konstrukts. Einen Hinweis darauf, inwiefern sich die Konstrukte der Länder unterscheiden könnten, liefert die Betrachtung der im Rahmen von Frage 1c erstellten testkulturellen Profile: So kommen einige Itemformate bei den Items mancher Länder gar nicht vor, wie beispielsweise das Format „Ordnen“. Hier ist möglicherweise davon auszugehen, dass sich das Konstrukt in Frankreich diesbezüglich von dem anderer Ländern strukturell unterscheidet, da dies das einzige Land ist, in dem das Itemformat überhaupt zur Testkonstruktion verwendet wird. In anderen Ländern scheint ein solches Format für das Testen fremdsprachlichen Leseverständnisses keinerlei Relevanz zu besitzen. Weitere solcher Beispiele finden sich auch im Hinblick auf andere Variablen und andere Länder.

Unterschiede existieren jedoch nicht nur bezogen auf Variablen bzw. Itemmerkmale, die—wie soeben beschrieben— entweder häufig oder gar nicht vorkommen in den unterschiedlichen Ländern und bei denen daher die Möglichkeit besteht, dass sich das Konstrukt der Länder diesbezüglich unterscheidet. Darüber hinaus existiert offenbar in den Ländern hinsichtlich einiger Variablen bzw. Itemmerkmale auch eine unterschiedliche Schwerpunktlegung. Dies drückt sich darin aus, dass diese Itemmerkmale zwar bei den Items aller Länder vorkommen und daher in allen Ländern bei der Testkonstruktion eine Rolle zu spielen scheinen, jedoch in signifikant

6.3.4. Relevanz für Theorie und Forschung im Bereich der Validität


Da Differentielle Item Funktionen ein Hinweis auf die Unterschiedlichkeit eines Konstrukts bei zwei oder mehr untersuchten Gruppen sein können, können sie auch als eine Einschränkung der Konstruktvalidität betrachtet werden.


Nach Messick (1996) existieren zwei mögliche Quellen von Validitätseinschränkung, nämlich zum einen die Einführung konstruktirrelevanter Varianz (das bedeutet, es wird mindestens eine nicht intendierte, konstruktirrelevante Dimension mit erfasst), und zum anderen eine Unterrepräsentanz des Konstrukts (das Konstrukt wird nicht vollständig erfasst). Unter 2.4 dieser Arbeit war die Überlegung angestellt worden, dass diese beiden Quellen von Invalidität mit den beiden Ansätzen zur Erklärung von DIF, nämlich DIF als „Nuisance Dimension“ (z.B. Ackerman, 1992) vs. DIF als „differentielle Stärken und Schwächen der Gruppen“ (z.B. Scheuneman & Gerritz, 1990) gleichgesetzt werden können: Wird DIF als „Nuisance Dimension“ betrachtet, als eine zu-
sätzlich erfasste, konstruktirrelevante Dimension, dann sollte in diesem Fall die Validitätsminde-
runung durch die Einführung konstruktirrelevanter Varianz bedingt sein. Wird DIF hingegen als ein
Ausdruck differentieller Stärken und Schwächen betrachtet, dann lässt das vermuten, dass nicht
alle Teile des Konstrukts gleichermaßen in den unterschiedlichen Gruppen repräsentiert sind. In
diesem Falle ist von einer differentiellen Unterrepräsentanz des Konstrukts in den unterschiedli-
chen Gruppen auszugehen.

Die Ergebnisse dieser Arbeit weisen darauf hin, dass vermutlich beide Ansätze und beide Quel-
len der Validitätsminderung bei der Entstehung von DIF eine Rolle spielen: Die Tatsache, dass
DIF Zusammenhänge mit den aufgrund der Testkultur erwarteten Stärken und Schwächen der
Gruppen aufweist, spricht dafür, dass der zweite Ansatz, nämlich die Betrachtung von DIF als
ebensolche systematischen, differentiellen Stärken und Schwächen von Gruppen hinsichtlich
eines bestimmten Konstrukts und einer damit einhergehenden differentiellen Unterrepräsentanz
des Konstrukts in den Gruppen, für die Entstehung von DIF eine Rolle spielt. Bezogen auf die
vorliegende Arbeit lässt sich diese Überlegung wie folgt übertragen: Weist eine Gruppe eine Stär-
ke hinsichtlich des Beantwortens von Items mit einem bestimmten Itemmerkmal (wie beispiels-
weise „komplexen grammatistischen Strukturen“) auf und geht dies mit der jeweiligen Testkultur
einher (d.h. die Items dieses Landes weisen signifikant häufiger dieses Itemmerkmal auf als die
aus dem Land der Vergleichsgruppe stammenden Items), dann kann erstens angenommen wer-
den, dass dieser Teil des Konstrukts in diesem Land repräsentiert ist, und dies, zweitens, stärker
als in der Vergleichsgruppe.

