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

Zusammenfassung

Die vorliegende Arbeit und die in ihr erörterten Fragestellungen wurden primär durch europäische Entwicklungen im Bereich der Fremdsprachenforschung und -diagnostik sowie durch die aktuelle, bildungspolitische Relevanz dieses Forschungsbereichs motiviert. Seit einigen Jahren sind sowohl in den mit dieser Frage befassten Institutionen der Europäischen Union als auch in den einzelnen Mitgliedsländern verstärkt Aktivitäten im Bereich des Lernens, Lehrens und Testens von Fremdsprachenkenntnissen zu beobachten. 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. Vor dem Hintergrund der wachsenden Anzahl länderübergreifender Vergleiche ist gleichzeitig das Interesse an der Frage der Fairness solcher Vergleiche gewachsen. Fairness in länderübergreifenden Vergleichen bedeutet hier, dass Personen auf dem gleichen Niveau hinsichtlich einer zu messenden latenten Variable unabhängig von Nationalität oder Gruppenzugehörigkeit dieselbe Wahrscheinlichkeit aufweisen, ein Testitem zu lösen. Sollte das nicht gegeben sein, ist dies ein Hinweis darauf, dass das zugrunde liegende Konstrukt nicht in allen Ländern oder Gruppen dasselbe ist. Die Validität des Tests wird dadurch verringert, und die Testaufgaben können nicht oder nur eingeschränkt für einen fairen Vergleich von Fähigkeiten verwendet werden. Das Phänomen von unterschiedlichen Item-Lösungswahrscheinlichkeiten bei gleicher Fähigkeit wird auch als Differentielle Item Funktion (DIF, engl.: „Differential Item Functioning", z.B. Holland & Wainer, 1993) bezeichnet.

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. Für den Bereich der Mathematik konnte die Existenz solcher Testkulturen von Klieme und Baumert (2001) und von Dogan, Guerrero und Tatsuoka (2005) anhand von TIMSS-

Daten gezeigt werden.

Die vorliegende Dissertation richtet also den Fokus auf die Analyse von Items zur Messung von fremdsprachlichem Leseverständnis 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.

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 Leseverständnissen 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 Gruppenzugehörigkeit verursachte Varianz erklärt werden. Der Arbeit werden drei Theoriebereiche zugrunde gelegt, nämlich Theorien und Modelle hinsichtlich Differentieller Item Funktionen, der Fremdsprachenforschung und angewandten Linguistiksowie der interkulturellen Vergleichbarkeit von Testergebnissen.

Hinsichtlich der Erklärung von DIF werden im Rahmen der Arbeit zwei Ansätze betrachtet. Der erste besagt, dass DIF aufgrund von konstruktirrelevanter Varianz zustande kommt, wobei diese auf die nicht-intendierte Messung von mindestens einer weiteren Dimension zurückführbar ist (z.B. Ackerman, 1992). Der zweite Ansatz besagt, dass DIF ein Ausdruck unterschiedlicher Stärken und Schwächen von Gruppen im Hinblick auf unterschiedliche Konstruktkomponenten ist (z.b. Scheuneman & Gerritz, 1990; Dogan, Guerrero & Tatsuka, 2005). Im Rahmen dieser Arbeit wird angenommen, dass DIF zumindest teilweise auch auf Letzteres zurückführbar ist. Ferner wird angenommen, dass diese Stärken und Schwächen von Gruppen auf unterschiedliche Testkulturen der Länder zurückzuführen sind, die sich wiederum auf den in den Ländern konstruierten Items abbilden. Darauf weist auch eine Reihe empirischer Studien hin (Artelt & Baumert, 2004; Klieme & Baumert, 2001; Dogan, Guerrero & Tatsuoka, 2005; Scheuneman & Gerritz, 1990).

Ein weiterer, der Arbeit zugrundeliegender Theoriebereich ist der Bereich der fremdsprachlichen Diagnostik welcher ein Teilbereich der Fremdsprachenforschung und angewandten Linguistik ist. Der vorliegenden Arbeit liegt theoretisch der Gemeinsame Europäische Referenzrahmen für Sprachen (GERS; Europarat, 2001) zugrunde, welcher ein Referenzsystem für das länderübergreifende Lehren, Lernen und Beurteilen von Sprachen ist. Der GERS beschreibt kriterienorientiert, was ein Sprachenlerner mindestens können muss, um ein bestimmtes Kompetenzniveau zu erreichen. Das Instrument sowie dessen Konstruktion und Validierung wird dargestellt.

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

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-Kategorisierungs-System „Dutch-Grid" (Alderson et al., 2006) zugrundegelegt, welches zum einen auf dem GERS basiert, und zum anderen eine Merkmalskategorie 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.

Da es sich bei Differentiellen Item Funktionen um eine Einschränkung der Validität von Testergebnissen handelt, wurde außerdem als Rahmenkonzept der Arbeit das Validitätsmodell von Messick gewählt. Dieses betrachtet soziale Werte und Folgen von Testwertinterpretation als Teil der Konstruktvalidität. Von Relevanz für die Arbeit ist außerdem der Umstand, dass im Rahmen dieses Modells zwei Quellen von Invalidität betrachtet werden, nämlich die Einführung konstruktirrelevanter Varianz und die Konstruktunterrepräsentation. Diese beiden Ursachen lassen sich mit den oben dargestellten Annahmen über die Entstehung Differentieller Item Funktionen gleichsetzen, nämlich mit der Betrachtung von DIF als konstruktirrelevante „Nuisance Dimension" sowie als „Stärken und Schwächen von Gruppen".

Mit Hilfe dieses Modells lassen sich außerdem Überlegungen hinsichtlich des Zustandekommens unterschiedlicher Testkulturen in unterschiedlichen Ländern anstellen. Darüberhinaus zeigen sich Berührungspunkte zu den drei der Arbeit zugrundeliegenden Theoriebereichen.

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.

- Zusammenfassung

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.

Der erste Fragenkomplex „Voraussetzungen und Skalierbarkeit" 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 ein eindimensionales IRT-Modell (Rasch-Modell; z.B. Embretson & Reise, 2000) gewählt wurde, handelt es sich bei der ersten zu überprüfenden Voraussetzung um die Rasch-Modell-Konformität der Items innerhalb der Länder. Zum Zweiten wurden die Items auf das Vorhandensein paarweiser Differentieller Item Funktionen unter Annahme des Rasch-Modells hin überprüft. Die dritte zu überprüfende Voraussetzung bezieht sich auf die Existenz unterschiedlicher Testkulturen in den Teilnehmerländern. Hier wurde untersucht, inwieweit anhand der eingereichten Items der Länder differentielle Testkulturen und somit unterschiedliche erwartete Stärken und Schwächen der Länder festgestellt werden können.
Zur Bearbeitung dieses ersten Fragenbereichs wurden die Daten zunächst unter Annahme eines einparametrischen Rasch-Modells erneut skaliert, um die Modellkonformität der Items zu überprüfen. Es zeigt sich, dass alle Items dem Modell entsprechen, legt man die Kriterien von Adams und Khoo (1996) zugrunde.
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% .

- Zusammenfassung

Im dritten Schritt des ersten Fragenbereichs wurde überprüft, ob sich aufgrund der Analyse der in den verschiedenen Teilnehmerländern konstruierten Items differentielle Testkulturen feststellen lassen. Dazu wurden die Items von Experten mit Hilfe des „Dutch Grid"-Item-Kategorisierungssystems eingeordnet. Es wurde überprüft, ob sich die Items der Länder hinsichtlich der Häufigkeit des Vorkommens der kognitiv-linguistischen Item-Anforderungsmerkmale unterscheiden. Signifikante Unterschiede wurden als Unterschiede hinsichtlich der Testkultur gewertet. Es wurde angenommen, dass Unterschiede der Testkultur auch zu differentiellen Lerngelegenheiten der aus den unterschiedlichen Kulturen stammenden Schüler führen und dass sich daher aufgrund der Testkulturen auch Hypothesen über die zu erwartenden Stärken und Schwächen von Schülergruppen hinsichtlich der Lösung von Items mit bestimmten Anforderungsmerkmalen aufstellen lassen. Es zeigten sich signifikante Unterschiede hinsichtlich der Testkultur-Profile der Länder. Darauf basierend wurden Hypothesen bezüglich der zu erwartenden Stärken und Schwächen der Gruppen aufgestellt, welche für die Analysen im dritten Fragenbereich, der sich mit der Erklärung von DIF befasst, benötigt wurden. Diese signifikanten Unterschiede hinsichtlich des Vorkommens von Item-Anforderungsmerkmalen wurden als Indikatoren nationaler Testkulturen interpretiert. Die Voraussetzung für die weiteren Analysen waren somit erfüllt.

Im zweiten Fragenbereich, „Erklärung von Itemschwierigkeit", wurde mit Hilfe von korrelativen und regressionsanalytischen 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 einzelnenLändern 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- Ausprä-

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.

Der dritte Komplex von Fragen, „Erklärung von differentiellen Item Funktionen", befasste sich mit der Untersuchung von Zusammenhängen zwischen DIF und Testkulturen, also kulturell bedingten Unterschieden der Itemschwierigkeiten zwischen den Ländern und aufgrund der Testkulturen zu erwartenden differentiellen Stärken und Schwächen der verschiedenen Gruppen.
Dazu wurden in einem ersten Schritt Korrelationen zwischen der Itemherkunft und paarweisen DIF-Parametern (zwischen jeweils 2 Ländern) berechnet. Die Annahme war hier, dass der Umstand, 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.
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

Im nächsten Schritt dieses Fragenbereichs 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 Stärken und Schwächen der Gruppen und somit auch bezüglich der Richtung der Korrelationen aufgestellt worden. Kommt ein Item-Anforderungsmerkmale bei den Items der einen Gruppe signifikant häufiger vor als bei den Items der Vergleichsgruppe, sollte diese Korrelation signifikant negativ ausfallen, da die Anwesenheit des Itemmerkmals mit einer geringeren Itemschwierigkeit (bzw. einer vorteilhaften Differentiellen Item Funktion) für diese Gruppe einhergeht, und umgekehrt. 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 entsprachen 23 der aufgrund der erwarteten Stärken und Schwächen der Gruppen angenommenen Richtung. Die Variablen „Itemtyp" und „Abstraktheit des Inhalts" wiesen bei den Englisch-Items insgesamt die meisten signifikanten Zusammenhänge mit DIF-Parametern auf. Ferner spielen auch noch die Itemmerkmale „Schwierigkeit des Vokabulars" und „Informationsgewinn 1" eine Rolle.
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 tedtkulturell 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 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 Ergebnisse weisen darauf hin, dass ein Zusammenhang zwischen Differentiellen Item Funktionen und Testkulturen besteht, weshalb die Hauptfragestellung insgesamt positiv beantwortet werden kann. Gerade auch die Replizierbarkeit der Ergebnisse in zwei Sprachen weist darauf hin. Vorbehalte sind etwa die Nicht-Repräsentativität der Stichproben und Stabilitäts- und Multikollinearitätsprobleme bei den Regressionsanalysen.
Die Relevanz der Ergebnisse für die drei Theoriebereiche wird diskutiert. Auch die Bedeutung der Ergebnisse für die Konstruktion und Validität von Tests für internationale Leistungsvergleiche wird dargelegt. Die Ergebnisse der Arbeit zeigen existierende Zusammenhänge zwischen DIF und Testkulturen in internationalen Leistungsvergleichen auf. Damit kann die Arbeit dazu beitragen, einen Teil der Ursachen von Differentiellen Item Funktionen aufzuzeigen. Dies ist ein wichtiger Schritt dafür, den Zielen „Fairness", „Validität" und „internationaler Vergleichbarkeit von Testergebnissen" näher zu kommen. Es wird aufgezeigt, dass es bei internationalen Leistungsvergleichen unverzichtbar ist, den Umstand, dass ein Einfluss der Testkultur auf die Itemschwierigkeit existiert, bei der Konstruktion von Testitems und der Zusammenstellung von Testverfahren zu berücksichtigen. Des weiteren werden mögliche Vorbehalte des gewählten

methodisch-theoretischen Ansatzes der Arbeit sowie der verwendeten Daten, wie etwa die Nicht-Repräsentativität der Stichproben, behandelt. Abschließend wird ein Ausblick auf zukünftige Forschungsperspektiven gegeben, die sich aus den Ergebnissen der Arbeit ableiten. Beispielsweise sollten die Ergebnisse an einer repräsentativen Stichprobe validiert und repliziert werden. Auch sollten dazu zusätzliche Kompetenzniveaus mit einbezogen werden, um die Generalsierbarkeit der Ergebnisse dahingehend zu überprüfen. Ferner wäre eine längsschnittliche Untersuchung des Einflusses sozialer Werte und der Interpretation von Testwertergebnissen auf die Veränderung von Testkultur von Interesse.

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 Vaildität von Testaufgaben zu minimieren, und die Fairness und auch Aussagekraft solcher Leistungsvergleich zu erhöhen.

1. Einleitung

Die vorliegende Arbeit und die in ihr erörterten Fragestellungen wurden primär durch europäische Entwicklungen im Bereich der Fremdsprachenforschung und -diagnostik sowie durch die aktuelle, bildungspolitische Relevanz dieses Forschungsbereichs motiviert: Seit einigen Jahren sind sowohl in den mit dieser Frage befassten Institutionen der Europäischen Union als auch in den einzelnen Mitgliedsländern verstärkt Aktivitäten im Bereich des Lernens, Lehrens und Testens von Fremdsprachenkenntnissen zu beobachten. Der Ursprung dieser vermehrten Aktivitäten liegt mutmaßlich in der stetig wachsenden Anzahl von EU-Mitgliedsstaaten und den damit einhergehenden transnationalen Tendenzen, durch die Fremdsprachenkenntnisse in verschiedensten Lebensbereichen, nicht zuletzt für Beruf, Berufswahl und Wahrnehmung von Karrierechancen im Ausland, zunehmend an Bedeutung gewinnen.

Das Jahr 2001 wurde von der Europäischen Union gemeinsam mit dem Europarat zum Jahr der Fremdsprachen in Europa ausgerufen. Dies hatte zum Ziel, sowohl die sprachliche und kulturelle Vielfalt Europas zu erhalten, als auch die Bürgerinnen und Bürger der Europäischen Union dazu anzuregen, sich neue Fremdsprachenkenntnisse anzueignen sowie bereits vorhandene zu erweitern. Vor diesem Hintergrund hat die Europäische Kommission einen Aktionsplan zur Förderung des Sprachenlernens und der Sprachenvielfalt beschlossen und vorgestellt (Europäische Kommission, 2003). Als überaus wichtig werden in diesem Zusammenhang vor allem die Förderung der Kommunikation über die Ländergrenzen hinweg, die Förderung von individueller Mobilität und Informationszugang sowie die Förderung von gegenseitigem Verständnis und Toleranz erachtet. Die Europäische Kommission ist der Auffassung, dass die Umsetzung dieser Ziele in der EU in Anbetracht von etwa 450 Millionen EU-Bürgern und 23 offiziell als Amtssprachen gesprochenen Sprachen ein erhöhtes Maß an Mehrsprachigkeit erfordert.

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.

In diesem Zusammenhang ist ein wichtiges Instrument der Gemeinsame Europäische Referenzrahmen für Sprachen (Europarat, 2001), auf dem gegenwärtig ein stetig wachsender Anteil der zunehmend internationalisierten europäischen Fremdsprachenforschung basiert. Beim europäischen Referenzrahmen für Sprachen handelt es sich um ein Instrument, welches Fremdsprachenkenntnisse in sechs unterschiedliche Niveaus, von A1-C2, einteilt und diese mit Hilfe sogenannter Can-Do-Statements kriterienorientiert beschreibt (siehe auch 2.2.3).

In dem Forschungsumfeld dieses Referenzrahmens kam es in den letzten Jahren immer wieder zu wissenschaftlichen Debatten. So befasst sich eine von mehreren aktuellen Forschungsdebatten mit der Frage: „Ist mein B1 auch Dein B1?" (Fandel et al., 2007), das heißt mit anderen Worten: „Messen wir das gleiche Konstrukt?" – „Ist unsere Wahrnehmung von Fremdsprache die gleiche?" Hintergrund ist hier die Unklarheit darüber, ob die verschiedenen Niveaus des GERS in verschiedenen Ländern tatsächlich als identisch wahrgenommen und in der gleichen Weise gehandhabt werden, d.h. konkret, ob dasselbe Konstrukt gemessen wird (und auch der Konstruktion von Fremdsprachentests zugrunde liegt), oder ob Items als unterschiedlich leicht oder schwer wahrgenommen werden. Die vorliegende Arbeit ist auch vor dem Hintergrund dieser Debatte zu betrachten.

Unter Anderem um Unklarheiten wie den soeben beschriebenen auf den Grund zu gehen bedient sich die Forschung seit einigen Jahren vermehrt der Möglichkeiten von Large Scale Assessments und interkulturellen Kompetenzvergleichen. Deren Ziel ist es unter anderem, bei der Realisierung bestimmter gesellschaftlicher und politischer Zielvorstellungen Hilfestellung zu leisten, indem eine wissenschaftlich valide Grundlage hinsichtlich des Wissens über Leistung und Kompetenzen zur Verfügung gestellt wird. Die vorliegende Arbeit basiert auf einem solchen Ansatz.

Bereits in der Vergangenheit haben solche internationalen Vergleiche Einfluss auf Bildungssysteme und Lehrpläne genommen. So wurde beispielsweise in Deutschland von der Kultusministerkonferenz in Folge von PISA-Ergebnisse eine systematisierteDarstellung der Maßnahmen der Länder zur Qualitätsentwicklung und Qualitätssicherung gefordert. Ferner wurden in vielen deutschen Bundesländern fortlaufende Lernstandsermittlungen initiiert (Avenarius et al., 2003). Derartige Prozesse, bei denen Vergleichsstudien Einfluss auf die Wissensvermittlung haben, werden auch als Washback-Effekt bezeichnet und können gesellschaftlich hoch relevant sein (Klein, Hamilton, McCaffrey & Stecher, 2000 ; siehe auch 2.4). Aus diesem Grund ist es, gerade vor dem Hintergrund von kulturübergreifenden Vergleichen, umso wichtiger, dass die Vergleiche, auf deren Ergebnisse sich in Diskussionen, Debatten und Aktionen gestützt wird, valide und fair sind. Hierzu soll die vorliegende Arbeit einen Beitrag leisten, indem untersucht wird, welche Faktoren bei der länderübergreifenden Erfassung von Fremdsprachenkompetenzen Unfairness bedingen könnten.

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 angeeigneten 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

vergleichbare Ergebnisse liefern können.

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.

Die Europäische Union hat im Rahmen des Arbeitsprogramms „Allgemeine und berufliche Bildung" im Jahr 2002 beschlossen, bis 2010 für wichtige Fächer europaweit gültige Indikatoren für den länderübergreifenden Vergleich von Kompetenzen zu schaffen (Europäische Kommission, 2002). In diesem Zusammenhang wurde gleichfalls die Entwicklung eines europäischen Indikators für Fremdsprachenkompetenzen geplant.

Vor dem Hintergrund der wachsenden Anzahl länderübergreifender Vergleiche und der oben genannten kritischen Punkte ist gleichzeitig das Interesse an der Frage der Fairness solcher Vergleiche gewachsen. Fairness in länderübergreifenden Vergleichen bedeutet hier, dass alle Personen, die sich hinsichtlich einer zu messenden Fähigkeit auf dem gleichen Niveau befinden, in der Lage sein sollten, ein Testitem mit der gleichen Wahrscheinlichkeit korrekt lösen zu können, und zwar unabhängig von Nationalität oder Gruppenzugehörigkeit. Sollte das nicht gegeben sein, ist dies ein Hinweis darauf, dass das zugrunde liegende Konstrukt nicht in allen Ländern oder Gruppen dasselbe ist; die Validität des Tests wird dadurch verringert, und die Testaufgaben können nicht oder nur eingeschränkt für einen fairen Vergleich von Fähigkeiten herangezogen werden. Das Phänomen von unterschiedlichen Item-Lösungswahrscheinlichkeiten wird auch als „Differentielle Item Funktion" (kurz: DIF, engl.: „Differential Item Functioning", beispielsweise Holland & Wainer, 1993; siehe auch 2.1) bezeichnet.

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.

Für den Bereich der Mathematik konnte die Existenz solcher Phänomene von Klieme und Baumert (2001) anhand von TIMSS-Daten gezeigt werden. Dort führten verschiedene Unterrichtsstile in sechs Ländern dazu, dass Items mit bestimmten Eigenschaften von Schülern aus Ländern, in denen diese Komponenten besonders gefördert wurden, auch eine höhere Wahrscheinlichkeit aufwiesen, das Item korrekt zu lösen. In diesem Fall ist anzunehmen, dass unterschiedliche Bildungskulturen zu unterschiedlichen Test- oder Unterrichtsstilen und letztlich zu unterschiedlichen Ausprägungen von Teilaspekten mathematischer Kompetenzen und somit zu unterschiedlichen Kompetenzprofilen führten. Dies konnte einen Teil der in TIMSS gefundenen Differentiellen Item Funktionen erklären. Auch im Zusammenhang mit PISA-2000-Ergebnissen bei der Messung von Leseverständnis konnte von Artelt und Baumert (2004) gezeigt werden, dass es einen kulturspezifischen Einfluss auf die Itemschwierigkeiten zu geben scheint: Items aus dem eigenen Sprach- und Kulturraum verringerten die Testschwierigkeit, und die gemittelte Differentielle Item Funktion eines Tests (Differentielle Test Funktion) wurde in vielen Fällen zugunsten der jeweiligen Gruppe verschoben. In dieser Arbeit wird nun der Frage nachgegangen, ob Ansätze wie die gerade beschriebenen auch auf Items zur Messung fremdsprachlichen Leseverständnisses übertragbar sind.

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.

Das Vorhandensein von DIF bedeutet im Kontext dieser Dissertation, dass Schüler *mit gleichen Fremdsprachenkompetenzen* aufgrund ihrer Länderzugehörigkeit unterschiedlich hohe Wahrscheinlichkeiten haben, ein Item korrekt zu beantworten. DIF sind daher ein Ausdruck für durch Gruppenzugehörigkeit verursachte Varianz der Itemschwierigkeiten. Ein Item, welches DIF aufweist, sollte nicht oder nur eingeschränkt für einen validen Vergleich von Leistung verwendet werden.
Grundidee dieser Dissertation ist, dass sich diese, durch kulturelle Zugehörigkeit verursachte Varianz 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 Sprachitems abbilden. Das heißt, unterschiedliche Schwerpunkte der Länder hinsichtlich dieser Aspekte (im Folgenden auch als Testkulturen bezeichnet) können durch die Analyse von Fremdsprachenitems festgestellt werden und zugleich kann damit die durch Gruppenzugehörigkeit verursachte Varianz erklärt werden. Zur Operationalisierung dieser Testkulturen werden die aus den unterschiedlichen Teilnehmerländern der EBAFLS-Studie stammenden Items dahingehend analysiert, ob signifikante Unterschiede hinsichtlich der Häufigkeit des Vorkommens von schwierigkeitsdeterminierenden, kognitiv-linguistischen Item-Anforderungsmerkmalen existieren. Das Vorhandensein solcher

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.

Eine weitere zentrale Aufgabe dieser Dissertation besteht in der Identifikation von schwierigkeitsbestimmenden Itemeigenschaften, die einen relevanten Beitrag zur Erklärung der in der EBAFLS-Studie gefundenen Differentiellen Item Funktionen liefern können. Zur Kategorisierung der Items hinsichtlich solcher Anforderungsmerkmale wird das in der angewandten Linguistik verwendete Item-Kategorisierungssystem „Dutch Grid" (Alderson et al., 2006) verwendet. Da es sich bei Differentiellen Item Funktionen um Unterschiede hinsichtlich der Itemschwierigkeit in unterschiedlichen Gruppen handelt, besteht der erste Schritt, und somit auch die erste Fragestellung der Arbeit darin, zu überprüfen, inwieweit die dort verwendeten kognitiv-linguistischen Anforderungsmerkmale dazu geeignet sind, Itemschwierigkeiten *innerhalb* der Länder zu erklären. Damit soll untersucht werden, inwieweit das verwendete Kategorisierungssystem bzw. die dort verwendeten Itemeigenschaften überhaupt mit der Itemschwierigkeit zusammenhängen. In einem zweiten Schritt sollen die Itemeigenschaften, die sich als geeignet herausstellen, auch zur Analyse von Zusammenhängen mit Differentiellen Item Funktionen herangezogen werden. Auf diesem Wege soll geklärt werden, inwieweit sich Differentielle Testkulturen und dadurch bedingte länderspezifische Stärken und Schwächen aufweisen lassen und inwieweit sich diese zur Erklärung von DIF eignen. Die vorliegende Arbeit ist der pädagogisch-psychologischen Diagnostik im Bereich der internationalen Leistungsvergleiche, sowie der interkulturellen Psychologie zuzuordnen. Ergänzend werden Theorien der angewandten Linguistik verwendet. Die aus den unterschiedlichen Disziplinen stammenden, theoretischen Grundlagen werden im nächsten Abschnitt dargelegt.

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.

Der zweite Abschnitt des Theorieteils legt den Fokus auf das Testen fremdsprachlichen Leseverständnisses. Zunächst wird ein Überblick über Modelle und Konstrukte fremdsprachlichen Leseverständnisses gegeben. Ein besonderes Augenmerk wird dabei auf die Konstruktion und die theoretischen und empirischen Grundlagen des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GERS; Europarat, 2001) gelegt. Dieser stellt die theoretische und praktische Grundlage der EBAFLS-Studie und somit auch dieser Dissertation dar. Weiterhin wird dargelegt, welche Charakteristika eines Items für die Itemschwierigkeit bzw. die Unterschiedlichkeit von Itemschwierigkeiten bei Items zur Messung fremdsprachlichen Leseverständnisses ursächlich sein könnten. Diesbezüglich wird gleichfalls ein Überblick über relevante empirische Ergebnisse und den Stand der Forschung gegeben. In diesem Rahmen soll außerdem das Itemkategorisierungssystem „Dutch Grid" (Alderson et al., 2006) vorgestellt werden. Dieses basiert auf dem GERS, und wurde im Rahmen der EBAFLS Studie zur Kategorisierung der Items hinsichtlich ihrer kognitiv-linguistischen Merkmale verwendet.

Da es sich bei der EBAFLS Studie um eine multinationale Studie handelt, werden dann im dritten Abschnitt einige für diese Dissertation relevante Theorien und Methoden für die Überprüfung der interkulturellen Vergleichbarkeit von Testverfahren dargestellt. Diese stammen aus der Disziplin der interkulturellen Psychologie. Auch und vor allem in diesem Bereich der Psychologie spielen DIF eine große Rolle. Außerdem werden Methoden zur Überprüfung der Äquivalenz von Konstrukten über verschiedene Kulturen hinweg vorgestellt.

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

Rolle, vor allem für die Übertragung von sozialen Werten einer Gesellschaft auf Testitems und Testkulturen. In diesem vierten Abschnitt wird zunächst das Validitätskonzept vorgestellt und dann in einem zweiten Schritt auf den Hintergrund dieser Dissertation übertragen.

2.1. Differentielle Item Funktionen

Die Modellierung sowie mögliche Ursachen für Differentielle Item Funktionen (z.B. Holland & Wainer, 1993; Camilli & Shepard, 1994) im Bereich des fremdsprachlichen Leseverständnisses sind der Ausgangspunkt dieser Dissertation. Innerhalb dieses ersten Theorieabschnitts wird zunächst auf die Definition und die historischen Hintergründe von DIF eingegangen. In einem zweiten Teil werden die klassischen Methoden zur Analyse von DIF dargestellt, und in einem dritten Teil wird die Bedeutung von DIF für die Validität diskutiert. Ferner wird darauf eingegangen, welche empirischen Ergebnisse hinsichtlich der Modellierung von DIF existieren und welche möglichen Ursachen für DIF genannt werden.

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, gruppenspezifische 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).

Ein Problem hinsichtlich solcher Analysen stellte sich zunächst jedoch darin dar, dass ein externes Validitätskriterium zur Überprüfung einzelner Items fehlte. Aus diesem Grund wurde das Gesamttestergebnis zugrunde gelegt, um die beiden Gruppen hinsichtlich ihrer tatsächlichen Fähigkeiten wenigstens annähernd zu matchen (Angoff, 1993). Der Ausdruck „Fähigkeit" bezeichnet daher im Zusammenhang mit DIF häufig den Gesamt-Test-Score. Ein weiteres Problem stellte die unscharfe Verwendung von Begrifflichkeiten dar. So wurde zu dieser Zeit und vor dem damaligen politischen Hintergrund der Ausdruck „Bias" sowohl im Sinne einer statistischen Abweichung als auch als Ausdruck für soziale Vorurteile, Ungerechtigkeit und Diskriminierung gebraucht. Damit existierte einerseits eine soziale, andererseits aber auch eine technische Konnotation des Wortes, was eine präzise Erörterung und Bewertung von Bias erschwerte (Cole, 1993). Um diese beiden Bedeutungen klar zu trennen, wurde der Terminus „Differentielle Item Funktion" für den technischen, statistischen Bereich eingesetzt. Gleichwohl wird auch heute der Ausdruck „Item Bias" in der Literatur häufig noch synonym mit dem Ausdruck DIF verwendet.

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?

Es haben sich insgesamt mindestens drei verschiedene Methodengruppen zur Analyse von DIF entwickelt: Das Modellieren von Itemantworten mit Hilfe von Kontingenztafeln (Mantel-Haentzel-Methode; z.B. Holland & Thayer, 1988) und/oder Regressionsmodellen, Modelle im Rahmen der Item Response Theorie, sowie mehrdimensionale Modelle (Zumbo, 2007). Diese drei methodischen Ansätze werden im folgenden Abschnitt besprochen. Dabei werden der Mantel-Haentzel-Ansatz sowie die logistische Regression nur in Kürze dargestellt, da in dieser Dissertation ausschließlich Modelle im Rahmen der Item Response Theorie zur Analyse und Modellierung von DIF verwendet werden. Auch auf die mehrdimensionalen Modelle wird nur in Kürze eingegangen, da in dieser Arbeit aus theoretischen Gründen, die unter 4.2 genauer erläutert werden, eine eindimensionale Skalierung zur Anwendung kommt.

Methoden zur Analyse von Differentiellen Item Funktionen. Die Gruppen, die Gegenstand der DIF-Analyse sind, werden klassischerweise in eine *Referenz-* und eine *Fokalgruppe* unterteilt. Die Fokalgruppe ist jene Gruppe, deren Performanz hinsichtlich eines Items von Interesse ist und deren Leistung mit der Leistung der jeweiligen Referenzgruppe(n) verglichen wird.

Die erste der angesprochenen Richtungen von Methoden zur Analyse von DIF folgt dem varianzanalytischen Gedanken, nämlich dass es sich bei DIF um die Untersuchung des Haupteffekts der Gruppenzugehörigkeit, bzw. des Interaktionseffektes von Gruppenzugehörigkeit und Leistung (oder, um in der eher statistischen Konnotation zu bleiben, des Test Scores) handelt. Daraus sind zwei breite Ansätze, Items mit DIF zu identifizieren (DIF-Detektion), hervorgegangen: erstens der sogenannte *Mantel-Haenszel-Ansatz*, (MH; Holland & Thayer, 1988) und zweitens die *logistische Regression.* Der MH-Ansatz beruht auf einer 3-fach-Kontingenztabelle, die drei Dimensionen beinhaltet. Die erste bezieht sich darauf, ob ein Item richtig oder falsch beantwortet wurde. Die zweite bezieht sich auf die Gruppenzugehörigkeit, und die dritte auf den Gesamttestscore, welcher als ein Maß für die Fähigkeit verwendet wird (siehe oben). Üblicherweise werden mehrere solcher Kontingenztabellen generiert, die jeweils die Mitglieder der beiden Gruppen, die sich im gleichen Gesamttestscore-Intervall der Verteilung befinden, beinhalten. Anhand dieser Prozedur wird die Wahrscheinlichkeit geschätzt, dass ein Mitglied der Fokalgruppe mit einem bestimmten Gesamttestscore das Item korrekt beantwortet. Es wird dann überprüft, ob sich ein statistisch signifikanter Unterschied zu Mitgliedern der Referenzgruppe mit demselben Gesamttestscore finden lässt. Im Rahmen des MH-Ansatzes wird davon ausgegangen, dass über alle Kontingenztabellen, d.h. über die verschiedenen Leistungsintervalle hinweg, der gefundene Unterschied gleich groß ist („uniform DIF").

Der Ansatz der *logistischen Regression* (Swaminathan, & Rogers, 1990) beruht auf der Verwendung von binären Itemscores und beinhaltet die Durchführung einer Regressionsanalyse für jedes einzelne Item. Dabei wird der statistische Einfluss der Gruppenzugehörigkeit, der Interaktion der Gruppenzugehörigkeit und der Leistung überprüft. Als Leistungsmaß dient der Gesamttestscore, nachdem die Gruppen diesbezüglich abgeglichen wurden. Der Unterschied der beiden Ansätze besteht vor allem darin, dass im Rahmen des MH-Ansatzes die konditionierende Variable diskret ist, was bei der logistischen Regression nicht der Fall ist. Auf der anderen Seite nimmt der MH-Ansatz im Gegensatz zur logistischen Regression keine Interaktion an (Zumbo, 2007), weshalb mit Hilfe der logistischen Regression auch sogenannter „non-uniform" DIF untersucht werden kann. Auf die Bedeutung von „uniform" versus „non-uniform" DIF wird im folgenden Abschnitt im Rahmen der Item Response Modelle genauer eingegangen. Beide Ansätze, MH und logistische Regression, legen den Analysen die beobachteten Werte zugrunde.

Ein weiterer Analyseansatz besteht in der Verwendung von Modellen der *Item Response Theorie (IRT).* Diese werden auch als „latent ability models" bezeichnet. Hier wird, im Gegensatz zu den oben beschriebenen Modellen, die geschätzte latente Fähigkeit als Matching-Variable und somit zur Beurteilung des Unterschieds zwischen Bias bzw. DIF und Impact verwendet. Im Rahmen von Item Response Modellen werden die itemcharakteristischen Funktionen (engl.: Item-Characteristik-Curves (ICC)) zweier oder mehr Gruppen hinsichtlich eines Items verglichen. Sie zeigen die Relation zwischen der Wahrscheinlichkeit, ein Item korrekt zu beantworten, und der vom Test gemessenen Fähigkeit.
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.

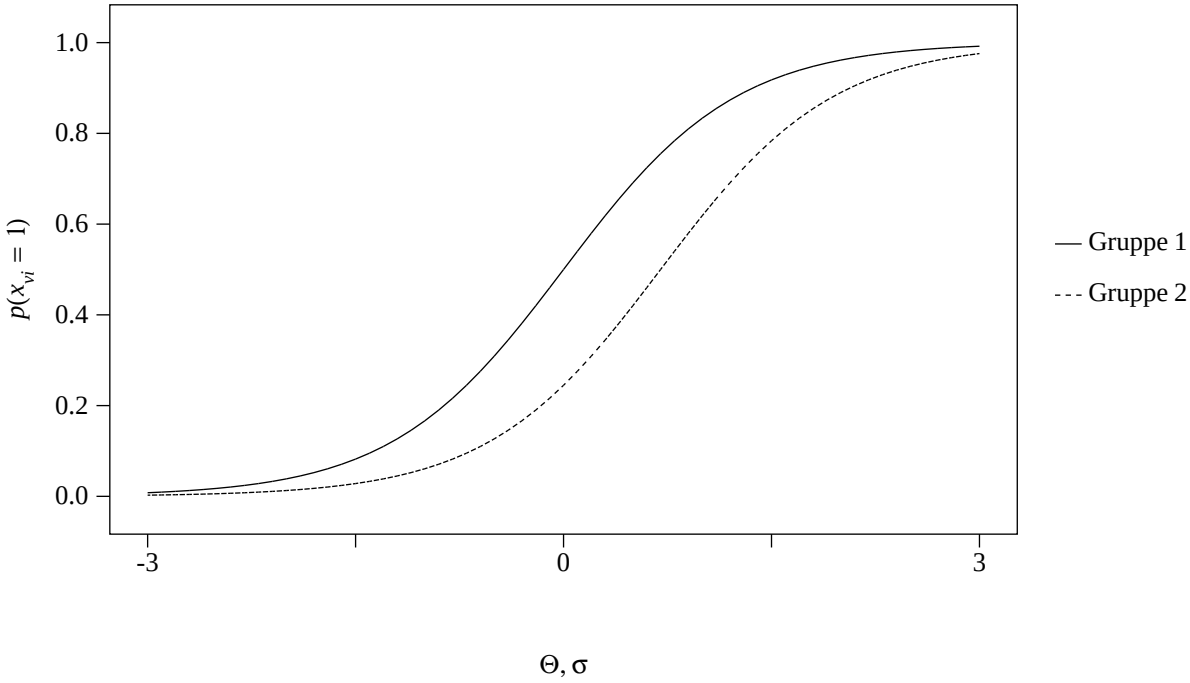

Abbildung 2.1. Differentielle Item Funktion in einem 1PL-IRT-Modell (uniform DIF)

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 Haupteffekt der Gruppenzugehörigkeit (Zumbo, 2007). Dieses Modell wird auch als

1PL-Modell (One-Parameter Logistic Model) bezeichnet. Den in dieser Arbeit durchgeführten DIF-Analysen wird ein solches 1-Parameter-Modell zugrunde gelegt (2.1). Die diesem Modell zugehörige Modellgleichung sowie eine detaillierte Beschreibung der in dieser Dissertation verwendeten Methoden werden im Rahmen des Methodenteils unter 4.2 dargelegt.

ICCs können sich jedoch nicht nur hinsichtlich ihrer Position auf der X-Achse, sondern auch hinsichtlich ihrer Steigung, d.h. hinsichtlich des Diskriminationsparameters, unterscheiden. Die Steigung bzw. Item-Diskrimination sagt etwas darüber aus, wie gut das Item zwischen den Schülern mit niedriger und hoher Fähigkeit trennen kann. Dieser Parameter entspricht dem klassischen Trennschärfe-Parameter. IRT-Modelle, die sowohl den Schwierigkeits- als auch den Diskriminations-Parameter berücksichtigen, werden als 2PL-Modelle bezeichnet. Die beiden Kurven können sich in einem solchen Modell zusätzlich auch überschneiden. Diese Art von DIF wird auch als „non-uniform" DIF bezeichnet, und entspricht im Prinzip der Interaktion von Gruppe und Fähigkeit. Ein solcher Fall wird in Abbildung 2.2 dargestellt:

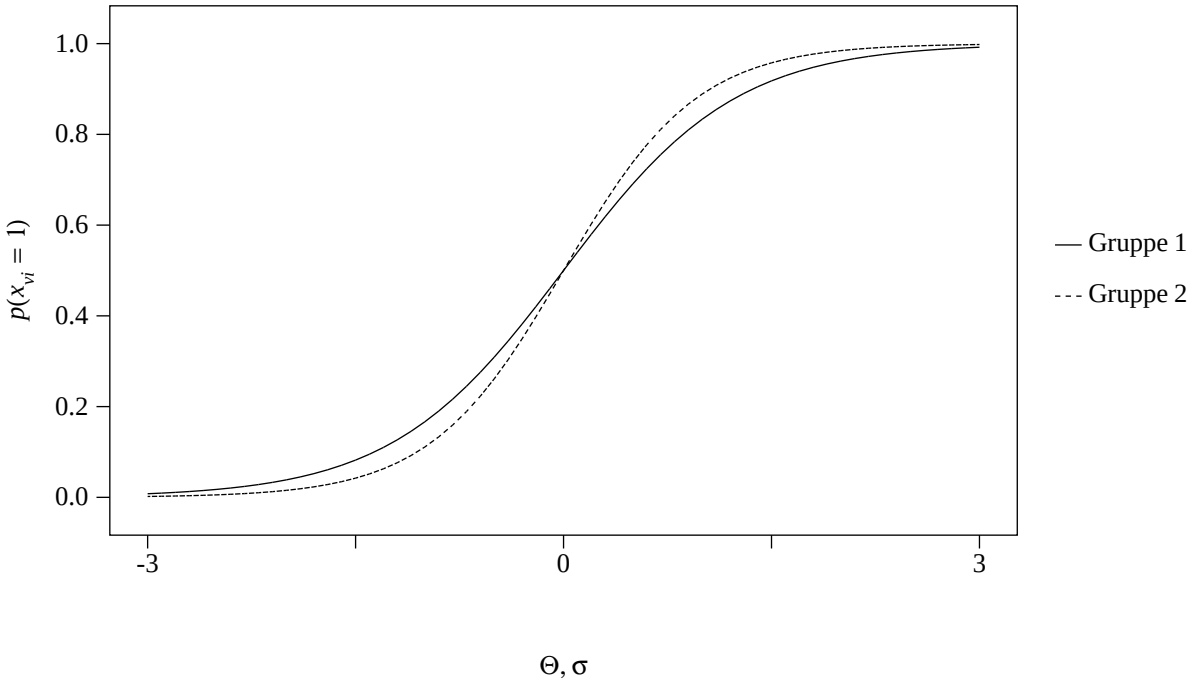

Abbildung 2.2. DIF in einem 2PL-IRT-Modell(non-uniform DIF)

Abb. 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. DIF-Methoden im Rahmen der Item Response Theorie zielen darauf ab, den Flächeninhalt zwischen den zwei Kurven zu analysieren und zu beschreiben, bzw. die Item-Parameter der beiden Gruppen zu vergleichen. Ferner werden die beiden Gruppen bei dieser Methode nicht explizit hinsichtlich ihres Gesamt-Testscores gematcht; vielmehr liegt den IRT-Modellen die Annahme zugrunde, dass die Fähigkeitsverteilung „herausintegriert" wurde (Zumbo, 2007): die Fläche zwischen den ICCs wird über die Verteilung des gesamten Kontinuums, Theta, berechnet.

Ein weiterer Analyseansatz, welcher die dritte Gruppe von DIF-Methoden darstellt, bezieht sich auf die Verwendung mehrdimensionaler IRT-Modelle. Dort wird davon ausgegangen, dass DIF ein Ausdruck der Mehrdimensionalität eines Items ist (z.B. Roussos & Stout, 1996; Gierl, 2005). Diese Modelle können auch zum Testen von Hypothesen im Hinblick auf die Ursache von DIF verwendet werden (Zumbo, 2007). Die Modelle bedienen sich unterschiedlicher Ansätze, so beispielsweise der „simultaneous item bias test" (Shealy & Stout, 1993), aber auch Ansätze basierend auf Strukturgleichungsmodellen und Item Response Theorie (Muthén, 1988; Muthén & Lehman, 1985). Auf die Methodik dieser Modelle soll hier jedoch nicht näher eingegangen werden, da die Verwendung mehrdimensionaler Modelle in dieser Arbeit nicht angestrebt wird. Dafür existieren mehrere Gründe: Zum Einen geht der in dieser Arbeit als theoretische Grundlage verwendete „Gemeinsame Europäische Referenzrahmen für Sprachen" (Europarat, 2001; siehe auch 2.2.3) von einer Eindimensionalität der Fremdsprachenfähigkeit aus. So wurden die dort verwendeten Skalen-Deskriptoren im Rahmen eines Rasch-Modells eindimensional skaliert (North, 2000), weshalb dieses auch in der vorliegenden Arbeit zugrunde gelegt wird. Zum Zweiten wurden die in dieser Arbeit verwendeten Daten auch in der ursprünglichen EBAFLS-Studie bereits eindimensional skaliert. Für eine tiefer gehende Erörterung multidimensionaler Modelle wird beispielsweise auf Roussos & Stout, (1996) und auf Zumbo (2007) als Übersicht verwiesen.

Mittlerweile werden häufig die auf der Item Response Theorie basierenden Methoden zur Detektion 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).

Eine weitere, vor allem für die Erstellung von Testverfahren relevante Frage bezieht sich darauf, wann und ob DIF als substantiell betrachtet werden sollte. Laut Camilli und Shepard (1994) existieren diesbezüglich zwei verwandte Methoden. Die eine basiert auf einem Index für die Größe von DIF, die andere auf einem Maß der statistischen Signifikanz. Häufig ist die Größe des Effekts eher von praktischer Relevanz als die statistische Signifikanz wenn es das Ziel einer Studie ist, mögliche Ursachen für DIF zu untersuchen. Die Frage nach der Signifikanz hingegen wird für Testkonstrukteure dann relevant, wenn ein DIF-freier Test entwickelt werden soll. ETS (Educational Testing Service), das größte private Testinstitut der USA, haben beispielsweise eine eigene, 3-stufige Klassifikation im Bezug darauf, ob DIF als substantiell betrachtet werden sollte oder nicht, aufgestellt. Diese besteht aus einer Kombination von Signifikanztest und Effektstärke (Dorans & Holland, 1993). Bezogen auf IRT-Modelle argumentieren Thiessen, Steinberg & Wainer (1993), dass DIF in dem Moment vorhanden ist, in dem sich die Itemcharakteristischen Funktionen zweier Gruppen bezüglich eines Items unterscheiden. Die Frage sei nun, inwieweit sie sich signifikant unterscheiden.

Im Rahmen dieser Dissertation werden DIF, wie im nächsten Abschnitt dargelegt wird, hingegen als unterschiedliche Stärken und Schwächen der Gruppen interpretiert, und es sollen mögliche Ursachen für DIF untersucht werden. Aus diesem Grund spielt die Frage nach der Signifikanz der Unterschiede in der vorliegenden Arbeit insofern keine Rolle, als dass es hier nicht das Ziel ist, einen DIF-freien Test zu entwickeln. Daher werden solche Maße im Rahmen dieser Dissertation zwar verwendet, spielen jedoch keine Rolle im Hinblick auf die Frage, ob ein Item in die weiteren Analysen eingeschlossen wird oder nicht. Scheuneman und Gerritz (1990), die einen ähnlichen Ansatz verfolgen, empfehlen sogar, DIF unabhängig von Größe und Signifikanz in solche Analysen mit einzubeziehen, da es ansonsten aufgrund von Varianzeinschränkung schwieriger sei, mögliche Ursachen zu entdecken.

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 unbeabsichtigterweise 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.

Eine andere, jedoch seltener geäußerte Betrachtungsweise besteht darin, DIF nicht als etwas Konstruktirrelevantes, als „ärgerliche" Dimension („Nuisance Dimension"), sondern als Ausdruck differentieller Profile von Stärken und Schwächen der Gruppen hinsichtlich des gleichen Konstrukts zu betrachten (z.B. Scheuneman & Gerritz, 1990; Klieme & Baumert, 2001; Dogan, Guerrero & Tatsuoka, 2005). Diesbezüglich äußern sich Scheuneman und Gerritz (1990) wie folgt:

„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).