Gegenteiliges gilt für den Fall, dass eine Gruppe bezüglich eines bestimmten Itemmerkmals (und
somit auch hinsichtlich der diesem Itemmerkmal zugrundeliegenden kognitiven Prozesse) eine
Schwäche aufweist: Geht dies damit einher, dass dieses Itemmerkmal bei den Items eines Landes
signifikant seltener als bei den Items der Vergleichsgruppe vorkommt, ist davon auszugehen, dass
bei dieser Gruppe diese Konstrukt komponente im Vergleich mit der anderen Gruppe unterreprä-
sentiert ist. Durch Testkultur und differentielle Lerngelegenheiten bedingte Schwächen können
somit als ein Ausdruck von Unterrepräsentanz des Konstrukts bzw. bestimmter Konstrukt kom-
ponenten in einer Gruppe interpretiert werden.

Die Ergebnisse dieser Arbeit weisen im Rahmen von Fragen 3b und 3c jedoch auch darauf hin,
dass mit Hilfe von differentiellen Stärken und Schwächen von Gruppen nur ein Teil der DIF-
Varianz aufgeklärt werden kann. Dies kann möglicherweise dahingehend interpretiert werden,
dass neben einer differentiellen Unterrepräsentanz des Konstrukts in verschiedenen Gruppen
Diskussion 6.3 Relevanz der Ergebnisse


Neben der Frage, welche Quellen der Invalidität für die Entstehung von DIF relevant sein könnten, bezieht sich der zweite Schwerpunkt in Messicks Modell auf die Bedeutung sozialer und kultureller Werte für die Interpretation und Konsequenzen von Testwerten (Messick, 1989; 1996). Im Rahmen dieser Arbeit ist dies vor allem im Hinblick auf Überlegungen hinsichtlich der Entstehung der beobachteten Testkulturen, also letztlich die Übertragung sozialer und bildungskultureller Werte auf die Testaufgaben eines Landes, von Relevanz. Eine wichtige Rolle spielt dabei das Einbeziehen der Interpretation von Testwerten und deren Konsequenzen („consequential validity“) in das Konzept der Konstruktvalidität.


Inwieweit die Konsequenzen von Testergebnissen und der Testwertinterpretation („consequential validity“) bei dem Entstehen dieser Testkulturen in den einzelnen Ländern eine Rolle spielen, lässt sich im Rahmen dieser Arbeit, basierend auf Messicks Überlegungen, nur vermuten:


Es ist also durchaus denkbar, dass das Konstrukt „fremdsprachliches Leseverständnis“ durch die bei der jeweiligen Testkonstruktion und dem jeweiligen Curriculum zugrunde liegenden sozialen Werte einer Kultur zunächst festgelegt, und durch „Teaching to the Test“ gefestigt wird.

Da wiederum die Ergebnisse dieser Arbeit auf Unterschiede zwischen den betrachteten Ländern hinsichtlich des Konstrukts hinweisen, ist davon auszugehen, dass —aus gesamteuropäischer Sicht— das Konstrukt innerhalb der einzelnen Länder nicht in der ganzen denkbaren Breite unterrichtet und getestet wird. Wie oben bereits angesprochen, findet also in dieser Hinsicht eine teilweise Einschränkung der Konstruktvarianz statt.

Ausdruck „Washback” für eine durch das Testen bedingte Verbesserung der Konstruktvalidität verwendet wird.


Denkbar ist, dass bereits geplante internationale Vergleichstests eine Chance dazu bieten könnten: In den nächsten Jahren startet der auf langfristige Ländervergleiche angelegte Europäische Sprachenindikator, dessen Zweck es ist, die Fremdsprachenkenntnisse von Schülern europäischer Länder zu erfassen und zu vergleichen.

Häufig sind die Ergebnisse solcher internationalen Vergleichstests zwar nicht für die einzelne Person relevant, umso mehr jedoch für die Gesellschaft eines Landes: Ein schlechtes Gesamtabschneiden, beziehungsweise ein als verhältnismäßig schlecht wahrgenommenes Ergebnis, kann zu Diskussionen hinsichtlich des eigenen Curriculums und des eigenen Lehrstoffs führen. Es wird analysiert, hinsichtlich welcher Aufgaben die eigenen Schüler kein gutes Ergebnis erreichen konnten. Diese werden dann möglicherweise als ein Teil des Konstrukts wahrgenommen, den zu beherrschen wünschenswert wäre, und die entsprechende Art von Aufgaben wird in den nationalen Lehrplan und die nationalen Testverfahren mit aufgenommen. Wenn diese wiederum in nationalen „high stakes”-Tests angewandt werden, „komplettiert” sich auf oben dargelegtem Wege das unterrichtete und getestete Konstrukt innerhalb der einzelnen Länder und möglicherweise auch im gesamteuropäischen Rahmen.

Betrachtet man allerdings die Ergebnisse der vorliegenden Arbeit hinsichtlich Frage 1c, dann zeigt sich auch, dass beinahe alle Länder unterschiedliche Stärken oder Schwächen aufweisen. Für eine europaweite ”Komplettierung” des Konstrukts wäre daher gleichzeitig vonnöten, dass Aufgaben aus allen Ländern in den Test mit einfließen.
Darüberhinaus müsste der Indikator in allen Ländern eine hohe Relevanz besitzen, so dass das Ergebnis der jeweils eigenen Nation als genügend wichtig wahrgenommen wird, um bei schlechtem Abschneiden eine Diskussion hinsichtlich der Lehrinhalte und der verwendeten Testaufgaben anzustoßen.


Obgleich die Ergebnisse der Arbeit mit Vorsicht zu interpretieren sind, ist jedoch insgesamt zu vermuten, dass die im Messick´schen Modell als Teil der Konstruktvalidität einbezogene „soziale“ Validität eine Rolle bei der Übertragung von kulturellen Normen und Werten auf die Testkultur und die Definition des Konstrukts „Fremdsprachliches Leseverständnis” innerhalb eines Landes und somit auch für die Entstehung nationaler Testkulturen spielt.

6.4. Grenzen der Arbeit und zukünftige Forschungsperspektiven

Wie im Rahmen der Ergebnisdiskussion bereits angesprochen wurde, sind die Ergebnisse der Arbeit hinsichtlich ihrer Generalisierbarkeit mit einer gewissen Vorsicht zu interpretieren. Dies ist zum einen auf die Nicht-Repräsentativität der Stichproben zurückzuführen. Zweitens ist im Rahmen dieser Arbeit nicht überprüfbar, ob die von den Ländern eingereichten Items tatsächlich repräsentativ für die in den jeweiligen Ländern verwendeten Test-Items sind. Ein weiterer Vorbehalt gegenüber den Daten betrifft die Einordnung der Items hinsichtlich ihrer schwierigkeitsbestimmenden Merkmale. Anlässlich der teilweise geringen Zusammenhänge, die sich zwischen Itemschwierigkeiten bzw. DIF und den Itemmerkmalen zeigt, die für die Art der für eine korrekte Lösung notwendigen Informationen stehen (Variablen Informationsgewinn 1-3), stellt sich die Frage, ob diese Itemmerkmale im Rahmen des „Dutch Grid“ ausreichend genau beschrieben sind, um es den Experten zu ermöglichen, Items diesbezüglich einzustufen. Als Grundlage für die in der Arbeit durchgeführte Untersuchung von Items hinsichtlich ihrer Item-Anforderungsmerkmale wurde die Expertengruppe ausgewählt, die sich größtenteils aus
den Autoren des „Dutch Grid“ zusammensetzte. Allerdings existieren nur niedrige Inter-Rater-
Reliabilitäten zwischen deren Einordnung und der von zusätzlich rekrutierten Experten. Hier
wäre es wünschenswert, bei zukünftigen Vorhaben zum einen die Beschreibung der Itemmerk-
male zu verbessern und zum anderen möglicherweise einen längeren Rating-Prozess durchzufüh-
ren, in dem die Experten während der Schulung und des späteren Ratings immer wieder Feed-
back hinsichtlich ihrer Übereinstimmung erhalten. Denkbar wäre auch, die Gruppe der Rater aus
Experten unterschiedlicher Länder zusammensetzen. So ließen sich länderspezifische Verzerr-
rungen bei der Einordnung von Items vermeiden.