Dabei wird davon ausgegangen, dass sich diese Stärken und Schwächen hinsichtlich der Performanz aufgrund der Gruppenzugehörigkeit manifestieren, beispielsweise durch differentielle Lerngelegenheiten in den unterschiedlichen Systemen. Hier bedeutet DIF auch nicht zwingend eine Verringerung der Konstruktvalidität des Tests insgesamt, sondern kann Hinweis auf eine zwischen den Gruppen unterschiedliche Ausprägung relevanter Aspekte des Konstrukts sein. Begreift man DIF in diesem Sinne als diagnostisches Instrument, kann DIF zur Analyse differentieller Stärken und Schwächen genutzt werden. Auch empirische Ergebnisse von Klieme und Baumert (2001) weisen darauf hin, dass unterschiedliche Stärken und Schwächen von Gruppen einen systematischen Einfluss auf Unterschiede hinsichtlich der Itemschwierigkeit zu haben scheinen. Dieser Argumentationslinie folgt auch die vorliegende Dissertation.

2.1.2. Die Erklärung von Differentiellen Item Funktionen

In diesem Abschnitt wird auf die Verwendung unterschiedlicher Methoden sowie auf empirische Ergebnisse hinsichtlich der Modellierung Differentieller Item Funktionen eingegangen. Dabei wird der Fokus auf Studien gelegt, die sich mit der Erklärung und Vorhersage von DIF anhand von Itemmerkmalen beschäftigen. Mit der Suche nach Ursachen für DIF wurde erst in der jüngeren Vergengenheit begonnen. Zumbo (2007) bezeichnet die Beschäftigung mit den Ursachen als die dritte Generation von DIF-Analysen, nach der Konzeptentwicklung als erster und der Entwicklung von Methoden zur empirischen Untersuchung von DIF als zweiter Generation. Nach Zumbo (2007) gibt es fünf unterschiedliche Gründe für die Verwendung von DIF-Analysen der dritten Generation: 1) Verbesserung von *Fairness und Äquivalenz* beim Testen, 2) Beschäfti-

gung mit der Bedrohung der *Test-Validität*, 3) Untersuchung der *Vergleichbarkeit* von übersetzten oder überarbeiteten Testverfahren, 4) das Verstehen von *Prozessen bei der Item-Beantwortung* und 5) die Untersuchung von *Gründen für fehlende Invarianz* zwischen Gruppen. Diese Arbeit befasst sich vornemlich mit den Punkten vier und fünf.

Klassischerweise wurden DIF immer als ein Problem des Items an sich betrachtet (Item Bias) und nicht der Personen bzw. Gruppen, zwischen denen DIF beobachtet wurden. Mittlerweile jedoch wurden auch Anstrengungen unternommen, Personenmerkmale zur Erklärung von DIF mit einzubeziehen (z.B. Van den Noortgate & de Boeck, 2005), und zwar teilweise mit Hilfe von „Logistic Mixed Models" , die zur Gruppe der sogenannten „Explanatory Item Response Models" (Wilson & deBoeck, 2004) gehören. Kausal- und Zusammenhangsmodelle werden in den letzten Jahren immer häufiger angewandt (Zumbo, 2007). Der Versuch, den Ursachen für Differentielle Item Funktionen auf den Grund zu gehen, fand in verschiedensten Kontexten statt, klassischerweise vorrangig in der Bildungsforschung, und dort teilweise auch im Rahmen von Large Scale Studien wie PISA (Programme for International Student Assessment; z.B. Artelt & Baumert, 2004), oder TIMSS (Trends in International Mathematics and Science Study, z.B. Klieme & Baumert, 2001; Klieme & Bos, 2000).

Für die Modellierung von DIF sind unterschiedliche Methoden angewandt worden. Die wohl häufigste ist eine *zweischrittige* Methode, in der in einem ersten Schritt DIF-Analysen durchgeführt und die so gewonnenen DIF-Parameter in einem zweiten Schritt mit Hilfe verschiedener, itemseitiger Prädiktoren per multipler linearer Regression vorhergesagt (z.B. Scheuneman & Gerritz, 1990) oder korreliert werden (z.B. Klieme & Baumert, 2001). Weitere Möglichkeiten bestehen in der Modellierung von DIF anhand eines *Linearen Logistischen Test Modells* (LLTM; Fischer, 1973) oder auch mit Hilfe der oben bereits erwähnten *erklärenden Item Response Modelle.* In einem Methodenvergleich von Hartig, Frey, Nold & Klieme (2010) konnte gezeigt werden, dass die zweischrittige Methode und ein im Vergleich dazu aufwändigeres LLTM+e-Modell (Janssen, Schepers & Peres, 2004) zur Erklärung von Itemschwierigkeit als gleichwertig betrachtet werden können, wobei beide Modelle sich als geeigneter herausstellten als ein „normales" LLTM (siehe auch 4.2). Für eine tiefergehende Erörterung dieser Modelle wird auf Wilson & deBoeck (2004) verwiesen.

Besonders interessieren im Kontext dieser Arbeit empirische Studien mit dem Ziel, *DIF anhand von Itemmerkmalen zu erklären.* Hier werden insbesondere vier Arbeiten hervorgehoben, die auch die methodischen Grundlagen dieser Arbeit darstellen, und im Folgenden beschrieben werden.

Scheuneman und Gerritz (1990) beispielsweise untersuchten DIF bei Leseverständnis-Items jeweils zwischen Männern und Frauen bzw. Schülern mit unterschiedlichem ethnischen Hintergrund. Sie gingen von differentiellen Lerngelegenheiten der Gruppen als Ursachen für DIF aus (ohne diese jedoch genauer zu definieren) und versuchten, *unterschiedliche Stärken und Schwächen* der Gruppen anhand von Itemeigenschaften wie der Itemstruktur, Indikatoren der kognitiven Anforderungen und semantischen Strukturen/Inhalten darzustellen. Sie fanden unter Verwendung einer multiplen Regression geringe bis moderate signifikante Zusammenhänge zwischen dem Mantel-Haenzel-DIF und den Indikatoren, teilweise jedoch höhere mit Interaktionseffekten. Die Autoren kommen zu dem Schluss, dass sich dadurch die differentielle Performanz der beiden Gruppen beschreiben lasse. Ferner ließen sich auf diese Art die Itemeigenschaften, die diese unterschiedliche Performanz verursachen, untersuchen (Scheuneman & Gerritz, 1990).

Auch Klieme und Baumert (2001) fanden bei einem Vergleich von Schülern verschiedener Länder im Rahmen der internationalen TIMS-Studie, dass sich *differentielle Profile von Stärken und Schwächen* der Gruppen ergaben, wenn Itemcharakteristika mit DIF korreliert wurden. Dazu wurden Items zur Messung der mathematischen Kompetenz hinsichtlich der Existenz von DIF untersucht. Es wurde davon ausgegangen, dass Schüler in unterschiedlichen Ländern testkulturbedingt differentiellen Lerngelegenheiten ausgesetzt sind, weshalb DIF ein Ausdruck unterschiedlicher Muster von länderspezifischen, durch die nationalen Curricula und Tests bedingten Stärken und Schwächen der Schülergruppen ist. Untersucht wurde diese Annahme, indem Items von Experten hinsichtlich schwierigkeitsdeterminierender Anforderungsmerkmale eingeordnet wurden. Es wurden paarweise DIF zwischen fünf ausgewählten Ländern berechnet. Diese wurden dann mit den Itemeigenschaften korreliert, um relative Stärken und Schwächen der Länder aufzudecken und hinsichtlich ihrer Hypothesenkonformität zu überprüfen. Es fanden sich Zusammenhänge in prognostizierter Richtung. Ein externes Kriterium für Testkultur wurde jedoch nicht definiert, da die Items, die hinsichtlich testkultureller Variablen eingeordnet wurden, gleichzeitig auch für die DIF-Analysen verwendet wurden.

Gleichfalls im Rahmen von TIMSS untersuchten Klieme und Bos (2000) anhand von Videodaten Unterrichtsstile in Deutschland und Japan. Sie fanden, dass sich bei den Schülergruppen gefundene unterschiedliche Stärken und Schwächen beim Bearbeiten von Aufgaben nicht unbedingt auf die verschiedenen Unterrichtsstile und inhaltliche Themenschwerpunkte zurückführen ließen, sondern eher auf unterschiedliche kognitive Anforderungen im Unterricht in den beiden Kulturen, die sich durch die Analyse von im Unterricht verwendeten Aufgaben feststellen ließen. Diese wiesen einen Zusammenhang zu DIF in erwarteter Richtung auf.

Auch Artelt und Baumert (2004) entdeckten Hinweise darauf, dass unterschiedliche Testkulturen einen Einfluss auf Differentielle Item Funktionen von Items zur Messung von Leseverständnis haben. So fanden sie anhand von Daten der PISA-2000-Studie heraus, dass sich die mittleren DIF des Tests zugunsten einer Gruppe verändern, je mehr Items aus der eigenen Testkultur der Gesamttest beinhaltet. Es wurde überprüft, ob die Übersetzung von Aufgaben und sprach- und kulturbedingte Eigenschaften der Testitems einen Vorteil für die Schüler aus dem Sprachraum darstellen, aus dem die Items jeweils stammten. Die Items wurden dazu Rasch-skaliert und die Differenz der Logit-Werte als DIF-Parameter und als direktes Maß der Effektstärke d verwendet. Dann wurden Gruppen von Ländern gebildet, in denen die gleiche Sprache gesprochen wird (Deutsch, Französisch, Englisch, Spanisch). Bei aus dem französischen Sprachraum stammenden Items zeigte sich ein klarer Vorteil für die französischsprachige Personengruppe (d = .21 Logits). 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.

Um herauszufinden, ob sich auch die Leistung der Gruppen bei Items aus dem eigenen Sprachraum verbesserte bzw. die Nicht-Berücksichtigung eigener Items verschlechtert, wurden Tests aus Items gebildet, die aus dem eigenen Sprach- und Kulturraum stammten, bzw. nicht aus dem Sprachraum stammten und in den Gruppen neu skaliert. Besonders bei den französischen Schülern zeigte die Nicht-Berücksichtigung eine Verschlechterung der Leistung auf der PISA-Skala. Insgesamt wurden jedoch keine signifikanten Effekte festgestellt. Ein für die vorliegende Arbeit besonders relevantes Ergebnis ist, dass sich innerhalb der englischen Sprachgruppe, in der die Länder nochmals differenziert betrachtet wurden, differentielle Schwierigkeiten zeigten. Dies lässt darauf schließen, dass hier nicht nur die Ursprungssprache einen Einfluss auf die Itemschwierigkeit hat, sondern noch weitere bildungskulturelle Variablen.

Neben den oben besonders hervorgehobenen Studien existieren noch einige weitere Untersuchungen, die sich mit der Erklärung Differentieller Item Funktionen beschäftigten und hier nicht unerwähnt bleiben sollten. Li, Cohen und Ibarra (2004) entdeckten bei Mathematik-Items, dass Struktur-Charakteristika der Items im Zusammenhang mit zur Lösung notwendigen Strategien stehen. Diese Struktur-Charakteristika verwenden sie zur Vorhersage von Geschlechter-DIF in einem 3PL-Modell. Die verwendeten Itemcharakteristika waren zur Prädiktion von DIF geeignet und entsprechen den Hypothesen hinsichtlich unterschiedlicher Stärken und Schwächen der beiden Geschlechter.

Ferner fand Abbott (2004), bei der Analyse eines Leseverstehenstests mit Hilfe von Differential Bundle Functioning (DBF), einer Methode zur Überprüfung von Hypothesen zu den Ursachen von DIF heraus, dass sich bei unterschiedlichen kulturellen Gruppierungen -in diesem Fall Chinesen und Araber- unterschiedliche Lesestrategien feststellen lassen. Andere Studien vermuten Zusammenhänge zwischen DIF und unterschiedlichen sprachlichen Hintergründen (Uiterwijk & Vallen, 2005; Baron, Curley& Feigenbaum, 2000), mit Kontext-Effekten wie der Reihenfolge der vorgegebenen Items (Ryan & Chiu, 1997) oder mit dem Alter (Geranpayeh & Kunnan, 2007).
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).

Insgesamt sind die erwähnten empirischen Ergebnisse, und dabei besonders die Ergebnisse der Studien von Artelt und Baumert (2004), Scheuneman und Gerritz (1990) und Klieme und Baumert (2001), insgesamt hoch relevant für die vorliegende Dissertation, da sie die zentralen Grundannahmen dieser Arbeit unterstützen. Die Ergebnisse von Artelt und Baumert (2004) weisen darauf hin, dass zum einen die Verwendung von Items aus dem eigenen Land einen systematischen Vorteil für die jeweilige Gruppe darstellen kann, und dass dies nicht nur auf die Sprache zurückzuführen ist, sondern dass vermutlich noch weitere kulturelle Faktoren eine Rolle spielen. Scheuneman & Gerritz zeigen, dass DIF teilweise auf unterschiedliche Stärken und Schwächen von Gruppen rückführbar ist. Die Ergebnisse von Klieme & Baumert (2001) deuten darauf hin, dass auf Testitems unterschiedlicher Länder unterschiedliche Schwerpunkte von Testkulturen abgebildet werden, welche wiederum systematische Zusammenhänge zu DIF in Richtung der erwarteten Stärken und Schwächen der Gruppen aufweisen. Diese Methoden sollen nun in der

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 konstruktirrelevanten 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

Nachdem in den vorhergegangenen Abschnitten der Ausgangspunkt dieser Dissertation, nämlich das Vorhandensein und die Bedeutung von Differentiellen Item Funktionen, betrachtet wurden, wird im Folgenden der eigentliche Anwendungsbereich dargestellt. Gegenstand dieser Arbeit sind Items zur Messung fremdsprachlichen Leseverständnisses. Die Erforschung von Fremdsprachenkompetenzen fällt unter anderem in den Bereich der angewandten Linguistik. Die aus dieser Disziplin verwendeten theoretischen Grundlagen sowie relevante empirische Forschungsergebnisse werden in diesem Teil der Arbeit erläutert.

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

Definition von Sprachenkompetenz. Der Europarat definiert Sprachkompetenz wie folgt: „Sprachliche Kompetenzen sind die Summe des (deklarativen) Wissens, der (prozeduralen) Fertigkeiten und der persönlichkeitsbezogenen Kompetenzen und allgemeinen kognitiven Fähigkeiten, die es einem Menschen erlauben, Handlungen auszuführen. (…) Kommunikative Sprachkompetenzen befähigen Menschen zum Handeln mit Hilfe spezifischer sprachlicher Mittel" (Europarat, 2001, S. 21). Hier wird hervorgehoben, dass Sprache die Funktion besitzt, Personen dazu zu ermächtigen, bestimmte situative Anforderungen zu bewältigen.

Disziplinen der Fremdsprachenforschung. Hinsichtlich des Konstrukts und der Struktur der fremdsprachlichen Kompetenz wird seit Langem eine verschiedene Forschungsdisziplinen umfassende Debatte geführt. So befasst sich beispielsweise der Bereich der allgemeinen Linguistik mit den systemischen Eigenschaften von Sprache, die vergleichenden Sprachwissenschaften mit dem Vergleich von Einzelsprachen oder Sprachgruppen, die Fremdsprachendidaktik mit dem Lehren und Lernen fremder Sprachen, die Psycholinguistik mit den Mechanismen der kognitiven Sprachverarbeitung, und die psychologisch-pädagogische Diagnostik mit der Erfassung und Messung von Sprache. Die Frage nach dem zugrunde liegenden Konstrukt und der Struktur von Fremdsprachenkompetenzen wird in den verschiedenen Disziplinen diskutiert. Für eine tiefer gehende Debatte diesbezüglich wird hier auf Jude (2008) verwiesen.

Die europäische Fremdsprachenforschung ist Teil des Forschungsfeldes der *angewandten Linguistik* (engl.: Applied Linguistics). Dabei handelt es sich um ein interdisziplinäres Forschungsfeld, unter anderem bestehend aus den Disziplinen der Linguistik, Pädagogik, Psychologie, Didaktik, Anthropologie (bzw. Ethnologie im deutschsprachigen Raum) und Soziologie. Die Gesellschaft für Angewandte Linguistik (2009) definiert die Forschungsdisziplin wie folgt:
„Die Angewandte Linguistik ist eine der großen Arbeitsrichtungen innerhalb der Linguistik. Sie untersucht das sprachlich-kommunikative Handeln in allen Feldern der gesellschaftlichen Praxis, und zwar unter dem Aspekt der Anwendung ihrer Ergebnisse in der Praxis. Zu den klassischen Aufgaben der Angewandten Linguistik gehört die Beschäftigung mit Spracherwerb und Sprachtherapie, Fremdsprachenvermittlung und interkultureller Kommunikation (…)."
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.
Kognitions- und entwicklungspsychologische Theorien der Sprach- und der Lesekompetenz sowie Theorien aus dem Bereich der Psycholinguistik werden in dieser Arbeit nicht vertieft behandelt. Dies hat zwei Gründe: Zum einen basiert der Gemeinsame Europäische Referenzrahmen, der aufgrund seines zentralen Stellenwerts in der EBAFLS-Studie hier die zentrale theoretische Basis dieser Arbeit sein muss, auf einem eher pragmatischen Kompetenzmodell. Innerhalb der Tradition dieser Kompetenzmodelle werden zugrunde liegende kognitive Prozesse zwar vorausgesetzt und auch beschrieben, deren Untersuchung jedoch nicht fokussiert. Alderson (2000) formuliert diesen Standpunkt folgendermaßen:

> „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 kognitive 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.

Geschichte der Sprachmessung und -modellierung. Bereits seit den frühen Jahren des 20. Jahrhunderts haben sich Forscher mit Faktoren der Sprachkompetenz befasst, dabei häufig im Rahmen von Intelligenztheorien. So sind beispielsweise zwei von Thurstones (1938) sieben „primary mental abilities" von verbaler Natur, nämlich „verbal comprehension" und „verbal fluency". Häufig wurden diese Intelligenzmodelle, aber auch spezifischere Modelle zur Modellierung von Sprachkompetenz, mit Hilfe explorativer Faktorenanalysen untersucht, anhand derer Aussagen hinsichtlich des Konstrukts und der Dimensionalität von Sprachkompetenz gemacht wurden. An diesem Umgang mit der Modellierung von Sprache wurde jedoch verschiedentlich Kritik geübt, etwa in Bezug auf Anzahl und Replizierbarkeit der Dimensionen sowie die fragwürdige Eindeutigkeit der Lösungen (De Jong & Verhoeven, 1992). In den 1950er Jahren mit Erreichen der kognitiven Ansätze gab es zunächst eine Tendenz, anhand von diskreter Punktmessung, also des Testens von einzelnen Sprachelementen in den vier Grundfähigkeiten der Sprache, nämlich Lesen, Schreiben, Sprechen und Hören, die Sprachfähigkeit zu erfassen (McNamara, 1996). Damit wurden Profile der Getesteten hinsichtlich bestimmter Komponenten linguistischer Kenntnisse und Fähigkeiten, d.h. von diskreten Sprachelementen, erstellt.

Eine weitere Unterscheidung ergibt sich durch die oben bereits angesprochene Unterscheidung von diskreter Punktmessung und integrativer, eher globaler Messung. Ersteres bezieht sich auf das Erstellen eines Profils des Getesteten hinsichtlich bestimmter Komponenten linguistischer Kenntnisse und Fähigkeiten, d.h. von diskreten Sprachelementen. Bei Letzterer handelt es sich um eine globalere Messung und Interpretation von Sprachkompetenz. Hinsichtlich der diskreten Punktmessung entstand eine Gegenbewegung, die eine eher globale und integrative Messung und Interpretation von Sprachkompetenz beinhaltete.

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

Englisch als Fremdsprache. Innerhalb dieses Rahmens wurde der TOEFL-Test (Test of English as a Foreign Language; McNamara, 1996; Brown & Hudson, 2002) vom „Educational Testing Center" (ETS) der USA entwickelt. Dieser Test wird heutzutage noch immer häufig als Zugangstest verwendet und liegt mittlerweile als computerbasierte und online-Version vor. Somit erlangte das Performanz-basierte Testen eine immer größere Bedeutung.

In den 1970er Jahren entwickelte sich dann die *„kommunikative Bewegung"* (communicative movement; McNamara, 1996). Diese wurde eingeleitet durch Hymes' (1971) Theorie der kommunikativen Kompetenz, welche als Ursprung des heutzutage häufig verwendeten kommunikativen Ansatzes in der Lehre und dem Testen von Sprache gelten kann. Die Jahre 1980 bis 1990 waren von zwei Bewegungen gekennzeichnet (De Jong & Verhoeven, 1992). Die erste stammte aus dem Bereich des Assessments von Sprachkompetenzen und beschäftigte sich vor allem mit der Erstellung von *Skalen und Referenzsystemen.* Das Ziel der zweiten Forschungsrichtung war es, ein *Modell der kommunikativen Kompetenz* zu erschaffen, indem Hymes' (1971) Konzept der kommunikativen Kompetenz in Theorien mit eingearbeitet wurde (De Jong & Verhoeven, 1992). Daraus entstanden teilweise relativ ausführliche Modelle und Konzepte des Sprachverhaltens (z.B. Canale& Swain, 1980; Bachman, 1990), die weiter unten noch detaillierter dargestellt werden. Performanz-basiertes Assessment fand so Eingang in die Theorien der kommunikativen Kompetenz, und die kommunikative Kompetenz wurde damit auch für das Testen von Sprachen relevant.

Die Entwicklung von Modellen der Sprachkompetenz und die Kompetenz-Performanz-Debatte. Diese beiden Bereiche werden gemeinsam dargestellt, da die Entwicklung von Modellen der Sprachkompetenz von der Debatte um Kompetenz und Performanz beeinflusst wurde. Auch die Psycholinguistik hat bereits früh theoretische Beiträge zur Sprachmodellierung und zur Sprachverarbeitung geleistet. In der ersten Hälfte des 20. Jahrhunderts wurden eher behavioristische Theorien über den Erwerb und die Verarbeitung von Sprache entwickelt. Dort vertrat man die Position, dass Syntax ein Produkt aus Intelligenz und sozialen Faktoren sei, die die Sprachentwicklung im behavioristischen Sinne fördere (z.B. Skinner, 1957). In den 50er/60er Jahren des 20. Jahrhunderts begann sich dann hinsichtlich der Theorien des Spracherwerbs und der Sprachfähigkeiten eine Gegenbewegung zu den bisherigen, eher behavioristisch orientierten Modellen zu manifestieren. Zu einem der wohl prägendsten Werke der *kognitiv orientierten Sprachkompetenztheorien* gehört das Werk von Noam Chomsky (1965). Er betrachtet die Syntax und Semantik in der menschlichen Sprache als das Unterscheidungsmerkmal der menschlichen von der tierischen Kommunikation. Chomsky postuliert, dass die Fähigkeit zum Verstehen und

der Anwendung von Syntax in allen Menschen angeboren und somit universal gültig seien. Diese Abkehr von behavioristischen Spracherwerbsvorstellungen leitete die kognitive Wende in der Psychologie ein, und in den folgenden Jahrzehnten begann eine Debatte über die Rolle von Kompetenz und Performanz (Competence & Proficiency; North, 2000), die noch bis heute anhält.

Eines der ersten und auch wohl einflussreichsten Modelle in der Debatte über die Art des Zusammenhangs von Kompetenz und Performanz ist daher nach wie vor das oben erwähnte Modell von Noam Chomsky (1965). Er knüpft damit an die Theorien von Ferdinand de Saussure (1916) an, der Sprache dichotomisiert in *langue* (abstraktes Regelsystem) und *parole* (Sprechen) (Dresselhaus, 1979). Chomsky trennt klar zwischen beiden Konzepten: „The term 'competence' entered the technical literature in an effort to avoid the slew of problems relating to 'knowledge', but it is misleading in that it suggests 'ability'— an association I would like to sever." (Chomsky, 1980, s. 59; zitiert aus Brown, G., Malkjoer, K. & Williams, J. (Eds.), 1996). Chomsky macht hier eine klare Unterscheidung zwischen *Wissen* auf der einen Seite, und, andererseits, der *Fähigkeit* Sprache anzuwenden sowie diese Fähigkeit in letzter Konsequenz auch tatsächlich auszuüben. Unterschiedliche Leistungen einer Person kommen hier demnach nur durch unterschiedliches Handeln in unterschiedlichen Situationen zustande, nicht aber aufgrund von Unterschieden und Veränderlichkeit der zugrunde liegenden Kompetenz. Er betrachtet Syntax als die *invariable und universell gültige* Grundlage der menschlichen Sprache. Die Hauptfunktion der Syntax sieht er in ihrem Wirken als Vehikel der kognitiven Weiterentwicklung (Brown, 1996).

Demgegenüber entwickelten sich Konzepte, die Sprachfähigkeit und Leistung als eine innerhalb einer einzelnen Person beziehungsweise personenübergreifend (in Gruppen) verfügbare/vorhandene *variable Fähigkeit* betrachten. Hierbei handelt es sich um Kompetenzmodelle, die *Unterschiede in den zugrunde liegenden Kompetenzen* postulieren. Unterschiede in der Performanz seinen demnach auf tatsächliche Unterschiede in der zugrunde liegenden Kompetenz zurückzuführen. 1961 formulierte Carroll beispielsweise ein Modell der Fremdsprachenkompetenz, in dessen Rahmen er zehn relevante Aspekte der Sprachkompetenz auflistete. Diese lassen sich in drei Gruppen einteilen: erstens Aspekte in Bezug auf linguistische Kenntnisse (structure & lexicon), zweitens Aspekte des sogenannten „channel control" , der den vier Grundfähigkeiten der Sprache entspricht, (auditory discrimination of speech sounds, oral production of speech sounds, technical reading & technical writing) und drittens Aspekte der „performance", das heißt der tatsächlichen Ausführung von Sprache, die er als die jeweiligen Wechselwirkungen der ersten beiden Gruppen definiert (beispielsweise 'rate and accuracy of reading comprehension'). Diese Ansätze lassen sich auch in neueren Kompetenzmodellen wiederfinden.

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).

In den 1970er Jahren wurden noch weitere Faktoren in die Sprachtheorien aufgenommen. So integriert Hymes (1971) beispielsweise außerdem noch *„ability to use"*, die Fähigkeit der Sprachverwendung, in sein Konzept. Von der linguistischen Anthropologie kommend, untersucht er Kommunikation in unterschiedlichen Kulturen mit ethnografischen Methoden. Dies kennzeichnet den Beginn einer weiteren Bewegung hinsichtlich der Modellierung der Fremdsprachenkompetenzen, die *pragmatische Wende* (Schneider & North, 2000). In diesem Rahmen führte Hymes (1971) im Gegensatz zu den rein kognitiven Modellen Chomskys den Begriff der *kommunikativen Kompetenz* ein. Später definierte Bachman (1990) dann den Begriff der *kommunikativen Sprachfähigkeit* (Communicative Language Ability). Auch die *kommunikativen Tests* setzten sich – im Gegensatz zu den vorher vorherrschenden, oben beschriebenen *diskreten Test* und den *integrativen Tests* (allen Fertigkeiten liegt ein Generalfaktor zugrunde; z.B. Oller, 1976) – in den 80er Jahren durch (Schneider & North, 2000). Diesen Tests liegt zugrunde, dass die Sprachfähigkeiten Lesen, Sprechen, Hören und Schreiben nicht mehr getrennt erfasst werden sollten, sondern dass es insgesamt auf die natürliche Verwendung von Sprache und die *Bewältigung der kommunikativen Aufgabe* (Schneider & North, 2000) ankomme. Diese Aufgaben zeichnen sich

durch einen sehr realitätsnahen Charakter aus. Nach Schneider und North (2000) waren vor allem drei Modelle der kommunikativen Kompetenz für den Bereich der Fremdsprachenforschung und -didaktik besonders relevant, nämlich die Modelle von Canale und Swain (1980), Van Ek (1986) und Bachman /Bachman & Palmer (1990/1996). Modelle, die die kommunikative Sprachkompetenz mit einschließen, sind im Kontext dieser Arbeit besonders wichtig, da sie dem GERS zugrunde liegen. Die eben erwähnten Modelle von van Ek (1986), Bachman (1990/1996) sowie Canale und Swain (1980) werden daher in Tabelle 2.1 gegenübergestellt (North, 2000).

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

Tabelle 2.1. Modelle der Kommunikativen Sprachkompetenz, entnommen aus North (2000)

In Tabelle 2.1 werden die Gemeinsamkeiten dieser drei Modelle deutlich: Alle gehen von unterschiedlichen, zur Sprachkompetenz beitragenden Kompetenzbereichen aus. Dabei beinhalten alle Modelle einen Kompetenzbereich, der sich speziell auf die linguistischen Fähigkeiten bezieht, nämlich „Grammatical Competence", „Linguistic Competence" und „Language knowledge", die wiederum jeweils Komponenten der Grammatik wie Struktur und Syntax beinhalten; teilweise beziehen sie sich auch auf das Vokabular. Ferner beziehen alle drei eine soziolinguistische Kompetenz in ihr Modell mit ein, sowie einen Kompetenzbereich, der sich auf den jeweiligen Diskurs bzw. textliche Komponenten bezieht. Darüber hinaus sprechen alle eine strategische bzw. kompensatorische Kompetenz an, was jeweils der oben beschriebenen kommunikativen Kompetenz entspricht. Van Ek und Bachman integrieren zusätzlich jeweils noch einen

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.

Versuche, die Struktur und Komponenten dieser Modelle in Tests zu operationalisieren, haben bislang nur sehr eingeschränkt funktioniert (North, 2000). Das Modell von Canale und Swain (1980) beispielsweise konnte anhand konfirmatorischer Faktorenanalysen empirisch nicht validiert werden.
Andere Autoren nahmen zusätzliche sozio-kulturelle Faktoren mit in ihre Modelle auf, wie etwa Etikette, Situationssprache und Diskurs (z.B. Loveday, 1982). Widdowson beispielsweise betrachtet die beiden Konzepte der linguistischen Kompetenz (systemisch) und der soziokulturellen Kompetenz (schematisch) als sich gegenseitig ergänzend (Widdowson, 1983, zitiert nach North, 2000).

Die Kompetenz-Performanz-Debatte schlägt sich auch in der Entwicklung von Theorien und Modellen des Spracherwerbs nieder. Es werden dabei häufig dichotome Modelle der Sprachfähigkeit formuliert. Diese folgen meist der Unterteilung Chomskys (1980) zwischen *grammatikalischer und pragmatischer Kompetenz* sowie Carrolls (1961) Unterscheidung zwischen *Wissen und Fähigkeiten* (knowledge and skills; z.B. Bialystok, 1981; Ellis, 1990). 1982 entwickelt Anderson ein Rahmenmodell des Spracherwerbs, in dem von *zwei kognitiven Entwicklungsstufen* ausgegangen wird: Einer *deklarativen,* in der das Fakten- und Regelwissen codiert ist, und einer *prozeduralen,* in der es sich es um die Performanz, also die Umsetzung der Fähigkeit, handelt. Rumelhart & McClelland (1986) hingegen schlagen vor, dass es sich beim Spracherwerb nicht direkt um das Erlernen von Regeln, sondern um die Stärke der neuronalen Verknüpfungen handelt, die es ermöglichen, Parallelprozesse ablaufen zu lassen. „Leaving aside the differences between the learning models proposed by Anderson (1982, 1983) and by Rumelhart et al. (1986), both theories agree in that they offer a process model that can serve to explain language acquisition by implying a distinction between knowledge possession and its use, from the learning and the performance point of view." (De Jong & Verhoeven, 1992, S. 7).

Ein weiteres zweistufiges Modell schlägt beispielsweise Bialystok (1981) vor. Sie unterscheidet zwei Dimensionen: erstens die des Wissens, und zweitens die der Kontrolle. Die Wissensdimension unterscheidet analysierte von nicht-analysierten mentalen Repräsentationen. Dies bezieht sich auf die Unterscheidung von bewusster Verbalisierung linguistischer Regeln und

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).

North (2000) beschreibt die derzeitige Situation hinsichtlich der Kompetenz-Performanz-Debatte wie folgt: „The current position in relation to the development of a model of communicative language use, or in relation to the acceptance of a standard interpretation of competence, proficiency and performance, remains confused. The complexity of the number of factors and variables involved in Communicative Language Use make an approach based either on a competence or a performance view difficult." (S. 53). Dies zeigt, dass die Debatte um Kompetenz und Performanz bis dato anhält. Der vorliegenden Arbeit liegt der Gemeinsame Europäische Referenzrahmen zugrunde, der Sprachkompetenz kriterienorientiert auf unterschiedlichen aufeinanderfolgenden Niveaus beschreibt. Insofern wird dort, wie im nächsten Abschnitt deutlich wird, im Prinzip Sprachkompetenz als Entwicklungsprozess und somit als variabel behandelt. Die vorliegende Dissertation folgt dahingehend dieser Annahme. Für eine ausführliche Behandlung der Kompetenz-Performanz-Debatte wird auf Brown, Malkjoer & Williams (1996) verwiesen.

Sprachkompetenz als Entwicklungsprozess: niveau- und kriterienorientiertes Testen. Sprachkompetenz kann in ihrem Entwicklungsprozess betrachtet werden. Aufbauend auf Andersons Prozessmodell (1982) des Spracherwerbs von deklarativem und prozeduralem Wissen schlagen Glaser, Lesgold und Lajoie (1987) ein prozessorientiertes Messmodell vor.

> „(…) 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

Kompetenz zugrunde, die eine wichtige Grundlage für *niveau- und kriterienorientiertes* Testen ist. Obgleich kriterienorientiertes Testen bereits in den 1960er Jahren erwähnt wurde (z.B. Glaser & Klaus, 1962), wurde diesem Testkonzept erst in den 1980er Jahren mehr Beachtung zuteil (Brown & Hudson, 2002). Besonders für die unten beschriebenen Skalen und Referenzsysteme der Sprachkompetenz ist dieses Konzept relevant.

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 Fremdsprachen*kenntnisse* 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

Im Folgenden werden zunächst die im Bereich der Fremdsprachenmessung existierenden Large Scale Studien kurz beschrieben. Dabei handelt es sich vor allem um die in den 1970er Jahren durchgeführte 10-Länder-Studie der IEA (International Association for the Evaluation of Educational Achievement, Lewis & Massad, 1975), um die in Deutschland und Südtirol durchgeführte DESI-Studie (Deutsch Englisch Schülerleistungen International, Beck & Klieme, 2007; DESI-Konsortium, 2008) sowie die EBAFLS-Studie (European Bank of Anchor Items for Foreign Language Skills, z.B. Fandel et al., 2007), welche die Grundlage der Dissertation darstellt. Darauf folgend wird auf unterschiedliche Ansätze hinsichtlich des Testens von Fremdsprachen und insbesondere auf Fremdsprachenskalen und Referenzsysteme eingegangen.

Internationale Studien zur Messung und zum Vergleich von Fremdsprachenkenntnissen. Bei der ersten der hier zu beschreibenden Studien handelt es sich um „The Teaching of English as a Foreign Language in Ten Countries" (Lewis & Massad, 1975). Diese Untersuchung war die erste große, internationale Studie im Bereich des Fremdsprachentestens und ist Teil des „Six Subject Survey" der IEA (Lewis & Massad, 1975) in den 1960er und 1970er Jahren. Es wurden Fremdsprachenkompetenzen in Englisch und Französisch als Fremdsprache in zehn Ländern miteinander verglichen. Dies kann als die erste internationale Large Scale Untersuchung im Bereich von fremdsprachlichen Kompetenzen angesehen werden.

Die zweite Studie, „Deutsch Englisch Schülerleistungen International" (DESI; Beck & Klieme, 2007; DESI-Konsortium, 2008), untersucht die sprachlichen Leistungen und die Unterrichtswirklichkeit in den Fächern Deutsch und Englisch. Die Studie wurde von der deutschen Kultusministerkonferenz in Auftrag gegeben. Es wurden Testverfahren zur Messung von Sprachkompetenz in Deutsch und Englisch entwickelt. Zu Beginn des Schuljahres 2003/04 wurden repräsentativ ca. 11.000 Schülerinnen und Schüler der neunten Jahrgangsstufe getestet. Zusätzlich wurden Videostudien sowie Befragungen von Lehrern und Eltern durchgeführt. Die Studie ermöglichte differenzierte Aussagen unter anderem über den Erwerb sprachlicher Kompetenzen (Klieme et al., 2006). Für die vorliegende Arbeit ist besonders diese Studie relevant, da die aus Deutschland stammenden Items, die in den EBAFLS-Itempool eingefügt wurden, aus DESI stammen.

Die dritte hier dargestellte Studie liefert den Rahmen und die Daten für die vorliegenden Arbeit. Dabei handelt es sich um die Studie „European Bank of Anchor Items for Foreign Language Skills" (EBAFLS; Gille & Sluiter, 2005; Fandel et al., 2007). Als Grundlage der vorliegenden Arbeit werden die Daten der EBAFLS-Studie herangezogen, welche mit Unterstützung seitens der Europäischen Kommission und acht europäischer Partnerländer seit November 2004 durchgeführt wurde. Diese Studie beschäftigte sich mit der Machbarkeit eines Kultur-fairen, länderübergreifenden Vergleichs von Fremdsprachenkompetenzen. Es wurden die Sprachen Englisch, Deutsch und Französisch als Fremdsprachen an insgesamt ca. 10.500 Schülern zwischen der 9. und 11. Jahrgangsstufe erhoben. Bei den in der Studie verwendeten Items handelte es sich um erprobte und validierte Items aus den Teilnehmerländern. Ziel war es zu überprüfen, ob die Items geeignet sind, Fremdsprachenkompetenzen von Schülern in unterschiedlichen europäischen Ländern fair miteinander zu vergleichen und somit als Ankeritems zur Verlinkung nationaler Fremdsprachentests zu dienen. Eine Überprüfung der Verwendbarkeit der Items erfolgte anhand von DIF-Analysen. Im Rahmen der EBAFLS-Studie zeigte sich, dass nur wenige Items keine

DIF aufzeigten. Der Datensatz der Studie dient als Grundlage dieser Dissertation. Die EBAFLS Studie wird daher bezüglich der Stichprobe, des Designs, der Instrumente sowie der Ergebnisse im Methodenteil unter 4.1 ausführlicher dargestellt.

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 Kommunikations*kanals* (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 Kommunikations*modus*, der in produktive (Sprechen, Schreiben) vs. rezeptive (Hören, Lesen) Fähigkeiten unterteilt wird. (siehe Tabelle 2.2).

	Kommunikationsmodus	
Kommunikationskanal	*rezeptiv*	*produktiv*
visuell	Lesen	Schreiben
auditiv	Hören	Sprechen

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.

De Jong und Verhoefen (1992) definieren die durch indirekte, diskrete Messinstrumente erfasste Kompetenz als die Summe der Ergebnisse von Subtests, die sich wiederum auf bestimmte grammatische oder linguistische Fähigkeiten beziehen. Diese Tests werden einesteils als objektiv und einfach umsetzbar betrachtet (Morrow, 1981), andererseits wurden die Technisierung und die Unnatürlichkeit dieser Testverfahren auch kritisiert (Spolsky, 1985). Integrative, globalere Sprachtests hingegen, so de Jong & Verhoeven (1992), sind durch die fehlende Differenzierbarkeit von sprach- und fähigkeitsbestimmenden Subskills in ihrer Validität eingeschränkt.
Ein weiterer Messansatz ist die Entwicklung von *Bezugssystemen oder Referenzrahmen*, welche Skalen zur Erfassung von Sprachkompetenz beinhalten. Diese verfolgen typischerweise einen verhaltens- und kriterienorientierten Ansatz. Dieser Messansatz ist für die vorliegende Arbeit der relevanteste. Eines der möglicherweise einflussreichsten Bezugssysteme, vor allem bezüglich der europäischen Fremdsprachenforschung und -politik, ist der *Gemeinsame Europäische Referenzrahmen für Sprachen* (Europarat, 2001). Im Folgenden werden nun zunächst die für diese Arbeit wichtigen verhaltens- und kriterienorienterten Rating-Skalen allgemein beschrieben.

Bezugssysteme und Referenzrahmen zur Beschreibung von Sprachkompetenz.
Es exisieren 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 der Sprach- und Fremdsprachenfähigkeit entstammen teilweise unterschiedlichen Traditionen, die North (2000) als „Rating Scales", „Examination Levels" und „Stages of Attainment" bezeichnet (S. 13ff). Im Rahmen dieser Dissertation interessieren vor allem die *verhaltensorientierten Rating Scales*. Eine wichtige Gemeinsamkeit dieser Rating-Skalen ist, dass sie einem *Outcome-orientierten* Ansatz folgen. Das bedeutet, Unterschiede bei beobachteter Leistung werden auf Unterschiede einer *variablen, zugrunde liegenden Kompetenz* zurückgeführt. So wird davon ausgegangen, dass durch das Testergebnis auch ein Rückschluss auf die zugrunde liegende

Kompetenz gezogen werden kann. Die erste wichtige Rating-Skala der Sprachkompetenz war die in den 50er Jahren in den USA entwickelte Rating-Skala des US Foreign Institute. Diese war direkte Vorgängerin weiterer Skalen, wie der Australian Second Language Proficiency Ratings - Skala (ASLPR-Skala), der ILR (Interagency Language Roundtable) und der unten detaillierter beschriebenen ACTFL-Skala (ACTFL, 1983).

Vor- und Nachteile verhaltensorientierter Rating-Skalen. Verhaltens- und kriterienorientierte Skalen können für unterschiedliche Zwecke verwendet werden. Die Getesteten haben die Möglichkeit, die eigene Leistung bzw. das eigene Sprachverhalten anhand der unterschiedlichen Niveaus einer Skala zu beschreiben sowie sich selbst eigene, klare Lernziele zu setzen. Andererseits bieten Skalen verschiedentlich auch einen Rahmen für die Entwicklung von (vergleichbaren) Testverfahren, die Grundlage für Curricula bilden, Lernfortschritte sichtbar machen und einen gemeinsamen Standard für unterschiedliche Organisationen, Systeme und Gruppen gestalten (North, 2000, S. 12).

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.

Ein weiterer Kritikpunkt ist die *normorientierte Bewertung.* Das bedeutet hier, dass die Leistung von Fremdsprachenlernern oft an den *Leistungen von Muttersprachlern* gemessen wird. Muttersprachler stellen in diesem Fall die Norm dar, und die Skalen beinhalten dementsprechende Deskriptoren. Die meisten Fremdsprachenlerner werden jedoch mutmaßlich nie dazu in der Lage sein, eine Sprache tatsächlich genauso gut wie ein Muttersprachler zu sprechen. Darüber hinaus unterscheiden sich Muttersprachler auch untereinander hinsichtlich ihrer Sprachkompetenz , so dass eine allgemeingültige Norm „Muttersprachler" kaum definiert werden kann. Daher stellt sich in solchen Fällen die Frage nach der Angemessenheit der verwendeten Norm.

Ferner wird eine *Interaktion von Fremdsprachenkompetenz und dem Grad der akademischen Ausbildung* bei der Beschreibung von Skalenniveaus kritisiert. Alderson (2000) merkt dazu an, dass Niveaus für Fremd- oder Zweitsprachenkompetenz für belesene, akademisch erfolgreiche Erwachsene sich beispielsweise von Niveaus für Lernende, die noch nicht einmal in der Erstsprache belesen oder akademisch nicht erfolgreich sind, unterscheiden.

Im Folgenden wird beispielhaft zunächst die in den USA entwickelte ACTFL-Skala zur Beschreibung von Fremdsprachenfähigkeit (American Council for Teaching of Foreign Languages: ACTFL, 1983) beschrieben. Als zweite Skala wird der Gemeinsame Europäische Referenzrahmen für Sprachen, der die theoretische Grundlage dieser Arbeit darstellt, hinsichtlich seiner Entwicklung und Validierung besprochen.

1983 verfasste der ACTFL Richtlinien zur Beschreibung von Fremdsprachenkompetenz (ACTFL, 1983). Diese beinhalten Beschreibungen für die vier Sprachfähigkeiten Lesen, Schreiben, Hören und Sprechen auf vier Hauptstufen (Novice, Intermediate, Advanced, Superior). Zusätzlich werden diese Hauptstufen in Zwischenstufen (high, medium, low) unterteils, so dass unterschiedliche Niveaus existieren, auf denen der jeweils erreichte Entwicklungsstand eines Lernenden auf diesen Niveaus beschrieben wird. Jedes Niveau beinhaltet auch immer das niedrigere, einfachere Niveau darunter:

> „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 Entwicklungsniveau 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; Europarat, 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.

Beim GERS handelt es sich um ein vom Europarat in Auftrag gegebenes Produkt. Dieses hatte formal seinen Ursprung im Jahre 1991, als der Europarat zu folgender Erkenntnis gelangte: „The mutual recognition of qualifications, and communication concerning objectives and achievement standards would be greatly facilitated if they were calibrated according to agreed common reference standards, purely descriptive in nature." (Trim, 2001, S. 5, zitiert aus Morrow, 2004, S. 6). Hieraus resultierte die Entwicklung des GERS mit der Zielvorstellung, ein Hilfsmittel zur *Entwicklung des Fremdsprachenunterrichts* in Europa bereitzustellen sowie die Ziele und *Leistungsstandards* der Länder und der Lernenden in unterschiedlichen nationalen Kontexten *untereinander zu vergleichen* (Morrow, 2004). Der GERS soll als grenzübergreifender *Leitfaden zum Lernen, Lehren und Beurteilen von Fremdsprachenkompetenzen* dienen. Dabei handelt es sich um einen Referenzrahmen, mit dessen Hilfe beschrieben werden kann, über welche Kompetenzen ein Sprachenlerner auf den verschiedenen Niveaus jeweils mindestens verfügt. Somit handelt es sich auch bei diesem Referenzrahmen um eine *verhaltensbasierte, kriterienorientierte Skala für Sprachkompetenzen*, die verschiedene *Niveaus* anhand von Deskriptoren, sogenannten „Can-Do-Statements", beschreibt.

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 Muttersprachler 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:

> „Kann die Hauptpunkte verstehen, wenn klare Standardsprache verwendet wird und wenn es um vertraute Dinge aus Arbeit, Schule, Freizeit usw. geht. Kann die meisten Situationen bewältigen, denen man auf Reisen im Sprachgebiet begegnet. Kann sich einfach und zusammenhängend über vertraute Themen und persönliche Interessengebiete äußern. Kann über Erfahrungen und Ereignisse berichten, Träume, Hoffnungen und Ziele beschreiben und zu Plänen und Ansichten kurze Begründungen oder Erklärungen geben." (GERS, S. 35)

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:

A		**B**		**C**	
Elementare		**Selbständige**		**Kompetente**	
Sprachverwendung		**Sprachverwendung**		**Sprachverwendung**	
/	****	**/**	****	**/**	****
A 1	**A 2**	**B 1**	**B 2**	**C 1**	**C 2**
(Break-through)	*(Waystage)*	*(Threshold)*	*(Vantage)*	*(Effective Operational Proficiency)*	*(Mastery)*

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:

	Verstehen		**Sprechen**		**Schreiben**
C2 bis A1	Hören	Lesen	An Gesprächen teilnehmen	Zusammenhängendes Sprechen	Schreiben

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, Sprachenkönnen detaillierter zu beschreiben. Für die Deskriptorskalen von „Hörverständnis" und „Leseverständnis" werden im Folgenden einige Beispiele gegeben:

Für das Niveau B1 im Leseverständnis lautet die Beschreibung des notwendigen Könnens auf der Selbsteinschätzungsskala wie folgt: „Ich kann Texte verstehen, in denen vor allem sehr gebräuchliche Alltags- oder Berufssprache vorkommt. Ich kann private Briefe verstehen, in denen von Ereignissen, Gefühlen und Wünschen berichtet wird" (GERS, S. 36).