Neben diesen Faktoren führt außerdem die Beschränkung der EBAFLS-Studie (CITO, 2008)
auf Schüler und auch Items, die sich in etwa auf dem Niveau B1 des GERS (Europarat, 2001)
beenden sollten, dazu, dass sich die Ergebnisse hinsichtlich der Testkulturen nicht auf andere
Kompetenzniveaus generalisieren lassen. Auch lassen sich keine Aussagen hinsichtlich der Gene-
ralisierbarkeit der Ergebnisse auf Items anderer Niveaustufen machen. Ferner lässt sich anhand
der Daten keine Aussage darüber treffen, ob die Ergebnisse auch auf andere Teilfähigkeiten von
Fremdsprachenkompetenzen wie etwa die produktiven Fähigkeiten „Sprechen“ und „Schreiben“
übertragbar sind.

Eine der Grenzen dieser Arbeit betrifft daher die Generalisierbarkeit der Ergebnisse. Aus diesem
Grunde wäre es sinnvoll, die Ergebnisse im Rahmen weiterer Studien, etwa an den Daten des
europäischen Sprachenindikators, zu überprüfen. Da dieser auch zusätzliche Kompetenzniveaus
erfassen soll, wäre hier auch möglicherweise ein Vergleich von testkulturellen Einflüssen über
unterschiedliche Niveaus hinweg denkbar.

Aufgrund des Datensatzes und dessen Struktur sind neben der Generalisierbarkeit der Ergebnis-
se auch Grenzen hinsichtlich darauf basierender Forschung und Forschungsmethoden gesetzt.
Da es sich um ein Multi-Matrix-Design handelt, in dessen Rahmen nicht alle Schüler alle Items
bearbeiteten, ist es beispielsweise nicht möglich, die genaue Struktur des Konstrukts innerhalb
der Länder sowie dahingehende Unterschiede zwischen den Ländern, beispielsweise anhand
von Mehrgruppenanalysen, zu erforschen. Aufgrund des Multi-Matrix-Designs haben Schüler je-
weils nur eine oder zwei Aufgaben aus einigen unterschiedlichen Ländern bearbeitet. Dies führt
dazu, dass nicht alle Items eines Landes für eine Überprüfung der Faktorenstruktur innerhalb der
unterschiedlichen Länder herangezogen werden können, weshalb anhand der hier vorhandenen
Daten keine stabilen Aussagen diesbezüglich gemacht werden können. Dies einzubeziehen wäre
für zukünftige Forschungsprojekte wünschenswert.


Ein letzter Vorbehalt betrifft die Einordnung der Items. Es wurde zwar der Versuch durchgeführt, die Ratings der ursprünglichen Experten zu validieren, jedoch waren die Inter-Rater-Korrelationen zwischen den ursprünglichen und den neu hinzugezogenen Experten auf niedrigem Niveau. Es wurde infolgedessen beschlossen, die ursprünglichen Ratings für die Einordnung der Items beizubehalten, da diese von langjährigen Experten stammten, bei denen es sich zum einen um die Autoren des „Dutch Grid“ handelt und diese zum anderen bereits über einen längeren Zeit-
Diskussion

6.4 Grenzen der Arbeit und zukünftige Forschungsperspektiven

raum hinweg gemeinsam an der Einordnung von Items gearbeitet hatten. Dennoch bleibt die Möglichkeit, dass aufgrund ihrer unterschiedlichen Herkunft länderspezifische Verzerrungen in die Bewertung der Items hinsichtlich ihrer Merkmale und ihrer Schwierigkeit eingeflossen sind. Dies lässt sich im Rahmen dieser Arbeit jedoch nicht eingehender überprüfen.