Ein Beispiel für einen Deskriptor des Hörverständnisses auf Niveau B1 wäre etwa: „Kann im Allgemeinen den Hauptpunkten von längeren Gesprächen folgen, die in seiner/ihrer Gegenwart geführt werden, sofern deutlich artikuliert und in der Standardsprache gesprochen wird" (Europarat, 2001, S. 72). Ein Beispiel für A1 — das leichteste Niveau: „Kann Anweisungen, die langsam und deutlich an ihn/sie gerichtet werden, verstehen und kurzen, einfachen Wegerklärungen folgen" (Europarat, 2001, S. 73). Diese *kriterienorientierte Beschreibung der Niveaus* stellt den Hauptvorteil des GERS dar, da so von vornherein festgelegt wird, was eine Person mindestens können muss um einem bestimmten Niveau zugeordnet zu werden. Kriterienorientiertes Testen kann zu objektiveren Testergebnissen führen, da diese sich dann nicht mehr, wie häufig im Fremdsprachenunterricht üblich, nach dem Niveau der anderen Schüler in einer Klasse oder eben dem gebildeten Muttersprachler richten (normorientiertes Testen), sondern es ermöglichen, auf allgemeine und übergreifend gültige Bewertungskriterien zurückzugreifen. Dies kann zu einer besseren Transparenz und Vergleichbarkeit der Testergebnisse auch über verschiedene Länder hinweg führen.

Einige Länder verwenden die verschiedenen Niveaus des GERS auch bereits als Basis für das schulische Curriculum für Fremdsprachen sowie als Grundlage für Testverfahren und Assessments. So werden im finnischen, italienischen, ungarischen, und französischen Schulwesen (Morrow, 2004) die Levels und Skalen des GERS als curriculare Basis eingesetzt, und es werden GERS-Niveaus als zu erreichendes Ziel zu bestimmten Zeitpunkten einer Ausbildung oder für bestimmte Schulabschlüsse definiert.

Theoretische Grundlagen: Kompetenz und Performanz im GERS. Den GERS-Skalen liegt eine in den 90er Jahren in der Schweiz durchgeführte Studie zugrunde. Das erklärte Ziel des dem GERS zugrunde liegenden schweizer Projekts (Schneider & North, 2000, siehe auch unter „Validierung") sowie bei der Entwicklung des GERS war es, Aussagen über die Leistungen und die Performanz eines Fremdsprachenlerners im Rahmen einer Theorie der kommunikativen Sprachkompetenz machen zu können (North, 2000). In diesem Zusammenhang wurde und wird nach wie vor insbesondere das Verhältnis von Kompetenz und Performanz, das sich als außerordentlich komplex darstellt, diskutiert. So sieht North (2000) einen Zusammenhang der Kompetenz-Performanz-Frage mit der Unterscheidung von theoretischen und operationalen Modellen. Skehan (1995, zitiert aus North, 2000) schlägt vor, anstatt des Begriffs „Kompetenz" den Begriff „Ability to use" (Fähigkeit zur Sprachverwendung) zu verwenden:

> „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.).

Das Vorgehen im GERS hinsichtlich Kompetenz und Performanz beschreibt North (2000). Demnach ist der dort gewählte Ansatz der, dass Kategorien der Kompetenz von Kategorien für kommunikative Aktivitäten separiert werden. Als Verbindung zwischen beiden dient die sogenannte Strategieverwendung (Strategy Use). Die Kompetenz-Kategorien werden dabei eher zur Beschreibung der zugrundeliegenden Kompetenz verwendet, wohingegen die Performanz-Kategorien verwendet werden, um Leistung, die mit der zugrundeliegenden Fähigkeit zusammenhängt, zu definieren. North beschreibt die Schwierigkeiten, ein Kompetenzmodell mit Leistungsdeskriptoren zu verknüpfen, wie folgt:

> „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).

Wie außerdem aus diesem Zitat hervorgeht, ist die theoretische Grundlage des GERS eher von Pragmatik gekennzeichnet. Dies ist aber nicht zwingend ein Anzeichen für einen laxen Umgang im Bezug auf die theoretische Basis des Instruments, sondern im Gegenteil Ausdruck des Versuchs, aus unterschiedlichen Modellen der kommunikativen Kompetenz ein neues zusammenzusetzen, das dem hohen Anspruch einer Verknüpfbarkeit von Theorie und Praxis, Kompetenzmodellierung und Kompetenzmessung gerecht werden soll. Zur Entwicklung und Validierung sowie den theoretischen Grundlagen des GERS äußern sich Schneider & North (2000) dahingehend, dass die Entwicklung der Skalen und von Sprachtests im Allgemeinen sich auf eine Theorie des sprachlichen Handelns stützen sollten, bisher jedoch noch kein Konsens bezüglich eines allgemeingültigen Modells, das der Komplexität des Themas gerecht werden könne, vorhanden sei. Diesbezüglich verweisen die Autoren auf folgende Aussage des Europäischen Rates: „The description also needs to be based on theories of language competence, although the available

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). ”

Gemeinsam ist den im GERS *als Basis verwendeten Modellen*, dass es sich um Modelle der *kommunikativen Kompetenz* handelt. Die Basis stellen die drei oben bereits dargestellten Modelle von Canale & Swain (1980), Van Ek (1986) und Bachman/Bachman und Palmer(1990/1996) dar (siehe auch 2.2.1).
Für den GERS wurden aus diesen Modellen der kommunikativen Kompetenz die Gemeinsamkeiten herausgearbeitet. Basierend auf diesen Modellen wurden von Experten die den GERS-Skalen zugrunde liegenden Deskriptoren entwickelt und validiert. Die dazugehörige Studie wird im Folgenden beschrieben.

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).
Die Studie begann, basierend auf oben beschriebenen Modellen, mit der Entwicklung kriterienorientierter Skalen für verschiedene Niveaus der fremdsprachlichen Kompetenz. Die Deskriptoren der Skalen selbst wurden zunächst von Experten formuliert, dann in 32 Workshops (siehe unten) mit bzw. von Lehrern validiert und die Ratings letztendlich IRT-skaliert. Die Skalen wurden mit Hilfe von verhaltensbasierten Deskriptoren dessen, was eine Person mindestens können muss, um sich auf einem bestimmten Niveau des GERS zu befinden, definiert („behaviorally anchored rating scales”; BARS). Die Verwendung dieser verhaltensbasierten Deskriptoren sollte dabei helfen, das oben beschriebene Problem von Theorie und Praxis in der Kompetenzmessung und -modellierung zu lösen. Auch das Geben und Verstehen von Feedback sollte durch ihre Anwendung transparenter und einfacher werden (North, 2000).

In einem ersten Schritt entwickelten Experten einen Pool von Deskriptoren, die man zunächst in provisorische Schwierigkeitskategorien und Levels einordnete. Diese Deskriptoren basierten auf den oben dargestellten Modellen von Bachman (1990), Bachman & Palmer (1996), Swain (1983) und van Ek (1986). Insgesamt wurden 1000 solcher Deskriptoren in einem ersten Durchlauf von Lehrern in speziellen Workshops getestet, wobei sie verschiedenen Niveaus zugeordnet wurden. Die Deskriptoren, die dort genügend Übereinstimmung aufweisen konnten, wurden in einem zweiten Schritt im Rahmen eines Matrix-Designs in sieben „überlappenden" Fragebögen eingesetzt („vertical equating"). Mit deren Hilfe wurden von Fremdsprachenlehrern sowohl die Leistungen der Schüler der jeweils eigenen Lerngruppen, als auch auf Video aufgezeichnete Dialoge zwischen schülern, die den beurteilenden Lehrern fremd waren, den verschiedenen definierten Niveaus zugeordnet. Damit wurde erstens die Übereinstimmung der Lehrer hinsichtlich der Zuordnung der Deskriptoren zu Niveaus und zweitens die Verwendbarkeit der Deskriptoren zur Einstufung von Schülern überprüft. Dies fand innerhalb der vier verschiedenen Sprachregionen der Schweiz gleichermaßen statt. Alle geeigneten Deskriptoren wurden in einem letzten Schritt mit einem einparametrischen IRT-Modell Rasch-skaliert; aus diesem Grund gilt für die Deskriptor-Skalen des GERS die Annahme der Eindimensionalität. Zur Bildung der Skalen für die sechs GERS-Niveaus wurden jeweils Deskriptoren verwendet, die ein homogenes Schwierigkeitsniveau aufwiesen. Für eine detailliertere Beschreibung der Studie wird auf Schneider und North (2000) verwiesen.

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 Bidlungskontexte 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.

Kritikpunkte. Der GERS hat seit seiner Veröffentlichung im Jahre 2001 als Grundlage für einige Fallstudien gedient, er wurde in einigen Schulsystemen als curriculare Grundlage verwendet (z.B. Polen; Komorowska, 2002), und es wurden sogar Versuche gemacht, bereits bestehenden Skalen (z.B. ALTE: Association of European Language Testers) sowie kommerzielle Testsysteme (z.B. Test-DAF, TOEFL) mit den GERS-Skalen zu verknüpfen (Jones, 2002; Wertenschlag, Müller & Schmitz, 2002). Gleichwohl wird das Instrument immer wieder auch kritisiert. So wurde er von Bausch et al. (2003) beispielsweise hinsichtlich seiner Verwendung im Rahmen der Bildungsstandards, hinsichtlich der durch den Referenzrahmen entstehenden Wünsche nach einem kulturübergreifenden Vergleich von Unterricht und Leistung, hinsichtlich des Vorgehens bei der Konstruktion und Skalierung der Skalen sowie hinsichtlich des verwendeten Spracherwerbs- und fremdsprachenlerntheoretischen Ansatzes kritisiert. Andere Autoren kritisieren vor allem die geringe Verwendbarkeit des Instruments für die Konstruktion von darauf basierenden Testverfahren, bzw. die Einordnung von bereits konstruierten Testaufgaben in die Niveaus des GERS (Alderson et al., 2006). Letztgenannter Punkt ist für diese Arbeit bedeutend, da auf dieser Kritik die Entwicklung des Dutch Grid-Kategorisierungsinstruments, mit dessen Hilfe die Items in dieser Studie hinsichtlich ihrer inhaltlichen und kognitiv-linguistischen Charakteristika eingeordnet wurden, basiert. Dies soll im nächsten Teil weiter ausgeführt werden.

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).

Testaufgaben und insbesondere deren Anforderungsmerkmale sind ein Bindeglied zwischen der zu messenden Fremdsprachenkompetenz und der Testleistung. Das von Bachman und Palmer (1996) entwickelte Modell hinsichtlich der Determinanten von Testleistung in Fremdsprachentests wird in Abbildung 2.3 dargestellt. Es stellt den Einfluss von persönlicher Situation und

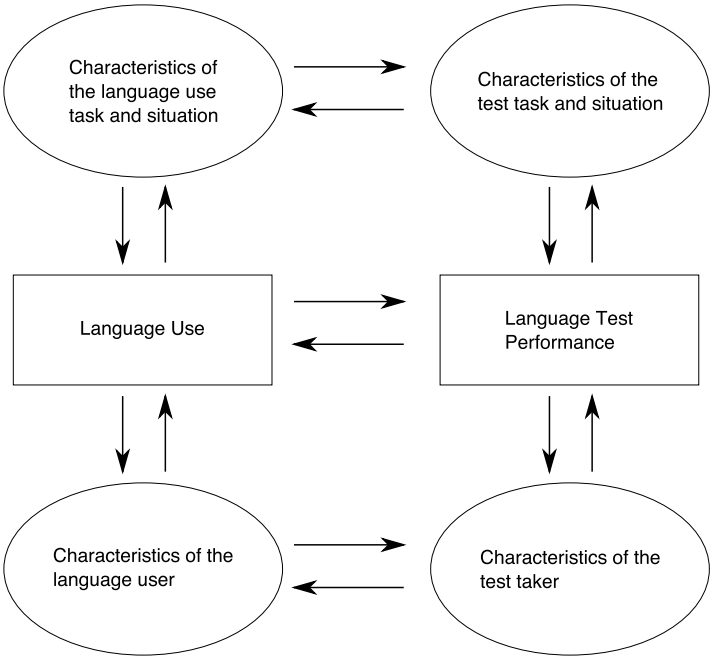

Abbildung 2.3. Korrespondenz zwischen Sprachgebrauch und Sprachtestleistung (entnommen aus Bachman & Palmer (1996), S. 12.)

individuellen Merkmalen des Getesteten auf die Sprachleistung dar. Die linke Seite der Abbildung bezieht sich dabei auf das „reale" Leben, die rechte auf die Testsituation. Dabei wird deutlich, dass neben der Analyse des Verhaltens des Lesers bzw. der getesteten Person also noch ein weiterer Analysebereich, nämlich der des Textes bzw. der zu bearbeitenden Aufgabe, existiert. Dies ist insbesondere für den Bereich des Assessments von Bedeutung, sei es nun formativ oder summativ: Beim Assessment bzw. der Diagnose von Leseverständnis existiert immer eine Wechselwirkung zwischen der Schwierigkeit der zu bearbeitenden Aufgabe und der Fähigkeit des Lesers. So ist beispielsweise die von Bachman (1990) bzw. Bachman & Palmer (1996) aufgestellte Taxonomie für Sprachtestaufgaben primär auf die Analyse von Items und Aufgaben, auch in Interaktion mit der Situation, in der sie gelöst werden, ausgerichtet.

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

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).

Im Rahmen dieser Arbeit soll zur Kategorisierung von Sprachtestaufgaben das Instrumtent „Dutch Grid" (Alderson et al., 2006) verwendet werden. Dieses basiert zum einen auf den Sprachkompetenzskalen des GERS, zum anderen aber auch auf dem Modell von Bachman & Palmer (1996) bzw. der Taxonomie für Sprachtestaufgaben von Bachman (1990). Aus diesem Grund soll auf diese Taxonomie im Folgenden kurz eingegangen werden —bereits inklusive einiger Hinweise auf die Berührungspunkte mit dem „Dutch Grid"—bevor das in dieser Arbeit verwendete Kategorisierungssystem „Dutch Grid" dargestellt wird.

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).

Im Rahmen dieses Modells werden ähnliche Charakteristika wie im „Dutch Grid" als relevant für die Schwierigkeit von Fremdsprachenitems angesehen. Dies ist vor allem der Fall bezüglich des Vokabulars, der Textlänge, der Sprachcharakteristika, des Antworttyps und der Beziehung zwischen Input und Output. Dabei betonen Bachman und Palmer (1996) ausdrücklich, dass sie hier keinen Anspruch auf die Vollständigkeit dieses Modells hinsichtlich die Testleistung beeinflussender Itemcharakteristika erheben.

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-

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

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

man und Palmer (1996) verwiesen.

Das Item-Kategorisierungssystem „Dutch Grid". Das Item-Kategorisierungsinstrument „Dutch Grid" (Alderson et al., 2006) geht insoweit über die Taxonomie von Bachman und Palmer (1996) hinaus, als dass es auf dem Gemeinsamen europäischen Referenzrahmen für Sprachen basiert. Da dieser, wie oben bereits ausgeführt, für die theoretische Grundlage und Ausrichtung dieser Arbeit einen zentralen Stellenwert einnimmt, wurde der „Dutch Grid" als Instrument zur Kategorisierung von Items hinsichtlich ihrer schwierigkeitsdeterminierenden Merkmale gewählt. Anhand der Taxonomie für Sprachtestaufgaben von Bachmann (1990) kann dargestellt werden, dass die im „Dutch Grid" verwendeten Kategorien auch in weiteren relevanten Sprachtestsystemen zu finden sind und damit die Verwendung des „Dutch Grid" als Kategoriensystem im Rahmen dieser Arbeit zusätzlich legitimieren. Ferner erlauben die von Bachman und Palmer getroffenen Aussagen eine grundlegende Annahme dieser Arbeit, nämlich dass die Leistung und auch differentielle Leistungsprofile mit Anforderungsmerkmalen von Items zusammenhängen.

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-

ger jedoch die *Art von Aufgaben*, die jeweils eines der GERS- Kompetenzniveaus repräsentieren und erfassen können. Hinsichtlich der tatsächlichen Verwendbarkeit des GERS zur Konstruktion und Einordnung von Testaufgaben zur Messung von Fremdsprachkompetenzen kristallisierten sich daher verschiedene Schwierigkeiten heraus. Dazu gehören beispielsweise eine zu abstrakte Beschreibung der einzelnen Niveaus (Figueras, North, Takala, Verhelst & Van Avermaet, 2005), Inkonsistenzen zwischen den Niveaus hinsichtlich ihrer Beschreibung sowie Inkonsistenzen hinsichtlich der Synonymie verwendeter Begriffe. Ferner gibt es hinsichtlich der Beschreibungen und Begriffsdefinitionen Auslassungen auf einigen der GERS-Niveaus (Alderson et al., 2006; Noijons & Kuijper, 2006). Weir (2005) kommentiert die Eignung des GERS für den Bereich des Testens von Sprache wie folgt: „Though also containing much valuable information on language proficiency and advice for practitioners, in its present form the CEFR is not sufficiently comprehensive, coherent or transparent for uncritical use on language testing" (S. 281). Einen Verbesserungsansatz stellt die Entwicklung des sogenannten „Dutch Grid" dar (Dutch CEFR Construct Project; Alderson et al., 2006). Hierbei handelt es sich um ein Instrument zur Kategorisierung von Items hinsichtlich verschiedener inhaltlicher und kognitiv-linguistischer Charakteristika. Dessen Entwicklung geht aus einem ursprünglich von der holländischen Regierung beauftragten Projekt hervor. Ziel des Projekts war, mit Hilfe von Fremdsprachenexperten aus unterschiedlichen Ländern den GERS daraufhin zu überprüfen, ob die in den Skalen vorhandenen Informationen zur Entwicklung von Testitems für die unterschiedlichen Niveaus ausreichen. Die Autoren kommen zu dem Schluss, dass eine ausschließlich auf den vorhandenen Informationen basierende Test- und Itementwicklung so nicht möglich ist (Alderson et al., 2006).

Das Projekt „Dutch Grid" stellt daher den Versuch dar, die für die Testentwicklung relevanten Dimensionen des GERS auf eine systematischere Art und Weise zu beschreiben (Noijons & Kuijper, 2006). Es handelt sich um ein Instrument, welches es ermöglichen soll, Testaufgaben hinsichtlich bestimmter Aufgaben-, Text- und Itemmerkmale einzuordnen. Damit soll es Testautoren erleichtert werden, diese den Niveaus des GERS zuzuordnen, beziehungsweise Items für ein bestimmtes Niveau zu konstruieren.

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 Itemcharakteristika 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.

	Itemcharakteristik	**Beschreibung**
Inhalt des Textes	Textquelle	(Magazin, Zeitung, etc.)
	Diskurs	Narrativ / beschreibend / Instruktion, etc.
	Domäne	Persönlich, beruflich, öffentlich, etc.
	Thema	Tägliches Leben, Reisen, Gesundheit, etc.
Linguistische und kognitive Komplexität des Textes	Textlänge	Anzahl der Wörter
	Abstraktheit des Inhalts	Konkret vs. abstrakt (ausschließlich konkret; hauptsächlich konkret; teilweise abstrakt; abstrakt)
	Authentizität des Texts	(angepasst / authentisch)
	Schwierigkeit des Vokabulars	Einfach / häufig vs. schwer / selten (ausschließlich häufig / einfach; hauptsächlich häufig / einfach; teilweise erweitert / selten; erweitert / selten)
	Komplexität grammatischer Strukturen	Einfache Strukturen vs. komplexe Strukturen (ausschließlich einfach; hauptsächlich einfach; teilweise komplex; komplex
	Itemtyp	Auswahl vs. Konstruktion einer Antwort (z.B. Multiple Choice / offene Antwort)
	Informationsgewinn 1	Information durch Erkennen / Schlussfolgern / Bewerten
	Informationsgewinn 2	Information explizit / implizit
	Informationsgewinn 3	Hauptidee vs. spezifisches Detail

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

(Noijons & Kuijper, 2006) zeigte sich beispielsweise, dass Texte mit höherem Abstraktionsgrad auch höheren GERS Niveaus zugeordnet wurden. Auch die Schwierigkeit des Vokabulars sowie die grammatikalische Komplexität erhöhten sich von den niedrigeren hin zu den höheren GERS-Niveaus. Obgleich dies ein Hinweis darauf ist, dass solche Aufgaben- und Itemeigenschaften einen Beitrag zur Schwierigkeit eines Items leisten, gibt es jedoch bisher noch keine allgemeingültigen Aussagen dazu, welchem GERS-Niveau welche Eigenschaften in welcher Kombination zuzuordnen sind.

Acht der im Dutch Grid verwendeten kognitiv-linguistischen Itemcharakteristika werden in dieser Arbeit als Prädiktoren der Itemschwierigkeiten verwendet. Im Bezug auf die Textmerkmale werden im Rahmen dieser Arbeit vier Merkmale für Leseverständnis-Aufgaben aus dem Dutch Grid übernommen. Dabei handelt es sich erstens um die Abstraktheit des Inhalts („Nature of Content": Behandelt der Text ein inhaltlich abstraktes oder konkretes Thema? Bezieht sich auf das erforderte Ausmaß abstrakten Denkens), zweitens um die Authentizität des Textes im Sinne genuiner, nicht veränderter versus pädagogischer Texte (authentisch/angepasst/vereinfacht), drittens um die Schwierigkeit und Häufigkeit des Vokabulars sowie, viertens, die Komplexität grammatischer Strukturen. Die Länge des Textes wird im Rahmen dieser Arbeit nicht berücksichtigt. Obgleich die Textlänge auch bereits als Prädiktor der Itemschwierigkeit eingesetzt wurde, gibt es Hinweise darauf, dass gefundene Zusammenhänge nicht auf die Textlänge als schwierigkeitsdeterminierende Itemcharakteristik zurückzuführen sind, sondern dies eine Moderatorvariable für grammatische Strukturen darstellt (z.B. Grotjahn, 2000).

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 mentalen 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?).

Zugrundeliegende kognitive Prozesse. Auf die kognitiven Prozesse, von denen angenommen wird, dass sie den Itemcharakteristika zugrunde liegen, wird leider im Rahmen des Dutch Grid außer den relativ kurzen, oben dargestellten Beschreibungen nicht weiter eingegangen. Daher sollen im Folgenden die diesbezüglichen Annahmen von Grotjahn (2000) dargestellt werden, der sich im Rahmen der Konstruktion von Items zum Test DAF (Projektgruppe TestDaF, 2000: Test Deutsch als Fremdsprache) mit den Zusammenhängen zwischen schwierigkeitsdeterminierenden Itemcharakteristika und zugrundeliegenden kognitiven Prozessen, teilweise im Rahmen von Leseprozess-Modellen, auseinandergesetzt hat.

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 hypothesentestenden Lesen, während die datenbasierten Prozesse eher auf der unteren Ebene anzusiedeln sind und zum großen Teil parallel und automatisiert ablaufen.
Grotjahn (2000) geht von einer Interaktion bzw. Parallelität beider Prozessarten aus: „Individuelles Verstehen resultiert danach letztlich aus der komplexen Wechselwirkung von Verarbeitungsfaktoren auf allen Ebenen und führt zur Konstruktion einer kohärenten, mehrstufigen mentalen Textrepräsentation" (S. 10). Er argumentiert weiter, dass sich bei der Konstruktion von Leseverständnis-Aufgaben dann auch die Frage nach der jeweiligen mentalen Repräsentationsebene stellt. Diesbezüglich gehen beispielsweise Van Dijk und Kintsch (1983, zitiert aus Grotjahn, 2000, S. 12) davon aus, dass bei dem Verstehen von Texten drei Typen mentaler Repräsentationen gebildet werden können, nämlich die der Textoberfläche im Rahmen der semantisch-syntaktischen Verarbeitung, eine propositionale Textbasis im Rahmen von semantischen und textbasierten Kohärenzprozessen, und ein mentales Modell über eine Integration von Textinformationen und individuellem Vorwissen.
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):

- Es ist der Interaktion zwischen Leser und Aufgabe Rechnung zu tragen.
- Es ist die Abhängigkeit der Lese- und Verstehensprozesse von interindividuell variierenden, L2-unspezifischen Rezipientenmerkmale 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).

Als wichtig erachtet Grotjahn (2000) außerdem die Frage nach der Authentizität einer Aufgabe. Dies ist hier zum einen bezogen auf „authentisch" im Sinne von aus einer genuinen Quelle stammend und im Gegensatz zu „pädagogisch", d.h. zum Lernen konstruiert oder abgeändert, zu betrachten. Zum anderen hat „Authentizität" jedoch auch die Bedeutung im Sinne vom „Grad der Übereinstimmung zwischen den Merkmalen einer gegebenen Testaufgabe und den Merkmalen der jeweiligen Zielsprache" (S. 13). Dabei bezieht sich Letzteres auf die Frage nach der Konstruktvalidität. Neben der Authentizität der Aufgabe betrachtet der Autor die Vertrautheit von Probanden mit bestimmten Aufgaben- und Itemformen als möglicherweise noch wichtiger, da eine unterschiedliche Vertrautheit zu differentiellen Aufgabenschwierigkeiten führen könne. Als weitere wichtige Variablen sieht er zum einen das Hintergrundwissen und zum anderen die Bedeutung der L1-Lesefertigkeiten beim Lesen der L2; darauf soll im Rahmen dieser Arbeit jedoch nicht weiter eingegangen werden.

Mit den kognitiven Prozessen, die einer Itemeigenschaft zugrunde liegen, beschäftigten sich auch Buck, Tatsuoka & Kostin (1997). Die Autoren bezeichnen die zugrundeliegenden Prozesse als Fähigkeit. Im Bezug auf Multiple-Choice Items konnten sie insgesamt 24 Itemcharakteristika entdecken, die zur Schwierigkeit der Items beitrugen, wobei einige der Charakteristika den im Dutch Grid und dieser Arbeit verwendeten Itemmerkmalen entsprechen. So beschreiben sie beispielsweise die der Itemcharakteristik „Schlussfolgerungen" zugrundeliegende Fähigkeit als die Fähigkeit, Informationen im Gedächtnis zu behalten und diese für die Schlussfolgerung zu verwenden. Wenn es sich hingegen um ein Item handelt, zu dessen korrekter Lösung es notwendig ist, die Kernaussagen eines Textes zu verstehen („main idea"), dann beschreiben sie den diesem Merkmal zugrundeliegenden Prozess als die Fähigkeit, die Kernaussage einer Textpassage identifizieren zu können. Auch Fortus, Coriat & Fund (1998) analysierten Multiple-Choice-Items für das Leseverständnis in englischer Sprache. Hinsichtlich der in dieser Arbeit verwendeten

Itemcharakteristika fanden sie Korrelationen zwischen Itemschwierigkeit und Schwierigkeit des Vokabulars (.86), grammatischer Komplexität (.86), und Abstraktheit (.23). Grotjahn (2000) zieht den Schluss, dass sich eine Taxonomie von Item- und Textmerkmalen aufstellen lässt, deren Effekte einerseits auf der Basis eines kognitiven Aufgabenverarbeitungsmodells interpretierbar sind, und die sich darüberhinaus empirisch als schwierigkeitsdeterminierend erwiesen haben. Von den in dieser Liste enthaltenen Variablen werden diejenigen, die auch den in dieser Arbeit verwendeten Dutch-Grid-Charakteristika entsprechen, in Tabelle 2.7 dargestellt.

Die beiden weiteren, in dieser Arbeit verwendeten Itemcharakteristika, „Authentizität" und „Item Type", sind nicht in dieser Liste enthalten. Dies ist mutmaßlich darauf zurückzuführen, dass im Kontext des Aufsatzes von Grotjahn „Authentizität" im Sinne von valide verwendet wird. Gleichwohl führt der Autor zu Beginn seines Artikels aus, dass auch die andere Betrachtungsweise von „Authentizität" möglich ist, nämlich im Sinne von genuin vs. pädagogisch/adaptiert/vereinfacht. Die Autorin der vorliegenden Arbeit ist der Ansicht, dass diese Unterscheidung nicht notwendig ist, da die Verwendung einer authentischen Quelle im Sinne von genuin, wie beispielsweise einem Zeitungsartikel, auch üblicherweise einem Text entspricht, der authentisch im Sinne von konstruktvalide ist: Bei solchen Texten handelt es sich immer um Texte, die auch von Muttersprachlern in der Zielsprache genauso verwendet werden. Die Konstruktvalidität sollte daher hoch sein. Es ist anzunehmen, dass es sich bei den der Bearbeitung von authentischen Texten zugrundeliegenden kognitiven Prozessen nicht um besondere, eigenständige zugrundeliegende Prozesse für genuine Texte handelt, sondern möglicherweise um eine bestimmte Kombination von zugrundeliegenden Prozessen, die bestimmten Textsorten in bestimmten Zielsprachen eigen ist (Grotjahn, 2000). Es ist jedoch davon auszugehen, dass genuine Texte üblicherweise schwieriger sind als vereinfachte oder adaptierte Texte, da die Vereinfachung von Texten für den Unterricht einzig und allein mit dem Ziel geschieht, die Schwierigkeit der Aufgabe gegenüber einem (häufig genuinen) Ursprungstext zu verringern. Daher wird dieses Merkmal als schwierigkeitsdeterminierendes Merkmal in dieser Arbeit mit einbezogen.

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

Grotjahn (2000)	Entspricht Dutch Grid Itemcharakteristik
Zahl der schwierigen / unvertrauten Wörter	Schwierigkeit des Vokabulars
Abstraktheit des Inhalts	Abstraktheit des Inhalts
Grammatische Komplexität	Grammatische Komplexität
Die für eine Antwort benötigte Information ist über den Text verteilt	Erkennen
Es ist kein unmittelbares Scannen der relevanten Information möglich	Schlussfolgern
Grad der Implizitheit	Implizit / explizit
Das Item erfragt eine Hauptinformation („main idea item")	Hauptidee / detail

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

die Unterschiedlichkeit der in den verschiedenen Teilnehmerländern häufig verwendeten Itemtypen so augenscheinlich, dass überprüft werden muss, ob die Vertrautheit mit dem Itemtyp nicht einen Einfluss auf die Unterschiedlichkeit der Itemschwierigkeiten zwischen Ländern hat. Wenn außerdem, wie bei Rauch und Hartig (in Vorbereitung) berichtet, mit unterschiedlichen Itemtypen unterschiedliche Teile des Konstrukts „Leseverstehen" gemessen werden und unterschiedliche Länder unterschiedliche Itemtypen bei der Konstruktion ihrer Testaufgaben verwenden, dann könnte dies durchaus mit verantwortlich für das Entstehen Differentieller Item Funktionen sein. Daher wird in dieser Arbeit auch die Variable „Itemtyp" als schwierigkeitsdeterminierende Variable einbezogen.

Determinanten der Itemschwierigkeit: empirische Ergebnisse. Im letzten Teil dieses Theorieabschnitts sollen nun noch einige weitere, die bisherigen theoretischen Überlegungen unterstützende relevante empirische Ergebnisse hinsichtlich der Modellierung von Itemschwierigkeiten bei Aufgaben zur Messung von Fremdsprachenfähigkeiten dargestellt werden. Dabei wird ein besonderes Augenmerk auf die schwierigkeitsdeterminierenden Itemcharakteristika gelegt. Zur Modellierung von Itemschwierigkeit bei fremdsprachlichen Leseverständnis-Items findet sich in der Literatur eine Reihe von relevanten Studien. Diese zeigen auf, dass Zusammenhänge zwischen kognitiv-linguistischen Itemcharakteristika und der Itemschwierigkeit existieren.

Perkins & Linville (1987) untersuchten beispielsweise in einer Studie den Zusammenhang von Itemcharakteristika und Itemschwierigkeit bei einem Englisch-Vokabeltest mit Hilfe von multipler Regression. Insgesamt konnten hier zwischen 30% und 80% der Varianz der Itemschwierigkeit aufgeklärt werden. Als beste Prädiktoren zeigten sich die Worthäufigkeit (in einer Sprache oft vorkommende Worte erleichterten die Lösung), Wortlänge und Abstraktheit.

Freedle & Kostin (1993) konnten mit Hilfe von sowohl subjektiven (z.B. Abstraktheit) als auch objektiven (z.B. Satzlänge, Anzahl der Wörter) Item-Charakteristika zwischen 30% und 52% der Itemschwierigkeitsvarianz bei fremdsprachlichen Leseverständnis-Items aufklären. Bachman, Davidson & Milanovic (1996) untersuchten den Zusammenhang zwischen der Itemschwierigkeit bei fremdsprachlichen Leseverständnis-Items und den aus Bachman's Framework (1990) entnommenen Itemcharakteristika. Dort wurden beispielsweise Charakteristika wie Häufigkeit des Vokabulars, Abstraktheit, Itemtyp, Grammatik, Syntax, Strategie und Organisation als Prädiktoren der Itemschwierigkeit in einem 2-PL IRT-Modell verwendet. Auch hier zeigt sich mit 44%-68% eine Varianzaufklärung in ähnlicher Größenordnung wie bei den vorherigen Studien. Alderman & Holland (1981) fanden ferner, dass Experten in der Lage waren, Item-Performanz durch linguistische Charakteristika der Items vorherzusagen.

Die letzte hier dargestellte Studie stammt von Hartig, Frey, Nold & Klieme (2010) und wird hier ausführlicher beschrieben, da sie gleichzeitig einen Methodenvergleich beinhaltet und somit eine Unterstützung für die in dieser Arbeit verwendeten Methoden zur Vorhersage von Itemschwierigkeiten darstellt. Die Autoren verwendeten im Rahmen von Auswertungen zur DESI-Studie (DESI-Konsortium, 2008) unterschiedliche Methoden zur Vorhersage von Itemschwierigkeiten mit Hilfe von apriori definierten schwierigkeitsdeterminierenden Itemeigenschaften. Dazu wurden drei Methoden hinsichtlich ihrer Eignung verglichen: ein Linear Logistisches Testmodell (LLTM; Fischer, 1973), ein LLTM +e, welches zufällige Item Effekte (random item effects) mit berücksichtigt (Janssen, Schepers & Peres, 2004), sowie drittens um eine sogenannte zweischrittige Prozedur. In Letzterer wurden in Schritt eins die Items zunächst IRT-skaliert. Danach wurden in einem zweiten Schritt in einer gängigen Analysesoftware wie SPSS die in Schritt eins berechneten Parameter zu Fällen bzw. abhängigen Variablen, und wurden nun per multipler linearer Regression mit Hilfe der Itemeigenschaften als Prädiktoren vorhergesagt.

Als Prädiktoren verwendeten die Autoren Item- und Aufgabencharakteristika, die bestimmte kognitive Operationen abbilden, wie beispielsweise die Globalität der zur Lösung des Items benötigten Information (global/lokal), oder auch die Schwierigkeit bzw. Häufigkeit des Vokabulars und das Item-Antwortformat. Auch hier wurden bei der Einordnung der Items hinsichtlich ihrer Charakteristika teilweise der GERS und der Dutch Grid zugrunde gelegt, genauer gesagt die Kategorien global/lokal (entspricht „Hauptidee/detail") sowie Kategorien der Textkomplexität (grammatische Komplexität, Schwierigkeit des Vokabulars).

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).

Zusammenfassung und Relevanz von Theorien und Modellen der Fremdsprachenforschung und angewandten Linguistik für die vorliegende Arbeit. Im Rahmen dieses Theorieabschnitts wurde zunächst auf themenrelevante Modelle und Theorien der Fremdsprachenforschung und ˇmessung eingegangen. Da diese Arbeit sich mit dem Thema Fremdsprachenkenntnisse befasst, sind diese als theoretische Basis relevant. Dabei war die Entwicklung von Modellen teilweise eng verzahnt mit der dargestellten Kompetenz-Performanz-Debatte. Auch die Darstellung von Methoden zur Messung und zum Testen von Fremdsprachenkompetenz sowie die Darstellung verhaltensorienterter Rating-Skalen spielen insoweit eine wichtige Rolle, als der „Gemeinsame europäische Referenzrahmen für Sprachen", der die theoretische Grundlage für diese Arbeit darstellt, mittels solcher Skalen Fremdsprachenkompetenz beschreibt.
Insbesondere die detaillierte Beschreibung der Validierung der Skalen ist relevant um zu belegen, dass die vorliegende Arbeit auf soliden theoretischen und empirisch überprüften Annahmen beruht. Ferner wurde in diesem Teil auf itemseitige Schwierigkeitsdeterminanten von Items zur Messung von fremdsprachlichem Leseverständnis eingegangen. Es wurde dargestellt, dass kognitiv-linguistische Itemcharakteristika einen Einfluss auf die Schwierigkeit eines Items besitzen. Hier wurden unterschiedliche Kategorisierungsmethoden präsentiert, sowie die Gemeinsamkeiten hervorgehoben. Das Kategorisierungssystem „Dutch Grid" wurde explizit vorgestellt und hervorgehoben, da die Kategorien dieses Systems im Rahmen der vorliegenden Arbeit für die Kategorisierung von Items Verwendung finden. Ferner wurde auf die den schwierigkeitsdeterminierenden Itemmerkmalen zugrundeliegenden kognitiven Prozesse eingegangen, die für die inhaltliche Interpretation der Ergebnisse dieser Arbeit von Bedeutung sind. Darüber hinaus wurden empirische Ergebnisse hinsichtlich der Erklärung von Itemschwierigkeiten mit Hilfe von Itemmerkmalen vorgestellt.
Insgesamt weisen die vorliegenden empirischen Ergebnisse darauf hin, dass kognitiv-linguistische Anforderungsmerkmale von Items, wie sie im „Dutch Grid" verwendet werden, als Prädik-

toren für die Itemschwierigkeit geeignet zu sein scheinen. Dies ist hoch relevant, umd rechtfertigt infolgedessen die Verwendung der im „Dutch Grid" definierten Itemmerkmale bei der Beantwortung der Hauptfragestellung dieser Dissertation (siehe auch 3.2) beruht.

2.3. Interkulturelle Vergleichbarkeit von Testergebnissen

Neben dem Bereich der Differentiellen Item Funktionen und dem Bereich des fremdsprachlichen Leseverständnisses wendet sich die vorliegende Arbeit in einem dritten Schwerpunkt der Erklärung und Deutung *kultureller Unterschiede* zu. Somit stellt die *interkulturelle Vergleichbarkeit von Testergebnissen* den dritten Theorieschwerpunkt dieser Dissertation dar. Die Befassung mit der interkulturellen Vergleichbarkeit von Testergebnissen ist der Disziplin der interkulturellen Psychologie zuzuordnen. Neben theoretischen Ansätzen aus dieser Disziplin greift die vorliegenden Arbeit vor allem auf die dort verwendeten Methoden zurück. Die nachfolgenden Abschnitten werden auf Theorien hinsichtlich der interkulturellen Vergleichbarkeit von Konstrukten sowie auf die dazugehörigen Methoden eingehen.

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.

In dieser Debatte nimmt diese Dissertation eine Zwischenposition ein. Vor dem Hintergrund der vorliegenden Arbeit wird angenommen, dass die Existenz von DIF ein Hinweis auf das Vorhandensein von schwierigkeitsbestimmenden Variablen ist, die in unterschiedlichen Kulturen eine unterschiedliche Wichtigkeit besitzen. Beispielsweise kann ein Item-Anforderungsmerkmal, obgleich es nicht konstruktirrelevant ist, in einer Kultur völlig unwichtig sein für das dort gemessene Konstrukt, da es bei der Konstruktion von Items keine Rolle spielt. Die gleiche Variable kann in einer anderen Kultur jedoch ein wichtiger Teil des gemessenen Konstrukts sein. Dies hat zur Folge, dass eine Gruppe bei der Bearbeitung ein und derselben Aufgabe eine höhere Itemschwierigkeit erfährt als eine andere Gruppe. Diese Beobachtung lässt die Annahme gerechtfertigt erscheinen, dass es in solchen Fällen zumindest marginale Unterschiede hinsichtlich der Struktur des Konstrukts gibt. Andererseits kann auch davon ausgegangen werden, wie es beispielsweise bei kognitiven Theorien des Spracherwerbs der Fall ist, dass Sprachentwicklung zumindest teilweise als universell betrachtet werden kann (Trautner, 1997). Daher kann auch die Annahme gerechtfertigt werden, dass die Struktur der untersuchten Konstrukte in den unterschiedlichen Ländern zumindest teilweise die gleiche ist, und Unterschiede durch Niveauunterschiede zustande kommen. Unterschiede können vermutlich also sowohl auf Struktur- als auch Niveau-

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.

Taxonomie interkultureller Studien. Es existieren unterschiedliche Arten empirisch-interkultureller Studien, die Van de Vijver und Leung (1997) in einer Taxonomie gegenüberstellen. Im Folgenden wird diese Taxonomie kurz dargestellt mit dem Ziel, die EBAFLS-Studie sowie diese Dissertation innerhalb der Forschungsansätze der kulturellen Psychologie einzuordnen, beiden innerhalb dieser Disziplin einen Rahmen zu geben und um klarzustellen, in welcher Form sich die vorliegende Arbeit von der zugrundeliegenden EBAFLS-Studie unterscheidet.

Nach Van de Vijver und Leung (1997) existieren zwei Dimensionen interkultureller Studien. Die erste Dimension betrifft die Orientiertung 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.

Die zweite Dimension der Taxonomie bezieht sich auf die kulturellen Kontextfaktoren für die Erklärung von beobachteten kulturellen Ähnlichkeiten oder Unterschieden. Die verwendeten Kontextvariablen können von demografischer Natur sein, beispielsweise das Alter oder das Bildungsniveau. Ferner können auch eher psychologische Variablen verwendet werden, wie etwa Werte, Persönlichkeitseigenschaften oder Einstellungen (Van de Vijver & Leung, 1997). Die EBAFLS Studie bezieht keinerlei Kontextvariablen außer der Länderzugehörigkeit heran. Infolgedessen finden sich in der Studie auch keine Erklärungen oder Deutungen kultureller Unterschiede. Sie ist also eher eine „Psychological Differences Study" (Van de Vijver & Leung, 1997), die sich weitgehend mit der reinen Feststellung interkultureller Unterschiede begnügt und sich nicht um Erklärungen oder Deutungen bemüht.

Auch hinsichtlich dieser Dimension findet bei der vorliegenden Dissertation eine Verschiebung zum anderen Pol statt: Es wird die Annahme getroffen, dass die unterschiedlichen Testkulturen und somit differentielle Lerngelegenheiten in den verschiedenen Kulturen zumindest zum Teil für die gefundenen Unterschiede verantwortlich sind. Es werden Kontextvariablen zur Erklärung von Unterschieden herangezogen, allerdings basieren diese Informationen auf den Testitems und nicht, wie klassischerweise üblich, auf den Merkmalen von Schülern der Länder. Diese Dissertation ist tendenziell eher den theoriebasierten Studien zuzuordnen.

Stuktur-Orientierung vs. Level-Orientierung bei interkulturellen Studien. Neben oben beschriebenen Ausrichtungen von interkulturellen Studien beschreiben Van de Vijver und Leung (1997) zusätzlich die Ausprägungen „strukturorientiert"- oder „level-orientiert". Diese sind mit dem oben beschriebenen emischen obzw. dem etischen Ansatz der interkulturellen Psychologie zu vergleichen. In struktur-orientientierten Studien wird davon ausgegangen, dass sich die verschiedenen Kulturen hinsichtlich der grundlegenden Struktur des Konstrukts unterscheiden. Im Prinzip sind diese kulturellen Gruppen dann nicht miteinander vergleichbar. Bei der Level- oder Niveau-Orientierung wird hingegen davon ausgegangen, dass die Struktur der Konstrukte insgesamt gleich ist und beobachtete Unterschiede zwischen den Gruppen durch Unterschiede hinsichtlich des Niveaus auf den relevanten Dimensionen eines Konstrukts bedingt sind.

Es wird davon ausgegangen, dass beobachtete Unterschiede sowohl durch Niveau- als auch durch Strukturunterschiede verursacht werden können: Zum einen wird angenommen, dass in allen Ländern im Prinzip die gleichen kognitiv-linguistischen Itemanforderungen die Schwierigkeit der Items bestimmten, nur eben mehr oder weniger stark bzw. in unterschiedlicher Richtung. Dies entspricht eher der Annahme, dass Unterschiede niveaubedingt sind. Jedoch kann das Vorhandensein von Differentiellen Item Funktionen auch auf Strukturunterschiede hinsichtlich des zu messenden Kontrukts zurückzuführen sein. Dies könnte dann der Fall sein, wenn eine Variable wie beispielsweise ein bestimmter Itemtyp in einer Kultur überhaupt nicht zum Testen oder Unterrichten verwendet wird. Diese Dissertation nimmt daher diesbezüglich, wie oben bereits dargestellt, eine Zwischenposition ein und ist somit keiner der beiden Ausprägungen konkret zuzuordnen. Bei welchen der beobachteten Unterschiede es sich um Strukturunterschiede und bei welchen es sich um Niveauunterschiede handelt, ist im Rahmen dieser Arbeit aufgrund der Datenstruktur (Multi-Matrix-Design) nicht überprüfbar, da eine Anwendung von Strukturgleichungsmodellen oder Mehrgruppenanalysen nicht möglich ist.

2.3.1. Validität in interkulturellen Studien: Konzept der Äquivalenz und Methoden zur Überprüfung von Invarianz

Die Überprüfung von Items hinsichtlich ihrer Fairness für unterschiedliche Gruppen hat in der interkulturellen Psychologie eine lange Tradition. Die Begriffe Bias, Fairness, Impact und DIF wurden bereits unter 2.1 im Rahmen der Darstellung von Differentiellen Item Funktionen beschrieben. Diese Konzepte stellen im Kontext der interkulturellen Vergleichbarkeit von Testverfahren einen wichtigen Teil des Validitätskonzepts und von Validitätsüberprüfungen dar. In diesem Rahmen ist zusätzlich auf die Konzepte der Äquivalenz und der Verzerrung (Bias) sowie deren Zusammenhang hinzuweisen:

> „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).

Die beiden Konzepte „Äquivalenz" und „Bias" sind demnach im Prinzip zwei Seiten einer Medaille. DIF-Analysen sind eine mögliche Methode, die An- bzw. Abwesenheit von Äquivalenz zu überprüfen.

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: Konstruktbias, 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.

Explorative und konfirmatorische Faktorenanalysen zur Überprüfung von Äquivalenz und Messinvarianz. Weitere klassische Methode zur Überprüfung von Konstruktäquivalenz sind Faktorenanalysen. Diese sollen hier nur der Vollständigkeit halber kurz vorgestellt werden, da sie, wie oben bereits dargelegt, aufgrund der Datenstruktur im Rahmen der vorliegenden Arbeit nicht angewandt werden können. Ursprünglich setzte man in der interkulturellen Psychologie häufig explorative Faktorenanalysen ein. Diese wurden jedoch aufgrund ihrer bekannten Unzulänglichkeiten — zu nennen sind hier Probleme mit der Reproduzierbarkeit und Uneindeutigkeit der Lösung — kritisiert und durch konfirmatorische Faktorenanalysen und Mehrgruppenmodelle ersetzt bzw. ergänzt.

Bei dieser Methode wird schon vorher definiert, welche Items der Theorie nach auf welchem Faktor laden sollten. Dies ermöglicht hypothesengeleitetes Testen. Innerhalb des Modells kann nun im Rahmen von Mehrgruppenmodellen außerdem überprüft werden, ob die Gruppen sich hinsichtlich der Passung der faktoriellen Struktur und so auch hinsichtlich des zugrunde gelegten Konstrukts unterscheiden oder ob die Struktur in allen Gruppen hinreichend gleich ist. Zur Überprüfung der Abweichungshypothese existieren verschiedene Modelltests. Klassischerweise wird ein Chi^2-Abweichungstest berechnet. Ist dieser signifikant, unterscheiden sich die Gruppen. Allerdings stellt die Abhängigkeit dieses Tests von der Stichprobengröße gerade bei etwas größeren Studien ein Problem dar, weshalb mittlerweile üblicherweise eher auf praktische Modelltests wie RMSEA (Root Mean Square Error Approximation) und CFI (Comparative Fit Index) oder TLI (Tucker Lewis Index) zurückgegriffen wird. Für eine ausführlichere Darstellung der Methoden von konfirmatorischer Faktorenanalyse wird auf Jöreskog & Sörbom (1979) und Jöreskog & Moustaki (2001) verwiesen.

Die Ergebnisse der mittels konfirmatorischer Faktorenanalysen, Mehrgruppenmodellen oder Strukturgleichungsmodellen durchgeführter Analysen werden hinsichtlich der Frage der Äquivalenz unterschiedlichen Stufen zugeordnet. Die niedrigste Stufe ist die *konfiguralen Äquivalenz.* Hier wird überprüft, ob die Faktorenstruktur in den unterschiedlichen Gruppen die gleiche ist. Das bedeutet, dass die Items zwar auf den gleichen Faktoren laden, die Richtung und Größe der Ladungen jedoch in den unterschiedlichen Gruppen voneinander abweichen können. Die zweite Stufe der Äquivalenz ist die *metrische Äquivalenz.* Damit diese zutreffend ist, müssen nicht nur die Faktorstruktur, sondern auch die Größe und Richtung der Faktorladungen übereinstimmen.

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

Aufgrund der Datenstruktur und des Multi-Matrix-Designs der EBAFLS-Studie kann im Rahmen der vorliegenden Arbeit leider nicht auf die Methode der konfirmatorischen Faktorenanalyse und der Mehrgruppenanalysen zurückgegriffen werden. Auf mögliche Forschungsprojekte unter Anwendung dieser Methoden wird im Rahmen der Ergebnisdiskussion eingegangen.

Zusammenfassung & und Relevanz für die vorliegende Arbeit. Die Frage nach der interkulturellen Validität von Testverfahren ist als einer ihrer Teilbereiche der Disziplin der interkulturellen Psychologie zuzuordnen. Es wurde dargestellt, dass Theorien und Methoden der interkulturellen Psychologie maßgeblichen Anteil an dem theoretisch-methodischen Rahmen dieser Arbeit haben. Es wurde eine Taxonomie interkultureller Studien (Van de Vijver & Leung, 1997) dargestellt. Unter Bezugnahme auf die Taxonomie interkultureller Studien von van de Vijver und Leung (1997), die in ihren wesentlichen Ansätzen kurz vorgestellt wurde, kann die vorliegende Arbeit als eher hypothesenbasiert charakterisiert werden und es werden Kontextfaktoren zur Erklärung von beobachteten Unterschieden herangezogen. Die Einbeziehung der genannten Taxonomie interkultureller Studien hat ferner den Zweck, die vorliegende Arbeit hinsichtlich ihrer Ausrichtung von der zugrundeliegenden EBAFLS-Studie abzugrenzen. Weiter wurde dargestellt, dass beobachtete Unterschiede entweder durch Struktur- oder Niveauunterschiede hinsichtlich des gemessenen Konstrukts zustande kommen können. Im Rahmen dieser Arbeit wird angenommen, dass beides ursächlich sein kann, was aufgrund der vorliegenden Datenstruktur allerdings nicht mit Hilfe von konfirmatorischen Faktorenanalysen und Strukturgleichungsmodellen überprüft werden kann.

Die in diesem Teil angesprochenen und erörterten Theorien und Methoden sind für diese Arbeit hoch relevant, da es mit ihrer Hilfe möglich wird, Antworten auf die Frage nach interkultureller Validität von Testverfahren zu formulieren. Mit anderen Worten: Sie stellen einen für die Arbeit relevanten, rahmengebenden theoretisch-methodischen Hintergrund dar. Die interkulturelle Vergleichbarkeit von Testergebnissen ist Gegenstand dieser Arbeit, weshalb diese Disziplin im Rahmen der theoretischen Grundlagen behandelt wurde.

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

Ziel dieses Theorieabschnitts ist es, dieser Dissertation und den aus unterschiedlichen Disziplinen stammenden Theorien ein Rahmenkonzept zuzuweisen. Darüberhinaus sollen hier die oben diskutierten Theoriestränge der Arbeit in dieses Rahmenkonzept eingeordnet werden.

Da es sich bei dem Ausgangsphänomen DIF um ein Validitätsproblem zwischen verschiedenen Gruppen handelt, bietet sich als Rahmenmodell ein Validitätsmodell an, das gleichzeitig *kulturelle und soziale Faktoren und Auswirkungen* beachtet. Um ein solches handelt es sich beim dem Validitätsmodell von Messick (1989). Es hat seinen Ursprung im Bereich des Sprachentestens und wurde in diesem Forschungsbereich bereits viel diskutiert (McNamara, 2006).

2.4.1. Samuel Messicks Validitätskonzept

Im Folgenden sollen Messicks (1989) Validitätskonzept genauer vorgestellt sowie dort wichtige Begrifflichkeiten und Konzepte erläutert werden. Deren Bedeutung wird speziell im Hinblick auf interkulturelle Vergleiche und Fremdsprachenkompetenzen diskutiert. Verschiedene, miteinander verwandte Teilkonzepte der Validität müssen im Kontext dieser Arbeit betrachtet werden. Darüberhinaus soll auf zwei Ursachen der Validitätseinschränkung eingegangen werden, nämlich konstruktirrelevante Varianz und Varianzeinschränkung. Des Weiteren werden die damit verwandten Konzepte des *„Teaching to the Test"* und *„Washback"* dargestellt. Ziel der Vorstellung dieses Konzepts ist, die Argumentationskette dieser Arbeit im Hinblick auf die Existenz unterschiedlicher Testkulturen und deren Einfluss auf DIF zu unterstützen. Ein weiterer Grund für die Darstellung des Konzepts ist die Zuordnung der Arbeit zu einem der wichtigsten Forschungsgebiete der Psychologie und der pädagogisch-psychologischen Diagnostik, nämlich dem der Validität.

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-

terpretation) eine hoch relevante Rolle spielen. Wie bereits angesprochen wurde, ist darüberhinaus 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 Validität und den Folgen von Invalidität das Konzept von Samuel Messick (1989) zugrunde zu 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 Annahmen die Adäquatheit und Angemessenheit von Interpretationen und Folgen von Testergebnissen unterstützen. Validität ist keine Eigenschaft eines Tests an sich, sondern der Bedeutung und Interpretation von Testergebnissen.

Dies bedeutet, dass in Messicks Verständnis von Validität vor allem die Bedeutung und Interpretation 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 and fairness." (Messick, 1995, S. 741).

Fairness beinhaltet hier beispielsweise Fairness hinsichtlich der Testverwendung, die Angemessenheit der dem Test und seiner Auswertung zugrunde gelegten Regeln, der verwendeten Konstrukte 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.

Messick setzt Validität gleich mit Konstruktvalidität. Diese beinhaltet sowohl den Testinhalt als auch Kriterium und Konsequenzen der Interpretation von Testwerten. Konstruktvalidität wird hier betrachtet als ein Set von Indikatoren, die Hinweise auf die Natur des zugrunde liegenden psychologischen Konstrukts geben können. Klieme (1989) beschreibt das Konzept der Konstruktvalidierung als „Meta-Programm für ganze Familien technologischer und Domain-spezifischer Forschungsprogramme". Die Gleichsetzung von Validität mit Konstruktvalidität wird schon seit den 50er Jahren des 19. Jahrhunderts diskutiert und umgesetzt (Klieme, 1989). Die von Messick definierten sechs Aspekte der Konstruktvalidität sind folgende:

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

An dieser Stelle sollen nicht alle Einzelaspekte des Konzepts besprochen werden, da viele auch bereits Teil der klassischen Validitätstheorien sind. Für eine detailliertere Diskussion von Messicks Validitätskonzept wird hier auf Messick (1989; 1996a) und McNamara (2006) verwiesen.

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 Konstrukts 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).

Da es sich in dieser Arbeit um die Analyse und Erklärung von Differentiellen Item Funktionen handelt, die ein Ausdruck von Invalidität eines Tests oder von Items über verschiedene Gruppen hinweg sind, ist es wichtig, mögliche Quellen von Invalidität zu identifizieren. Im Rahmen von Messicks Validitätskonzept werden verschiedene genannt. Zwei davon sind für diese Dissertation relevant und werden im Folgenden genauer beschrieben.

2.4.2. Quellen der Invalidität

Es existieren nach Messick zwei Ursachen für die Einschränkung der Konstruktvalidität: eine *Unterrepräsentanz des Konstrukts* auf der einen Seite und die *Einführung von konstruktirrelevanter* Varianz auf der anderen. Bei Ersterer erfasst ein Test nicht alle Komponenten des Konstrukts, bei Letzterer werden hingegen — möglicherweise zusätzliche — Komponenten erfasst, die nicht dem fraglichen Konstrukt zugehörig sind. Eine Kombination aus beidem wäre gleichfalls denkbar: Dies wäre beispielsweise dann der Fall, wenn ein Konstrukt anhand der Aufgaben eines Tests nicht vollständig erfasst wird, diese gleichzeitig jedoch auch konstruktirrelevante Varianz durch Erfassung einer konstruktirrelevanten Dimension erzeugen.

Hinsichtlich konstruktirelevanter 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 Ernschei-

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 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.

Im Kontext dieser Arbeit wird angenommen, dass es sich bei DIF zumindest nicht ausschließlich um einen „Nuisance Dimension", also konstruktirrelevante Varianz handelt, sondern um (aus unterschiedlichen Bildungskulturen hervorgegangene) unterschiedliche Stärken und Schwächen von Gruppen. Diese Betrachtungsweise entspricht der zweiten von Messick genannten Quelle für die Einschränkung von Konstruktvalidität, nämlich der Konstruktunterrepräsentation: Werden DIF als ein *Ausdruck unterschiedlicher Stärken und Schwächen* von Gruppen hinsichtlich des gleichen Konstrukts definiert, dann kann dies auch als eine *differentielle Unterrepräsentanz des Konstrukts innerhalb der unterschiedlichen Gruppen* betrachtet werden: Vorstellbar wäre, dass in unterschiedlichen Gruppen, bedingt durch differentielle Lerngelegenheiten, der Schwerpunkt auf unterschiedlichen Konstruktkomponenten liegt. Dies führt zwar zu unterschiedlichen Schwierigkeiten in den Gruppen hinsichtlich dieser Konstruktkomponenten, d.h. zu unterschiedlichen Stärken und Schwächen, ist jedoch nicht Ausdruck konstruktirrelevanter Varianz. Diese Überlegungen werden im Rahmen der Ergebnisdiskussion in Abschnitt 6 erneut aufgegriffen.

2.4.3. „Consequencial 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-

ness 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.

Messick geht in seinem Konzept ferner davon aus, dass das zu messende Konstrukt von den einer Gesellschaft zugrundeliegenden sozialen Werten mit definiert wird. Aufgrund dessen sind in seinem Konzept gesellschaftliche und soziale Werte ein Teil der Konstruktvalidität. Das hängt damit zusammen, dass die Testwertinterpretation und deren Folgen, von Messick *„consequential aspect of construct validity"* genannt, durch die sozialen Werte einer Gesellschaft mit definiert werden:

> „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.

Um diesen Zusammenhang genauer zu verdeutlichen, sollen an dieser Stelle zwei weitere Begrifflichkeiten eingeführt werden: der sogenannte „Washback" Effekt" (z.B. Alderson & Wall, 1993) und das Phänomen des „Teaching to the Test" (z.B. Koretz, 2005). Die von Messick (1996) verwendeten Ausdrücke „positive washback" und „negative washback" sind synonym dazu zu verwenden. Beides sind letztlich Wege der Übertragung gesellschaftlicher

Normen und Werte, also der Messick'schen „social values", auf das getestete Konstrukt und somit auf Testaufgaben: Haben Testwertergebnisse und deren Interpretation relevante Folgen, beispielsweise für die berufliche Zukunft von Schülern, dann — so die Annahme — wird im Unterricht ein erhöhter Wert auf das Üben und die Verwendung ebendieser im Test verwendeten Aufgaben und Inhalte gelegt. Bei den beiden Effekten handelt es sich im Prinzip um zwei Aspekte desselben Sachverhalts: „Teaching to the Test" wird tendenziell negativ wahrgenommen. Bei diesem Konzept wird davon ausgegangen, dass aufgrund der hohen Relevanz der Testwertergebnisse die Form von im „high stakes"-Test verwendeten Testaufgaben häufiger unterrichtet und geübt wird. Die ursprünglichen Testaufgaben jedoch erfassen nicht das gesamte zu vermittelnde Konstrukt. Aufgrund der hohen Relevanz der Testergebnisse nähert sich das im Unterricht vermittelte Konstrukt nunmehr dem getesteten, jedoch nicht vollständigen Konstrukt an: Aufgrund von durch die Testergebnisse drohenden Konsequenzen für Schüler und teilweise auch Lehrer (in den USA beispielsweise wurden Lehrer teilweise aufgrund der Testergebnisse ihrer Schüler bewertet, und Schulen wurden finanziell belohnt oder bestraft) kann es geschehen, dass der Unterricht des entsprechenden Fachs nur noch auf den Testinhalt, der in diesem Falle nicht das gesamte zu vermittelnde Konstrukt repräsentiert, ausgerichtet ist.

Damit soll sichergestellt werden, dass die Testwertergebnisse gut sind und keine negativen Folgen für die Beteiligten erwachsen. Daher findet durch „Teaching to the Test" letztlich eine habitualisierte Unterrepräsentanz des Konstrukts statt, und relevante Lerninhalte werden im Unterricht teilweise zu Gunsten der Testinhalte vernachlässigt. Koretz & Barron (1998) haben sich mit diesem Phänomen in den USA, und speziell in Kentucky im Rahmen des „No Child Left Behind"-Programms, über mehrere Jahre hinweg beschäftigt. Das Problem, das sich dort darstellte, war eine annähernd ausschließliche Fokussierung des Unterrichts auf die Testinhalte. Die Autoren konnten zeigen, dass auf diese Weise eine starke Einschränkung des Konstrukts stattfinden kann und dass Inhalte verstärkt gelehrt werden, die nicht Teil der zu erfassenden latenten Fähigkeit sind. Es zeigte sich das Phänomen der „score inflation", der Inflation von Testwerten, bei dem die Schüler zunächst durch „high stakes-testing" im US-Bundesstaat Kentucky deutlich bessere Testwerte erreichten. Allerdings zeigten spätere landesweite Schulleistungstests (NAEP; National Assessment of Educational Progress), dass dieses Phänomen ausschließlich auf ganz bestimmte Aufgaben beschränkt war und die Schüler im Mittel nicht besser als ihre Mitschüler aus anderen Staaten geworden waren. Lehrer berichteten außerdem, sich in ihrem Unterricht deutlich mehr auf den getesteten Inhalt konzentriert zu haben. Durch diese Konzentration auf bestimmte Testinhalte wird also hier Invalidität im Sinne einer Konstruktunterrepräsentanz gefördert.

Bei dem zweiten Konzept handelt es sich um den „Washback Effect" (Alderson & Wall, 1993). In diesem Fall fördert das Testen das Erlernen von genau den gewünschten Inhalten und Konzepten.
Washback stellt also den anderen Aspekt des Sachverhalts dar. Das zugrundeliegende Konzept ist im Prinzip dasselbe: Zur Vermeidung negativer Folgen der Testwertinterpretation, wie sie oben bereits dargelegt wurden, wird der Unterricht stärker auf die Testinhalte ausgerichtet. Allerdings wird in diesem Fall das zu erfassende Konstrukt komplett von den Testinhalten abgebildet. Darüberhinaus wurden zuvor im Unterricht nicht das gesamte zu vermittelnde Konstrukt beziehungsweise viele konstruktirrelevante Inhalte berücksichtigt. Dieses Phänomen beschreibt also die Tatsache, dass durch die Folgen der Testwertinterpretationen die Schüler mehr oder konstruktrelevantere Inhalte im Unterricht erlernen, indem sich das vermittelte Konstrukt dem getesteten (vollständigen) Konstrukt annähert. Verstärkend wirkt hier die Anwendung von „high stakes"-Tests, da die Folgen dieser Tests für die Beteiligten relevant sind. Damit der „Washback-Effekt" zum Tragen kommen kann, müssen Test- und Konstruktvalidität unbedingt gegeben sein. Ansonsten handelt es sich wiederum um „Teaching to the Test" mit den dazugehörigen negativen Folgen für das zu vermittelnde Konstrukt. Neben Messick beschäftigen sich noch weitere Forscher mit dem Konzept des „Washback" (z.B. Alderson & Wall, 1993). Einen Überblick gibt Bailey (1996).

2.4.4. Zusammenfügung der Theoriestränge im Kontext der Validität

Die drei diskutierten Theoriebereiche, nämlich erstens „Interkulturelle Vergleichbarkeit von Testergebnissen", zweitens „Differentiellen Item Funktionen" und drittens „Fremdsprachenforschung", sollen im Folgenden zu Messicks (1989) Ansatz in Beziehung gesetzt werden. Ferner soll anschließend der Stellenwert von Messicks Ansatz im Rahmen dieses Dissertationsvorhabens geklärt werden.

Die Frage nach der Vergleichbarkeit von Testwertergebnissen sowie von Konstrukten ist eine klassische Fragestellung der interkulturellen Vergleichbarkeit von Testergebnissen. Auch in Messicks Validitätsansatz haben, wie oben ausgeführt, die Bereiche von von Vergleichbarkeit und Fairness einen zentralen Stellenwerden, da die Nicht-Vergleichbarkeit von Testergebnissen eine Einschränkung der Konstruktvalidität darstellt.

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

Varianz, also als konstruktirrelevante Leichtigkeit oder Schwierigkeit interpretiert werden. Zum anderen können DIF auch als Konstruktunterrepräsentation betrachtet werden, wobei differentielle Stärken und Schwächen bzw. differentielle Relevanz einzelner Konstruktkomponenten bei unterschiedlichen Gruppen beim Bearbeiten von Testitems zum Tragen kommen. Daher kann es sich bei DIF auch um eine differentielle Unterrepräsentanz des zu messenden Konstrukts in den unterschiedlichen Gruppen handeln, wie oben bereits dargelegt wurde. Beide Quellen für Konstruktvaliditäts-Minderung spielen eine zentrale Rolle in Messicks Theorie.

Primär wichtig wird das Messick'sche Validitätskonzept jedoch bei der Frage nach dem *Entstehen von differentiellen Stärken und Schwächen.* Bei Anwendung von Messicks Rahmenkonzept könnten diese wie folgt zustande kommen, wobei zunächst von zwei Grundsachverhalten ausgegangen wird: Erstens, die Interpretationen und Folgen von Testwertergebnissen sind durch die sozialen Werte einer Gesellschaft geprägt. Diese sind, zweitens, zumindest teilweise gruppen- bzw. kulturspezifisch. Anhand der oben beschriebenen Übertragungsmechanismen von „Washback" und „Teaching to the Test" werden diese wiederum Teil des Unterrichts und der jeweiligen Testkultur. Aufgrund der unterschiedlichen, zugrunde liegenden sozialen Werte der verschiedenen Kulturen bilden sich so auf Dauer Schwerpunkte, die sich auf den vor dem Hintergrund der Bildungskultur konstruierten Items abbilden. Somit entsteht auch eine differentielle Relevanz einzelner Konstruktkomponenten. Bei einem Gruppenvergleich kommen diese besonders zum Tragen, da ein Item einfacher (bzw. schwerer) wird, wenn es der eigenen Testkultur entspricht (bzw. nicht entspricht). Es entstehen also differentielle Stärken und Schwächen. Im Rahmen der Ergebnisdiskussion (siehe Abschnitt 6) wird diskutiert, wie sich diese Überlegungen auf dieses Dissertationsvorhaben übertragen lassen können.

Die in dieser Arbeit verwendeten Daten stammen aus dem Bereich der *Fremdsprachenforschung*, daher wird der Fokus speziell auf Konstrukte dieses Bereichs gelegt. Es handelt sich bei dem Konstrukt „fremdsprachliches Leseverständnis" um das Konstrukt, bei dem aufgrund der Existenz differentieller Item Funktionen eine Einschränkung der Konstruktvalidität zu beobachten ist. Alle drei der besprochenen Theoriebereiche weisen somit Berührungspunkte zu Messicks Validitätsansatz auf und lassen sich dort einordnen.

2.4.5. Zusammenfassung und Relevanz für die Arbeit

Das Validitätskonzept von Messick ist aufgrund der Berücksichtigung des Einflusses von sozialen und kulturellen Werten auf Testwertinterpretationen und deren Folgen, auf die Konstruktdefinition und somit auch auf die Validität, für die Erklärung von beobachteten Unterschieden in interkulturellen Studien gut geeignet. Es dient daher als Rahmenkonzept für die vorliegende Arbeit. Es wurde dargelegt, auf welche Weise Konstruktrepräsentation, Folgen von Testwertinterpretationen und soziale Werte nach Messicks Konzept zusammenhängen. In diesem Rahmen wurden Quellen für Invalidität sowie die Konzepte „Washback" und „Teaching to the Test" dargelegt.

Es wurden, basierend auf Messicks Konzept, Überlegungen hinsichtlich des Zustandekommens von DIF aufgrund der beiden Quellen für Konstruktvaliditätsminderung, nämlich die Einführung konstruktirrelevanter Varianz und Konstruktunterrepräsentation, angestellt. Bezüglich Letzterem wird im Rahmen dieser Arbeit angenommen, dass es mit der Existenz von Differentiellen Item Funktionen zusammenhängt. Dies wäre im Rahmen von Messicks Konzept demnach so zu erklären, dass zwar in allen analysierten Gruppen dasselbe zugrundeliegende Konstrukt gültig ist, durch die unterschiedlichen, von sozialen Werten der verschiedenen Gesellschaften geprägten Testkulturen jedoch bei jeweils den Konstruktkomponenten, die eben gerade keinen Schwerpunkt der Test- und Bildungskultur darstellen, eine Konstrukt*unterrepräsentation* stattfindet. Die Einschränkung der Validität ist in diesem Falle also (zumindest teilweise) dadurch bedingt und nicht unbedingt nur, wie häufig in der DIF-Forschung postuliert (siehe auch 2.1), durch die zweite mögliche Quelle von Invalidität, nämlich die Konstruktirrelevanz. DIF entstehen demnach vor allem dann, wenn eine Konstruktkomponente in einer Gesellschaft einen Schwerpunkt darstellt, in einer anderen hingegen nicht. Inwieweit die Ergebnisse dieser Arbeit dazu beitragen, diese Überlegungen zu unterstützen, wird im Rahmen des Diskussionsteils dargelegt. Da Validität und interkulturelle Vergleichbarkeit von Testverfahren das Hauptthema dieser Arbeit darstellt, ist die Anwendung von Messick's Validitätskonzept hier relevant und sinnvoll: Es erlaubt, Vermutungen darüber anzustellen, wodurch Konstruktunterschiede in unterschiedlichen Gesellschaften und Kulturen möglicherweise bedingt sein könnten.

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 Bezug auf die Frage nach Itemsschwierigkeits-Determinanten und die Erklärung von Itemschwierigkeit sind dabei die in dem Modell von Bachman & Palmer (1996, siehe 2.2) entwickelten theoretischen Überlegungen von Relevanz. Im Rahmen dieses Modells wird die Annahme getroffen, dass die Itemschwierigkeit durch bestimmte Anforderungsmerkmale der Items mitbedingt wird. Diese werden auch bereits spezifiziert, wobei insbesondere Item-Anforderungsmerkmale aus der kognitiv-linguistischen Kategorie hervorgehoben werden. Für die Erklärung von Itemschwierigkeit kommt weiterhin den in dem auf dem GERS aufbauenden Itemkategorisierungsinstrument „Dutch Grid" (Alderson et al., 2006) verwendeten, kognitiv-linguistischen Anforderungsmerkmalen besondere Bedeutung zu. Auch die Überlegungen von Grotjahn (2000) hinsichtlich der kognitiven Prozesse, die den einzelnen Item-Anforderungsmerkmalen zugrunde liegen, sind vor allem für die Interpretation von Itemschwierigkeit und diesbezüglich beobachteten Unterschieden interessant.

Des Weiteren sind vor allem vier der unter 2.1.2 und 2.2.4 vorgestellten empirischen Arbeiten von besonderer theoretischer, empirischer und methodischer Wichtigkeit für diese Dissertation. Die erste dieser Arbeiten stammt von Hartig, Frey, Nold und Klieme (2010) und beschäftigt sich mit der Erklärung von Itemschwierigkeiten durch (aus der kognitiv-linguistischen Kategorie stammende) Item-Anforderungsmerkmale bei Englisch-Items. Darüber hinaus nehmen die Autoren einen Vergleich sowie eine Einschätzung der Eignung unterschiedlicher zu diesem Zweck verwendbarer Methoden vor. Bei der zweiten Arbeit handelt es sich um eine Studie von Artelt und Baumert (2004). Die Autoren untersuchten PISA 2000-Leseverständnis-Daten und fanden heraus, dass die Verwendung von Items aus dem eigenen Sprachraum für die Schüler aus Ländern, in denen die Items ursprünglich konstruiert wurden, einen Vorteil darstellte. Ferner wurde gefunden, dass diese Vorteile innerhalb unterschiedlicher Länder der gleichen Sprachgruppe sich unterschiedlich darstellten. Dies lässt darauf schließen, dass hier nicht nur die Ursprungssprache und mögliche Übersetzungsfehler einen Einfluss auf die Itemschwierigkeit haben, sondern noch weitere länderspezifische bildungskulturelle Variablen. Dies wird im Zusammenhang mit der Fragestellung dieser Dissertation als ein Hinweis gewertet, dass Items offenbar auch bildungs- und testkulturelle Faktoren des Herkunftslandes abzubilden scheinen. Welche genau das sind, wurde jedoch in der Arbeit von Artelt & Baumert (2004) nicht ausdifferenziert.

Eine differenziertere Betrachtung möglicher kultureller Einflussfaktoren auf DIF unternahmen hingegen Klieme und Baumert (2001) in der dritten relevanten Studie. Die Autoren konnten für Mathematik-Items zeigen, dass Testitems tatsächlich länderspezifische, curriculare Schwerpunkte und Testkulturen widerspiegeln, welche korrelative Zusammenhänge zu Differentiellen Item Funktionen in der erwarteten Richtung aufwiesen. Die Hypothesen hinsichtlich der zu erwartenden Richtung der Zusammenhänge basierten auf der Häufigkeit des Vorkommens bestimmter Anforderungsmerkmale bei den Items der unterschiedlichen Teilnehmerländer. Diese wurden von den Autoren als zu erwartende Stärken und Schwächen der Gruppen definiert. Diese Ergebnisse sind für die vorliegende Arbeit von zentraler Bedeutung, da sie eine ihre Grundannahmen unterstützen: Auf Testitems unterschiedlicher Länder werden unterschiedliche Schwerpunkte von Testkulturen abgebildet; diese wiederum weisen systematische Zusammenhänge zu DIF in Richtung der erwarteten Stärken und Schwächen der Gruppen auf.

In dem vierten, für die vorliegende Arbeit bedeutsamen 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 unterschiedliche Profile von Stärken und Schwächen der unterschiedlichen Gruppen sind, die sich wiederum 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 sprechen neben den theoretischen Überlegungen von Bachman & Palmer (1996) und Alderson (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 ähnlichen 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 eingesetzten Itemschwierigkeits-Determinanten und der Kategorien, hinsichtlich derer die Items einzuordnen 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 Referenzrahmen ist, der die angenommene Vergleichbarkeit ja bereits namentlich impliziert.

Im Rahmen dieser Arbeit wird angenommen, dass in allen Ländern zumindest teilweise die gleichen 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

Eine weitere in dieser Arbeit getroffene Annahme ist, dass die in einem Land konstruierten Testitems jeweils länderspezifische, curriculare Schwerpunkte der Unterrichts- und Testkultur abbilden. Diese wiederum werden durch die Häufigkeit des Vorkommens von Item-Anforderungsmerkmalen abgebildet, wie sie etwa im „Dutch Grid" beschrieben werden. Diese Annahme wird durch die oben dargelegten Arbeiten von Klieme und Baumert (2001) sowie Artelt und Baumert (2004) unterstützt. Weiterhin wird angenommen, dass dies durch die einer jeweiligen Gesellschaft zugrundeliegenden sozialen Werte (mit)bedingt wird, die wiederum in den einzelnen Ländern durch „Washback" bzw. „Teaching to the Test" (siehe auch 2.4) auf die Testaufgaben und die Testkultur an sich „übertragen" werden. Letzteres kann im Rahmen dieser Arbeit jedoch nicht empirisch überprüft werden.

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

Die Arbeiten von Klieme und Baumert (2001), sowie von Dogan, Guerrero und Tatsuoka (2005) weisen darauf hin, dass sich die testkulturellen Schwerpunkte (im Sinne von Häufigkeiten von Item-Anforderungsmerkmalen) in den unterschiedlichen Ländern zumindest teilweise unterscheiden. Es ist anzunehmen, dass dies differentielle Lerngelegenheiten bedingt und dadurch bei einem Leistungsvergleich von Gruppen mit unterschiedlichen testkulturellen Schwerpunkten Unterschiede hinsichtlich der Itemschwierigkeit zumindest teilweise verursacht werden. Demnach sollten die Testkulturen und die zu erwartenden Stärken und Schwächen der Gruppen bei der Beantwortung von Items mit bestimmten Item-Anforderungsmerkmalen dadurch operationalisiert werden können, dass die den unterschiedlichen Ländern bzw. Kulturen entstammenden Items hinsichtlich der Häufigkeit des Vorkommens dieser Anforderungsmerkmale analysiert werden. Signifikante Unterschiede sollten hier ein Hinweis auf unterschiedliche zugrundeliegende Testkulturen sein und somit auch Hinweise auf die zu erwartenden Stärken und Schwächen der einzelnen Gruppen geben. Das bedeutet, dass der Umstand, dass ein Item ein Anforderungsmerkmal enthält, das in der eigenen Testkultur ein häufig vorkommendes Anforderungsmerkmal von Items ist, die Itemschwierigkeit verringern sollte im Vergleich zu Gruppen, in denen dies nicht der Fall ist. Umgekehrt gilt natürlich: Wenn Schüler aus einem Land bestimmten Itemmerkmalen oder Ausprägungen dieser Itemmerkmale weniger häufig exponiert sind und so in der Testkultur dieser Gruppe schwerpunktmäßig andere zugrundeliegende kognitive Prozesse gefördert werden, sollte die Itemschwierigkeit für diese Schüler höher sein. Diese Arbeit

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

Die bisher getroffenen Annahmen lassen sich wie folgt zusammenfassen: Bestimmte Anforderungsmerkmale von Items bilden zur Lösung des Items notwendige kognitive Prozesse ab. Das bedeutet, diese spiegeln einen Teil der Anforderungen für eine korrekte Lösung des Items wider. Diese Anforderungsmerkmale bestimmen daher zum Teil die Schwierigkeit eines Items.

Die Schwerpunktlegung diesbezüglich wiederum, das heißt die Häufigkeit des Vorkommens der unterschiedlichen Anforderungsmerkmale bei den Items eines Landes, ist durch zugrundeliegende soziale und bildungskulturelle Werte geprägt. Die Unterrichts- und Testkultur eines Landes sollte sich zum Teil durch die Analyse von repräsentativen Items dieses Landes operationalisieren lassen.

Solche Schwerpunkte hinsichtlich des Vorkommens bestimmter Item-Anforderungsmerkmale bei den Items eines Landes führen zu häufigeren Lerngelegenheiten und somit auch zu Stärken und Schwächen der aus dem Land stammenden Schüler hinsichtlich der Beantwortung von Items mit diesen Anforderungsmerkmalen. Es ist ferner davon auszugehen, dass sich unterschiedliche Länder bezüglich der Schwerpunkte von Anforderungsmerkmalen aufgrund unterschiedlicher zugrundeliegender sozialer und bildungskultureller Werte voneinander unterscheiden.

Die kulturell bedingte Varianz der Itemschwierigkeit zwischen unterschiedlichen Ländern, d.h. DIF, sollte sich demnach zumindest teilweise durch diese unterschiedlichen erwarteten Stärken und Schwächen erklären lassen. Die zu erwartenden Stärken und Schwächen der Gruppen wiederum sollten sich durch die Analyse möglichst repräsentativer, aus den jeweiligen Ländern stammenden Items herleiten lassen.

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.

Insgesamt ergeben sich aus den oben dargelegten Annahmen drei aufeinander aufbauende Komplexe von Fragen, deren sukzessive Bearbeitung zur Beantwortung der Hauptfragestellung notwendig ist. Diese werden im Folgenden jeweils gemeinsam mit den dazugehörigen Hypothesen formuliert. Im Hinblick auf die Richtung der Zusammenhänge zwischen DIF und Indikatoren nationaler Testkulturen können ferner spezifische Hypothesen aufgestellt werden. Dies ist jedoch nicht möglich hinsichtlich der Höhe der Zusammenhänge. Daher ist diese Arbeit in dieser Hinsicht eher von explorativer Natur.
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 zweite Fragenkomplex bezieht sich auf die Erklärung von Itemschwierigkeit anhand kognitiv-linguistischer Item-Anforderungsmerkmale der Items innerhalb der Länder. In diesem Rahmen soll darüberhinaus festgestellt werden, ob sich die Zusammenhänge zwischen Itemschwierigkeit und Itemmerkmalen in den verschiedenen Ländern unterscheiden. Dieser Fragenkomplex dient zur Überprüfung und zur Feststellung der Eignung des Itemkategorisierungsinstruments Dutch Grid zur Kategorisierung von Items hinsichtlich ihrer kognitiv-linguistischen Anforderungsmerkmale.
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

Eine Voraussetzung für die Anwendung des Rasch-Modells ist die Modellkonformität der Items innerhalb der Länder. Diese soll daher an erster Stelle überprüft werden. Dahinter steht die Frage, ob die Items innerhalb der einzelnen Länder das gleiche Konstrukt erfassen. Dies ist eine Voraussetzung dafür, überhaupt Analysen zwischen den Ländern durchführen zu können und diese zu interpretieren (Maris, Bechger & Veldhuijzen, 2006). Ferner werden die im Rahmen

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 defniniert. 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 unter 2.2 diskutierten Theorien von Bachman (1990; Bachman & Palmer, 1996), des Dutch Grid (Alderson et al., 2006) bzw. des Europäischen Referenzrahmens (Europarat, 2001), sowie weitere empirische Befunde (siehe 2.2.4) weisen darauf hin, dass die aus dem „Dutch Grid" verwendeten, kognitiven Anforderungsmerkmale der Items die Itemschwierigkeit zumindest teilweise determinieren sollten, da die Schwierigkeit eines Items durch zwei Sets von Charakteristika determiniert werden: Eigenschaften auf Seiten des Getesteten und Eigenschaften auf Seiten des Items/Texts. In dieser Arbeit werden nur Eigenschaften aus dem Charakteristika-Set der Items herangezogen, da die Verwendung beider Charakteristika-Sets mit den in der EBAFLS-Studie erhobenen Daten nicht möglich ist.
Die Beibehaltungg 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:

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

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

auch im Hinblick auf regressionsanalytische Zusammenhänge bearbeitet werden, da auf diese Weise einerseits ein Maß für den Zusammenhang der Itemschwierigkeit mit den einzelnen Anforderungsmerkmalen, unter Berücksichtigung weiterer Prädiktoren, andererseits auch ein Maß für den Gesamtzusammenhang und den Anteil der durch die Prädiktoren erklärten Varianz der Itemschwierigkeit innerhalb der Länder entsteht. Diese Vorgehensweise empfehlen auch Scheuneman & Gerritz (1990).

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:

> Hypothese 2c: Anhand der verwendeten Prädiktoren lässt sich ein Teil der Varianz der Itemschwierigkeiten innerhalb der Länder aufklären. Hypothese 2d: Die kognitiv-linguistischen Item-Anforderungsmerkmale des „Dutch Grid"-Kategoriensystems weisen einen regressionsanalytischen Zusammenhang zu den Itemschwierigkeiten innerhalb der Länder auf.

Über die Richtung und Stärke dieser Zusammenhänge innerhalb der Länder sollen an dieser Stelle keine Annahmen getroffen werden, da die Möglichkeit besteht, dass innerhalb der Länder der Einfluss von Testkultur und der Einfluss der kognitiv-linguistischen Schwierigkeit der Items miteinander konfundiert sind. Daher erfolgt die Überprüfung dieser Hypothesen hinsichtlich der Höhe und Richtung der Prädiktoren sowie der Größe des Anteils der aufklärbaren Varianz explo-ratorisch.

3.5. Fragenkomplex 3: Erklärung von Differentiellen Item Funktionen

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

Nachdem in Frage 2 auf die Zusammenhänge von Item-Anforderungsmerkmalen und Itemschwierigkeiten und auf die Erklärung der Varianz der Itemschwierigkeiten innerhalb der Länder anhand der Anforderungsmerkmale eingegangen wurde, wird nun der Fokus auf Unterschiede hinsichtlich der Itemschwierigkeiten *zwischen* den Ländern gelegt. Dabei wird hier ausschließlich die durch Gruppenzugehörigkeit und somit kulturelle Unterschiede verursachte Varianz, DIF, betrachtet. Um einen Eindruck hinsichtlich der Größe und Richtung des Einflusses der *einzelnen Indikatoren* zu erhalten, werden in einem ersten Schritt korrelative Zusammenhänge analysiert. Wie oben bereits dargestellt, konnte in empirischen Studien (Klieme & Baumert, 2001; Klieme & Bos, 2000; Scheuneman & Gerritz, 1990) gezeigt werden, dass signifikante korrelative Zusammenhänge zwischen Testkultur-Indikatoren, d.h. unterschiedlichen erwarteten Stärken und Schwächen der Länder, und differentiellen Item-Funktionen, also der kulturell bedingten Varianz der Itemschwierigkeiten *zwischen* den Ländern, bei Mathematikitems bzw. Leseverständnisitems, existieren. Daher wird hier überprüft, ob ähnliche Ergebnisse auch für den Bereich des fremdsprachlichen Leseverständnisses festzustellen sind. Dazu wird folgende Hypothese formuliert:

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

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-

gezogen. Auf diese Weise ließen sich beispielsweise in der Studie von Scheuneman & Gerritz (1990) Hinweise auf Zusammenhänge zwischen DIF und differentiellen Stärken und Schwächen von Gruppen finden. Es wird daher angenommen, dass auch in der vorliegenden Arbeit mit Hilfe der hier verwendeten Item-Anforderungsmerkmale ein Teil der kulturell bedingten Varianz aufzuklären ist.

Da aufgrund der fehlenden bzw. unvollständigen Erfassung von Hintergrundvariablen und personenbezogenen Variablen der Stichproben keine Personenmerkmale als Prädiktoren mit in das Modell aufgenommen werden können, diese jedoch vermutlich auch einen Teil der Varianz zwischen den Ländern mit verursachen (z.B. Van den Noortgate & de Boeck, 2005; Bachman & Palmer, 1996), muss davon ausgegangen werden, dass nicht die komplette Varianz mit Hilfe von itembasierten Testkultur-Indikatoren aufgeklärt werden kann. Die bei Scheuneman und Gerritz (1990) erklärten Anteile der Varianz bewegen sich zwischen 25% und 45%. Es ist zu vermuten, dass sich der mit Hilfe der Itemeigenschaften erklärte Anteil an Varianz auch in dieser Dissertation in diesem Rahmen bewegt; Da jedoch die Autoren des oben genannten Artikels andere DIF-Parameter (MH) verwendet haben, ist nicht klar, ob sich die Ergebnisse auf IRT-basierte Parameter übertragen lassen. Aus diesem Grund können hier keine expliziten Hypothesen bezüglich der Höhe der aufgeklärten Varianz aufgestellt werden. Daher wird diese Fragestellung exploratorisch behandelt, und es wird folgende Hypothese aufgestellt:

> 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 Theorieteil 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.

Danach werden die Methoden zur Bearbeitung von Fragenkomplex 2 „Erklärung von Itemschwierigkeiten" sowie für Fragenkomplex 3 „Erklärung von Differentiellen Item Funktionen" beschrieben.

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.
Hintergrund des Projekts war der Umstand, dass die Relevanz von Fremdsprachenkenntnissen in der multilingualen EU seit geraumer Zeit stetig zunimmt. In zahlreichen europäischen Ländern existieren daher bereits nationale Fremdsprachenzertifikate und -diplome. Zugleich und infolgedessen erhält dadurch die Frage nach der europaweiten Vergleichbarkeit dieser Fremdsprachenkompetenzen und -zertifikate eine immer größere Bedeutung. Ziel des dem Lingua2-Programm der europäischen Kommission zugeordneten EBAFLS-Projekts war, einen Beitrag zur Beantwortung dieser Frage zu leisten.
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

Europäischen Referenzrahmen für Sprachen (GERS; Europarat, 2001) zusammenzubringen. EBAFLS wurde in Kooperation acht europäischer Länder durchgeführt; teilnehmende Länder waren Deutschland, Frankreich, Luxemburg, die Niederlande, Schottland, Schweden, Spanien und Ungarn. Der deutsche Teil der EBAFLS-Studie wurde vom Bundesministerium für Bildung und Forschung finanziert und am Deutschen Institut für Internationale Pädagogische Forschung koordiniert und durchgeführt. Die offizielle Laufzeit des Projekts war von Januar 2005 bis Dezember 2007, in Deutschland wurde die Studie aufgrund der Zusatzstudie bis Ende 2008 verlängert. Deutschland beteiligte sich im internationalen Teil der Studie mit der Testung von Lese- und Hörverständnisaufgaben in den Sprachen Englisch und Französisch. Vorrangiges Ziel des Projekts war es, eine Itemdatenbank mit Ankeritems zur Überprüfung und zum länderübergreifenden Vergleich von Fremdsprachenkompetenzen (Leseverständnis und Hörverständnis) in den drei europäischen Verkehrssprachen Deutsch, Englisch und Französisch zusammenzustellen.
Die dort verwendeten Items sollten über verschiedene europäische Länder hinweg zu fairen und vergleichbaren Testergebnissen führen. Das bedeutet, dass die Items möglichst frei von differentiellen Item Funktionen sein sollten. Um innerhalb von Europa bereits bestehende nationale Fremdsprachentests und -zertifikate vergleichbar zu machen, sollte europäischen Ländern so die Möglichkeit geboten werden, ihre nationalen Tests mit der Itembank zu verknüpfen. Ziel war es, einen Beitrag zu einem transparenteren und valideren Sprachentesten zu leisten.

4.1.1. Herkunft der Items

Relevant für diese Dissertation ist vor allem die Tatsache, dass es sich bei den in der EBAFLS-Studie verwendeten Items um bereits überprüfte und für geeignet befundene Aufgaben aus den unterschiedlichen Teilnehmerländern handelt. Dahinter stand ursprünglich die Absicht zu überprüfen, ob in unterschiedlichen Ländern konstruierte Items als Ankeritems zum Verknüpfen verschiedener nationaler Examen geeignet sind. Die Items wurden außerdem den Niveaus des Gemeinsamen Europäischen Referenzrahmens für Sprachen (GERS; Europarat, 2001) zugeordnet, welcher gleichzeitig die theoretische Basis des Projekts darstellt. In der EBAFLS-Studie wurden das GERS - Niveau B1 fokussiert und zusätzlich die Niveaus A2 bis B2 analysiert (für eine theoretische Beschreibung des GERS siehe auch 2.2.3). Eine erste Zuordnung der Items wurde mit Hilfe des im unter 2.2.4 beschriebenen „Dutch Grid" (Alderson et al., 2006) in den Herkunftsländern der Testaufgaben durch Fremdsprachenexperten durchgeführt. Jedes Land stellte Items für die Testung von fremdsprachlichem Lese- und Hörverständnis zur Verfügung. Die Herkunft der das Leseverständnis betreffenden Items ist in Abbildung 4.1 dargestellt.

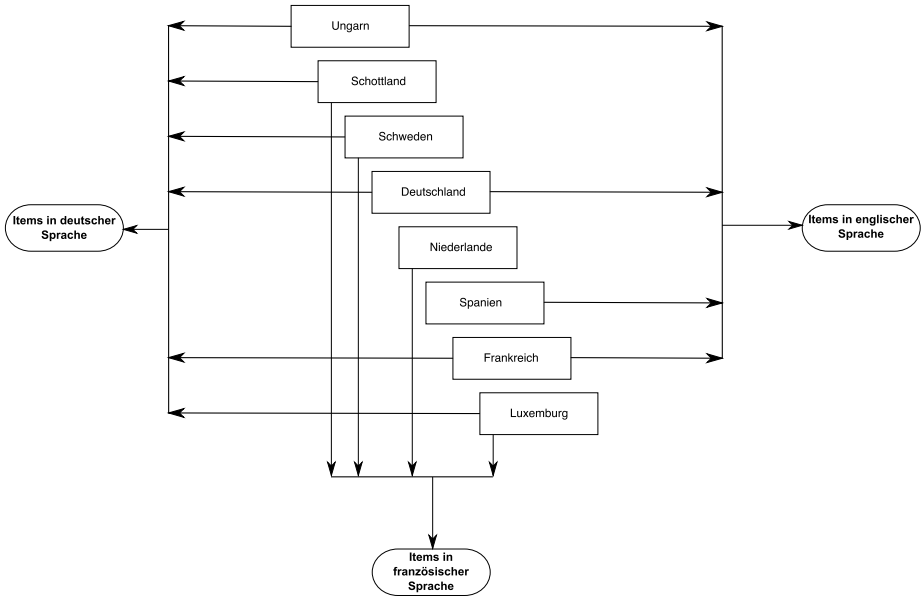

Abbildung 4.1. Herkunft der Items zur Messung fremdsprachlichen Leseverständnisses

4.1.2. Analysen in der EBAFLS Studie

Die Überprüfung der Testaufgaben erfolgte in den acht europäischen Ländern an Schülern der 9.-11. Klasse. Ziel dabei war es festzustellen, welche der in der Studie verwendeten Testitems zur Einspeisung in die Itemdatenbank geeignet sind. Hauptkriterium dafür ist, dass diese Items es ermöglichen, Sprachkenntnisse auf eine Kultur-faire Art und Weise zu messen, um diese damit auch vor dem Hintergrund verschiedener europäischer Bildungskulturen vergleichbar zu machen. Die Items wurden hinsichtlich ihrer Vergleichbarkeit vom federführenden Institut CITO in den Niederlanden analysiert.