Die Ergebnisse dieser Arbeit weisen darauf hin, dass bei zukünftigen internationalen Leistungsvergleichen darauf geachtet werden sollte, gleich viele Items aus allen teilnehmenden Ländern in die Tests aufzunehmen. So könnten die durch die Verwendung bestimmter Aufgabenformate entstehenden Vor- und Nachteile der Gruppen zumindest teilweise über den Gesamttest hinweg ausgeglichen werden.

Auch das Konstrukt an sich könnte anhand eines anderen Forschungsdesigns untersucht werden. Denkbar wäre hier ein vollständiges Forschungsdesign, um beispielsweise konfirmatorische Faktorenanalysen und Mehrgruppenmodelle anwenden zu können. Mit Hilfe dieser Methoden könnte das Konstrukt hinsichtlich seiner Äquivalenz in den unterschiedlichen Ländern untersucht werden. So könnte eine Überprüfung der Äquivalenz des Konstrukts stattfinden, und eine
Aussage dahingehend gemacht werden, bezüglich welcher Konstruktkomponenten die Länder sich unterscheiden. Auch könnte untersucht werden, ob es sich dabei um Struktur- oder Niveau-unterschiede handelt.

Um die angesprochenen Forschungsfragen realisieren zu können, wäre ein Forschungsdesign notwendig, welches

1. zur Replikation der Ergebnisse mindestens die Länder erfasst, die auch in der vorliegenden Arbeit in die Analysen mit einbezogen wurden;

2. repräsentative Stichproben beinhaltet, um Aussagen über die Generalisierbarkeit der Ergebnisse machen zu können;

3. mehr Kompetenzniveaus beinhaltet als es in dieser Arbeit der Fall ist, weshalb mindestens die Jahrgangsstufen 5 bis 12, oder aber auch Studenten, beispielsweise der Anglistik, mit einbezogen werden sollten;

4. Items zur Erfassung aller Kompetenzstufen beinhaltet. (Diese sollten möglichst, wie auch in der vorliegenden Arbeit, aus den Teilnehmerländern stammen, damit überprüft werden kann, ob sich die Ergebnisse hinsichtlich der Testkulturen replizieren lassen und ob dies auf andere Kompetenzniveaus generalisiert werden kann);

5. vollständig ist, so dass beispielsweise Mehrgruppenanalysen durchgeführt werden können um die Konstrukstrukturen der Länder zu vergleichen;

6. längsschnittlich ist, um mögliche Veränderungen bezüglich der Testkulturen und des Konstrukts „fremdsprachliches Leseverständnis“ messen zu können und so möglicherweise die in Messicks Modell gemachten Annahmen hinsichtlich des Einflusses sozialer Werte und Konsequenzen von Testwertergebnissen auf das Konstrukt untersuchen zu können.

Es wurden ferner Möglichkeiten diskutiert, die Validität interkultureller Vergleiche im Bereich des fremdsprachlichen Leseverständnisses zu erhöhen.

Darüber hinaus wird auch ersichtlich, dass die Erforschung Differentieller Item Funktionen in der Domäne „Fremdsprachenkompetenzen“ sowie deren Ursachen einen wichtigen Forschungsbereich darstellt, in dem allerdings noch viele Fragen offen sind, deren Beantwortung anhand weiterer Forschungsvorhaben verfolgt werden sollte.
Anhang A

Im Falle von Anfragen bezüglich der in der Dissertation verwendeten Daten und Items sowie bezüglich der in der Arbeit als im Anhang gekennzeichneten zusätzlichen Analysen kann die Autorin unter jurecka@em.uni-frankfurt.de kontaktiert werden.
Literaturverzeichnis


Bonnet, G., Braxmeyer, N., Hornet, S., Lappalainen, H-P., Levasseur, J., Nardi, E., Remond,


bench_ed_trai_en.pdf. Zugriff am 26.04.07


nonlinear approach. Springer, New York


242


245


246


Zumbo, B.D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.
Danksagung