Dazu wurde das One-Parameter Logistic Model (Verhelst, Glas & Verstralen, 1995) zugrunde gelegt. Obgleich dieses Modell als einparametrisch bezeichnet wird, handelt es sich dabei letztlich formal um ein 2-parametrisches IRT-Modell, da sowohl der Itemschwierigkeits- als auch der Itemdiskriminationsparameter berücksichtigt werden. Dabei wird in diesem Modell zweischrittig vorgegangen: der Diskriminationsparameter wird imputiert und nicht frei geschätzt und kann dabei ganzzahlige Werte von 1-15 annehmen. Der Schwierigkeitsparameter wird hingegen frei geschätzt. Dieses Modell wird hauptsächlich in der CITO angegliederten Forschergruppe angewandt. Für eine detaillierte Abhandlung dieses Modells wird auf Verhelst, Glas & Verstralen (1995) verwiesen. In der EBAFLS-Studie wurden nun die Items hinsichtlich der Existenz von Differentiellen Item Funktionen analysiert; In der vorliegenden Arbeit hingegen soll einen Schritt weiter gegangen werden mit dem Versuch, diese auch zu erklären.

4.1.3. Design und Stichprobe

Das EBAFLS-Projekt hatte eine offizielle Laufzeit von 3 Jahren, von November 2004 bis September 2007, und wurde in Deutschland aufgrund nationaler Zusatzerhebungen bis Ende 2008 verlängert. Die Erhebung des fremdsprachlichen Leseverständnisses wurde im Juni 2006 in den acht EBAFLS-Partnerländern durchgeführt. Die getesteten Schüler und Schülerinnen waren zu diesem Zeitpunkt meist zwischen 15 und 17 Jahre alt. Grund für das unterschiedliche Alter der Schülerinnen und Schüler war die projektinterne Festlegung, dass sich alle getesteten Schüler in etwa auf dem Niveau B1 des GERS befinden sollten. Je nach Beginn oder Intensität des Fremdsprachenunterrichts kann dies in unterschiedlichen Ländern zu einem unterschiedlichen Zeitpunkt der Fall sein. Insgesamt wurden 11712 Schüler getestet (Englisch: 4276; Französisch: 2670; Deutsch: 3894; gültig insgesamt 10841). Die für die Tests ausgewählten Items (Englisch: 122; Französisch:71; Deutsch: 104) wurden in einem Multi-Matrix-Design in jeweils 15 Testheften pro Sprache überprüft. In jedem der acht Länder wurden, wie in Abb. 4.2 dargestellt, zwei Sprachen getestet: Ziel war es, ca. 200 Antworten pro Item und Land zu erhalten. Die Testhefte enthielten, je nach Länge und Itemblocks, zwischen 20 und 30 Items (CITO, 2008). Beispielitems sind im Anhang einsehbar.

Design

Bei dem Design der Studie handelte es sich um ein sogenanntes „complete balanced block design". Dieses wird in Tabelle 4.1 dargestellt. Ein solches Design wird häufig im Rahmen großer Large-Scale-Studien eingesetzt. Der Vorteil dieses Vorgehens ist, dass eine größere Menge Items getestet werden kann, ohne dass sich die *Anzahl der Testitems* und somit auch die *Testzeit pro Schüler* erhöht. Das wird umgesetzt, indem nicht jeder Schüler jedes Item beantworten muss. Die Items werden dazu zunächst in sogenannte Itemblocks von etwa gleicher Testzeit eingeteilt (*block design, in Tabelle 4.1 dargestellt von A-F*). Davon werden wiederum jeweils 2 der Blocks zu unterschiedlichen Testheften (hier: 15 Testhefte aus 6 Itemblocks) zusammengesetzt. Die Items kommen in unterschiedlichen Testheften in unterschiedlichen Kombinationen vor. Auf diese Art und Weise ist jedes mögliche Itempaar in einem der Testhefte gemeinsam vorhanden (*complete*), und jedes Item kommt in der gleichen Anzahl von Testheften vor (*balanced*). In diesem Fall ist jeder Itemblock in 5 Testheften vorhanden, damit kommt jedes Item in 5 Testheften vor. Jedes Testheft wird der gleichen Anzahl von Personen vorgegeben. Das heißt, von jedem Item existiert die gleiche Anzahl von Beobachtungen. Das Design hat außerdem den Vorteil, dass mögliche Reihenfolgeeffekte analysiert und ggf. ausgeschlossen werden können. Dieses Design

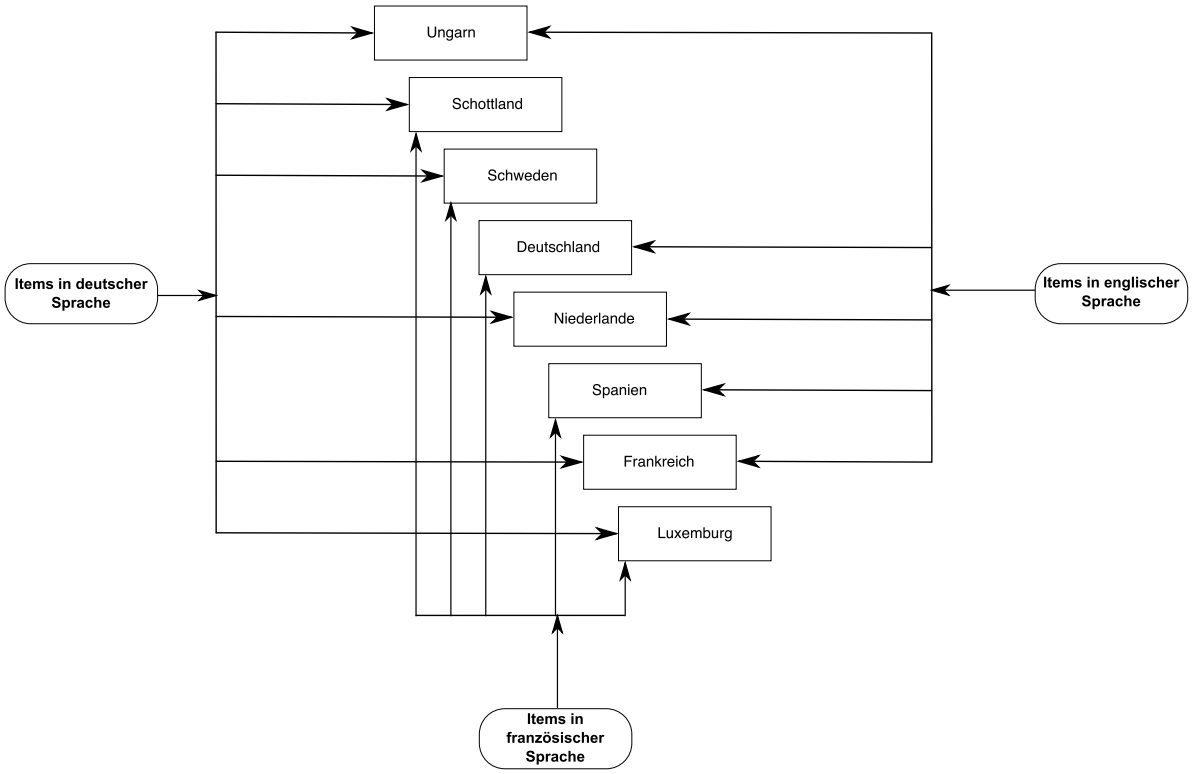

Abbildung 4.2. Getestete Sprachen in den Teilnehmerländern

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.

	Position			
Testheft	Position 1	Position 2	Position 3	Position 4
1	A1	A2	B1	B2
2	A2	A1	C1	C2
3	D1	D2	A1	A2
4	E1	E2	A2	A1
5	A1	A2	F1	F2
6	B1	B2	C1	C2
7	B2	B1	D1	D2
8	E2	E1	B1	B2
9	F1	F2	B2	B1
10	D2	D1	C2	C1
11	C1	C2	E1	E2
12	C2	C1	F2	F1
13	D1	D2	E2	E1
14	F2	F1	D2	D1
15	E1	E2	F1	F2

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

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

Tabelle 4.2. Anzahl der Testhefte pro Sprache pro Land; Anzahl der Beobachtungen pro Item insgesamt (Internationaler EBAFLS-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 Personenfähigkeiten 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, stichprobenunabhängig 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 teilgenommenen 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.
Für den Bereich des Leseverständnisses in deutscher Sprache werden die Schüler-Stichproben aus Frankreich, Ungarn, den Niederlanden und Schweden in dieser Arbeit verwendet. Die schottische Stichprobe war mit nur 249 Schülern zu klein. Auch die luxemburgische Stichprobe wird nicht verwendet. Grund hierfür ist die Tatsache, dass die von Luxemburg eingereichten Items nicht hinsichtlich ihrer Anforderungsmerkmale eingeordnet wurden und daher auch hier keine Hypothesen hinsichtlich der Testkultur aufgestellt werden können.

Alter der Schüler/innen

Insgesamt wurden von 705 deutschen Schülerinnen und Schülern gültige Datensets gesammelt. 667 Schülerinnen und Schüler beantworteten die Frage nach ihrem Alter, davon sind 13 als ungültig (entweder sehr alt oder sehr jung) auszuschließen. In Frankreich beantworteten 1199 Schülerinnen und Schüler die Frage, in Ungarn 512. Wie oben bereits erwähnt, wurde in Spanien aus politischen Gründen kein Schülerfragebogen vorgegeben. Tabelle 4.3 gibt die Altersverteilung nach Geburtsjahrgängen der deutschen, französischen und ungarischen Schüler wieder. Für die Englisch-Stichprobe gilt, dass die deutschen Schülerinnen und Schüler deutlich jünger sind als die aus Frankreich und Ungarn. Während dort der Großteil der Schüler in den Jahren 1988-1990 (Ungarn) bzw. 1988-1989 (Frankreich) geboren wurden, entstammen die deutschen Schüler/innen hauptsächlich den Jahrgängen 1990-1991. Die Schüler aus Deutschland waren somit zum Testzeitpunkt deutlich jünger als die Schüler aus Frankreich und Ungarn.

	Schüler F		Schüler D		Schüler U	
Geburtsjahr	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**
≤ **1984**	1	.1	2	.2	7	13
1985	2	.2	—	—	18	3.4
1886	8	.6	—	—	12	2.3
1987	41	3.3	—	—	59	111
1988	291	23.6	1	.1	135	25.5
1989	808	65.5	25	3.6	133	25.1
1990	47	3.8	336	47.7	120	22.6
1991	2	.2	287	40.8	29	5.5
1992	—	—	4	.6	—	—
1993	—	—	1	.1	—	—
1994	—	—	—	—	—	—
1995	—	—	—	—	—	—
≥ **1996**	3	.2	10	14	2	.4
Keine Angabe	30	2.4	38	5.4	—	—
Fragebögen gesamt	**1233**		**704**		**530**	

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

Für die Deutsch-Stichprobe gilt, dass hier von den französischen Schüler/innen der Großteil im Jahr 1989 geboren wurde. Die Schüler/innen aus Ungarn wurden hauptsächlich 1988 und 1989 geboren, und die niederländischen Schüler/innen entstammen zu einem großen Teil den Jahrgängen 1989 und 1990. Der größte Teil der schwedischen Schülerinnen und Schüler wurde 1988 und 1990 geboren.

	Schüler F		Schüler U		Schüler NL		Schüler SW	
Geburtsjahr	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**
≤ **1984**	—	—	8	13	—	—	—	—
1985	1	.1	8	12	—	—	—	—
1886	2	.1	26	3.9	—	—	—	—
1987	11	.8	99	15.0	10	2.4	8	14
1988	217	15.4	221	33.4	51	12.1	467	78.9
1989	1026	72.9	167	25.3	158	37.5	85	14.4
1990	107	7.6	75	113	184	43.7	3	.5
1991	1	.1	29	4.4	3	.7	—	—
1992	—	—	—	—	—	—	—	—
1993	—	—	—	—	—	—	—	—
1994	1	.1	—	—	—	—	—	—
1995	—	—	—	—	—	—	—	—
≥ **1996**	7	.5	3	.5	2	.5	—	—
Keine Angabe	35	2.5	25	3.8	13	3.1	29	4.9
Fragebögen gesamt	**1408**		**661**		**421**		**592**	

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-

	Schüler F		Schüler D		Schüler U	
Geschlecht	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**
Männlich	493	40	274	38.9	267	50.4
Weiblich	723	58.6	412	58.5	245	46.2
Keine Angabe	17	14	18	2.6	18	3.4
Fragebögen gesamt	**1233**	—	**704**	—	**530**	—

Tabelle 4.5. Geschlecht der französischen (F), deutschen (D) und ungarischen (U) Schülerinnen und Schüler in den nationalen Englisch-Stichproben

	Schüler F		Schüler U		Schüler NL		Schüler SW	
Geschlecht	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**
Männlich	588	418	290	43.9	179	42.5	249	42.1
Weiblich	796	56.5	355	53.7	230	54.6	301	50.8
Keine Angabe	24	17	16	2.4	12	2.9	42	7.1
Fragebögen gesamt	**1408**		**661**		**421**		**592**	

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

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-

	Schüler F		Schüler D		Schüler U	
Herkunftsland	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**
Testland	1140	92.5	657	93.3	493	93
Anderes Land	71	5.7	34	4.9	22	4.2
Keine Angabe	22	18	13	18	15	2.8
Fragebögen gesamt	**1233**		**704**		**530**	

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 F		Schüler U		Schüler NL		Schüler SW	
Herkunftsland	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**	**Anzahl**	**(%)**
Testland	1338	95	626	94.7	381	90.5	526	88.9
Anderes Land	40	2.9	19	2.9	32	7.6	42	7.0
Keine Angabe	30	2.1	16	2.4	8	19	24	4.1
Fragebögen gesamt	**1408**		**661**		**421**		**592**	

Tabelle 4.8. Herkunft der französischen (F), ungarischen (U), niederländischen (NL) und schwedischen (SW) Schülerinnen und Schüler in den nationalen Deutsch-Stichproben

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

Eine große Anzahl von Items zur Messung fremdsprachlichen Leseverständnisses weisen DIF auf (Maris, Bechger & Veldhuijzen, 2006; CITO, 2008), obgleich keine Angaben darüber gemacht werden, wie viele der DIF-Parameter auch tatsächlich signifikant sind. Zur Überprüfung der Modellfits wurde die im Rahmen des OPLM-Modells (Verhelst, Glas & Verstralen, 1995) R1c- Statistik verwendet. Dabei handelt es sich um ein Maß der Abweichung des beobachteten vom erwarteten Wert. Die Leseverständnis-Items wiesen innerhalb der Länder insgesamt OPLM-

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.
Die Einzelanalyse der Items wies ferner darauf hin, dass sich der Gesamt-Modellfit nicht durch das Herausnehmen einzelner Items verbessern lässt, sondern dass eher alle Items einen leichten Misfit aufweisen. Für eine genauere Beschreibung der Analysen und Ergebnisse wird auf Fandel und Kollegen (2007) bzw. den Internationalen Projektbericht (CITO, 2008) verwiesen.
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?"

Zunächst wird eine Überprüfung der Rasch-Modellkonformität innerhalb der einzelnen Länder durchgeführt. Dies ist eine Voraussetzung für die Anwendung des Rasch-Modells innerhalb der Länder zur Berechnung der Itemschwierigkeiten und für deren Verwendung als abhängige Variablen im Rahmen von Fragenkomplex 2. Den ursprünglichen EBAFLS-Analysen liegt formal ein 2-Parameter Item Response Modell zugrunde (OPLM; Verhelst, Glas & Verstralen, 1995). Das bedeutet, dass sich die Item-Charakteristik-Kurven (ICC) sowohl hinsichtlich ihrer Position auf der X-Achse, das heißt hinsichtlich ihrer Schwierigkeit, unterscheiden können, als auch hinsichtlich ihrer Steigung (Trennschärfeparameter; siehe auch 2.1). Für diese Dissertation wurde die Skalierung der Daten erneut in einem einparametrischen-IRT-Modell (Rasch-Modell) durchgeführt, das heißt die ICC unterscheiden sich nur hinsichtlich ihrer Position auf der X-Achse. Die Wahl des Rasch-Modells hat hat mehrere Gründe: Erstens sollte die Aufklärung von DIF (siehe Fragenkomplex 3) in dieser Dissertation auf die Erklärung des Schwierigkeitsparameters begrenzt werden. Zweitens handelt es sich bei dem von CITO verwendeten OPLM-Modell um ein Modell, das kaum über die Forschergruppe, die es entwickelt hat, hinaus Verwendung findet. Es wurde daher entschieden, dieses Modell in dieser Dissertation nicht zu verwenden, sondern auf ein in der Literatur gängigeres und besser erforschtes, nämlich das Rasch-Modell, zurückzugreifen. Der dritte, möglicherweise relevanteste Grund besteht darin, dass der Leseverständnis-Skala des GERS, d.h. den dort verwendeten Deskriptoren, ein 1PL-Rasch-Modell zugrunde liegt (siehe auch 2.2.3). Da wiederum die hier verwendeten Item-Anforderungsmerkmale zum großen Teil auf den Deskriptoren des GERS basieren, schafft die Verwendung des Rasch-Modells daher eine größere theoretische und methodische Nähe zum der Arbeit zugrundeliegenden GERS als es letztlich im ursprünglichen EBAFLS-Projekt durch die Verwendung des OPLM der Fall war. Die Wahl des Rasch-Modells stellt bezüglich der Interpretation der DIF-Parameter eine Vereinfachung dar, die hier jedoch als sinnvoll und zulässig angesehen wird, da die „zentrale Tendenz" der Schüler der unterschiedlichen Länder untersucht werden soll. Eine mögliche Umkehr der Schwierigkeitsabfolge in einem begrenzten Teil des Fähigkeitensprektrums soll an dieser Stelle nicht berücksichtigt werden. Die Grundlagen des Rasch-Modells werden im Folgenden kurz dargelegt.

Diese Formel (Abbildung 4.3) sagt aus, dass die Wahrscheinlichkeit dass eine Person v, ein

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

$p(X_{vi} = 1)$	die Wahrscheinlichkeit, dass Person v Item i korrekt beantwortet
θ_v	Personenparameter der Person v, d.h. die Position der Person hinsichtlich der latenten Fähigkeit
σ_i	Itemschwierigkeit von Item i
$\exp(\theta_v - \sigma_i)$	die Exponentialfunktion der Differenz des Personenparameters und der Itemschwierigkeit

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 ≤ 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-

teten Wert 1 ein 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). Darüberhinaus wird auch aufgrund der Stichprobenabhängigkeit der T-Werte der absolute WMNSQ-Wert in die Interpretation einbezogen. Nach Adams & Khoo (1996) sollte dieser absolute WMNSQ-Wert nicht die Werte 0.75 unter- bzw. 1.33 überschreiten; Werte innerhalb dieses Bereichs werden als akzeptabel angesehen. Zur Schätzung der Itemparameter werden in ConQuest MML-Schätzer (marginal maximum likelihood estimates) verwendet. Die genaue Beschreibung des Schätz-Verfahrens ist bei Wu, Adams und Wilson (1998) nachzulesen. Zur Beantwortung dieser Fragestellung werden alle getesteten Items einer Sprache (Deutsch bzw. Englisch) unabhängig von ihrer Herkunft in jedem der in die Analysen einbezogenen EBAFLS-Teilnehmerländer hinsichtlich ihrer Rasch-Konformität überprüft. Im Folgenden wird eine Beispielsyntax dargestellt: Die Bedeutung des constraints-Befehls wurde

```
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.

Da es sich um binäre Items handelt, existieren hier im Datensatz die Codierungen 1 und 0 für richtig bzw. falsch beantwortete Items. Der „Model Item" Befehl schätzt nun die Itemschwierigkeiten. Für eine ausführlichere Darlegung der ConQuest-Syntax wird auf Wu, Adams und Wilson (1998) verwiesen. Die Analysen erfolgen bezüglich der englischen Sprache separat in den Ländern Deutschland, Frankreich, Spanien und Ungarn. In Bezug auf die deutsche Sprache handelt es sich um die Länder Frankreich, die Niederlande, Schweden und Ungarn, in denen die Rasch-Analysen jeweils seperat durchgeführt werden. In jedem Land werden dazu alle Schüler, die ein gültiges Testheft abgegeben haben, in die Analyse mit einbezogen. Es wird erwartet, dass die Items innerhalb der Länder Rasch-konform sind. Einen Gesamt-Modelltest stellt ConQuest nicht zur Verfügung. Eine Möglichkeit zur Überprüfung der Modellpassung wäre der Vergleich zu einem anderen, z.B. mehrparametrischen Modell; da ein solches hier jedoch aus oben genannten theoretischen Gründen sowie aus Gründen der Interpretierbarkeit der Daten bewusst nicht verwendet werden soll, wird die Modellanpassung lediglich mit Blick auf die einzelnen Items durchgeführt. Sollten Items in mehreren Ländern nicht dem Rasch-Modell entsprechen, werden diese aus der weiteren Analyse ausgeschlossen.

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

Gruppenvariable und konstruiert daraus zwei (beziehungsweise je nach Anzahl der Gruppen auch mehr) sogenannte generalisierte Items („generalised Items"; Wu, Adams & Wilson, 1998, S. 90) pro analysiertem Item. Mit Hilfe dieses Befehls berechnet Conquest die Wahrscheinlichkeit einer korrekten Antwort auf diese Items unter Berücksichtigung eines Item-Haupteffekts, eines Gruppen-Haupteffekts und einer Interaktion zwischen Item und Gruppe. Der Ausdruck „Item" schätzt die Itemschwierigkeit, der Ausdruck„Gruppe" die mittlere Fähigkeit der jeweiligen Gruppen, und der Interaktionsterm „Item*Gruppe" schätzt den Unterschied hinsichtlich der Itemschwierigkeit für die Gruppen. Der DIF-Parameter stellt also eine Schätzung des Unterschiedes der Schwierigkeiten dieser Items (in Logits) für die Gruppen dar.

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).
Die DIF-Analysen erfolgen jeweils zwischen zwei Ländern. Für die getesteten Items werden die Schülerantworten beider Länder gemeinsam skaliert. Ferner werden zusätzlich DIF-Parameter geschätzt, die, wie oben dargestellt, der Differenz der Logit-Werte der generalisierten Items entsprechen. Im Ergebnisteil wird die Anzahl von Items mit signifikanten DIF-Parametern bezüglich der jeweiligen Länder-Analysepaare berichtet. Für die Analysen in den Fragekomplexen zwei und drei werden die DIF-Parameter aller Items mit einbezogen, unabhängig davon, ob diese signifikant sind oder nicht. Dies empfehlen Scheuneman und Gerritz (1990) vor allem dann, wenn ein Interesse an der Analyse von Ursachen für DIF und an Stärken und Schwächen der jeweils betrachteten Gruppen besteht, wie dies in der vorliegenden Arbeit der Fall ist.

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

Die Items wurden bereits im Rahmen der EBAFLS-Studie in ihren Herkunftsländern von Fremdsprachenexperten eingeordnet. Dazu wurde das auf dem GERS basierende Kategoriensystem „Dutch Grid" (Alderson et al., 2006, siehe 2.2.4) verwendet. In den Ländern waren üblicherweise jeweils 2 Experten mit der Einordnung betraut. Zu Beginn des EBAFLS Projekts einigten sich die Teilnehmerländer darauf, dass die Items entweder von einer Person eingeordnet und dies dann von einer zweiten Person überprüft werden sollte, oder dass die beiden Personen die Einordnung aller Items gemeinsam vornehmen. Es handelt sich also bei den Itemmerkmalen üblicherweise um Konsens-Urteile von zumeist zwei Ratern. Bei den Experten handelte es sich zum größten Teil um die Autoren des „Dutch Grid", die bereits gemeinsam eine große Anzahl von Items hinsichtlich ihrer Eigenschaften eingeordnet hatten (Alderson et al., 2006). Es wird hier daher von einer gemeinsamen Schulung und somit einer gemeinsamen Basis der Experten ausgegangen.
Da die Einordnung der Items hinsichtlich ihrer Item-Anforderungsmerkmale bereits im Rahmen der oben dargestellten EBAFLS-Studie geschehen ist, wird im Ergebnisteil dieser Dissertation darauf nicht mehr gesondert eingegangen, sondern auf die Ergebnisse des EBAFLS-Projekts verwiesen (Fandel et al., 2007; CITO, 2008).

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

Diese im Rahmen des EBAFLS-Projekts erfolgte Bewertung der Itemmerkmale wurde für diese Dissertation erneut durch zwei zusätzliche Rater pro Sprache in Deutschland überprüft, indem sämtliche verwendete und eingereichte Items nochmals hinsichtlich ihrer kognitiv-linguistischen Anforderungsmerkmale kategorisiert wurden. Damit sollten die ursprünglichen Ratings validiert werden. Zusätzlich zu den Einzelurteilen fällten die Rater bei Nicht-Übereinstimmung ein Konsens-Urteil.
Die zur Überprüfung der Urteile hinzugezogenen deutschen Experten wurden im Sommer 2008 rekrutiert. Diese Experten hatten bereits im Jahr 2007 im Rahmen der internationalen EBAFLS-Studie an einer 2-tägigen Schulung zur Einordnung von Items in die Niveaus des GERS teilgenommen und waren so bereits mit dem Vorgehen vertraut. Es handelte sich dabei um Studenten der Anglistik und um Lehrer mit den Fächerschwerpunkten Englisch bzw. Deutsch.
Die Validierung der Itemmerkmale im Rahmen dieser Dissertation erfolgte für die Items in englischer und deutscher Sprache. Vor der eigentlichen Validierung erhielten die Experten nochmals Schulungsmaterial zur Vorbereitung. Zu Beginn fand dann eine halbtätige Schulung statt, in der mit Hilfe des im Internet erhältlichen elektronischen Trainingstools des „Dutch Grid" (Alderson et al., 2006) das Einordnen der Items hinsichtlich der kognitiv-linguistischen „Dutch

Grid"-Anforderungsmerkmale geübt wurde. Im Anschluss daran fand die Einordnung der Items statt. Dazu wurden die Experten gebeten, zuerst einzeln die Items einzuordnen. Im Anschluss daran sollten sie ihre Urteile vergleichen und bei Nicht-Übereinstimmung ein Konsens-Urteil fällen, so wie es bei den Experten aus den Herkunftsländern auch geschehen war. Als Maß für die Inter-Rater-Korrelation wird hier der Rangkorrelationskoeffizient nach Spearman (z.B. Bortz, 2005) gewählt, da von einer Ordinalskalierung der Variablen ausgegangen wird. Dabei werden jeweils zwei miteinander gepaarte Urteilswerte in Verbindung zueinander gesetzt.

Schritt 3: Auswahl der schwierigkeitsdeterminierenden Eigenschaften und Prädiktoren für Itemschwierigkeit und Differentielle 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 Differentiellen 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.

Eine Voraussetzung für die Verwendung eingereichter Items in der EBAFLS-Studie war eine im Herkunftsland durchgeführte Überprüfung der Items hinsichtlich Ihrer Güte. So entstammten die aus Deutschland eingereichten Englisch-Items der DESI-Studie (Deutsch-Englisch Schülerleistung International; Beck & Klieme, 2007). Ferner reichte Deutschland Items zur Messung von Deutsch als Fremdsprache in den EBAFLS Itempool ein. Diese entstammten dem TestDAF (Deutsch als Fremdsprache; Bolton, 2000). Auch die Items aus den restlichen Teilnehmerländern waren bereits in nationalen Testverfahren eingesetzt worden. Ferner sollten die Items hinsichtlich ihrer Form repräsentativ für die Testkultur des jeweiligen Landes sein. Diese Tatsache ist zwar in dieser Studie nicht empirisch überprüfbar; jedoch war das Einreichen typischer Items eine Voraussetzung für die Aufnahme der Items im EBAFLS-Itempool.

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 sicherstellen 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:

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

(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

Da zur Beantwortung der Fragestellungen in den Fragenkomplexen zwei und drei im Grundsatz dieselben Methoden und Prädiktoren verwendet werden, werden diese im folgenden Abschnitt zunächst gemeinsam dargelegt. Danach wird noch mal einzeln auf die für die spezifischen Fragestellungen der beiden Fragenbereiche verwendeten Methoden eingegangen.
Nachdem nun oben unter 4.2 die Methoden zur Überprüfung der Voraussetzungen und zur Analyse von Testkulturen dargestellt wurden, werden in diesem Teil *Methoden zur Analyse von Zusammenhängen* zwischen Item-Anforderungsmerkmalen und Itemschwierigkeiten innerhalb der Länder sowie zwischen Testkultur-Indikatoren und DIF dargestellt. Die Beantwortung der unter Abschnitt 3 aufgestellten Fragestellungen und Hypothesen zu den Fragenkomplexen 2 und 3 soll mit Hilfe korrelativer und regressionsanalytischer Methoden (z.B. Bortz, 2005) zur Vorhersage von Itemschwierigkeiten innerhalb der Länder (Fragenkomplex 2) bzw. zur Erklärung Differentieller Item Funktionen (Fragenkomplex 3) erfolgen. Für eine detaillierte Beschreibung korrelations- und regressionsanalytischer Methoden wird auf Bortz (2005) verwiesen.

Trotz der methodischen Ähnlichkeit von Korrelation und multipler Regression werden im Rahmen dieser Arbeit für beide Fragenbereiche beide Methoden angewandt. Grund dafür ist zum einen, dass beide Methoden in den unter 2.1.2 und 2.2.4 beschriebenen empirischen Studien verwendet wurden und darüberhinaus die Verwendung beider Methoden zur Analyse von DIF auch empfohlen wird (Scheuneman & Gerritz, 1990). Ferner existiert bei der multiplen Regression häufig das Problem der Multikollinearität, weshalb einige der Ergebnisse anhand von Einzelkorrelationen möglicherweise besser interpretierbar sind. Die multiple Regression wiederum ist ein unverzichtbares Maß, wenn es das Ziel ist, Hinweise auf die gemeinsame Wirkung der unabhängigen Variablen, das heißt der Anforderungsmerkmale beziehungsweise Testkultur-Indikatoren, auf die Itemschwierigkeit beziehungsweise den dahingehen existierenden Unterschied zu erhalten. Ferner wurden sowohl einschlägige korrelations- als auch regressionsanalytische Auswertungen im Kontext internationaler Large-Scale Studien durchgeführt. Wie im Fall vorliegender Studie wurden dort eindimensionale Item Response Modelle für die Kompetenzskalierung eingesetzt. Sämtliche korrelations- und regressionsanalytischen Berechnungen erfolgen in SPSS (SPSS Inc., 2009).

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?

Die Anwendung von Korrelationsanalysen zur Feststellung des Zusammenhangs zwischen Itemschwierigkeiten und Itemmerkmalen findet sich beispielsweise in den empirischen Studien von Scheuneman und Gerritz (1990) und Klieme und Baumert (2001).

Ziel ist es zu analysieren, inwieweit Korrelationen zwischen den Itemeigenschaften und den Itemschwierigkeiten *innerhalb* der Länder existieren. Es steht die Frage dahinter, inwieweit die verschiedenen Merkmalsausprägungen einen Zusammenhang zur Itemschwierigkeit in den einzelnen Ländern aufweisen. Über die Richtung des Zusammenhangs kann hier keine Aussage getroffen werden, da möglicherweise eine Konfundierung zwischen den Schwierigkeitsstufen der kognitiv-linguistischen Merkmale und der Testkultur stattfindet. So könnte beispielsweise eine negative Korrelation zwischen Items mit schwieriger Grammatik und der Itemschwierigkeit bedeuten, dass Items mit schwieriger Grammatik für die Schüler leichter sind, da sie an solche Items eher gewöhnt sind.

Auf der anderen Seite könnte genauso ein positiver Zusammenhang zwischen einer schwierigen grammatischen Struktur und der Itemschwierigkeit zu finden sein; dies würde dann darauf hindeuten, dass (wie theoretisch erwartet) Items mit einem schwierigeren Anforderungsmerkmal auch tatsächlich schwieriger für die Schüler zu lösen sind. Mit Hilfe dieser Fragestellung soll also zunächst auf die Einzelzusammenhänge zwischen den Item-Anforderungsmerkmalen und den Itemschwierigkeiten innerhalb der Länder eingegangen werden. Im Rahmen von Frage 2 werden dazu innerhalb der Länder die im Kategoriensystem „Dutch Grid" verwendeten, kognitiv-linguistischen Item-Anforderungsmerkmale jeweils mit der Itemschwierigkeit korreliert um festzustellen, ob die verwendeten Itemmerkmale einen Einfluss auf die Itemschwierigkeiten innerhalb der Länder besitzen, und wenn ja, wie groß dieser ist. Dabei wird wie folgt vorgegangen: Die in Frage 1.a erhaltenen Itemschwierigkeiten innerhalb der Länder werden zu Fällen. Diese werden dann mit den verschiedenen Item-Anforderungsmerkmalen korreliert (Pearson Produkt-Moment-Korrelation *r)* und auf ihre Signifikanz hin überprüft. Bei der Itemschwierigkeit handelt es sich um eine stetige, intervallskalierte Variable. Obgleich die Item-Charakteristika dagegen ordinalskaliert vorliegen, wird hier dennoch der Pearson Produkt-Moment-Korrelationskoeffizient verwendet, da hier angenommen wird, dass stetige Merkmale zugrunde liegen. Dies erfolgt für

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):

$$z = \frac{1}{2}\ln\frac{(1+r)}{(1-r)}$$

(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

$$\sigma_{(Z_1 - Z_2)} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

Es wird getestet, ob der so ermittelte z-Wert den kritischen Wert *zkrit* (a=5%) überschreitet. Ist dies der Fall, unterscheiden sich die Korrelationskoeffizienten zweier Grundgesamtheiten signifikant. In Standardlehrbüchern werden Transformationstabellen des Korrelationskoeffizienten r in Fisher-Z-Werte zur Verfügung gestellt (z.B. Bortz, 2005). Cohen (1988) stellt ferner Tabellen zur Überprüfung der Signifikanz des Unterschiedes zur Verfügung. Für eine detailliertere Darlegung des Vorgehens wird auf Bortz (2005) und Cohen (1988) verwiesen. Im Rahmen dieser Arbeit werden die Korrelationen der unterschiedlichen Länder zwischen Itemeigenschaften und Itemmerkmalen jeweils paarweise mit der oben dargestellten Hypothese überprüft. Zum Umgang mit der Inflation des alpha-Fehlers wird auf 4.4 verwiesen.

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 ?

Die Anwendung von regressionsanalytischen Methoden zur Erklärung der Itemschwierigkeit findet sich beispielsweise in den empirischen Studien von Scheuneman und Gerritz (1990) und Hartig, Frey, Nold und Klieme (2010). In letzerem wurden unterschiedliche Methoden hinsichtlich ihrer Eignung zur Modellierung von Itemschwierigkeit im Rahmen von Auswertungen zu DESI-Studie untersucht. Das Ziel dieser Arbeit war, Itemschwierigkeiten mit Hilfe von schwierigkeitsdeterminierenden Itemeigenschaften zu erklären. Dazu wurden drei Methoden hinsichtlich ihrer Eignung verglichen. Dabei handelte es sich um ein Linear Logistisches Testmodell (LLTM; Fischer, 1973), ein LLTM +e welches den Zufallseffekt e mit einbezieht (Janssen, Schepers & Peres, 2004), sowie um eine zweischrittige Prozedur. In Letzterer werden in Schritt 1 die Items zunächst IRT-skaliert. Danach werden in einem zweiten Schritt in einer gängigen Analysesoftware wie SPSS die in Schritt 1 berechneten Parameter zu Fällen bzw. abhängigen Variablen und werden nun per multipler linearer Regression mit Hilfe der Itemeigenschaften als Prädiktoren vorhergesagt.
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 und zwischen den Ländern angewandt.

Hier werden die Item-Anforderungsmerkmale dazu verwendet, innerhalb der einzelnen Länder die Itemschwierigkeiten vorherzusagen. Das dient dem Zweck festzustellen, ob die hier verwendeten Itemeigenschaften des „Dutch Grid"-Kategorisierungssystems dazu geeignet sind, Itemschwierigkeiten zu erklären. Untersucht wird, ob zumindest ein Teil der Varianz der Itemschwierigkeiten innerhalb der Länder durch die hier verwendeten Itemeigenschaften erklärt wird. Anhand dessen soll überprüft werden, ob die verwendeten kognitiv-linguistischen Anforderungsmerkmale zur Schwierigkeit beitragen. Auch die Überprüfung dieser Fragestellung erfolgt explorativ, da a priori keine Aussagen über die Größe und Richtung der Prädiktoren getroffen werden können. Methodisch wird wie folgt vorgegangen: Es werden die in Frage 1.a erhaltenen Itemschwierigkeitsparameter innerhalb der Länder verwendet. Diese werden zu Fällen, die Item-Anforderungsmerkmale zu Variablen. Es wird dann eine Regression der Itemschwierigkeiten (abhängige Variable) auf die Item-Anforderungsmerkmale (unabhängige Variablen) durchgeführt. Es wird das Einschluss-Verfahren gewählt, da es keine Hypothesen hinsichtlich der Größe der Zusammenhänge gibt. Es werden mehrere sukzessive Modelle gerechnet, die zunächst alle und zuletzt nur noch signifikante Prädiktoren in die Analyse mit einschließen. Es wird zum einen analysiert, wie groß der Anteil der durch die signifikanten Prädiktoren aufgeklärten Varianz (R^2) ist.

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.
Ferner behandelt diese Fragestellung auch Zusammenhänge zwischen aufgrund der Testkulturen der Länder erwarteten Stärken und Schwächen und den DIF-Parametern. Die Hypothese ist hier, dass die Anwesenheit eines Item-Anforderungsmerkmals, das bei den „eigenen" eingereichten Items einer Gruppe signifikant häufiger als bei der jeweiligen Referenzgruppe vorkommt, das Item für diese Gruppe im Vergleich mit der Referenzgruppe erleichtern und für die Referenzgruppe erschweren sollte. Negative Korrelationen bedeuten auch hier einen Vorteil für die Fokusgruppe. Die Korrelationen sollten den aufgrund der Analyse der Testitems erwarteten Stärken und Schwächen entsprechen. Die Hypothesen werden aus den unter Punkt 4.2.3 gewonnenen Erkenntnissen hinsichtlich der nationalen Testprofile abgeleitet. Auch hier werden die DIF-Parameter zu Fällen und die Item-Anforderungsmerkmale zu Variablen. Zur Berechnung der Korrelationen wird der Pearson Produkt-Momoent-Korrelationskoeffizient herangezogen.

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 Punkt 4.3.1, nicht die Varianz innerhalb, sondern DIF, d.h die *kulturell bedingte Varianz* der Itemschwierigkeiten *zwischen* den Ländern, mit Hilfe der kognitiv-linguistischen Item-Anforderungsmerkmale erklärt werden. Die Entscheidung, in dieser Arbeit regressionsanalytischen Methoden zur Erklärung von DIF zu verwenden, basiert auf der empirischen Arbeit von Scheuneman & Gerritz (1990), die DIF mit Itemmerkmalen per Regressionsanalyse vorhersagten, und der bereits dargestellten Arbeit von Hartig, Frey, Nold und Klieme (2008).
Auch hier werden wieder paarweise DIF-Parameter gebildet (zur Berechnung der Parameter siehe Frage 1b), die dann als abhängige Variable anhand der Item-Anforderungsmerkmale als Prädiktoren vorhergesagt werden. Die Regressionsanalysen erfolgen auch hier in SPSS, und werden für alle paarweise festgestellten Differentiellen Item Funktionen durchgeführt. Es wird die Hypothese aufgestellt, dass durch die Itemmerkmale ein Teil der Varianz aufgeklärt werden kann (Frage 3b). Ferner sollte die Richtung der Beta-Gewichte den nationalen Testkulturen und den erwarteten Stärken und Schwächen der Gruppen entsprechen (Frage 3c). Im Gegensatz zu den korrelativen Zusammenhängen werden hier ausschließlich die Item-Anforderungsmerkmale in die Analysen mit einbezogen, nicht die Itemherkunft. Dies hat den Grund, dass in einer multiplen Regressionsanalyse vermutlich die Itemherkunft einen großen Anteil gemeinsamer Varianz mit den Item-Anforderungsmerkmalen in den Ländern aufweist.
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, konstruktirrelevante 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.
Auch hier könnte argumentiert werden, dass bei einer Anzahl von 18 Korrelationen eine Korrektur des Alpha-Fehlers notwendig wird. Hier ist jedoch zu betonen, dass es sich bei den Korrelationsanalysen um Analysen von explorativer Natur handelt. Hier werden die einzelnen Korrelationen betrachtet, und jedwede Verringerung des Signifikanz-Niveaus würde dazu führen, dass möglicherweise testkulturelle Effekte unentdeckt bleiben. Aus diesem Grund wurde eine Korrektur des Alpha-Fehlers bewusst nicht durchgeführt.

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 Itemschwierigkeiten" (5.2) und „Erklärung von DIF" (5.3) formulierten Fragestellungen nacheinander dargestellt.

5.1. Ergebnisse Fragenkomplex 1: „Voraussetzungen und Skalierbarkeit"

Dieser Abschnitt der Arbeit wird sich mit den Ergebnissen hinsichtlich der für weitere Analysen notwendigen Voraussetzungen beschäftigen. Dabei wird zunächst auf die Frage nach der Rasch-Konformität der Items innerhalb der Länder eingegangen. Darauf folgt die Darstellung der Ergebnisse der DIF Analysen, darüberhinaus wird der Frage nach der Existenz von Testkulturen und deren Abbildung auf den Items der unterschiedlichen Länder nachgegangen.

5.1.1. Zu Frage 1a

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

Zunächst soll kurz auf die Reliabliltät eingegangen werden. Da es sich bei dem Datensatz aufgrund des dem EBAFLS-Projekt zugrunde liegenden Multi-Matrix-Designs um einen Datensatz handelt, der mehr als 10% Missing-Werte aufweist, werden einige klassische Parameter, wie die Reliabilität der Skala, nicht berechnet. Lediglich die EAP/PV-Reliabilität wird berechnet. Diese ist insgesamt eher als sehr niedrig einzustufen (zwischen .11 und .15, siehe Anhang). Allerdings bezieht sich diese Reliabilität auf den Personen- und nicht auf den Itemparameter. Damit wird erklärt, wie sehr sich die Vorhersage der Personenfähigkeit verbessert, im Gegensatz zu der Annahme, dass keine Itemantworten beobachtet werden. Die EAP-Reliabilität hängt nicht mit dem Item-Fit oder der Validität zusammen, und beschreibt auch nicht die Genauigkeit der Messwerte, sondern ist ein Maß dafür, wie sehr die Vorhersage verbessert wurde (Adams, 2006). Daher ist die Tatsache, dass dieser Wert nur relativ klein ist, für diese Arbeit nicht von großer Bedeutung. Die EAP-Reliabilitäten sind in Anhang gemeinsam mit dem Itemschwierigkeitsparametern und den übrigen Kennwerten dargestellt.

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

Land	Anzahl modellkonformer Items	Nicht modell-konforme Items	MNSQ (weighted) KI	T
Frankreich	119/122	REFR_4_2	0.90 (0.91, 1.09)	-2.2
		REHU_24_1	0.83 (0.90, 1.10)	-3.4
		REHU_25_6	1.09 (0.92, 1.08)	2.0
Deutschland	121/122	REFR_6_6	1.12 (0.89, 1.11)	2.1
Ungarn	121/122	REGE_9_10	1.17 (0.86, 1.14)	2.3
Spanien	119/122	REFR_4_7	0.90 (0.91, 1.09)	-2.1
		REFR_6_3	0,90 (0.91, 1.09)	-2.1
		RESP_18_4	1.11(0.93, 1.07)	2.7

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.

1.33 überschreiten; Werte innerhalb dieses Bereichs werden als akzeptabel angesehen. Dieser Vorschlag wird im Rahmen dieser Arbeit angewandt; es werden daher ausschließlich Items aus den Analysen entfernt, deren WMNSQ-Werte außerhalb der von Adams und Khoo (1996) genannten Werte liegen.

In Tabelle 5.1 zeigt sich ferner, dass es sich bei den möglicherweise vom Modell abweichenden Items in den verschiedenen Ländern um unterschiedliche Items handelt. Ferner gilt für keines der Items, dass alle Fit-Indices problematische Werte aufweisen. Daher werden keine der Items aus weiteren Analysen ausgeschlossen. Insgesamt deuten die Ergebnisse darauf hin, dass die Englisch-Items innerhalb der Länder dem eindimensionalen Rasch-Modell entsprechen.

Land	Anzahl modellkonformer Items	Nicht modell-konforme Items	MNSQ (weighted) KI	T
Frankreich	99/101	RGGE_27_3	1.08 (0.93, 1.07)	2.0
		RGGE_28_3	1.09 (0.92, 1.08)	2.3
Niederlande	101/101			
Schweden	100/101	RGNE_11_22	0.86 (0.87, 1.13)	-2.3
Ungarn	98/101	RGGE _27_4	0.88 (0.88, 1.12)	-2.1
		RGNE_10_3	1.16 (0.88, 1.12)	2.6
		RGNE_10_6	0,84 (0.85, 1.15)	-2.1
		RGSW_25_5	1.16 (0.85, 1.15)	19

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

Wie in Tabelle 5.2 zu erkennen ist, zeigt sich bezüglich der Deutsch-Items ein ähnliches Bild wie schon bei den Englisch-Items: nur einige wenige der 101 getesteten Items weisen als problematisch einzustufende T-Werte auf. Auch hier liegt keiner der WMNSQ-Werte außerhalb der von Adams und Khoo (1996) genannten Grenzen. Es wird daher keines der Items bezüglich der späteren Analysen ausgeschlossen. Ferner lässt sich auch hier beobachten, dass es sich bei den problematischen Items immer um unterschiedliche Items handelt; es lässt sich diesbezüglich über die Länder hinweg kein systematisches Muster feststellen. Auch hier deuten die Ergebnisse darauf hin, dass die Deutsch-Items innerhalb der Länder insgesamt dem Rasch-Modell entsprechen.

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

Die in Hypothese 1a getroffene Annahme kann insgesamt beibehalten werden. Nur wenige Items weisen problematische T-Werte auf, keines der Items problematische WMNSQ-Werte, legt man die empfohlenen Grenzen von Adams & Khoo (1996) zugrunde. Innerhalb der Länder kann daher insgesamt, sowohl bezüglich der deutschsprachigen als auch der englischsprachigen Lesekompetenz, von einer Eindimensionalität der mit den Items erfassten latenten Fähigkeit ausgegangen werden.

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

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

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.

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

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

Paarweise DIF-Analyse zwischen	Anzahl Items mit DIF (sig.)
Frankreich-Ungarn	50/101 (49.5%)
Frankreich-Niederlande	58/101 (57.4%)
Frankreich-Schweden	60/101 (59.4%)
Niederlande-Schweden	48/101 (47.5%)
Niederlande-Ungarn	48/101 (47.5%)
Schweden-Ungarn	40/101 (39.6%)

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.
In Schritt zwei wurden die Ratings aus Schritt 1, das heißt die Einordnung der Items hinsichtlich ihrer kognitiv-linguistischen Itemeigenschaften durch die Experten hinsichtlich deren Übereinstimmung mit zusätzlichen Ratern nochmals überprüft. Darüber hinaus wurde eine Entscheidung bezüglich der für die weiteren Analysen zu verwendenden Ratings getroffen. Die Überprüfung erfolgte durch zwei zusätzliche Rater, die jeweils zunächst Einzel- und danach Konsensurteile abgaben. Die Ergebnisse werden im Folgenden zusammenfassend berichtet, alle Inter-Rater-Übereinstimmungen sowohl zu den Original-Ratings aus den Herkunftsländern als auch bezüglich der deutschen Rater sind im Anhang einzusehen. Als Maß für die Inter-Rater-Korrelation wurde hier der Rangkorrelationskoeffizient nach Spearman (Bortz, 2005) gewählt, da von einer Ordinalskalierung der Variablen ausgegangen wird. Dabei werden jeweils zwei miteinander gepaarte Urteilswerte in Verbindung zueinander gesetzt. Insgesamt zeigte sich, dass die Zusammenhänge zwischen den Einordnungen der deutschen und der internationalen Experten, je nach eingschätzem Aufgabenmerkmal, tendenziell eher niedrige Werte aufwiesen (zwischen r = .11 und r = 1.0).
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 rekrutierten 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.

Aufgrund dieser teilweise unbefriedigenden Ergebnisse musste für die vorliegende Dissertation jedoch eine Entscheidung dahingehend getroffen werden, welchen Experten mehr Vertrauen bezüglich der Richtigkeit ihrer Urteile entgegengebracht werden kann. Für die Verwendung der ursprünglichen, aus den verschiedenen Ländern stammenden Experten sprechen drei starke Argumente: Erstens ist davon auszugehen, dass diese Experten die Items aus dem eigenen Land besser beurteilen und einschätzen können als dies ein Experte aus einem anderen Land könnte. Zweitens handelt es sich bei diesen Experten zu einem großen Teil um diejenigen Personen, die den „Dutch Grid", also das Kategoriensystem, auf dem die Urteile basierten, konstruiert haben. Daher ist davon auszugehen, dass diese Personen das System ausreichend kennen und somit auf dessen Basis Items besser beurteilen können, als es bei den deutschen Beurteilern nach einer kürzeren Schulung und mit weniger Erfahrung vermutlich der Fall ist. Drittens haben die Autoren lange Zeit gemeinsam an der Einordnung von Items gearbeitet. Es ist daher davon auszugehen, dass sie in ihren Urteilen insgesamt konsistenter sind, als es bei den deutschen Zusatzratern der Fall ist. Aus diesen Gründen wurden den späteren Analysen die Urteile dieser Experten zugrunde gelegt.

Schritt 3: Auswahl der schwierigkeitsdeterminierenden Eigenschaften und Prädiktoren für Itemschwierigkeit und Differentielle Item Funktionen. Nach Sichtung der unter 2.2.4 beschriebenen Literatur sowie weiterer empirischer Studien aus dem Bereich der Fremdsprachenforschung empfehlen sich Item-Anforderungsmerkmale, die kognitiv-linguistische Prozesse abbilden und somit schwierigkeitsdeterminierend sein sollten. Diese werden in den Analysen zu Fragenkomplex 2 als Prädiktoren für Itemschwierigkeit und DIF eingesetzt. Die Anforderungsmerkmale, deren Ausprägungsstufen sowie die Terminologie werden dem „Dutch Grid" (Alderson et al., 2006) entnommen. Für die Analysen dieser Arbeit wurden die Itemeigenschaften der kognitiv-linguistischen Kategorie des Dutch Grid ausgewählt. Die verwendeten Itemeigenschaften wurden unter 2.2.4 im Rahmen der Darstellung des „Dutch Grid" bereits dargestellt, daher wird an dieser Stelle darauf verwiesen.

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.

Die folgenden Ausführungen stellen die Ergebnisse der Analyse der Items, mit anderen Worten die in dieser Arbeit verwendete Operationalisierung von Testkultur, dar. Es werden jeweils die länderspezifischen Ausprägungen auf den kognitiv-linguistischen Anforderungsmerkmalen der Items berichtet. Je häufiger eine Ausprägung bei den Items eines Landes vorkommt, desto eher sollte diese Ausprägung relevant für die jeweilige Testkultur sein. Für jedes Itemmerkmal werden die jeweiligen Prozentanteile der verschiedenen Merkmalsausprägungen pro Land berichtet.

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.

	Multiple Choice	Multiple Matching	Richtig-Falsch	Zitieren	Ordnen	Lückentext	Kurz antwort
Frankreich	62.3	15.1	0	9.4	113	0	0
Deutschland	100	0	0	0	0	0	0
Spanien	100	0	0	0	0	0	0
Ungarn	0	43.33	23.3	0	0	33.33	0

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,

	Informationssgewinn 1		Informationsgewinn 2		Authentizität Text	
	Erkennen	*Schlussfolgern*	*Explizit*	*Implizit*	*angepasst/ vereinfacht*	*authentisch*
Frankreich	88.68	1132	100	0	20.8	79.2
Deutschland	63.0	37.0	50	50	100	0
Spanien	62.5	37.5	62.5	37.5	96.9	3.1
Ungarn	66.7	33.3	100	0	0	100

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

	ausschließlich konkret	hauptsächlich konkret	teilweise abstrakt
Frankreich	0	100	0
Deutschland	0	100	0
Spanien	90.6	9.4	0
Ungarn	16.7	66.7	26.7

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

	ausschließlich häufig/einfach	hauptsächlich häufig/einfach	teilweise erweitert/selten
Frankreich	45.3	34.0	20.8
Deutschland	0	78.3	21.7
Spanien	15.6	46.9	46.6
Ungarn	16.7	23.3	60

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

häufigsten ungarische Items eingeordnet, am zweit häufigsten spanische Items mit 46.6%, darauf folgen deutsche und französische Items.

	ausschließlich einfach	hauptsächlich einfach	teilweise komplex
Frankreich	13.2	28.3	58.5
Deutschland	0	0	100
Spanien	0	75.0	25.0
Ungarn	16.7	26.7	56.7

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

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" fallen 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

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

Zusätzlich zu der deskriptiven Beschreibung der Unterschiede stellt sich auch noch die Frage nach der Deutlichkeit und Relevanz der gefundenen Unterschiede. Dazu wurde im nächsten Schritt Cohen's *h* (Cohen, 1988, siehe auch 4.2.3), eine Effektgröße zur Überprüfung von Unterschieden bei Prozentwerten, berechnet. Zusätzlich zu den Effektstärken wurden darauf basierende Signifikanztests durchgeführt. Die Ergebnisse hinsichtlich der einzelnen Signifikanztests und Effektstärken werden aufgrund der großen Anzahl von Tests und aus Platzgründen im Anhang dargestellt. Alle im Folgenden berichteten und für Hypothesen über die unterschiedlichen Stärken und Schwächen der Länder verwendeten Unterschiede (bezüglich der Ausprägungen der Itemeigenschaften) sind signifikant.
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

Itemeigenschaft	Ausprägung	Rangreihe der Häufigkeiten
Itemtyp	Multiple Choice	F>U=SP>D
	Multiple Matching	D>SP=F>U
	Richtig-Falsch	U=SP>F=D
	Zitieren	F>U=D=SP
	Ordnen	F>U=D=SP
	Lückentext	U>F=D=SP
	Kurzantwort	
Informationsgewinn 1	Schlussfolgern/Erkennen	SP=D=U>F
Informationsgewinn 2	Implizit/Explizit	D=SP>F=U
Authentizität des Texts	Text authentisch/angepasst= vereinfacht	U>F>D=SP
Abstraktheit des Inhalts	Ausschließlich konkret	SP>U>D=F
	Hauptsächlich konkret	F=D>U>Sp
	Ziemlich abstrakt	U>D=F=SP
Vokabular	Ausschließlich häufig/einfach	F>U=Sp>D
	Hauptsächlich häufig/einfach	D>Sp=F>U
	Teilweise erweitert/selten	U=Sp>F=D
Grammatik	Ausschließlich einfache Strukturen	U=F>D=Sp
	Hauptsächlich einfache Strukturen	Sp>F=U>D
	Teilweise komplexe Strukturen	D>F=U>Sp

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.

	F-D	F-U	F-Sp	D-U	D-Sp	U-Sp
Grammatik ausschl. einfach	F > D	F = U	F > Sp	D > U	D = Sp	U > Sp
Grammatik haupts. einfach	F > D	F = U	F < Sp	D < U	D < Sp	U < Sp
Grammatik teilw. komplex	F < D	F = U	F > Sp	D > U	D > Sp	U > Sp
Vokabular ausschl. häufig	F > D	F > U	F > Sp	D < U	D < Sp	U = Sp
Vokabular haupts. häufig	F < D	F > U	F = Sp	D > U	D > Sp	U < Sp
Vokabular teilw. erweitert / selten	F = D	F < U	F < Sp	D < U	D < Sp	U = Sp
Informationsgewinn: Schlussfolgern	F < D	F < U	F < Sp	D = U	D = Sp	U = Sp
Informationsgewinn: implizit	F < D	F = U	F < Sp	D > U	D = Sp	U < Sp
Inhalt ausschl.konkret	F = D	F < U	F < Sp	D < U	D < Sp	U < Sp
Inhalt haupts. konkret	F= D	F > U	F > Sp	D > U	D > Sp	U > Sp
Inhalt meist abstrakt	F = D	F < U	F = Sp	D < U	D = Sp	U > Sp
Inhalt authentisch	F > D	F > U	F > Sp	D < U	D = Sp	U > Sp
Itemtyp Multiple Choice	F > D	F > U	F > Sp	D < U	D < Sp	U < Sp
Multiple Matching	F < D	F > U	F > Sp	D > U	D > Sp	U < Sp
Ordnen	F > D	F > U	F > Sp	D = U	D = Sp	U = Sp
Zitieren	F > D	F > U	F > Sp	D = U	D = Sp	U = Sp
Lückentext	F = D	F < U	F = Sp	D < U	D = Sp	U > Sp
Richtig - Falsch	F = D	F < U	F < Sp	D < U	D < Sp	U = Sp

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.

	Multiple Choice	Banked Multiple Choice	Richtig-Falsch	Multiple Matching	Zitieren	Kurzantwort	Lücken-text	Informations-transfer
Frankreich	86.8	0	0	0	13.2	0	0	0
Ungarn	0	0	26.5	14.7	0	0	0	58
Niederlande	45.2	0	0	0	0	38.1	16.7	0
Schweden	0	34.1	0	29.3	0	31.7	4.9	0

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.

	Informationsgewinn 1		Informationsgewinn 2		Informationsgewinn 3		Authentizität Text	
	Erkennen	Schlussfolgern	Explizit	Implizit	Hauptidee	Detail	angepasst/ vereinfacht	authentisch
Frankreich	76.3	23.7	100	0	10.5	89.5	0	100
Ungarn	41.2	0	100	0	100	0	0	100
Niederlande	54.8	45.2	100	0	19.0	81	14.3	85.7
Schweden	22.0	78.0	22	78	78	22	70.7	29.3

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.

	ausschließlich konkret	hauptsächlich konkret	teilweise abstrakt
Frankreich	73.7	26.3	0
Ungarn	14.7	58.8	26.5
Niederlande	35.7	64.3	0
Schweden	0	82.9	17.1

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.

	ausschließlich häufig/einfach	hauptsächlich häufig/einfach	teilweise erweitert/selten	erweitert/selten
Frankreich	0	68.4	31.6	0
Ungarn	0	14.7	58.8	26.5
Niederlande	0	4.8	95.2	0
Schweden	0	0	9.8	90.2

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.

	ausschließlich einfach	hauptsächlich einfach	teilweise komplex	komplex
Frankreich	97.4	0	0	2.6
Ungarn	14.7	0	58.8	26.5
Niederlande	71.4	28.6	0	0
Schweden	0	0	73.2	26.8

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).

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

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).

	FR-NL	FR-SW	FR-HU	NL-SW	HU-NL	SW-HU
Itemtyp Multiple Choice	F>NL	F>SW	F>U	NL>SW	U<NL	SW=U
Banked Multiple Choice	F=NL	F<SW	F=U	NLSW	U=NL	SW>U
Multiple Matching	F=NL	F<SW	F<U	NL<SW	U>NL	SW>U
Multiple Choice	F=NL	F=SW	F<U	NL=SW	U>NL	SW<U
Zitieren	F>NL	F>SW	F>U	NL=SW	U=NL	SW=U
Kurzantwort	F<NL	F<SW	F=U	NL>SW	U<NL	SW>U
Lückentext	F<NL	F<SW	F=U	NL>SW	U<NL	SW>U
Informationsgewinn: Schlussfolgern	F<NL	F<SW	F>U	SW>NL	U<NL	SW>U
Informationsgewinn: implizit	F>NL	F>SW	F=U	NL>SW	U>NL	SW<U
Informationsgewinn: Hauptidee	F<NL	F<SW	F<U	NL<SW	U>NL	SW<U
Inhalt authentisch	F>NL	F>SW	F=U	NL>SW	U>NL	SW<U
Inhalt ausschl.konkret	F>NL	F>SW	F>U	NL>SW	U<NL	SW<U
Inhalt haupts. konkret	F<NL	F<SW	U>F	NL<SW	U<NL	SW>U
Inhalt meist abstrakt	F=NL	F<SW	F<U	NL<SW	U>NL	SW<U
Vokabular ausschl. häufig/einfach	F=NL	F=SW	F=U	NL=SW	U=NL	SW=U
Vokabular meist häufig/einfach	F>NL	F>SW	F>U	NL>SW	U>NL	SW<U
Vokabular teilw. erweitert/selten	F<NL	F>SW	F<U	NL>SW	U<NL	SW<U
Vokabular erweitert/selten	F=NL	F<SW	F<U	NL<SW	U>NL	SW<U
Grammatik ausschl. einfach	F>NL	F>SW	F>U	NL>SW	U<NL	SW<U
Grammatik meist einfach	F<NL	F=SW	F=U	NL>SW	U<NL	SW=U
Grammatik teilw. komplex	F=NL	F<SW	F<U	NL<SW	U>NL	SW>U
Grammatik komplex	F=NL	F<SW	F<U	NL<SW	U>NL	SW=U

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?

Diese Frage hat zum Ziel zu analysieren, inwieweit Korrelationen zwischen den Itemeigenschaften und den Itemschwierigkeiten *innerhalb* der Länder existieren. Es steht die Frage dahinter, inwieweit die einzelnen Item-Anforderungsmerkmale überhaupt einen Zusammenhang zur Itemschwierigkeit in den einzelnen Ländern aufweisen. Über die Richtung des Zusammenhanges kann hier keine Aussage getroffen werden, da möglicherweise eine Konfundierung zwischen den theoretisch angenommenen Schwierigkeitsstufen der kognitiv-linguistischen Merkmale und der Testkultur besteht. So könnte beispielsweise eine negative Korrelation zwischen Items mit schwieriger Grammatik und der Itemschwierigkeit bedeuten, dass Items mit komplexen grammatischen Strukturen für die Schüler leichter sind, da sie solchen Items in der entsprechenden Testkultur häufiger ausgesetzt sind. Auf der anderen Seite könnte genauso ein positiver Zusammenhang zwischen einer komplexen grammatischen Struktur und der Itemschwierigkeit zu finden sein; dies würde dann darauf hindeuten, dass Items mit einem schwierigeren Anforderungsmerkmal auch tatsächlich schwieriger für die Schüler zu lösen sind.
Insgesamt wird hier davon ausgegangen, dass signifikante Korrelationen zwischen den Itemschwierigkeiten innerhalb der Länder und den kognitiv-linguistischen Itemeigenschaften 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-

Itemeigenschaft		Itemschwierigkeit			
		Frankreich	**Deutschland**	**Ungarn**	**Spanien**
Itemtyp	Multiple Choice	**-.241****	**-.300****	**-.229***	**-.213***
	Multiple Matching	.02	.004	.02	.043
	Ordnen	-.077	-.073	-.021	-.057
	Zitieren	.023	**.229***	**.238****	-.052
	Lückentext	**.431****	**.406****	**.213***	**.392****
Informationsgewinn 1	Schlussfolgern/Erkennen	**.212***	.165	.105	.141
Informationsgewinn 2	Implizit (vs. Explizit)	-.082	-.137	-.042	-.080
Informationsgewinn 3	Detail (vs. Hauptidee)	Durch Mehrfachnennungen in Englisch nicht auswertbar (*'Detail'* & *'main':* obgleich gegensätzlich, konnte beides angewählt werden)			
Authentizität	Authentisch (vs. angepasst/ vereinfacht)	.106	**.248****	.132	.112
Abstraktheit des Inhalts	ausschließlich konkret	-.136	**-.224***	**-.192***	**-.188***
	hauptsächlich konkret	-.134	-.038	.063	-.063
	teilweise abstrakt	**.431****	**.406****	**.213***	**.392****
Vokabular	ausschießlich häufig/einfach	**-.264****	**-.236****	**-.300****	**-.221***
	hauptsächlich häufig/einfach	.046	-.025	.138	.028
	teilweise erweitert/selten	**.199***	**.254****	.128	**.179***
	erweitert/selten	Items mit dieser Ausprägung kommen bei den Englisch-Items nicht vor			
Grammatik	ausschließlich einfache Strukturen	**.188***	**.184***	.137	**.184***
	hauptsächlich einfache Strukturen	-.168	-.149	-.127	**-.260****
	teilweise komplexe Strukturen	.039	.025	.032	.125
	komplexe Strukturen	Items mit dieser Ausprägung kommen bei den Englisch-Items nicht vor			

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

gen 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

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

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

Tabelle 5.20. Korrelationen zwischen kognitiv-linguistischen Anforderungsmerkmalen und Itemschwierigkeiten innerhalb der Länder bei Deutsch-Items

geringeren Itemschwierigkeit auf. 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.

Es fällt außerdem sowohl bei den Englisch- als auch bei den Deutsch-Items auf, dass gerade bei den drei- oder vierfach gestuften Kategorien meist eine der leichteren und eine der schwierigeren Kategorien signifikante Zusammenhänge aufweisen, während der Zusammenhang zwischen der Itemschwierigkeit und den anderen Abstufungen des jeweiligen Merkmals nahe Null liegt. Auf mögliche Gründe wird im Diskussionsteil (Abschnitt 6) eingegangen.

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.

Diese Hypothese kann insgesamt beibehalten werden. Obgleich nicht in allen Ländern die Korrelationen die gleiche Höhe aufweisen, bzw. in einigen wenigen Fällen in eine andere Richtung deuten oder aber auch signifikante Korrelationen zu unterschiedlichen Item-Anforderungsmerkmalen zu finden sind, zeigt sich doch insgesamt, dass alle Item-Anforderungsmerkmale in einem oder mehreren Ländern und in einer oder beiden Sprachen einen Zusammenhang mit der Itemschwierigkeit aufweisen. Daraus wird geschlossen, dass die in dieser Dissertation gewählten Item-Anforderungsmerkmale auch für weitere Analysen und für die Erklärung von Unterschieden hinsichtlich der Itemschwierigkeit herangezogen werden können. Ferner weisen die Korrelationen zwischen den Itemschwierigkeiten und den schwierigkeitsbestimmenden Merkmalen insgesamt darauf hin, dass das Kategorisierungsinstrument „Dutch Grid" in den unterschiedlichen Ländern in ähnlichem Maße zur Einordnung von Items geeignet zu sein scheint. Dem wird im Rahmen von Frage 2b nachgegangen.

5.2.2. Zu Frage 2b

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

Englisch-Items

Im Folgenden werden die Ergebnisse der Korrelationsvergleiche der unterschiedlichen Länder in den Korrelationen zwischen Itemschwierigkeit und den schwierigkeitsbestimmenden Itemmerkmalen dargestellt. Dies hat zum Ziel, zu einer Einschätzung dahingehend zu gelangen, ob das Item-Kategorisierungsinstrument „Dutch Grid" in allen Ländern ähnlich gut geeignet ist, d.h. ob die Itemmerkmale innerhalb der unterschiedlichen Länder einen vergleichbaren Zusammenhang mit der Itemschwierigkeit aufweisen.

Dazu wurden die Korrelationen zwischen Itemschwierigkeit und den verschiedenen Itemmerkmalen zunächst jeweils Fisher-z-transformiert. Dann wurden jeweils paarweise die Korrelationen zweier Länder mit dem gleichen Itemmerkmal mit Hilfe eines Signifikanztests daraufhin überprüft, ob sich diese Korrelationen signifikant voneinander unterscheiden. In Tabelle 5.21 werden die Ergebnisse der paarweise durchgeführten Korrelationsvergleiche dargestellt. Dabei ist in der ersten Spalte jeweils die Prüfgröße z abgetragen, in der zweiten Spalte p (z).

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.

Itemeigenschaft		Frankreich- Deutschland		Frankreich-Ungarn		Frankreich- Spanien		Deutschland- Spanien		Deutschland-Ungarn		Spanien-Ungarn	
		z	p (z)	z	p (z)	z	p (z)	z	p (z)	z	p (z)	z	p (z)
Itemtyp	Multiple Choice	.49	.623	.1	.922	.68	.82	.57	.566	.59	.556	.13	.897
	Multiple Matching	.12	.902	.0 1		.18	.859	.3	.763	.12	.902	.18	.859
	Ordnen	.03	.975	.43	66.5	.15	.877	.12	.901	.4	.866	.28	.819
	Zitieren	162	.105	1069	.09	.58	.563	**2.2****	.028	..07	.941	**2.27****	.023
	Lückentext	.23	.815	189	.059	.36	.717	.13	.898	165	.098	.153	.127
Informationsgewinn 1	Schlussfolgern / Erkennen	.38	.707	.85	.397	.57	.566	.19	.850	.47	.637	.28	.778
Informationsgewinn 2	Implizit (vs. Explizit)	.44	.657	.31	.757	.0	.988	.45	.656	.74	.46	.29	.656
Authentizität	Authentisch (vs. angepasst / vereinfacht)	113	.257	.2	.839	.05	.963	109	.277	.93	.35	.16	.876
Abstraktheit des Inhalts	ausschließlich konkret	.7	.483	.44	.657	.41	.68	.29	.772	.26	.79	.03	.974
	hauptsächlich konkret	.75	.455	153	.127	.75	.58	.19	.847	.78	.435	.97	.33
	teilweise abstrakt	.23	.815	**189***	.06	.36	.717	.13	.898	**165***	.1	153	.127
Vokabular	ausschießlich häufig / einfach	.23	.818	.3	.763	.35	.724	.12	.903	.53	.595	.65	.513
	hauptsächlich häufig / einfach	.55	.584	.72	.474	.14	.889	.41	.683	126	.206	.86	.392
	teilweise erweitert / selten	.45	.655	.56	.573	.16	.873	.61	.544	101	.312	.40	.687
Grammatik	ausschließlich einfache Strukturen	.03	.975	.4	.686	.03	.975	.0	10	.37	.71	.37	.710
	hauptsächlich einfache Strukturen	.15	.88	.32	.746	.74	.457	.89	.371	.17	.863	107	.286
	teilweise komplexe Strukturen	.11	.904	.05	.957	.67	.504	.78	.438	.05	.957	.72	.47

N=104; * $p \leq 0.1, \alpha = 10\%$; ** $p \leq 0.05, \alpha = 5\%$; *** $p \leq 0.01, \alpha = 1\%$

Tabelle 5.21. Überprüfung der Signifikanz der Unterschiedlichkeit von Korrelationen zwischen kognitiv-linguistischen Anforderungsmerkmalen und Itemschwierigkeiten innerhalb der Länder bei Englisch-Items

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).

Insgesamt zeigt sich bezüglich der Korrelationen, dass diese zwischen den verschiedenen Ländern hinsichtlich Größe und Richtung größtenteils vergleichbar zu sein scheinen. Das spricht dafür, dass die unterschiedlichen schwierigkeitsbestimmenden Merkmale zumindest innerhalb der verschiedenen Länder eine ähnliche Rolle zu spielen scheinen. Wie unter 2a bereits berichtet, entspricht die Richtung der Korrelationen, falls signifikant, bei den Englisch-Items in 20 von 20 Fällen der Richtung, die aufgrund der im „Dutch Grid" vorgenommenen Schwierigkeitsabstufungen der unterschiedlichen Itemmerkmale anzunehmen war. Bei den Deutsch-Items trifft dies auf 35 von 37 Korrelationen zu. Die Korrelationen mit den Itemtypen wurden dabei nicht mitgezählt, da es hier keine eindeutigen Hypothesen darüber gibt, welches der Itemformate schwieriger zu bearbeiten ist. Der im Rahmen dieser Fragestellung vorgenommene Vergleich der Korrelationen bestätigt, dass die im „Dutch Grid" vorgenommenen Abstufungen in den unterschiedlichen Ländern eine ähnliche Gültigkeit zu besitzen scheinen. Dies unterstützt zumindest teilweise die dort vorgenommenen Merkmalsabstufungen und somit auch die Güte des Instruments „Dutch Grid".

Itemeigenschaft		Frankreich-Niederlande		Frankreich-Ungarn		Frankreich-Schweden		Niederlande-Ungarn		Niederlande-Schweden		Ungarn-Schweden	
		z	p(z)	z	p(z)	z	p(z)	z	p(z)	z	p(z)	z	p(z)
Itemtyp	Multiple Choice	.33	.742	145	.146	103	.303	178	**.074 ***	.174	136	.43	.671
	Banked Multiple Choice	.19	.853	.39	.70	.32	.751	155	.121	.895	.13	.7	.482
	Multiple Choice	.61	.54	.24	.811	.2	.843	.85	.394	.418	.81	.04	.967
	Multiple Matching	.05	.962	186	**.063 ***	117	.243	1.9	**.057 ***	.224	122	.69	.492
	Kurzantwort	.49	.627	.43	.671	.2	.854	.91	.362	.772	.29	.62	.535
	Lückentext	.31	.757	.4	.688	.34	.734	.71	.478	.517	.65	.06	.951
Informationsgewinn 1	Schlussfolgern/Erkennen	.36	.722	.18	.860	.67	.502	.18	.858	.304	103	.85	.396
Informationsgewinn 2	Implizit(vs. Explizit)	.07	.947	.99	.322	147	.142	.92	.356	.162	14	.48	.634
Informationsgewinn 3	Detail (vs. Hauptidee)	.31	.189	.84	.40	.66	.511	**2.16****	.031	**.049****	197	.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	1.43	.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	ausschießlich 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/selten	.68	.497	.59	.557	.36	.716	.23	.817	.994	.01	.22	.823
	erweitert/selten	.27	.784	.74	.459	157	.117	.47	.641	.197	129	.82	.409
Grammatik	ausschließlich einfache Strukturen	.36	.719	124	.215	174	.081	.88	.380	.167	138	.5	.614
	hauptsächlich einfache Strukturen	.71	.479	.60	.548	.47	.63	.11	.914	.810	.24	.13	.895
	teilweise komplexe Strukturen	.39	.694	116	.245	.97	.333	156	.12	.173	.36	.19	.847
	komplexe Strukturen	.01	.993	.64	.522	.54	.587	.63	.528	.582	.55	1.18	.237

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

Tabelle 5.22. Überprüfung der Signifikanz der Unterschiedlichkeit von Korrelationen zwischen kognitiv-linguistischen Anforderungsmerkmalen und Itemschwierigkeiten innerhalb der Länder bei Deutsch-Items

5.2.3. Zu Fragen 2c und 2d

Frage 2c: Weisen die kognitiv-linguistischen Anforderungsmerkmale der Items regressionsanalytische Zusammenhänge mit den Itemschwierigkeiten innerhalb der Länder auf? Frage 2d: Wie 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 Itemschwierigkeiten vorherzusagen. Dies diente dem Zweck zu überprüfen, inwieweit die hier verwendeten Itemeigenschaften des Dutch-Grid-Kategorisierungssystems dazu geeignet sind, die Varianz der Itemschwierigkeiten 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 Itemschwierigkeiten innerhalb der Länder stellten jeweils die abhängige Variable dar.

Es ist zu beachten, dass die theoretisch einfachste Abstufung der Prädiktoren, das heißt beispielsweise im Falle von „grammatischen Strukturen" die Ausprägung „ausschließlich einfache Strukturen", jeweils als Kontrastvariable verwendet worden ist. Die Beta-Gewichte müssen daher immer im Verhältnis dazu interpretiert werden. Es werden für jedes Land mindestens zwei aufeinander folgende Modelle dargestellt. Modell 1 entspricht dem Modell mit allen Itemeigenschaften als Prädiktoren. Die Prädiktoren mit einem signifikanten Beta-Gewicht werden in ein zweites Modell übernommen. Dabei wurden drei Signifikanzniveaus festgelegt, nämlich 1%, 5% und 10%. Dies soll sicherstellen, dass auch kleine Zusammenhänge entdeckt werden, da auch diese im Rahmen der Arbeit als relevant erachtet werden. Die Betrachtung der einzelnen Prädiktoren bezieht sich auf die Beantwortung von Frage 2d.

Das zweite Modell zeigt dann letztendlich jeweils den Anteil der Varianz der Itemschwierigkeit 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 Itemschwierigkeit auf die Item-Anforderungsmerkmale in den einzelnen Län-

dern werden zunächst für die Englisch-, und danach für die Deutsch-Items dargestellt.

Ergebnisse für die Englisch-Items

	Modell 1 β (sig)	Modell 2 β (sig)
Schlussfolgern/Erkennen (Info 1)	.180 (.203)	
Implizit/explizit(Info 2)	-.116 (.281)	
Authentischer Text	.102 (.487)	
Inhalt hauptsächlich konkret	**.230 (.040)****	**.245 (.010)*****
Inhalt teilweise abstrakt	**.401(.024)****	**.508 (.000)*****
Vokabular hauptsächlich häufig/einfach	**.254 (.060) ***	**.200 (.066) ***
Vokabular teilweise ausgeweitet/selten	**.381(.003)*****	**.275 (.011)****
Haupts. einfache grammatikalische Strukturen	**-.477 (.002)*****	**-.408 (.003)*****
Teilweise komplexe grammatische Strutkuren	**-.602 (.001)****	**-.564 (.000)*****
Itemtyp Multiple Choice	.196 (.142)	
Itemtyp Ordnen	-.076 (.516)	
Itemtyp Zitieren	**.295 (.002)*****	**.256 (.001)*****
R^2	.387	.367

Abhängige Variable: Itemschwierigkeit Deutschland; Methode: Einschluß; Zelleninhalt: Standardisierte β-Gewichte; * p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 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 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/explizit".

	Modell 1 β (sig)	**Modell 2 β (sig)**
Schlussfolgern/Erkennen (Info 1)	-.046 (.763)	
Implizit/explizit(Info 2)	-.018 (.876)	
Authentischer Text	.016 (.919)	
Inhalt hauptsächlich konkret	.127 (.291)	
Inhalt teilweise abstrakt	**.446 (.021)****	**.383 (.000)*****
Vokabular hauptsächlich häufig/einfach	.312 (.115)	
Vokabular teilweise ausgeweitet/selten	.206 (.140)	
Haupts. einfache grammatische Strukturen	**-.449 (.006)*****	**-.448 (.001)*****
Teilweise komplexe grammatische Strukturen	**-.397 (.034)****	**-.316 (.023)****
Itemtyp Multiple Choice	.028 (.849)	
Itemtyp Ordnen	-.045 (.727)	
Itemtyp Zitieren	-.063 (.533)	
R^2	.273	.227

Abhängige Variable: Itemschwierigkeit Spanien; 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 5.24. 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.

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

Abhängige Variable: Itemschwierigkeit Frankreich; Methode: Einschluß; Zelleninhalt: Standardisierte β-Gewichte; * p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 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 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.

	Modell 1 β (sig)	Modell 2 β (sig)
Schlussfolgern/Erkennen (Info 1)	.078 (.614)	
Implizit/explizit (Info 2)	-.016 (.892)	
Authentischer Text	.031(.850)	
Inhalt hauptsächlich konkret	**.231(.061)***	**.221(.031)****
Inhalt teilweise abstrakt	**.319 (.10)***	**.339 (.003)*****
Vokabular hauptsächlich häufig/einfach	**.365 (.015)****	**.358 (.003)*****
Vokabular teilweise ausgeweitet/selten	**.340 (.017)****	**.301(.011)****
Haupts.einfache grammatikalische Strukturen	**-.331(.045)****	**-.283 (.052)***
Teilweise komplexe grammatische Strukturen	**-.519 (.007)*****	**-.471(.002)*****
Itemtyp Multiple Choice	.099 (.501)	
Itemtyp Ordnen	-.021(.871)	
Itemtyp Zitieren	**.263 (.011)****	**.232 (.006)*****
R^2	.254	.250

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%

Tabelle 5.26. Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Ungarn (Englisch-Items)

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 französischen Schülergruppe (Tabelle 5.27) können insgesamt 58.4% der Itemschwierigkeits-Varianz der Deutsch-Items auf Anforderungsmerkmale der Items zurückgeführt werden. Wie auch schon bei den Englisch-Items wurden auch bezüglich der Deutsch-Items das Modell und die Anzahl der Prädiktoren immer so weit reduziert, bis ausschließlich Prädiktoren mit signifikanten Beta-Gewichten im Modell verbleiben. Bei der französischen Gruppe bleiben letztendlich drei Prädiktoren mit signifikanten Beta-Gewichten übrig, die insgesamt 45.1% der Itemschwierigkeits-Varianz erklären. Es zeigt sich, dass sich der Anteil der aufgeklärten Varianz vom ersten zum dritten Modell verringert, obgleich die ausgeschlossenen Prädiktoren keine signifikanten Beta-Gewichte aufweisen. Dies könnte für eine hohe Multikollinearität der Variablen sprechen. Auch zeigen sich Beta-Koeffizienten > 1. Diese deuten in der Regel auf Suppressionseffekte hin (Kline, 2005; Smith, Ager & Williams, 1992). Darauf wird in Abschnitt 6 im Rahmen der Ergebnisinterpretation ausführlicher eingegangen. Um den Effekt der Einzelvariablen unabhängig von den übrigen Prädiktoren besser einschätzen zu können, sollten daher auch die Einzelkorrelationen betrachtet werden. Auch hier zeigt sich, dass die grammatischen Strukturen hinsichtlich der Schwierigkeitsabstufungen nicht den erwarteten negativen Zusammenhang zur Itemschwierigkeit aufweisen. Dies ist teilweise auch bei den weiter unten dargestellten Modellen der Fall. Auf mögliche Gründe wird in der Ergebnisdiskussion eingegangen.

	Modell 1 β (sig)	Modell 2 β (sig)	Modell 3 β (sig)
Itemtyp Multiple Choice	-.016 (.967)		
Itemtyp Banked Multiple Choice	-.140 (.456)		
Richtig/Falsch	-.589 (.087)	**-.220 (.019)****	**-.201(.030)****
Itemtyp Multiple Matching	-.047 (.894)		
Itemtyp Kurzantwort	.106 (.622)		
Schlussfolgern/Erkennen (Info 1)	-.102 (.531)		
Implizit/explizit(Info 2)	-.041(.828)		
Detail/Hauptidee	-.104 (.584)		
Authentizität	**-.441(.005)*****	-.133 (.186)	
Inhalt haupts. konkret	.113 (.439)		
Inhalt teilweise abstrakt	-.037 (.816)		
Vokabular teilweise häufig / einfach	**.847 (.005) ***	**.410 (.044)****	**.319 (.095)***
Vokabular hauptsächlich ausgeweitet/selten	**1105 (.001)*****	**.685 (.003)*****	**.567 (.006)*****
Vokabular ausgeweitet/selten	**1359 (.005)*****	**1023 (.000)*****	**.985 (.000)*****
Haupts. einfache grammatische Strukturen	**.626 (.001)*****	**.432 (.000)*****	**.450 (.000)*****
Teilweise komplexe grammatische Strukturen	.316 (.329)		
Komplexe grammatische Strukturen	-.326 (.417)		
R^2	.584	.463	.451

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

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

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 signifikantes Beta-Gewicht und klärt 22.8% der Varianz der Itemschwierigkeiten auf.

	Modell 1 β (sig)	Modell 2 β (sig)
(.202)		
Banked Multiple Choice -.128 -.129		
Richtig-Falsch	**-.527 (.011)****	**-.408 (.000)*****
Multiple Matching	**-.559 (.008)*****	**-.272 (.007)*****
Kurzantwort	.132 (.473)	
Lückentext	-.073 (.553)	
Schlussfolgern/Erkennen (Info 1)	-.199 (.233)	
Implizit/explizit(Info 2)	-.024 (.901)	
Detail/Hauptidee	-.310 (.111)	
Authentizität	-.199 (.204)	
Inhalt haupts. konkret	**.379 (.013)****	**.315 (.004)*****
Inhalt teilweise abstrakt	.042 (.795)	
Vokabular hauptsächlich häufig/einfach	**.759 (.013)****	**.566 (.005)*****
Vokabular teilweise ausgeweitet/selten	**.858 (.011)****	**.661 (.005)****
Vokabular ausgeweitet/selten	**1654 (.001)*****	**1021 (.000)*****
Haupts.einfache grammatische Strukturen	**.423 (.022)****	**.406 (.001)****
Teilweise komplexe grammatische Strutkuren	-.176 (.592)	
Komplexe grammatische Strukturen	-.660 (.108)	
R^2	.569	.494

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%***

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

In der schwedischen Gruppe (Tabelle 5.30) können mit Hilfe aller verwendeter Prädiktoren insgesamt 42.6% der Varianz aufgeklärt werden. Nach dem Ausschluss von Prädiktoren mit nicht-signifikanten Beta-Gewichten werden noch 20.8% der Varianz erklärt. Die starke Verringerung des Anteils der aufgeklärten Varianz durch die Herausnahme nicht-signifikanter Prädiktoren spricht auch hier für Multikollinearitätseffekte. Darüber hinaus deutet auch hier die Anwesenheit von Beta-Koeffizienten > 1 auf Suppressionseffekte hin. Hier tragen vor allem Abstufungen der Variablen „Komplexität der Grammatik", „Schwierigkeit des Vokabulars" und „Itemtyp" (Kurzantwort) zur Itemschwierigkeit bei.

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

	Modell 1 β (sig)	Modell 2 β (sig)
Itemtyp Multiple Choice	-.016 (.955)	
Banked Multiple Choice	.119 (.364)	
Multiple Matching	.037 (.899)	
Kurzantwort	.284 (.233)	
Lückentext	.107 (.471)	
Schlussfolgern/Erkennen (Info 1)	.004 (.985)	
Implizit/explizit(Info 2)	-.080 (.711)	
Detail/Hauptidee	-.254 (.248)	
Authentizität	-.104 (.557)	
Inhalt haupts. konkret	.232 (.170)	
Inhalt teilweise abstrakt	.149 (.422)	
Vokabular hauptsächlich häufig/einfach	.093 (.785)	
Vokabular teilweise ausgeweitet/selten	.253 (.500)	
Vokabular ausgeweitet/selten	**.933 (.091)***	**.478 (.000)*****
Haupts.einfache grammatische Strukturen	.202 (.327)	
Teilweise komplexe grammatische Strutkuren	-.322 (.387)	
Komplexe grammatische Strukturen	-.690 (.138)	
R^2	**.447**	**.228**

Abhängige Variable: Itemschwierigkeit Niederlande; 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 5.29. 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

	Modell 1 β (sig)	Modell 2 β (sig)	Modell 3 β (sig)
Banked Multiple Choice	.011(.931)		
Richtig-Falsch	-.365 (.120)		
Multiple Matching	-.383 (.111)		
Kurzantwort	**.354 (.097)***	**.413 (.007)*****	**.449 (.004)*****
Lückentext	.063 (.656)		
Schlussfolgern/Erkennen (Info 1)	-.270 (.162)		
Implizit/explizit(Info 2)	-.180 (.416)		
Detail/Hauptidee	-.305 (.173)		
Authentizität	-.166 (.357)		
Inhalt haupts. konkret	**.296 (.087)***	.563 (.575)	
Inhalt teilweise abstrakt	.170 (.368)		
Vokabular hauptsächlich häufig/einfach	.514 (.141)		
Vokabular teilweise ausgeweitet/selten	**.680 (.078)***	.161(.220)	
Vokabular ausgeweitet/selten	**1750 (.002)*****	**.482 (.006)*****	**.435(.002)*****
Haupts.einfache grammatische Strukturen	.272 (.196)		
Teilweise komplexe grammatische Strutkuren	-.342 (.367)		
Komplexe grammatische Strukturen	**-1240 (.010)*****	**-.399 (.035)****	**-.456 (.014)****
R^2	.426	.240	.208

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

Tabelle 5.30. Regression der Itemschwierigkeiten auf Itemanforderungs-Merkmale in Schweden (Deutsch-Items)

jedem der Länder oder in jeder der beiden hier überprüften Sprachen. Die Itemeigenschaften des „Dutch Grid" scheinen insgesamt einen Teil der Itemschwierigkeit länderübergreifend erklären zu können. Daher wird angenommen, dass die Merkmale auch zur Erklärung von Unterschieden zwischen Ländern hinsichtlich der Itemschwierigkeiten herangezogen werden können.

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?

Wenn es Indikatoren der nationalen Testkultur gibt, dann sollte sich das zunächst darin zeigen, dass „eigene" Items aus dem eigenen Land für Schüler leichter korrekt zu beantworten sind als für Schüler, die aus einer anderen Testkultur stammen. Die Berechnung einer Korrelation zwischen der Itemherkunft und DIF stellt eine Art von Screening-Prozedur dar, da angenommen wird, dass die Itemherkunft die übrigen kulturellen Merkmale der Items beinhaltet und somit eine Art Über-Kategorie für detailliertere kulturelle Merkmale darstellen sollte.
Zur Untersuchung dieser Annahme wurden jeweils paarweise DIF Parameter zwischen den Ländern berechnet (siehe auch 5.1.2) 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 (dem der jeweiligen Fokusgruppe) stammt, negativ ausfallen sollte, da ein niedriger Itemschwierigkeitswert (in Logits) auch für eine geringere Itemschwierigkeit steht. Der Umstand, dass ein Item aus dem eigenen Land stammt, sollte einen Vorteil für die Fokusgruppe

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.

	F-D	F-U	F-Sp	D-U	D-Sp	U-Sp
Itemherkunft Land A (für A-B)	**-.407****	**-.201***	-.004	**-.30****	**-.312***	**-.209***
Itemherkunft Land B (für A-B)	**.281***	**.251***	.054	**.249****	-.058	.11

D = Deutschland; F = Frankreich; SP = Spanien; U = Ungarn
**** p ≤ 0.05, α = 5%; ** p ≤ 0.01, α = 1%;***

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 ($p \leq 0.05$) 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.

	FR-NL	FR-SW	FR-U	NL-SW	U-NL	SW-U
Herkunft Land A (für A-B)	-.030	**-.486****	**-.218***	**-.197***	**-.260***	.172
Herkunft Land B (für A-B)	**.302***	**.395****	.135	**.242***	**.205***	-.177

F = Frankreich; NL = Niederlande; SW = Schweden; U = Ungarn
**** p ≤ 0.05, α = 5%; ** p ≤ 0.01, α = 1%;***

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

Tabelle 5.32 ist analog zu Tabelle 5.31 zu lesen und zu interpretieren. Die Ergebnisse der Korrelationen zwischen DIF-Parametern zweier Länder und der Itemherkunft von Deutsch-Items zeigen ähnliche Ergebnisse wie bei den Englisch-Items. Auch hier zeigen sich für die Analysen zwischen Frankreich und Schweden, den Niederlanden und Schweden sowie Ungarn und den Niederlanden erwartungsgemäß für die jeweilige Fokusgruppe signifikant negative Korrelationen zwischen DIF-Parametern und der Tatsache, dass ein Item aus dem eigenen Land stammt, sowie signifikant positive Korrelationen zwischen DIF und der Tatsache, dass das Item aus dem Land der jeweiligen Referenzgruppe stammt. Einzig zwischen Schweden und Ungarn zeigt sich kein signifikanter Zusammenhang zwischen DIF und der Itemherkunft. Dies ist eventuell mit einer relativ kleinen Varianz der Differentiellen Item Funktionen zwischen Schweden und Ungarn zu erklären: Hier findet möglicherweise eine Einschränkung der Varianz der abhängigen Variable statt, die zu einer Unterschätzung der Korrelationskoeffizienten führt. Die Varianzen der Differentiellen Item Funktionen sind im Anhang einzusehen.

Nachdem bei einem großen Teil der Länderpaarungen und bezüglich beider Sprachen Zusammenhänge zwischen der Itemherkunft und DIF in erwarteter Richtung gefunden wurden, wird im nächsten Schritt versucht, diese zu spezifizieren. Dazu werden korrelative Zusammenhänge zwischen den paarweise geschätzten DIF Parametern und den kognitiv-linguistischen Anforderungsmerkmalen eines Items analysiert. Die Hypothese ist hier, dass die Ausprägung einer Itemeigenschaft, die bei den „eigenen" eingereichten Items der einen Gruppe häufiger vorkommt, das Item für diese Gruppe im Vergleich mit der Referenzgruppe erleichtern und für die Referenzgruppe erschweren sollte. Negative Korrelationen bedeuten auch hier einen Vorteil für die Fokusgruppe.

Die Einzelhypothesen werden aus den unter 5.1.3 gewonnenen Erkenntnissen über die nationalen Testkultur-Profile abgeleitet, welche dort tabellarisch dargestellt wurden.
Kam dort ein Merkmal bei den Items eines Landes signifikant häufiger vor als bei den Items des anderen Landes mit dem diese jeweils verglichen wurden, wurde die Hypothese aufgestellt, dass ein Item mit dem Merkmal dann mit einer niedrigeren Itemschwierigkeit für dieses Land, verglichen mit dem anderen Land, zusammenhängen sollte. Bei den Ergebnissen sollte sich das in einer negativen Korrelation ausdrücken, da ein niedrigerer Logit-Wert mit einer niedrigeren Itemschwierigkeit einhergeht. Wie oben bereits dargestellt, werden ausschließlich signifikante Korrelationen interpretiert sowie lediglich Ergebnisse zu den Einzelhypothesen, die sich auf die Unterschiedlichkeit zweier Testkulturen beziehen, da sich nur diese eindeutig interpretieren lassen.
Die Ergebnisse der Korrelationsanalysen zwischen DIF und den kognitiv-linguistischen Itemmerkmalen werden im Folgenden dargestellt. Die nicht-signifikanten Korrelationen werden dabei aus Gründen der Übersichtlichkeit in der Tabelle nicht dargestellt. Die komplette Korrelationsmatrix ist im Anhang einsehbar.
Tabelle 5.33 ist wie folgt zu lesen: Links sind die Itemeigenschaften sowie deren Merkmalsausprägungen abgetragen und in den Spalten die Länder, zwischen denen jeweils paarweise DIF-Analysen durchgeführt wurden. Die Zellen beinhalten die Korrelationen zwischen den Merkmalsausprägungen und den DIF-Parametern aus Sicht der Fokusgruppe. Ein fett gedruckter Zelleninhalt bedeutet dabei, dass der Zusammenhang der durch die jeweilige Testkultur bewirkten und erwarteten Richtung entspricht. Ist der Zelleninhalt fett und kursiv, bedeutet dies, dass im Vorfeld keine spezifische Hypothese bezüglich dieses Zusammenhangs aufgestellt werden konnte, da bei der Analyse der Testkulturen kein signifikanter Unterschied hinsichtlich des Vorkommens des Itemmerkmals festgestellt wurde, hier jedoch ein signifikanter Zusammenhang zu DIF beobachtbar ist. Daher werden im Folgenden solche Zusammenhänge als „neutral” bezeichnet. Ein nicht hervorgehobener Eintrag, der mit einem „ungleich”-Zeichen versehen ist, weist auf einen den Hypothesen entgegen gesetzten Zusammenhang hin.
Es zeigt sich, dass 23 von 29 der signifikanten Korrelationen hinsichtlich ihrer Richtung den oben aufgestellten Hypothesen entsprechen. Einige wenige der signifikanten Korrelationen, nämlich fünf, weisen auf einen signifikanten Zusammenhang zwischen den DIF-Parametern zweier Länder und einem Item-Anforderungsmerkmal hin, obgleich sich diese Länder den Hypothesen nach nicht unterscheiden sollten. Lediglich der Zusammenhang zu den DIF zwischen Ungarn und Spanien für Items mit konkretem Inhalt entspricht nicht der aufgestellten Hypothese.

Itemeigenschaft		DIF					
		F-D	F-U	F-Sp	D-U	D-Sp	U-Sp
Itemtyp	Itemtyp Multiple Choice	**.194****	—	—	**-.256*****	***-.233****	—
	Banked Multiple Choice	—	—	—	—	—	—
	Richtig-Falsch	—	—	—	—	—	—
	Multiple Matching	—	—	—	—	—	—
	Zitieren	**-.396*****	**-.283*****	—	—	—	***.419*****
	Lückentext	—	**.400*****	**.207****	**.467*****	—	***-.276*****
Informationsgewinn 1	Schlussfolgern/Erkennen	—	**.198****	**.209****	—	—	—
Informationsgewinn 2	Implizit (vs. Explizit)	—	—	—	**-.194****	—	—
Informationsgewinn 3	Detail (vs. Hauptidee)	—	—	—	—	—	—
Authentizität	Authentisch (vs. angepasst/ vereinfacht)	**-.304*****	—	—	**.283*****	**.263*****	—
Abstraktheit des Inhalts	ausschließlich konkret	—	—	—	—	—	—
	hauptsächlich konkret	—	**-.296****	**-.183****	—	—	**.183**** ≠
	teilweise abstrakt	***-.217*****	**.40*****	***.207*****	**.467*****	***.198*****	**-.276*****
Vokabular	ausschießlich häufig / einfach	—	—	—	—	—	—
	hauptsächlich häufig / einfach	—	—	—	**-.216****	—	—
	teilweise erweitert / selten	—	—	—	**.297****	**.198****	—
	erweitert/selten	—	—	—	—	—	—
Grammatik	ausschließlich einfache Strukturen	—	—	—	—	—	—
	hauptsächlich einfache Strukturen	—	—	—	—	—	**.203****
	teilweise komplexe grammatische Strukturen	—	—	—	—	—	—

D = Deutschland; F = Frankreich; SP = Spanien; U = Ungarn
**** p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 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.

Im Folgenden werden nun die Ergebnisse für die Deutsch-Items berichtet.

Die zur Analyse der Deutsch-Items verwendeten kognitiv-linguistischen Itemeigenschaften weichen leicht von den für die Englisch-Items verwendeten ab. Das ist darauf zurückzuführen, dass die beiden Sprachen teilweise in unterschiedlichen Ländern getestet wurden, die teilweise Items mit unterschiedlichen Formaten bzw. Ausprägungen der Itemeigenschaften einreichten. So gibt es bei den Englisch-Items beispielsweise keine Kurzantwort-Items, da diese nicht im Itempool vorhanden waren, bei den Deutsch-Items hingegen kommen keine Items vom Typ „Zitieren" vor.

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 komplette 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.

Itemeigenschaft		DIF					
		FR-NL	FR-SW	FR-HU	NL-SW	HU-NL	SW-HU
Itemtyp	Multiple Choice	—	**-.264****	**-.369*****	**-.279*****	**.332*****	—
	Banked Multiple Choice	—	—	—	—	—	—
	Richtig-Falsch	—	—	—	—	—	—
	Multiple Matching	—	**.304*****	**.470*****	**.256****	**-.354*****	—
	Kurzantwort	—	—	—	—	—	—
	Lückentext	—	—	—	—	—	—
Informationsgewinn 1	Schlussfolgern/Erkennen	—	**.227****	—	**.256****	—	—
Informationsgewinn 2	Implizit(vs. Explizit)	—	**.396*****	***.275****	**.321*****	—	—
Informationsgewinn 3	Detail (vs. Hauptidee)	***.242****	—	—	**-.400*****	**.397*****	—
Authentizität	Authentisch (vs. angepasst / vereinfacht)	—	**-.235****	—	—	—	—
Abstraktheit des Inhalts	ausschließlich konkret	—	**-.230****	—	**-.245****	—	**-.328***** ≠
	hauptsächlich konkret	—	—	—	—	—	**.265**** ≠
	teilweise abstrakt	—	—	—	—	—	—
Vokabular	ausschießlich häufig/einfach	—	—	—	—	—	—
	hauptsächlich häufig/einfach	—	**-.258****	—	***-.231****	—	—
	teilweise erweitert/selten	—	—	—	—	—	—
	erweitert/selten	—	**.416*****	**.228****	**.306*****	—	—
Grammatik	ausschließlich einfache Strukturen	—	**-.452*****	—	**-.311*****	—	—
	hauptsächlich einfache Strukturen	—	—	—	—	—	—
	teilweise komplexe Strukturen	—	**.241***	**.287****	**.278****	**-.289****	—
	komplexe Strukturen	—	—	***-.330* ******	—	—	***.286****

F = Frankreich; NL = Niederlande; SW = Schweden; U = Ungarn
**** p ≤ 0.1, α = 10%; ** p ≤ 0.05, α = 5%; *** p ≤ 0.01, α = 1%***
≠ = entgegen der Hypothese; kursiv = neutral

Tabelle 5.34. Korrelation zwischen DIF und kognitiv-linguistischen Itemmerkmalen (Deutsch-Items)

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.

Insgesamt sprechen die Ergebnisse für einen Zusammenhang zwischen Testkulturen und Differentiellen Item Funktionen.
Nur einige wenige Korrelationen weisen auf Zusammenhänge hin, die nicht erwartet waren. Dies ist möglicherweise ein Hinweis darauf, dass zwar ein Unterschied in der Testkultur zweier Länder besteht, dieser hier jedoch aufgrund einer falschen Einordnung von Items durch die Experten oder aufgrund der Nicht-Repräsentativität der Items eines Landes bezüglich eines Merkmals möglicherweise nicht darstellbar ist.

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 (α = 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

Tabelle 5.35 ist wie folgt zu lesen: In der linken Spalte wurden die verwendeten Prädiktoren abgetragen. In den Spalten finden sich die (jeweils von Modell 1 ausgehend) sukzessive gerechneten Regressionsmodelle, und in der letzten Zeile ist jeweils der Anteil der durch die verwendeten Faktoren aufgeklärten Varianz der DIF-Parameter abgetragen. Die Zelleninhalte beinhalten standardisierte Beta-Gewichte. Wie hier ersichtlich wird, werden im ersten Modell sieben der Prädiktoren signifikant ($p \leq 0.1$). Hier werden unter Hereinnahme aller Prädiktoren 32.7% der Varianz der DIF-Parameter aufgeklärt. Im zweiten Modell werden dann nur diejenigen Prädiktoren aufgenommen, die Signifikanz aufweisen und der Richtung nach den oben aufgestellten Hypothesen entsprechen. Dies trifft beispielsweise nicht auf den Prädiktor „hauptsächlich einfache grammatische Strukturen" zu, weshalb er, obwohl siginifkant, aus den weiteren Analysen ausgeschlossen wird. Ferner werden in dem zweiten Modell auch diejenigen Prädiktoren aufgenommen, die als „neutral" bezeichnet werden, da sie der Hypothese zwar nicht entsprechen, dieser aber auch nicht zuwiderlaufen. Bei diesen Prädiktoren zeigt sich hier ein signifikanter Zusammenhang, der aufgrund der Testkultur-Unterschiede nicht erwartet war. In dem oben dargestellten Modell trifft dies beispielsweise auf den Prädiktor „Vokabular selten" zu. Wie oben bereits erwähnt, könnte dies jeweils auf fehlende Repräsentativität der Items aus dem entsprechenden Land (dieses eine

	Modell 1	Modell 2	Modell 3
	Beta(sig)	Beta(sig)	Beta(sig)
Itemtyp Multiple Matching	**-.240 (.032)****	-.090 (.312)	-.016 (.856)
Itemtyp Ordnen	-.175 (.186)		
Itemtyp Zitieren	**.351(.001)*****	**.414 (.000)*****	**.396 (.000)*****
Schlussfolgern/Erkennen (Info 1)	**.251(.090)*** ≠		
Implizit/explizit(Info 2)	-.145 (.200)		
Authentischer Text	.241(.118)		
Inhalt haupts. konkret	***.227*** **(.052)***	***.232*** **(.013)****	
Inhalt teilweise abstrakt	-.228 (.258)		
Vokabular häufig/einfach	-.037 (.790)		
Vokabular ausgeweitet/selten	***.374*** **(.006)*****	***.304*** **(.001)*****	
Haupts.einfache grammatikalische Strukturen	**-.311(.047)**** ≠		
Teilweise komplexe grammatische Strukturen	**-.298 (.097)***	-.138 (.133)	-.080 (.381)
R^2	.327	.251	.163

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 entsprechen. 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 diesem 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 Schülern um ca. 0.4 Logits. Gleiches gilt für den Prädiktor „Vokabular erweitert / selten" (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.

	Modell 1	**Modell 2**	**Modell 3**
	Beta(sig)	Beta(sig)	Beta(sig)
Itemtyp Multiple Matching	**.183 (.094)***	.099 (.638)	.086 (.978)
Itemtyp Ordnen	***.217* (.095)***	*.123* (.157)	
Itemtyp Zitieren	-.108 (.265)		
Schlussfolgern/Erkennen (Info 1)	-.227 (.118)		
Implizit/explizit (Info 2)	**.189 (.089)***	**.239 (.007)****	**.222 (.011)****
Authentischer Text	-.146 (.334)		
Inhalt haupts. konkret	-.128 (.263)		
Inhalt teilweise abstrakt	-.192 (.330)		
Vokabular hauptsächlich häufig/einfach	-.003 (.982)		
Vokabular teilweise ausgeweitet/selten	**-.264 (.046)****	**-.344 (.000)*****	**-.345(.000)*****
Haupts.einfache grammatikalische Strukturen	**.449 (.004)*****	**.201(.023)****	**.189 (.032)****
Teilweise komplexe grammatische Strukturen	**.440 (.013)***** ≠		
R^2	.352	.186	.171

Abhängige Variable: DIF Ungarn - Deutschland; 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 5.36. 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.

	Modell 1	**Modell 2**
	Beta(sig)	Beta(sig)
Itemtyp Multiple Matching	-.050 (.212)	
Itemtyp Ordnen	.053 (.695)	
Itemtyp Zitieren	**.246 (.017)****	**.260 (.002)*****
Schlussfolgern/Erkennen (Info 1)	.015 (.924)	
Implizit/explizit(Info 2)	.054 (.641)	
Authentischer Text	.091 (.566)	
Inhalt haupts. konkret	.096 (.423)	
Inhalt teilweise abstrakt	**-.442 (.035)****	**-.384 (.000)*****
Vokabular hauptsächlich häufig/einfach	-.042 (.773)	
Vokabular teilweise ausgeweitet/selten	.101 (.468)	
Haupts.einfache grammatikalische Strukturen	.163 (.311)	
Teilweise komplexe grammatische Strutkuren	.167 (.367)	
R^2	.277	.227

Abhängige Variable: DIF Ungarn - 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 5.37. 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).

	Modell 1	**Modell 2**	**Modell 3**
	Beta(sig)	Beta(sig)	Beta(sig)
Itemtyp Multiple Matching	**-.220 (.041)****	-.021(.812)	-.112 (.248)
Itemtyp Ordnen	-.196 (.122)		
Itemtyp Zitieren	***.416* (.000)*****	***.492* (.000)*****	
Schlussfolgern/Erkennen (Info 1)	***.333* (.020)****	***.216* (.014)****	
Implizit/explizit (Info 2)	-.161(.137)		
Authentischer Text	.142 (.337)		
Inhalt haupts. konkret	**.218 (.052)*** ≠		
Inhalt teilweise abstrakt	-.029 (.879)		
Vokabular hauptsächlich häufig/einfach	.140 (.298)		
Vokabular teilweise ausgeweitet/selten	**.367 (.005)*****	-.109 (.224)	**.200 (.029)****
Haupts.einfache grammatikalische Strukturen	-.246 (.100)		
Teilweise komplexe grammatische Strukturen	**-.501(.004)*****	**-.195 (.043)****	-.121(.209)
R^2	.381	.256	.056

Abhängige Variable: DIF Deutschland - Spanien; 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 5.38. Regression von DIF Deutschland - Spanien auf Anforderungsmerkmale (Englisch-Items)

Mit Hilfe des ersten Modells in Tabelle 5.38 können insgesamt 38.1% der kulturellen Varianz zwischen Deutschland und Spanien aufgeklärt werden. Fünf der Prädiktoren sind signifikant und entsprechen der Testkultur bzw. gehören zu den „neutralen" Prädiktoren. Diese werden in einem zweiten Modell wiederverwendet. Hier können 25.6% der Varianz erklärt werden. Da dieses Modell „neutrale" Prädiktoren beinhaltet, bei denen der Zusammenhang nicht eindeutig auf die Testkultur zurückzuführen ist, wird dies als die Obergrenze der kulturellen Varianz interpretiert, die anhand der hier verwendeten Testkultur-Indikatoren erklärt werden kann. Es wird hier angenommen, dass möglicherweise eigentlich ein Unterschied besteht, der hier durch eine falsche Einordnung der Items durch die Experten, oder aber wegen einer fehlenden Repräsentanz der Items des Landes nicht entdeckt wurde. In einem dritten Modell werden nur die signifikanten und Testkultur-hypothesenkonformen Prädiktoren verwendet. Es zeigt sich, dass dort noch 5.6% der Varianz aufgeklärt werden können. Allerdings ist hier nur einer der Prädiktoren, nämlich die Variable „Vokabular selten/schwierig", signifikant. Da alle Prädiktoren vorher hoch signifikant oder signifikant waren, ist hier zu vermuten, dass eine hohe Multikollinearität vorhanden ist. Inwiefern die Variablen untereinander zusammenhängen oder mögliche, hier nicht modellierte Wechselwirkungen eine Rolle spielen, kann im Rahmen dieser Arbeit nicht überprüft werden, sollte aber Forschungsgegenstand zukünftiger Studien sein.

Gemeinsam erklären alle Prädiktoren 25.1% der Varianz der Differentiellen Item Funktionen zwischen Frankreich und Spanien (Tabelle 5.39). Zwei der Prädiktoren werden in ein weiteres Modell übernommen, da sie beide signifikante Beta-Gewichte aufweisen, und den oben aufgestellten Hypothesen nicht widersprechen. In diesem zweiten Modell können nur noch 4.1% der Varianz erklärt werden. Ferner sind die Beta-Gewichte beider Prädiktoren deutlich kleiner als in Modell 1. Dies weist auf mögliche Kollinearitätsprobleme hin, oder auch auf Wechselwirkungen zwischen den Prädiktoren. Eine weitere Möglichkeit für die insgesamt relativ geringen Zusammenhänge in der Regressionsanalyse könnte die geringe Varianz der Differentiellen Item Funktionen zwischen Frankreich und Spanien sein (siehe Anhang).
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-

	Modell 1	**Modell 2**
	Beta(sig)	Beta(sig)
Itemtyp Multiple Matching	-.037 (.751)	
Itemtyp Ordnen	-.092 (.508)	
Itemtyp Zitieren	**.226 (.031) **≠**	
Schlussfolgern/Erkennen (Info 1)	.225 (.149)	
Implizit/explizit(Info 2)	-.074 (.532)	
Authentischer Text	-.107 (.509)	
Inhalt haupts. konkret	.053 (.663)	
Inhalt teilweise abstrakt	.290 (.172)	
Vokabular hauptsächlich häufig/einfach	**.309 (.038)* ***	.137 (.165)
Vokabular teilweise ausgeweitet/selten	.102 (.470)	
Haupts.einfache grammatikalische Strukturen	.022 (.892)	
Teilweise komplexe grammatische Strutkuren	**-.460 (.016)* ***	**-.215 (.030)* ***
R^2	.251	.041

Abhängige Variable: DIF Frankreich - Spanien; 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 5.39. 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 unterscheiden. 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 Länderpaarung mit beeinflusst zu sein: Paarungen betreffend, die sich aus den Ländern Deutschland, Frankreich und Ungarn konstituieren (R^2 = .163 bis R^2 = .227), scheinen besser geeignet zu sein als Paarungen, die die spanische Stichprobe mit einschließen (R^2 = .0 bis R^2 = .041). Ob dies lediglich auf die spezifische, im Rahmen der EBAFLS Studie verwendete spanische Stichprobe

	Modell 1	**Modell 2**
	Beta(sig)	Beta(sig)
Itemtyp Multiple Matching	-.080 (.463)	
Itemtyp Ordnen	-.010 (.936)	
Itemtyp Zitieren	***.428* (.000)***	***.419* (.000)***
Schlussfolgern/Erkennen (Info 1)	.180 (.216)	
Implizit/explizit(Info 2)	.004 (.971)	
Authentischer Text	.020 (.895)	
Inhalt haupts. konkret	.142 (.215)	
Inhalt teilweise abstrakt	-.262 (.188)	
Vokabular hauptsächlicht häufig/einfach	.182 (.168)	
Vokabular teilweise ausgeweitet/selten	.182 (.168)	
Haupts.einfache grammatikalische Strukturen	.191(.213)	
Teilweise komplexe grammatische Strukturen	-.157 (.372)	
R^2	.345	.176

Abhängige Variable: DIF Spanien - Ungarn; 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 5.40. 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 Kontrastvariablen 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-

dell, das Prädiktoren beinhaltet, die der Richtung nach den 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.
Es besteht die Möglichkeit, dass in solchen Fällen möglicherweise Fehler bei der Einordnung der Items durch die Experten zugrunde liegen, oder dass die von den Ländern eingereichten Items hinsichtlich dieses einen Merkmals nicht repräsentativ waren. Dies kann hier jedoch nur vermutet und nicht eindeutig auf Richtigkeit hin überprüft werden, da auch noch weitere, stichprobenspezifische, im Rahmen dieser Arbeit nicht mit erhobene kulturelle Faktoren eine Rolle spielen könnten. Daher wird für Modelle, die „neutrale" Prädiktoren beinhalten, auch hier der Anteil der dort aufgeklärten Varianz, R^2, als die Obergrenze der theoretisch aufgrund der Testkultur und der hier verwendeten Prädiktoren aufklärbaren Varianz betrachtet. In einigen Fällen folgt, ausgehend von Modell 1, noch ein drittes Modell, welches auch die „neutralen" Prädiktoren ausschließt und nur noch Prädiktoren beinhaltet, die hinsichtlich ihrer Richtung den aufgrund der Testkulturen erwarteten Stärken und Schwächen der Ländergruppen entsprechen. Hier ist dann der Anteil der aufgeklärten Varianz eindeutig auf durch unterschiedliche Testkulturen verursachte Stärken und Schwächen der Gruppen zurückzuführen.

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 den 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 den 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

	Modell 1	**Modell 2**	**Modell 3**
	Beta(sig)	Beta(sig)	Beta(sig)
Banked Multiple Choice	-.033(.772)		
Itemtyp Richtig-Falsch	-.195(.336)		
Itemtyp Multiple Matching	**.916 (.000)*** ≠**	**.770 (.000)*** ≠**	**.524(.000)*** ≠**
Itemtyp Kurzantwort	-.004 (.984)		
Itemtyp Lückentext	.138 (.264)		
Schlussfolgern/Erkennen (Info 1)	.150 (.367)		
Implizit/explizit(Info 2)	-.038 (.884)		
Hauptidee/Detail (Info 3)	**.343 (.079)***	.202(.145)	.026 (.854)
Authentischer Text	***-.508* (.002)*** ≠**	***-.430* (.000)*** ≠**	
Inhalt haupts. konkret	**-.445(.004) *** ≠**		
Inhalt teilweise abstrakt	-.147 (.396)		
Vokabular hauptsächlich häufig/einfach	.316 (.295)		
Vokabular teilweise ausgeweitet/selten	**.643 (.056)***	**.267 (.023)****	.026 (.807)
Vokabular ausgeweitet/selten	-.261(.589)		
Haupts.einfache grammatikalische Strukturen	**.475 (.011)****	**.238 (.020)****	**.337 (.002)*** ≠**
Teilweise komplexe grammatische Strukturen	**.924 (.006)*** ≠**	**.263 (.084)***	.163 (.193)
Komplexe grammatische Strukturen	.527 (.199)		
R^2	.567	.457	.347

Abhängige Variable: DIF Frankreich - Ungarn; 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 5.41. 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

	Modell 1	**Modell 2**	**Modell 3**
	Beta(Sig.)	Beta(Sig.)	Beta (Sig.)
Itemtyp Banked Multiple Choice	***-.375* (.015)****	*-.162* (.202)	
Itemtyp Richtig-Falsch	***-.850* (.002)****	***-.275* (.053)***	
Itemtyp Multiple Matching	-.122(.653)		
Itemtyp Kurzantwort	-.263(.277)		
Itemtyp Lückentext	-.157(.334)		
Erkennen/Schlussfolgern (Info 1)	-.153(.483)		
Implizit/explizit (Info 2)	-.058(.817)		
Hauptidee/Detail (Info 3)	.221(.384)		
Authentischer Text	**-.484 (.021)****	-.109 (.393)	-.086 (.474)
Inhalt haupts. konkret	-.177 (.366)		
Inhalt teilweise abstrakt	-.271(.209)		
Vokabular hauptsächlich häufig/einfach	**1086(.007)***≠**		
Vokabular teilweise ausgeweitet/selten	**1222(.006)*****	.034 (.784)	-.058 (.632)
Vokabular ausgeweitet/selten	.596 (.348)		
Haupts.einfache grammatikalische Strukturen	**.607(.013)****	**.269 (.054)***	.125 (.278)
Teilweise komplexe grammatische Strukturen	***.926* (.034)****	*.133* (.349)	
Komplexe grammatische Strukturen	.538 (.317)		
R^2	.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.

	Modell 1	**Modell 2**	**Modell 3**
	Beta(Sig.)	Beta(Sig.)	Beta(Sig.)
Itemtyp Banked Multiple Choice	**-.277(.030)**** ≠		
Itemtyp Richtig-Falsch	*-.511* **(.023)****	-.147(.240)	
Itemtyp Multiple Matching	**.569 (.014)****	**.394(.006)*****	**.331(.004)*****
Itemtyp Kurzantwort	**-.361(.074)*** ≠		
Itemtyp Lückentext	-.095(.477)		
Schlussfolgern/Erkennen (Info 1)	.249(.171)		
Implizit/explizit(Info 2)	.219(.294)		
Hauptidee/Detail (Info 3)	.306(.149)		
Authentischer Text	**-.579(.001)*****	**-.313(.018)****	**-.371(.004)*****
Inhalt haupts. konkret	**-.272(.096)*** ≠		
Inhalt teilweise abstrakt	**-.355(.049)**** ≠		
Vokabular hauptsächlich häufig/einfach	**.789 (.018)**** ≠		
Vokabular teilweise ausgeweitet/selten	**1014(.006)*****	.130(.283)	
Vokabular ausgeweitet/selten	-.270(.608)		
Haupts.einfache grammatische Strukturen	***.762(.000)******	***.381(.001)******	
Teilweise komplexe grammatische Strukturen	**1181(.001)*****	**.382(.009)*****	**.227(.047)****
Komplexe grammatische Strukturen	**1429(.002)*****	**.249(.054)***	.068(.566)
R^2	.487	.369	.259

Abhängige Variable: DIF Frankreich - 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 5.43. 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-

spricht. Daher kann man bei allen vier Variablen, die im Rahmen dieser Regressionsanalyse den Hypothesen entgegengesetzte Regressionskoeffizienten aufweisen, von Suppressionseffekten ausgehen. In einem zweiten Modell werden daher nur die Prädiktoren mit signifikanten, Testkultur-konformen bzw. „neutralen" Beta-Gewichten mit einbezogen. Hier lassen sich noch 36.9 % der Varianz erklären. Der Prädiktor „Hauptsächlich einfache grammatische Strukturen" hat hier ein signifikantes Beta-Gewicht, obwohl den aufgrund der Testkulturen aufgestellten Hypothesen nach die französischen Items sich diesbezüglich nicht signifikant von den schwedischen unterscheiden. Dieses Ergebnis ist nicht eindeutig interpretierbar, obgleich denkbar ist, dass die Items nicht ganz korrekt von den Experten eingeordnet wurden, und hier daher ein Testkultur-Effekt vorhanden ist, der aber nicht klar abgebildet werden kann. Dieses Modell stellt daher eine mögliche Obergrenze der anhand der verwendeten Testkultur-Indikatoren aufklärbaren Varianz dar. Es wird ferner ein drittes Modell gerechnet, in dem der nicht klar interpretierbare „neutrale" Prädiktor nicht mit einbezogen wird. Mit diesem Endmodell können dann noch 25.9% der DIF-Varianz aufgeklärt werden. Die Beta-Koeffizienten sind auch hier als die relativen Stärken und Schwächen der beiden Gruppen im Vergleich mit der jeweils anderen Gruppe zu interpretieren. Hier tragen vor allem der Itemtyp „Multiple Matching", die Variable „authentischer Text" sowie die Variable „teilweise komplexe grammatische Strukturen" zur differentiellen Schwierigkeit der Testitems bei. Die Ergebnisse werden so interpretiert, dass zwischen 36.9% (Obergrenze) und 25.6% (Untergrenze) der durch kulturelle Unterschiede bedingten Varianz zwischen Frankreich und Schweden auf die Unterschiedlichkeit der Testkulturen zurückführbar sind.

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-

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

Abhängige Variable: DIF Niederlande - Schweden; Methode: Einschluß; Zelleninhalt: β-Gewichte
**** $p \leq 0.1$, $\alpha = 10\%$; ** $p \leq 0.05$, $\alpha = 5\%$; *** $p \leq 0.01$, $\alpha = 1\%$***
$\neq$ = 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 hypothesenkonformen 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 hypothesenkonformen Prädiktoren mit aufgenommen, hier können noch 23.6% der Varianz erklärt werden. Die Ergeb-

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

Abhängige Variable: DIF Ungarn - 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 5.45. Regression von DIF Ungarn - Niederlande auf Anforderungsmerkmale (Deutsch-Items)

nisse 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 aller 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-

	Modell 1	**Modell 2**	**Modell 3**
	Beta(Sig.)	Beta(Sig.)	Beta(Sig.)
Itemtyp Banked Multiple Choice	**-.261(.072)***	-.160(.134)	**-.166(.095)***
Itemtyp Richtig-Falsch	-.339(.181)		
Itemtyp Multiple Matching	**-.368(.156)** ≠		
Itemtyp Kurzantwort	**-.383(.098)***	-.065(.683)	.154(.121)
Itemtyp Lückentext	-.249(.108)		
Schlussfolgern/Erkennen (Info 1)	.107(.606)		
Implizit/explizit(Info 2)	.275(.250)		
Hauptidee/Detail (Info 3)	-.038(.873)		
Authentischer Text	-.078(.690)		
Inhalt haupts. konkret	.184(.322)		
Inhalt teilweise abstrakt	-.223(.276)		
Vokabular hauptsächlich häufig/einfach	.509(.177)		
Vokabular teilweise ausgeweitet/selten	.401(.333)		
Vokabular ausgeweitet/selten	-.010(.987)		
Haupts.einfache grammatikalische Strukturen	.309(.175)		
Teilweise komplexe grammatische Strukturen	.279(.495)		
Komplexe grammatische Strukturen	***.976(.061)****	***.321(.046)****	
R^2	.329	.108	.055

Abhängige Variable: DIF Schweden - Ungarn; 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 5.46. 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.

Die Beibehaltung der Hypothese ist auch hier abhängig von der jeweils betrachteten Länderpaarung. Insgesamt werden in den meisten Modellen nur wenige der verwendeten Prädiktoren überhaupt signifikant. Diese entsprechen in den meisten Fällen 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

Basierend auf den im Ergebnisteil dargestellten Resultaten der Einzelfragestellungen soll in diesem Teil der Arbeit die Hauptfragestellung der vorliegenden Dissertation beantwortet werden. Dazu werden zunächst die Ergebnisse der Einzelfragen in Kürze zusammengefasst und interpretiert. In einem zweiten Schritt ist dann deren Bedeutung für die Hauptfragestellung zu eruieren. Darauf folgend soll die Relevanz der Ergebnisse der Haupt- und ggf. auch der Einzelfragestellungen für die unterschiedlichen Theorieströnge dieser Arbeit sowie für die Konstruktion von Testverfahren dargestellt und diskutiert werden. In einem letzten Teil soll ferner auf Kritik und Grenzen dieser Arbeit eingegangen sowie ein Ausblick auf zukünftige Forschungsfragestellungen gegeben werden.

6.1. Zusammenfassung und Interpretation der Ergebnisse

Im Folgenden werden die Ergebnisse der drei Fragekomplexe „Voraussetzungen und Skalierbarkeit", „Erklärung der Itemschwierigkeiten innerhalb der Länder" und „Erklärung von Differentiellen Item Funktionen" jeweils kurz zusammengefasst und dann bezüglich ihrer Bedeutung für diese Arbeit interpretiert. Dabei soll an dieser Stelle nicht mehr auf Details der Einzelergebnisse eingegangen werden, sondern diese in ihrer Gesamtheit dargestellt und interpretiert werden.

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

Dieser erste Komplex von Fragen befasste sich mit den zur Beantwortung der Hauptfragestellung notwendigen Voraussetzungen. Dabei wurde zunächst im Rahmen von Frage 1a die Rasch-Skalierbarkeit der Deutsch- bzw. Englisch-Items innerhalb der einzelnen Länder überprüft. Diese Prozedur basiert auf den Empfehlungen von Maris, Bechger & Veldhuizen (2006), welche besagen, dass zunächst die Modellpassung innerhalb der Länder überprüft werden sollte. Ferner wurde als Modell ein 1PL-IRT-Modell gewählt, da dieses auch den Skalen des GERS

zugrundeliegt (North, 2000; Schneider & North, 2000;Europarat, 2001).

Bezüglich der Englisch-Items zeigte sich, dass über die vier analysierten Länder hinweg nur insgesamt 9 der Items als möglicherweise problematisch angesehen werden können. Das heißt, dass sie entweder einen t-Wert von +/-2 über- bzw. unterschritten, oder aber signifikant von dem unter Annahme des Rasch-Modells erwarteten Parameter abwichen. Bezüglich der Entscheidung, ob ein Item aufgrund eines ungenügenden Modellfits für weitere Analysen ausgeschlossen wird, wurden die Kriterien von Adams und Khoo (1996) zugrunde gelegt, welche besagen, dass der absolute WMNSQ-Wert 0.75 bzw. 1.33 nicht unter- bzw. überschreiten sollte. Dies war bei keinem Item der Fall, somit konnten sämtliche Items für weitere Analysen verwendet werden.

Die Ergebnisse der Rasch-Analysen lassen sich dahingehend interpretieren, dass die anhand der Items erfasste Fähigkeit innerhalb der Länder insgsamt als eindimensional einzustufen ist. Das bedeutet, dass davon auszugehen ist, dass die Items, auch wenn sie aus unterschiedlichen Ländern stammen, innerhalb eines jeweiligen Landes dieselbe Dimension bzw. dieselbe Fähigkeit erfassen. Diese Voraussetzung für die Verwendung der Daten für weitergehende Analysen konnte demnach als gegeben angesehen, und die so gewonnenen Itemschwierigkeitsparameter für weitere Analysen verwendet werden. Die Annahme, dass die Items innerhalb der Länder dem Rasch-Modell entsprechen, konnte somit bestätigt werden. Diese Ergebnisse replizieren ferner die Ergebnisse der den Daten zugrundeliegenden EBAFLS-Studie (Fandel et al., 2007; CITO, 2008). 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

von Empfehlungen von Scheuneman und Gerritz (1990) auch die nicht-signifikanten Parameter mit ein.

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.

Die dritte wichtige Voraussetzung zur Beantwortung der Hauptfragestellung betraf die Existenz unterschiedlicher Testkulturen sowie deren Feststellbarkeit anhand der von den Ländern im Rahmen der EBAFLS-Studie eingereichten Items. Testkultur wurde dabei definiert als die Häufigkeit des Vorkommens schwierigkeitsbestimmender Itemeigenschaften bei den Items eines jeweiligen Landes. Diese Methode ist angelehnt an die von Klieme und Baumert (2001) verwendete Methode zur Analyse des Zusammenhangs zwischen Testkulturen und DIF. Zur Feststellung der Testkulturen wurden mehrere Schritte durchlaufen, beginnend bei der Einordnung der Items hinsichtlich ihrer schwierigkeitsbestimmenden Merkmale mit Hilfe von internationalen Experten. Im letzten Schritt wurde geprüft, ob signifikante Unterschiede hinsichtlich der Häufigkeit des Vorkommens eben dieser Itemmerkmale bei den Items unterschiedlicher Länder existieren.

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.
Die Ergebnisse hinsichtlich der Existenz von zu erwartenden Stärken und Schwächen von Gruppen sowie deren Analysierbarkeit anhand von Testitems gehen mit einschlägigen Forschungsarbeiten konform (Klieme & Baumert, 2001; Dogan, Guerrero & Tatsuoka, 2005). Somit kann die dritte Voraussetzung für die Beantwortung der Hauptfragestellung als erfüllt angesehen werden.

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-Profile 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.

Auch bezüglich der DIF-Analysen muss auf die Nicht-Repräsentativität der Länder-Stichproben hingewiesen werden, weshalb die Ergebnisse möglicherweise nur eingeschränkt generalisierbar sind und daher in weiteren Studien überprüft werden sollten. Darüberhinaus sollte auch hinsichtlich der Generalisierbarkeit der gefundenen Testkulturen und somit der von den Gruppen zu erwartenden Stärken und Schwächen eher vorsichtig umgegangen werden: Obgleich die Repräsentativität der Items bezüglich der in den unterschiedlichen Ländern üblicherweise verwendeten Testformate eine Voraussetzung für deren Verwendung in der EBAFLS-Studie war, ist im Rahmen der vorliegenden Arbeit eine letztendliche und valide Überprüfung nicht möglich. Auch hier wäre daher eine erneute Überprüfung der Ergebnisse angemessen.

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-

ser Methoden basiert auf den Forschungsarbeiten von Scheuneman & Gerritz (1990) sowie von Klieme und Baumert (2001). Zweck der Analysen innerhalb der Länder waren die Frage nach der Güte des in der vorliegenden Arbeit verwendeten Item-Kategorisierungsinstruments „Dutch Grid" (Alderson et al., 2006) sowie die Frage danach, inwieweit die innerhalb der Länder gefundenen korrelativen Zusammenhänge vergleichbar sind.

In einem ersten Schritt wurden für jedes Land Pearson Produkt-Moment-Korrelationskoeffizienten für den Zusammenhang zwischen den in Frage 1a berechneten Itemschwierigkeitsparametern 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 $r = -.30$ ($p \leq 0.01$) und $r = .431$ ($p \leq 0.01$) 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.

Bezüglich der Englisch-Items zeigt sich ferner, dass sich die Zusammenhänge zwischen der jeweiligen Itemschwierigkeit und den schwierigkeitsbestimmenden Itemmerkmalen über die Länder hinweg hinsichtlich Richtung und Größe ähneln. So lässt 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- Ausprägung „ausschließlich häufig" finden, was darauf hinweist, dass diese Itemmerkmale eher mit einer niedrigen Itemschwierigkeit einhergehen. Eine signifikant positive Korrelation zeigt sich hier hingegen in allen Ländern zwischen der Itemschwierigkeit und dem Itemtyp „Lückentext", der Ausprägung „hauptsächlich abstrakt" der Variablen „Abstraktheit des Inhalts" und der Ausprägung „ausschließlich einfache Strukturen" der Variablen „Grammatische Strukturen". Dies deutet darauf hin, dass das Vorhandensein dieser Variablenausprägungen bei allen Ländern mit einer höheren Itemschwierigkeit einhergeht.

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-

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

Ob diese Ähnlichkeit auch einer Überprüfung mit Hilfe von Signifikanztests standhält, wurde in Frage 1b beantwortet. Dort zeigte sich, dass sich, bis auf einige wenige Ausnahmen, die Korrelationen über die Länder hinweg nicht signifikant voneinander unterscheiden. Dies weist zum einen darauf hin, dass das gewählte Itemkategorisierungs-Instrument „Dutch Grid" in allen Ländern gleichermaßen zum Einordnen von Zusammenhängen geeignet zu sein scheint und dass zum anderen auch die dort verwendeten Schwierigkeitsabstufungen wie erwartet zur Schwierigkeit (bei „schwierigeren" Abstufungen wie beispielsweise „seltenes Vokabular") bzw. Leichtigkeit (bei „einfacheren" Abstufungen wie „ausschließlich einfaches Vokabular") beitragen. Bezüglich der schwierigkeitsbestimmenden Merkmale mit drei oder vier Ausprägungen zeigte sich außerdem, dass einige Kategorien in keinem der Länder einen Zusammenhang aufweisen. Das könnte möglicherweise mit einer zu feinen Abstufung der Merkmale erklärbar sein, so dass die Rater nicht mehr eindeutig zwischen den unterschiedlichen Stufen unterscheiden können.

Die Ergebnisse für die Deutsch-Items ähneln denen der Englisch-Items: Die Korrelationen liegen hier zwischen $r = -.444$ ($p \leq .01$) und $r = .505$ ($p \leq .01$) 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.

Bezüglich der Englisch-Items fällt auf, dass im Rahmen der multiplen Regressionsanalysen in allen Ländern ein signifikant negativer Zusammenhang zwischen der Itemschwierigkeit und dem Merkmal „teilweise komplexe grammatische Strukturen" zu finden ist. Möglicherweise handelt es sich hierbei um einen sprachspezifischen Effekt. Die Items wurden von den Ratern zwar als verhältnismäßig schwer hinsichtlich der Grammatik eingeschätzt, da jedoch die englische Sprache insgesamt eine verhältnismäßig einfache Grammatik im Gegensatz zu beispielsweise der deutschen Sprache besitzt, wäre es möglich, dass Schüler, die aus Ländern stammen, deren Landessprache eine schwerere Grammatik zugrunde liegt, auch die als schwer eingestuften Englisch-Items als leicht empfinden. Allerdings weisen die Einzelkorrelationen dieser Variablen keinen signifikanten Zusammenhang zur Itemschwierigkeit auf, was möglicherweise für einen Suppressionseffekt sprechen kann. Jedoch existiert auch in fast allen Ländern eine signifikant positive Einzelkorrelation zwischen der Itemschwierigkeit und der Variable „ausschließlich einfache grammatische Strukturen". Möglicherweise handelt es sich hier auch um einen durch das häufige Vorkommen von geschlossenen Antwortformaten bedingten Rateeffekt: Items mit schwieriger Grammatik sind auch tatsächlich schwerer zu lösen. Daher raten möglicherweise Schüler, denen das Item zu schwer ist bei der Beantwortung dieses Items häufiger. Das kann zu dem paradoxen Phänomen führen, dass bei einfachen Items aufgrund der Tatsache, dass diese meist ohne zu

raten bearbeitet werden mehr Fehler gemacht werden als bei den schweren Items. Die Variablen mit der größten Bedeutung für die Itemschwierigkeit scheinen insgesamt die Schwierigkeit der grammatischen Strukturen, die Häufigkeit des Vokabulars und teilweise der Itemtyp zu sein. Diese Ergebnisse gehen konform mit den Ergebnissen von einschlägigen Forschungsarbeiten zur Erklärung der Itemschwierigkeit bei fremdsprachlichen Leseverständnis-Items wie etwa von Bachman, Davidson & Milanovic (1996), oder Fortus, Coriat & Fund (1998, zitiert aus Grotjahn, 2000), obgleich dort der Anteil aufklärbaraer Varianz etwas höher lag.

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-siginifikanten 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 ($p \leq$) auf.

Eine Möglichkeit wäre, Variablen zusammenzufassen, oder von den Analysen auszuschließen. Dies ist im Rahmen dieser Arbeit nicht geschehen, da es ein Ziel der Arbeit war, Informationen über die unterschiedlichen Ausprägungen der Variablen und deren Zusammenhänge zur Itemschwierigkeit zu gewinnen. In weiteren Analysen sollte jedoch überprüft werden, ob das Zusammenfassen von Variablen dazu beiträgt, Suppressions- und Multikollinearitätseffekte zu vermindern, und ferner möglicherweise weniger Ausprägungen der Variablen ausreichen um die Schwierigkeit eines Items hinsichtlich dieses Merkmals zu beschreiben. Dafür spricht auch, dass sich innerhalb der Länder in der Regel nicht bei allen Merkmalsausprägungen signifikante Korrelationen mit der Itemschwierigkeit finden lassen.

Die Ergebnisse der multiplen Regressionsanalysen sind insgesamt ein weiterer Hinweis darauf, dass die im Dutch Grid verwendeten Merkmale zur Erklärung der Itemschwierigkeit zumindest zum Teil geeignet zu sein scheinen. Da jedoch mit Hilfe dieser Merkmale in allen Ländern nur ein Teil der Itemschwierigkeitsvarianz aufgeklärt werden konnte, ist zu vermuten, dass noch weitere schwierigkeitsbestimmende Merkmale existieren, die entweder der inhaltlichen Kategorie des „Dutch Grid" zugehörig, oder aber gar nicht Teil der „Dutch Grid"-Kategorien sind. Diese Frage sollte in weitergehender Forschung näher untersucht werden.

Insgesamt lässt sich für diesen Komplex von Fragen, der sich mit der Güte des zur Itemkategorisierung verwendeten Instruments „Dutch Grid" und der dort verwendeten Merkmale befasst, feststellen, dass das Instrument insegesamt für die Einordnung von Items hinsichtlich deren Anforderungsmerkmale geeignet zu sein scheint. So zeigen sich in allen Ländern und in beiden Sprachen Zusammenhänge zwischen der Itemschwierigkeit und den schwierigkeitsbestimmenden Merkmalen. Die wichtigsten Merkmale über alle Länder hinweg scheinen dabei Grammatik, Vokabular und Itemtyp zu sein, jedoch weisen auch die restlichen Merkmale jeweils in einzelnen Ländern Zusammenhänge zur Itemschwierigkeit 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-Variablen, 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, stichprobenspezifische 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 unterschätzte 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.

Bei der Paarung Ungarn-Schweden zeigen sich keine signifikanten Zusammenhänge. Da bei dieser Paarung 39.4% aller Items signifikante DIF-Parameter aufweisen (siehe Ergebnisse zu Frage 1b), wäre auch hier eine denkbare Erklärung, dass zwischen diesen beiden Ländern entweder andere, im Rahmen dieser Arbeit nicht mit erfasste testkulturelle Merkmale auf die Unterschiedlichkeit der Itemschwierigkeiten Einfluss nehmen, oder aber dass auch hier aufgrund der Nicht-Repräsentativität der Stichproben der Einfluss der testkulturellen Merkmale entweder unterschätzt wird oder aber andere, stichprobenspezifische Merkmale eine Rolle spielen. Ferner weisen die DIF-Parameter dieser Paarung insgesamt die geringste Varianz auf, daher könnte hier außerdem eine Varianzeinschränkung für die Unterschätzung von Zusammenhängen ursächlich sein. Eine weitere mögliche Erklärung ergibt sich aus der Betrachtung der Herkunft der Schüler. Sowohl in Schweden, als auch in den Niederlanden sind 7% der Schüler nicht in dem jeweiligen Land geboren. Möglicherweise hat daher bei diesen beiden Ländern die Testkultur einen geringeren Einfluß, vor allem wenn diese beiden Länder gemeinsam analysiert werden.

Für den größten Teil der Analysen lässt sich die Annahme, dass die Itemherkunft und damit auch die Testkultur einen signifikanten korrelativen Zusammenhang in erwarteter Richtung aufweist, bestätigen. Bei einem großen Anteil der Paarungen geht die Tatsache, dass ein Item aus dem eigenen Land stammt, signifikant mit einer im Vergleich zur jeweils anderen Gruppe geringeren Itemschwierigkeit einher und umgekehrt: Wenn ein Item aus dem jeweils anderen Land stammt, zeigt sich eine signifikant höhere Itemschwierigkeit. Bei einem weiteren Teil der Paarungen ist, bis auf wenige Ausnahmen, zumindest einer dieser beiden Zusammenhänge zu beobachten. Diese Ergebnisse können als ein erster deutlicher Hinweis auf den Einfluss von Itemherkunft und Testkultur auf die kulturell bedingte Varianz der Itemschwierigkeit interpretiert werden.

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-

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

Fünf der 29 Korrelationen sind signifikant, obgleich aufgrund der Testkulturen eigentlich keine Korrelation zu erwarten war, da sich die betroffenen Gruppen bezüglich des Vorkommens eines Merkmals bei den eigenen Items nicht unterscheiden. Eine mögliche Erklärung könnte eine Nicht-Repräsentativität der Items aus den jeweiligen Ländern hinsichtlich dieses einen Itemmerkmals sein, weshalb dieses im Rahmen dieser Arbeit nicht als testkulturspezifisches Merkmal mit aufgenommen wurde. Lediglich eine einzige Korrelation weist eine den Hypothesen gegenläufige Richtung auf (DIF Ungarn-Spanien mit der Variablen „Inhalt hauptsächlich konkret"). Dies könnte auch hier auf die Nicht-Repräsentativität der Stichprobe, die Nicht-Repräsentativität der Items des Landes hinsichtlich dieses Merkmals oder auf eine falsche Einordnung des Items bezüglich dieses Merkmals seitens der Rater zurückzuführen sein.
Die Variablen „Itemtyp" und „Abstraktheit des Inhalts" weisen bei den Englisch-Items insgesamt die meisten signifikanten Zusammenhänge mit DIF-Parametern auf. Ferner spielen auch noch „Vokabular" und „Informationsgewinn 1" eine Rolle. Nur jeweils eine Korrelation weisen die Variablen „Informationsgewinn 2" und „Grammatik" auf. Dies spricht möglicherweise dafür, dass die Schwierigkeit grammatischer Strukturen zwar für die Itemschwierigkeit innerhalb der Länder eine Rolle spielt, nicht so sehr jedoch für Unterschiede in der Itemschwierigkeit zwischen den Ländern. Wie im Anhang einsehbar ist, sind die Varianzen der DIF-Parameter (zwischen 0.07 und 0.23 Logits) insgesamt durchweg relativ niedrig. Daher wäre es auch hier möglich, dass vorhandene Zusammenhänge unterschätzt werden. Umso stärker sind daher jedoch auch die gefundenen Zusammenhänge zu interpretieren.

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.

Lediglich zwei der Korrelationen laufen hinsichtlich ihrer Richtung den aufgrund der Testkulturen getroffenen Annahmen entgegen. Beide finden sich bei der Länderpaarung Schweden-Ungarn. Auch hier könnten mögliche Ursachen die Nicht-Repräsentativität der Stichprobe, die Nicht-Repräsentativität der Items des Landes hinsichtlich dieses Merkmals oder eine falsche Einordnung seitens der Rater sein. Bei den beiden Länder-Paarungen Frankreich-Niederlande und Schweden-Ungarn zeigen sich insgesamt nur wenige signifikante Zusammenhänge zwischen den DIF-Parametern und den Itemeigenschaften. Ferner sind dort die gefundenen Korrelationen entweder, wie bereits erwähnt, entgegen der Hypothese oder aber sind signifikant, obgleich aufgrund der Testkulturen eigentlich kein Zusammenhang zu erwarten war. Neben den oben bereits genannten möglichen Erklärungen wie Varianzeinschränkung oder Nicht-Repräsentativität der Items oder der Stichproben könnte eine weitere Erklärung in nicht mit Hilfe des „Dutch Grid" erfassbaren, länderkombinationsspezifischen Ursachen liegen. Dies ist im Rahmen der vorliegenden Arbeit jedoch nicht überprüfbar. Diesbezüglich wären daher weitere Forschungsvorhaben notwendig.

Ein weiteres beachtenswertes Ergebnis ergibt sich aus der gemeinsamen Betrachtung der eben dargestellten Korrelationen sowie der Korrelationen der itemschwierigkeitsbestimmenden Merkmale mit der Itemschwierigkeit innerhalb der Länder: Wie oben bereits dargestellt wurde, unterscheiden sich die Korrelationen der Länder hinsichtlich der Zusammenhänge zwischen Itemschwierigkeit und Itemmerkmalen meist nicht und folgen größtenteils den theoretisch erwarteten Schwierigkeitsabstufungen der Variablen (siehe Ergebnisse zu Frage 2c). Werden hingegen kulturell bedingte Unterschiede der Itemschwierigkeit, nämlich DIF, in Zusammenhang mit den Itemmerkmalen gebracht, folgen diese hinsichtlich ihrer Richtung teilweise eher den aufgrund der Testkulturen erwarteten Stärken und Schwächen der Länder.
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-

gnifikant 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 ($r = -.304$; $p \leq .01$), 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 ($r = .213$; $p \leq .05$ bzw. $r = .392$; $p \leq .01$), 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 ($r = -.276$; $p \leq .01$). 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 Testkultur 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 Differentiellen 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"

Prädiktoren, die zwar Signifikanz aufweisen, jedoch aufgrund der testkulturellen Hypothesen eigentlich keinen Zusammenhang aufweisen sollten. Gleichfalls vom Basis-Modell ausgehend folgte in einigen Fällen noch die Berechnung eines dritten Modells. Dazu wurden ausschließlich die in Modell 1 signifikant gewordenen Prädiktoren, die der Richtung nach den Testkulturen entsprechen, verwendet. Somit kann mit dem Endmodell eine Aussage dahingehend gemacht werden, welcher Anteil der kulturell bedingten Varianz sicher auf die unterschiedlichen, testkulturell bedingten Stärken und Schwächen der Länder zurückführbar ist.

Bezüglich der Englisch-Items zeigte sich, dass —je nach Länderpaarung— mit Hilfe der Endmodelle zwischen 22.7 % und 4.1% der Varianz durch testkulturell bedingte Stärken und Schwächen derLänder aufgeklärt werden konnten. Bei Hinzunahme der hier aufgrund ihrer nicht eindeutigen Interpretierbarkeit als „neutral" bezeichneten Prädiktoren (d.h. ein Prädiktor wird signifikant, obgleich laut Hypothese hier eigentlich kein Zusammenhang bestehen sollte) erhöht sich teilweise der Anteil aufgeklärter Varianz deutlich (z.B. von 5.6% auf 25.6% für die Paarung Deutschland—Spanien). Da es möglich ist, dass es sich bei den „neutralen" Prädiktoren um einen verdeckten testkulturellen Effekt handelt, der aufgrund von Nicht-Repräsentativität der in EBAFLS verwendeten Items eines Landes bezüglich eines bestimmten Merkmals oder aber wegen einer ungenügend genauer Einordnung der Items seitens der Rater nicht in das Testkultur-Profil des Landes mit aufgenommen wurde, wird der Anteil der Varianz, der durch Modelle erklärt wird, welche die „neutralen" Prädiktoren beinhalten, als eine mögliche Obergrenze der aufgrund der Testkultur aufklärbaren Varianz interpretiert. Die Endmodelle mit ausschließlich den Hypothesen entsprechenden Prädiktoren stellen dann diesbezüglich hingegen die Untergrenze dar.

Einige wenige der zu Beginn signifikanten Prädiktoren widersprachen hinsichtlich ihrer Richtung den testkulturellen Hypothesen. Dies stellte jedoch im Vergleich zu der Anzahl der Prädiktoren, die der Hypothese entsprechen, den deutlich kleineren Anteil dar. Bezüglich der Englisch-Daten fiel insgesamt auf, dass ein deutlich geringerer Anteil der Varianz aufgeklärt werden konnte in Paarungen, an denen die spanische Stichprobe beteiligt war. Hier ist denkbar, dass stichprobenspezifische Effekte der spanischen Gruppe eine Rolle spielen, die testkulturelle Effekte überlagern, oder aber dass für die spanische Stichprobe testkulturelle Aspekte, die nicht im Rahmen des „Dutch Grid" erfasst werden, eine Rolle spielen. Hier wäre eine genauere Erforschung unter Hinzunahme etwa von inhaltlichen Merkmalen, wie beispielsweise das in der Aufgabe behandelte Thema, in der Zukunft wünschenswert.

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.

Ferner stellten sowohl bezüglich der Deutsch- als auch der Englisch-Daten nicht alle der in den Korrelationsanalysen signifikanten Zusammenhänge (Frage 3a) auch signifikante Prädiktoren in der multiplen Regression dar. Dies spricht für eine hohe Interkorrelation der Merkmale untereinander. Auch scheinen in einem Teil der Analysen Multikollinearitätsprobleme und Suppressionseffekte eine Rolle zu spielen. So werden häufig vormals signifikante Prädiktoren bei Ausschluss von nicht-signifikanten Variablen in weiteren Analysen nicht mehr signifikant. Auch zeigen sich bei den Regressionen bezüglich der Deutsch-Items beta-Koeffizienten größer 1.

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.

Der Anteil der durch testkulturell bedingte Stärken und Schwächen der Gruppen aufklärbaren DIF-Varianz variiert je nach Länderpaarung stark. Vermutlich kann nicht für alle Paarungen gleichermaßen DIF auf die hier verwendeten Variablen zurückgeführt werden. Die gefundenen Ergebnisse müssen aufgrund der Nicht-Repräsentativität der Stichproben mit Vorsicht interpretiert werden. Allerdings spricht für einen testkulturellen Einfluss, dass der größere Teil der signifikanten Prädiktoren auch den aufgrund der Testkulturen aufgestellten Hypothesen entsprechen. Bezüglich der Englisch-Items erwiesen sich über die Regressionsanalysen hinweg 25 der Prädiktoren als signifikant, 14 davon entsprachen den Hypothesen, sechs der Prädiktoren zeigten sich neutral und fünf Prädiktoren wiesen ein den Hypothesen entgegengesetztes Vorzeichen auf. Bei den Deutsch-Items wiesen insgesamt 39 der Prädiktoren Signifikanz auf, 19 davon entsprachen den Hypothesen, zehn der Prädiktoren zeigten sich neutral und zehn Prädiktoren zeigten sich den Hypothesen entgegengesetzt. Die Ergebnisse sprechen insgesamt dafür, dass DIF zumindest teilweise, je nach Länderpaarung mehr oder weniger gut, mit Hilfe von testkulturbedingten Stärken und Schwächen der Länder erklärt werden kann. Diese Ergebnisse gehen konform mit Ergebnissen der Forschungsarbeit von Scheuneman & Gerritz (1990) sowie Li, Cohen und Ibarra (2004), die gleichfalls DIF, letztere allerdings bei Mathematik-Items, anhand von Charakteristika von Items vorhersagen konnten. Die Ergebnisse entsprachen dort gleichfalls den erwarteten Stärken und Schwächen der Gruppen.

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 Testkultur-Profile 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.

Viertens: Obgleich lediglich ein Teil der Korrelationen und Prädiktoren Signifikanz aufweist, ist über sämtliche Analysen hinweg zu beobachtens, dass der größte Teil der signifikanten Korrelationen und Prädiktoren auch den aufgrund der Testkultur aufgestellten Hypothesen entsprechen. Dies verringert die Möglichkeit, dass es sich hier um zufällige Effekte in Richtung der Testkulturen handelt.

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.

Obgleich, wie oben dargestellt, der größte Teil der signifikanten Korrelationen zwischen DIF und den Testkultur-Indikatoren in die aufgrund der erwarteten Stärken und Schwächen erwartete Richtung deuten, gibt es jedoch einige wenige, die den Annahmen widersprechen. Gleiches gilt für die beta-Gewichte der multiplen Regressionen. Ferner konnte bei einigen der Länder-Paarungen nur ein sehr geringer Anteil der kulturell bedingten Varianz anhand der verwendeten Prädiktoren aufgeklärt werden. Daher ist für diese Länder-Paarungen die Fragestellung nicht eindeutig positiv zu beantworten. Aus diesem Grund müssen die Ergebnisse, wie bereits erwähnt, vorsichtig interpretiert werden. Es ist keine Generalsierbarkeit gegeben, und die Ergebnisse sollten anhand größerer, repräsentativer Stichproben, möglicherweise auch in weiteren Sprachen und Ländern, repliziert werden.

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.

Betrachtet man nun die Ergebnisse der vorliegenden Arbeit, so bestätigen sie diesen zweiten Ansatz zumindest teilweise. Die Korrelationen der schwierigkeitsbestimmenden Itemmerkmale in der aufgrund der Testkulturen angenommenen Richtung sprechen dafür, dass testkulturell bedingte Stärken und Schwächen der Gruppen bei der Entstehung von DIF durchaus eine Rolle spielen. Ferner liefern die multiplen Regressionen von DIF auf die Itemmerkmale in dieser Hinsicht einen weiteren, wenn auch mit Vorsicht zu interpretierenden Hinweis: Auch dort entsprechen die den testkulturellen Hypothesen entsprechenden Beta-Koeffizienten genau eben jenen Stärken und Schwächen der Gruppen. Das bedeutet, die aufgrund der Testkulturen erwarteten Stärken und Schwächen entsprechen hier den beobachteten.

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 konstruktrelevanten Dimension, wie beispielsweise grammatische Strukturen oder Häufigkeit des Vokabulars.

Dennoch kann DIF nur teilweise durch solche unterschiedlichen testkulturellen Prägungen der Gruppen erklärt werden. Dies mag durchaus auch darin begründet sein, dass auch konstruktirrelevante, 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 konstruktrelevanter Itemmerkmale für die Modellierung von DIF. Es kann gemutmaßt werden, dass letztendlich ein Teil von DIF durch konstruktrelevante, differentielle Stärken und Schwächen der Gruppen zustande kommt, ein weiterer Teil jedoch durch die oben genannte „Nuisance Dimension" und konstruktirrelevante Multidiminsionalität.

Obgleich dies hier nicht abschließend geklärt werden kann, so ist insgesamt an dieser Stelle aufgrund dieses Ergebnisses der diagnostische Nutzen von DIF doch weiter in den Vordergrund gestellt worden. Es bestätigt ferner die Ergebnisse von Scheuneman & Gerritz (1990), von Klieme und Baumert (2001) sowie Dogan, Guerrera und Tatsuoka (2005), die in ihren jeweiligen Studien gleichfalls DIF teilweise auf differentielle Stärken und Schwächen der untersuchten Gruppen zurückführen konnten. Diese Ergebnisse in Kombination mit den Ergebnissen dieser Arbeit sprechen klar dafür, dass DIF zumindest teilweise als unterschiedliche Profile von Stärken und Schwächen von Gruppen interpretiert werden können. Dogan, Guerrera und Tatsuoka (2005) beschreiben die Bedeutung von DIF hinsichtlich diagnostischer Überlegungen in den Fällen, in denen ein Zusammenhang zwischen Stärken und Schwächen von Gruppen besteht, wie folgt:

„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).

Dies weist auf die diagnostische Nutzbarkeit von DIF hin, vor allem mit Hinblick auf die Interpretation und Bewertung von Gruppenunterschieden. Gerade die Tatsa che, dass unter Einbezug der Ergebnisse dieser Arbeit gezeigt werden kann, dass die teilweise systematische Rückführbarkeit von DIF auf Stärken und Schwächen von Gruppen auch in unterschiedlichen Disziplinen wie Mathematik (Dogan, Guerrera & Tatsuoka, 2005; Klieme & Baumert, 2001) und Leseverständnis (Abbott, 2004; Artelt & Baumert, 2001) beziehungsweise fremdsprachlichem Leseverständnis (in dieser Arbeit) beobachtbar ist, deutet auf eine mögliche Generalisierbarkeit eines diagnostischen Nutzens von DIF sowie eines kulturspezifischen Einflusses auf die Itemschwierigkeit hin. Dieses Gesamtbild sollte möglichst durch eine Replikation der Ergebnisse in weiteren Disziplin-

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.

Auch die Frage nach der Ähnlichkeit einzelner Wörter zur Muttersprache (Chen & Henning, 1985) oder die Verwendung bestimmter umgangssprachlicher Ausdrücke (Sasaki, 1991) zur Erklärung von DIF werden verschiedentlich als relevant für die Erklärung Differentieller Item Funktionen erachtet. Diese Variablen wurden in der vorliegenden Arbeit nicht berücksichtigt. Abbott (2004) kritisiert das Nicht-Verwenden inhaltlicher Merkmale für die Erklärung von DIF.

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 van 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 Skalendeskriptoren 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.

Insbesondere jedoch sind die Ergebnisse der vorliegenden Arbeit für das auf dem GERS basierende Itemkategorisierungsinstrument „Dutch Grid” (Alderson et al., 2006) relevant. Zumindest hinsichtlich der darin zur Kategorisierung von Items verwendeten kognitiv-linguistischen Itemmerkmale zeigen die Ergebnisse, dass das Instrument für die Kategorisierung von Items geeignet zu sein scheint. Dafür spricht zum einen, dass innerhalb aller Länder Korrelationen zwischen einem Teil der Itemmerkmale des „Dutch Grid” und den jeweiligen Itemschwierigkeiten existieren. Zum anderen zeigt auch die Tatsache, dass die verwendeten Itemmerkmale zwischen 20% und 50% der Varianz der Itemschwierigkeit innerhalb der Länder aufklären können, dass die kognitiv-linguistischen Itemmerkmale insgesamt als Prädiktoren der Itemschwierigkeit geeignet sind. Bisher gab es zwar erste Versuche, einige „Dutch Grid”-Itemmerkmale mit höheren Schwierigkeitsgraden von Items in Verbindung zu bringen (Noijons & Kuijper, 2006), diese sind aber

eher von deskriptiver Natur. Ferner ist bis dato auch noch nicht der Versuch gemacht worden, die Itemschwierigkeit anhand der „Dutch Grid"-Merkmale per multiper linearer Regression vorherzusagen. Da die Ergebnisse dieser Dissertation zur Validierung des Instruments beitragen können, sind diese für das „Dutch Grid"-System und dessen künftige Verwendung bei der Einordnung und Konstruktion von Testitems fremdsprachlichen Leseverständnisses von Bedeutung.

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.

Da dem GERS —und somit auch dem „Dutch Grid"— ferner gängige Theorien aus dem Bereich der Fremdsprachenforschung (z.B. Bachman, 1990; Bachman & Palmer, 1996; Alderson, 2000) zugrunde liegen, die gleichfalls eine der Ursache für unterschiedliche Leistung auf Seiten der Items und deren Merkmalen sehen, sind die Ergebnisse außerdem, wenn auch indirekt, eine Unterstützung dieser Theorien im Bezug auf die Wichtigkeit dieser itemseitigen Kategorie für die Determination von Itemschwierigkeit und Schülerleistung.

Neben der Bedeutung der Ergebnisse für den GERS und den „Dutch Grid" kann die Arbeit darüber hinaus auch zur Beantwortung der Frage, welche Merkmale die Schwierigkeit eines Items im Bereich des Fremdsprachentestens mitbestimmen, einen Beitrag leisten. Bisherige empirische Ergebnisse im Bereich der Determination von Itemschwierigkeiten zeigten, dass zum Teil ähnliche wie die in der vorliegenden Arbeit verwendeten Prädiktoren wie beispielsweise Worthäufigkeit, Abstraktheit, Vokabular und Grammatik (Perkins & Linville, 1987; Freedle & Kostin, 1993; Bachman, Davidson & Milanovic, 1996) einen Zusammenhang zur Itemschwierigkeit aufweisen. Die Ergebnisse der vorliegenden Arbeit bestätigen dies. Auch bezüglich des Anteils der anhand dieser Prädiktoren aufklärbaren Itemschwierigkeits-Varianz weisen die Ergebnisse dieser Arbeit in eine ähnliche Richtung wie bisherige empirische Studien (siehe auch 2.2.4). Perkins & Lindville (1987) konnten anhand von Itemmerkmalen 30-80% der Itemschwierigkeit aufklä-

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 bestätigen 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 Curricula und Testkulturen der Länder einnehmen. Ob es sich dabei um Struktur- oder Niveauunterschiede handelt, ist dabei, wie oben bereits angesprochen wurde, nicht klärbar.

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 Itemmerkmale einen ähnlichen Zusammenhang mit den Itemschwierigkeiten aufweisen, und zwar sowohl hinsichtlich der Richtung als auch der Größe der Zusammenhänge. Es zeigten sich kaum signifikante 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 Dimensionalität 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 Niveauunterschiede 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.

Zusammenfassend kann konstatiert werden, dass die Arbeit in Bezug auf die oben angesprochenen Bereiche zur Klärung von verschiedenen Fragen hat beitragen können. Einschränkend ist allerdings zu konzedieren, dass keine Aussage darüber gemacht werden kann, ob bestimmte schwierigkeitsdeterminierende Itemmerkmale speziell mit bestimmten Kompetenzniveaus des GERS einhergehen (z.B. Noijons und Kuijper (2006)). Dies ist darauf zurückzuführen, dass in der ursprünglichen EBAFLS-Studie hauptsächlich Items auf Niveau B1, teilweise noch Items auf den Niveaus A2 und B2 verwendet wurden. Die Items sollten sich hinsichtlich ihres Niveaus möglichst nicht unterscheiden, weshalb nicht die gesamte Spannbreite der GERS-Niveaus abgedeckt wurde. Gerade für die Konstruktion von Tests für spezielle Niveaus wäre es jedoch bedeutsam, mehr Informationen über den Zusammenhang zwischen Itemschwierigkeitsdeterminanten und GERS-Niveau zu besitzen.

Auch die Verwendung des GERS als theoretische Grundlage der Arbeit wäre möglicherweise punktuell zu problematisieren. So wurde er von Bausch et al. (2003) beispielsweise hinsichtlich des Vorgehens bei der Konstruktion und Skalierung der Skalen kritisiert. Ferner handelt es sich bei der theoretischen Grundlage des GERS zwar um relevante Modelle und Theorien (z.B. Bachman & Palmer, 1996), jedoch kann der pragmatische Ansatz des GERS, die theoretische Grundlage des Instruments aus unterschiedlichen Theorien zusammenzustellen, durchaus auch mit Skepsis betrachtet werden. So kritisiert Christ (2003) beispielsweise den verwendeten Spracherwerbs- und fremdsprachenlerntheoretischen Ansatz des GERS. Darüberhinaus kritisiert er die ungenügende Validierung des Instruments sowie die Tatsache, dass bei der Erschaffung des Instruments beinahe ausschließlich englischsprachige Forschung berücksichtigt wurde.

Andererseits erhält das von den Autoren des GERS gewählte pragmatische Vorgehen hinsichtlich der theoretischen Grundlage des Instruments durch die Tatsache Unterstützung, dass aufgrund der Komplexität und des hohen Anspruchs, der in Bezug auf Ziel und Zweck des Instruments bei dessen Konstruktion gestellt wurde (auch und gerade sprachpolitisch), keine einzelne, auf Fremdsprachenkompetenz bezogene Theorie zur Verfügung stand, die alle unterschiedlichen Bereiche des GERS abdeckt, sondern dass eine theoretische Grundierung nur durch eine Kombination mehrerer Theorien möglich war. Ferner handelt es sich bei den ausgewählten theoretischen Grundlagen (Bachman & Palmer, 1996; Canale & Swain, 1980; van Ek, 1986) um relevante und anerkannte Theorien der angewandten Linguistik. Darüber hinaus wurden die Skalendeskriptoren des GERS empirisch untersucht und erwiesen sich als Rasch-modellkonform.

Ein weiterer möglicher Kritikpunkt an der vorliegenden Arbeit könnte den Umstand betreffen, dass zur Determination der Itemschwierigkeit ausschließlich Itemmerkmale der kognitiv-linguistischen und nicht der inhaltlichen Kategorie des „Dutch Grid" herangezogen wurden. So erachtet beispielsweise Abbott (2004) inhaltliche Itemmerkmale als relevant für die Itemschwierigkeit. Auch können so keine Aussagen über möglicherweise existierende Interaktionen zwischen inhaltlicher und kognitiv-linguistischer Merkmalskategorie gemacht werden.

Die Nichtverwendung dieser Kategorie ist zum einen darauf zurückzuführen, dass die Mehrheit empirischer Forschungsarbeiten die hohe Relevanz von kognitiv-linguistischen Merkmalen empirisch untermauert. Zum anderen waren die Items von den Fremdsprachenexperten der Länder nicht immer vollständig hinsichtlich dieser eher inhaltlichen Itemmerkmale des „Dutch Grid"-Instruments eingeordnet worden. Hätte man es unternommen, eine Vervollständigung durchzuführen, so hätte dies durch das Hinzuziehen zusätzlicher Rater einen Einfluss auf die Validität und Vergleichbarkeit der Ratings gehabt. Darüberhinaus war eine erneute Einordnung auch von dem dazu notwendigen Aufwand im Rahmen dieser Arbeit nicht leistbar. Dennoch ist dies ein Punkt, der prinzpiell zu kritisieren bleibt.

Trotz einiger Einschränkungen sind die Ergebnisse der vorliegenden Arbeit insgesamt für die Forschung im Bereich der fremdsprachlichen Diagnostik von Relevanz. Sie tragen zur Validierung des Instruments „Dutch Grid" bei, und es werden Hinweise auf die Unterschiedlichkeit des Konstrukts „fremdsprachliches Leseverständnis" in verschiedenen Ländern gegeben. Darüberhinaus tragen die Ergebnisse zur Forschung im Bereich der Determination von Itemschwierigkeit bei Items zur Messung fremdsprachlichen Leseverständnisses bei und unterstützen teilweise bisherige empirische Forschung in dieser Disziplin.

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

Wie unter 2.3 dargestellt wurde, ist ein Ziel der interkulturellen Psychologie, zusätzlich zur bloßen Feststellung von Unterschieden, wie es etwa in der EBAFLS-Studie getan wurde, diese auch zu erklären (Van de Vijver & Leung, 1997). Dies war eines der Ziele der vorliegenden Arbeit. Aus diesem Grunde wurde auch der diesbezüglich der interkulturellen Psychologie zugehörige Theoriebereich, nämlich der Bereich der interkulturellen Vergleichbarkeit von Testverfahren, als relevant für diese Dissertation erachtet.

DIF-Analysen sind dabei ein wichtiges Mittel zum Feststellen kultureller Unterschiede. Die Existenz kultureller Unterschiede im Bereich des fremdsprachlichen Leseverständnisses wurde bereits mit Hilfe von auf dem OPLM-Modell (Verhelst, Glas & Verstralen, 1995) basierenden DIF-Analysen in der EBAFLS-Studie (Fandel et al., 2007) festgestellt, welche im Rahmen dieser Arbeit auch mit Hilfe von auf dem Rasch-Modell basierenden Analysen bestätigt wurden. Diese Dissertation ist nun im Vergleich zur EBAFLS-Studie einen Schritt weiter gegangen und hat außerdem einen Fokus auf die Erklärung ebendieser Unterschiede gelegt.

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 = Niveauunterschiede; 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 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

unterschiedlicher Häufigkeit.

Dies lässt zwar auf unterschiedliche Schwerpunkte hinsichtlich einiger Itemmerkmale beziehungsweise Konstruktkomponenten in den unterschiedlichen Testkulturen schließen, es ist vermutlich jedoch in diesen Fällen nicht von einem grundsätzlich strukturellen Unterschied auszugehen. Vielmehr ist anzunehmen, dass diese Art von Unterschieden eher Niveauunterschiede abbildet. Dies sind jedoch nur Mutmaßungen. Wie bereits angesprochen wurde lässt die gegebene Datenstruktur weder die Berechnung von Strukturgleichungsmodellen noch von konfirmatorischen Faktorenanalysen oder Mehrgruppenanalysen zu, weshalb die Art der Unterschiedlichkeit des Konstrukts an dieser Stelle nicht detaillierter erforscht werden kann. Die Betrachtung der unterschiedlichen Testkultur-Profile in der oben beschriebenen Form gibt jedoch einen Hinweis darauf, welche Variablen in welchem Land eine mehr oder weniger große (oder womöglich gar keine) Rolle für das Konstrukt der fremdsprachlichen Lesekompetenz spielen könnte. Die Ergebnisse hinsichtlich der Zusammenhänge zwischen DIF als kulturell bedingten Unterschieden der Itemschwierigkeiten und den Testkulturen bestätigen diese testkulturellen Unterschiede teilweise.
Die vorliegende Arbeit kann insofern für den Bereich der interkulturellen Vergleichbarkeit von Testverfahren relevant sein, als sie aufzeigt, auf welche Art und Weise kulturelle Unterschiede bei Leistungstests möglicherweise untersucht und ihren Ursachen, gerade durch die Verwendung von Differentiellen Item Funktionen als diagnostisches Mittel, auf den Grund gegangen werden kann. Ferner können aufgrund der Testkultur-Profile Vermutungen hinsichtlich der Art der Konstruktunterschiede angestellt werden. Diese Hypothesen könnten anhand von Mehrgruppenanalysen und Strukturgleichungsmodellen im Rahmen zukünftiger Forschung überprüft werden.

Auch hinsichtlich dieses Bereichs muss angemerkt werden, dass die Arbeit nicht auf alle Fragestellungen in wünschenswerter Vollständigkeit antwortet. So können über die genauen Konstruktunterschiede keine Aussagen gemacht werden; die vorgelegten Ergebnisse geben nur gewisse Hinweise. Ein Kritikpunkt an der Arbeit hinsichtlich dieses Bereichs ist die obengenannte Tatsache, dass keine Aussagen über die genauen Konstruktunterschiede hinsichtlich Struktur und Niveau in den unterschiedlichen Ländern gemacht werden können, sondern dass die Ergebnisse dahingehend nur schwache Hinweise wie die oben dargelegten liefern. Auch hinsichtlich des Grades der Äquivalenz können keine Aussagen gemacht werden, obgleich dies gerade bei der Untersuchung von Itembias eine relevante Rolle spielt (Poortinga, 1989; Van de Vijver & Leung, 1997). Auch solche Analysen sind (wie schon an anderer Stelle hinsichtlich anderer Fragestellungen angemerkt) mit dem vorliegenden Datensatz bedauerlicherweise nicht durchführbar.

6.3.4. Relevanz für Theorie und Forschung im Bereich der Validität

Da es sich bei DIF, dem Hauptgegenstand dieser Arbeit, um eine Einschränkung der Validität von Testergebnissen und Testverfahren über unterschiedliche Gruppen hinweg handelt (Holland & Wainer, 1993; Li, Cohen & Ibarra, 2004), wurde als Rahmenthema dieser Arbeit der Bereich der Validität gewählt. Aufgrund dieser Gruppenkomponente ist für diese Arbeit mit dem Validitätsmodell von Messick (1989) die Wahl auf eine Validitätstheorie gefallen, welche gleichzeitig kulturelle und soziale Faktoren und Auswirkungen mit einbezieht. Darüber hinaus spielt dieses Modell im Bereich der Fremdsprachenforschung eine relevante Rolle (McNamara, 2006), weshalb es für diese Arbeit besonders gut geeignet scheint. In Messicks (1989, 1996) Verständnis von Validität müssen neben den klassischen Validitätskennwerten auch die Bedeutung und Interpretation von Testergebnissen sowie die daraus für eine getestete Person resultierenden Folgen valide sein. Messick (1989) setzt Validität insgesamt mit Konstruktvalidität gleich. Demzufolge beinhaltet Konstruktvalidität sowohl den Inhalt als auch Kriterium und Konsequenzen der Interpretation von Testwerten. Konstruktvalidität wird hier betrachtet als ein Set von Indikatoren, die Hinweise auf die Natur des zugrunde liegenden psychologischen Konstrukts geben können.

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.

Ein wichtiger Aspekt, sowohl Messicks Theorie als auch die Erklärung von DIF betreffend, ist die Konstruktrepräsentation. Eine der Hauptfragestellung dieser Arbeit zugrunde liegende Überlegung betrifft letztlich genau diesen. Dabei handelt es sich um die Frage danach, ob das Konstrukt des fremdsprachlichen Leseverständnisses in den Ländern zumindest teilweise unterschiedlich repräsentiert ist. Eine unterschiedliche Konstruktrepräsentation bedeutet eine Einschränkung der Validität interkultureller Vergleiche und eine Minderung der Konstruktvalidität.

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 Validitätsminderung 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 unterschiedlichen Gruppen auszugehen.

Die Ergebnisse dieser Arbeit weisen darauf hin, dass vermutlich beide Ansätze und beide Quellen 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 Stärke hinsichtlich des Beantwortens von Items mit einem bestimmten Itemmerkmal (wie beispielsweise „komplexen grammatischen 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 werden, 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 Konstruktkomponente im Vergleich mit der anderen Gruppe unterrepräsentiert ist. Durch Testkultur und differentielle Lerngelegenheiten bedingte Schwächen können somit als ein Ausdruck von Unterrepräsentanz des Konstrukts bzw. bestimmter Konstruktkomponenten 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

auch die Einführung konstruktirrelevanter Varianz bei der Entstehung von DIF eine Rolle spielt. In diesem Fall würden die Items eine zusätzliche, nicht dem Konstrukt zugehörige Dimension sowie die diesbezüglich existierenden Fähigkeitsunterschiede der Gruppen erfassen. Möglicherweise ist DIF daher durchaus auch teilweise als „Nuisance Dimension" zu betrachten. Dies ist im Rahmen dieser Arbeit jedoch nicht schlussendlich klärbar, da hier nur ein Teil des Konstrukts, nämlich der kognitiv-linguistische Bereich (und dieser nur itemseitig), behandelt wurde. Es ist daher auch denkbar, dass weitere, hier nicht untersuchte Konstruktbereiche eine Rolle bei der Entstehung von DIF spielen und auch dahingehend differentielle Stärken und Schwächen von Gruppen existieren. Es muss sich somit bei dem bisher nicht aufgeklärten Anteil der DIF-Varianz nicht zwingend um konstruktirrelevante Varianz handeln.

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.

Wie unter 2.4.3 bereits dargestellt wurde, geht Messick in seinem Validitätsmodell davon aus, dass ein zu messendes Konstrukt immer auch von den einer Gesellschaft zugrunde liegenden sozialen Werten mit definiert wird, weshalb demnach auch gesellschaftliche und soziale Werte Teil der Konstruktvalidität sind („social validity"; Messick, 1989; 1996).

Die Ergebnisse dieser Arbeit deuten darauf hin, dass solche zugrunde liegenden gesellschaftlichen Werte und Normen auch auf die Definition des Konstrukts „fremdsprachliches Leseverständnis" Einfluss nehmen könnten. Dafür spricht, dass überhaupt signifikant unterschiedliche Profile von Testkulturen in den untersuchten Ländern gefunden wurden. Es ist zu vermuten, dass dies — zumindest indirekt— auch auf unterschiedliche Curricula in den Ländern zurückzuführen ist, auch wenn dies im Rahmen der vorliegenden Arbeit nicht empirisch überprüfbar ist. Das Festlegen von Curricula basiert wiederum auf gemeinsamen sozialen Werten einer Gesellschaft. Dass eine unterschiedliche Leistung von Gruppen wiederum teilweise auf die Unterschiedlichkeit der Testkulturen zurückführbar ist, unterstützt diese Annahme.

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:
In vielen der an der Studie teilnehmenden Länder existieren bisher bereits zentrale, nationale Schulabschluss-Prüfungen, bei denen es sich um sogenannte „high stakes" Tests handelt. Die Ergebnisse und auch die Interpretation dieser Testkennwerte haben einen großen Impact auf das Leben der Getesteten. Dies zeigt sich beispielsweise darin, dass hier die Weichen hinsichtlich der späteren Berufswahl und der Möglichkeiten einer Person hinsichtlich des Besuchs höherer Bildungseinrichtungen gestellt werden. Es ist daher davon auszugehen, dass diese Tests sowie damit auch deren Inhalte für die jeweilige Gesellschaft und deren Mitglieder relevant sind. Diese Tatsache führt dann möglicherweise dazu, dass auch im Rahmen des Fremdsprachenunterrichts die Art von Testaufgaben, die den in den Abschlusstests verwendeten Aufgaben hinsichtlich Struktur, Inhalt und Schwierigkeit entsprechen, verstärkt verwendet und geübt werden. So sind beispielsweise in den Niederlanden die Abschlusstests im Bereich „Fremdsprachen" der vorhergegangenen Jahre öffentlich einsehbar und können somit zum Üben verwendet werden. Es ist hier daher davon auszugehen, dass zumindest in gewissem Maße „Teaching to the Test" (z.B. Korretz, 2005) stattfindet.

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.

Die oben angestellten Überlegungen hinsichtlich der Übertragung gesellschaftlicher und sozialer Werte auf die Testkultur legen nahe, dass in den Ländern für die Übertragung von gesellschaftlich-sozialen Normen auf die Testkultur und das Konstrukt eher ein „Teaching to the Test"-Effekt als ein „Washback"- Effekt (z.B. Alderson & Wall, 1996) verantwortlich ist. Dabei ist „Teaching to the Test" meist eher mit einer negativen Konnotation versehen und wird mit einer Einschränkung von Konstruktvalidität durch das Testen in Verbindung gebracht, während der

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

Eine wichtige und interessante Frage wäre, in welcher Form der Mechanismus der Übertragung sozialer Werte auf die Testkultur möglicherweise dazu genutzt werden könnte, statt eines „Teaching to the Test”-Effeks einen „Washback”-Effekt zu initiieren. Dies hätte den Vorteil, dass zum einen das Konstrukt möglichst vollständig unterrichtet und getestet würde und damit die Tests insgesamt auch innerhalb der Länder valider würden. Zum anderen würde dies aber auch zu einer Erhöhung der Validität der Tests bei Leistungsvergleichen zwischen den Ländern durch eine Verringerung von Differentiellen Item Funktionen beitragen: Durch eine „Vervollständigung” des Gesamtkonstrukts innerhalb der Länder würde sich vermutlich auch das Konstrukt zwischen den Ländern immer weniger unterscheiden, da auf diesem Wege eine differentielle Unterrepräsentanz des Konstrukts vermindert würde.

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.

Der oben anhand von Messicks (1989) Modell der Validität dargestellte Weg der „Übertragung" von sozialen und bildungskulturellen Werten und Normen in den unterschiedlichen Ländern auf das Konstrukt kann nur unter Vorbehalt angenommen werden, da im Rahmen der vorliegenden Arbeit nicht empirisch prüfbar ist, ob diese tatsächlich wie dargestellt auf die Implementation der verschiedenen Testkulturen wirken. Eine empirische Überprüfung wäre vermutlich nur im Rahmen einer Längsschnittuntersuchung zu realisieren (worauf im nächsten Abschnitt genauer eingegangen wird).

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 Itemmerkmale zu verbessern und zum anderen möglicherweise einen längeren Rating-Prozess durchzuführen, in dem die Experten während der Schulung und des späteren Ratings immer wieder Feedback hinsichtlich ihrer Übereinstimmung erhalten. Denkbar wäre auch, die Gruppe der Rater aus Experten unterschiedlicher Länder zusammenzusetzen. So ließen sich länderspezifische Verzerrungen 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) befinden sollten, dazu, dass sich die Ergebnisse hinsichtlich der Testkulturen nicht auf andere Kompetenzniveaus generalisieren lassen. Auch lassen sich keine Aussagen hinsichtlich der Generalisierbarkeit 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 Ergebnisse 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 jeweils 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.

Eine diesbezügliche Forschungsperspektive wäre beispielsweise, die unterschiedlichen Äquivalenzstufen des Konstrukts zwischen den verschiedenen Ländern zu überprüfen. Auf diesem Wege könnten auch Aussagen über die Art der Konstruktunterschiede gemacht werden, etwa inwieweit es sich um Struktur- oder Niveauunterschiede handelt. Anhand von Strukturgleichungsmodellen könnten ferner die jeweilige Struktur des Konstrukts und Zusammenhänge zwischen den schwierigkeitsbestimmenden Merkmalen untersucht werden. Wie bereits angesprochen wurde war dies anhand der Daten der EBAFLS-Studie nicht möglich.

Ferner sind im Rahmen dieser Arbeit bisher lediglich Itemmerkmale einbezogen worden, die sich auf kognitiv-linguistische Merkmale beziehen. Die Ergebnisse weisen jedoch darauf hin, dass diese nur einen Teil der Differentiellen Item Funktionen in einem Teil der Länderpaarungen erklären können. Daher wäre es wünschenswert, im Rahmen weiterer Forschung auch beispielsweise Itemmerkmale mit einzubeziehen, die sich eher der inhaltlichen Kategorie zuordnen lassen. Die Nicht-Verwendung inhaltlicher Merkmale zur Erklärung von DIF wird beispielsweise von Abbott (2004) kritisiert.

Neben den Überlegungen hinsichtlich der durch die Datenstruktur und die Stichproben gesetzten Grenzen und Vorbehalte sollen darüberhinaus auch kritische Überlegungen bezüglich des in dieser Arbeit gewählten methodischen Ansatzes dargelegt werden. In der vorliegenden Dissertation wurden zur Erklärung Differentieller Item Funktionen Testkultur-Indikatoren gewählt. Einen anderen, gleichfalls vielversprechenden Ansatz verfolgten beispielsweise Dogan, Guerrera und Tatsuoka (2005) im Rahmen von auf den Daten der TIMS-Studie basierenden DIF-Analysen. Auch diese Autoren stellten die Hypothese auf, dass DIF auf differentielle Stärken und Schwächen von Gruppen rückführbar ist. Jedoch unterscheidet sich die dort gewählte Methode zur Analyse der erwarteten Stärken und Schwächen. Die Autoren analysierten dazu zunächst, wie gut die Schüler verschiedener Länder Items mit bestimmten kognitiven Anforderungen (analysiert anhand von Itemmerkmalen) lösen konnten, unabhängig von der Herkunft der Items. Die Autoren bezogen dabei 10 Länder in ihre Analyse mit ein, zwischen denen jeweils paarweise DIF-Analysen durchgeführt wurden. Hypothese war, dass je mehr sich die Länder hinsichtlich der Wahrscheinlichkeit unterschieden, Items mit einem bestimmten Attribut zu lösen, und auf je mehr Attribute dies bei einem Item zutraf, desto höher sollte der DIF-Parameter dieses Items sein. Diese Zusammenhänge konnten die Autoren herstellen, und konnten zwischen 40% und 81% der DIF-Varianz auf diese Weise aufklären. Der so aufgeklärte Varianzanteil liegt zum Teil deutlich höher als es in der vorliegenden Arbeit der Fall ist. Dieses Beispiel soll darlegen, dass neben dem in dieser Arbeit gewählten Ansatz zur Analyse erwarteter Stärken und Schwächen der Grppen auch ein anderer

denkbar gewesen wäre. Andererseits können Dogan, Guerrera und Tatsuoka (2005) keine Aussage dahingehend machen, warum die Schüler eines Landes Items mit bestimmten Attributen mit einer höheren oder niedrigeren Wahrscheinlichkeit lösen können als die Schüler eines jeweils anderen Landes, wohingegen die Ergebnisse in der vorliegenden Arbeit darauf hinweisen, dass dies zumindest teilweise auf unterschiedliche Testkulturen zurückführbar ist.

Ein weiterer denkbarer Kritikpunkt am methodischen Ansatz ist die Wahl des Modells. Da DIF auf eine nicht-modellierte Mehrdimensionalität hinweist und auch mehrdimensionale Modelle zur Erklärung von DIF bereits angewandt wurden (für einen Überblick siehe Zumbo, 2007), wäre auch dies ein möglicher methodischer Ansatz gewesen. Dieser wurde hier nicht gewählt, da so die theoretische Nähe zum GERS und auch den in der EBAFLS-Studie gewählten Modellen verloren gegangen wäre.

Wie oben bereits dargestellt, hätten zur Modellierung von DIF auch zusätzliche inhaltliche Prädiktoren mit einbezogen werden können (z.B. umgangssprachliche Ausdrücke; Chen & Henning, 1985). Darüberhinaus hätten auch Personeneigenschaften möglicherweise zur Erklärung von DIF beitragen können, wie es beispielsweise van den Noortgate und deBoeck (2005) vorschlagen. Durch das Einbeziehen weiterer konstruktrelevanter Variablen könnten dann auch möglicherweise deutlichere Aussagen dahingehend gemacht werden, welcher Anteil von DIF auf ausschließlich konstruktrelevante Faktoren rückführbar ist und welcher Teil auf „Nuisance-Dimensionen".

Zur Erklärung von DIF wurden ferner multiple Regressionsanalysen gewählt. Diese sind jedoch aufgrund von Multikollinearitätsproblemen nicht immer eindeutig und stabil interpretierbar, weshalb die Interpretation der Ergebnisse hier immer nur unter Vorbehalt geschehen kann. Koeffizienten können immer nur im Kontext anderer Koeffizienten interpretiert werden. Andererseits wurden nicht zuletzt aus diesem Grund auch Einzelkorrelationen berechnet, die einen Eindruck bezüglich des Zusammenhangs der einzelnen Testkultur-Indikatoren bzw. schwierigkeitsdeterminierenden Itemmerkmale und DIF geben.

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-

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.

Einige der aufgezeigten Forschungsperspektiven ließen sich möglicherweise anhand einer Längsschnittuntersuchung realisieren. Dies wäre beispielsweise für die Untersuchung des oben bereits angesprochenen Einflusses kultureller Werte und Normen auf die Definition des Konstrukts „fremdsprachliches Leseverständnis" sowie eine möglicherweise durch internationale Leistungsvergleiche stattfindende Veränderung der Testkulturen (und somit indirekt auch der Konstruktdefinition) denkbar. Eine diesbezüglich interessante Forschungsfrage wäre etwa, ob sich die in den nationalen Tests eingesetzten Item- und Testformate bei Ländern mit „schlechteren" Ergebnissen verändern, ob sich über die Zeit beobachten lässt, dass Testformate aus anderen Ländern für nationale Examina verwendet werden, und ob sich Testkulturen möglicherweise über die Zeit angleichen.

Ein weiteres denkbares Forschungsfeld ist mit der Frage verknüpft, ob die Ergebnisse auch auf andere Weise dazu beitragen könnten, zwischen den Ländern existierende Differentielle Item Funktionen zu verringern und so die Validität von internationalen Leistungsvergleichen zu erhöhen. So könnten beispielsweise die bei den Gruppen diagnostizierten Stärken und Schwächen bei der Interpretation von Testergebnissen eine Rolle spielen: Sollten sich die Ergebnisse dieser Arbeit stabil replizieren lassen, wäre es möglicherweise denkbar, die Testwertergebnisse von Gruppen entsprechend ihrer auf den Testkulturen basierenden Stärken und Schwächen zu gewichten. Auch dies könnte zu einer erhöhten Fairness von Gruppenvergleichen beitragen.

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 Niveauunterschiede 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.

Abschließend lässt sich konstatieren, dass die Ergebnisse dieser Arbeit darauf hinweisen, dass die Unterschiedlichkeit von Testkulturen und damit einhergehende systematische Stärken und Schwächen von Gruppen zumindest teilweise eine Rolle bei der Entstehung Differentieller Item Funktionen zu spielen scheinen. Die Arbeit leistet somit durch das Aufzeigen möglicher Gründe für kulturell bedingte Varianz einen wichtigen Beitrag zur Erforschung und Verbesserung von Validität und Fairness in internationalen und interkulturellen Leistungsvergleichen.

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

Abbott, M.L. (2004). *The Identification and Interpretation of Group Differences on the Canadian Language Benchmarks Assessment Reading Items* . Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA, April 15, 2004.

Ackerman, T.A. (1992). A Didactic explanation of item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

American Council for the Teaching of Foreign Languages (1983). *ACTFL Proficiency Guidelines. Revised 1985* . Hastings-on-Hudson, NY: ACTFL Materials Center.

Adams, R.J. (2006). *Reliability and Item Response Modelling: Myths, Observations and Applications* . URL *Retrieved June24th, 2009, from http://bearcenter.berkeley.edu/IOMW2006/presentations/Opening/Ray%20Adams%20-%20IOMW2006%20Opening.ppt.* Paper presented at the 13th IOMW.

Adams, R.J. & Khoo, S.T. (1996). *Quest* . Melbourne, Australia: Australian Council for Educational Research.

Alderman, D.L. & Holland, P.W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language. TOEFL Research Report 9* . Princeton, NJ: Educational Testing Service.

Alderson, J.C. (2000). Assessing Reading. In Alderson, J.C. & Bachman, L.F., *The Cambridge Language Assessment Series.* Cambridge University Press

Alderson, J.C., Figueras, N., Kuijper, H.,; Nold, G., Takala, S, & Tardieu, C. (2006). Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly, 3 (1),* 3–30.

Alderson, J.C. & Wall, D. (1993). Does Washback Exist?. *Applied Linguistics, 14 (2),* 115–129.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89,* 369–406.

Angoff, W.H. (1993). Perspectives on Differential Item Functioning Methodology. In Holland, Paul W. & Wainer, Howard (Eds.), *Differential Item Functioning.* Lawrence Erlbaum Associates, Hillsdale, NJ

Artelt, C. & Baumert, J. (2004). Zur Vergleichbarkeit von Schülerleistungen bei Aufgaben

unterschiedlichen sprachlichen Ursprungs. *Zeitschrift für Pädagogische Psychologie, 18 (34)*, 2–35.

Avenarius, H., Ditton, H., Döbert, H., Klemm, K., Klieme, E., Rürup, M., Tenorth, H.-E., Weishaupt, H. & Weiß, M. (2003). *Bildungsbericht für Deutschland: Erste Befunde (Zusammenfassung)* . Frankfurt am Main / Berlin: Federführung: Deutsches Institut für Internationale Pädagogische Forschung (DIPF). URL *http://www.kmk.org/fileadmin/pdf/PresseUndAktuelles/2003/bb_zusammenfassung.pdf* . Entnommen am 3.2.2010

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing* . Oxford: Oxford University Press.

Bachman, L.F., Davidson, F. & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13 (2,* 125–150.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice* . Oxford: Oxford University Press.

Bailey, K.M. (1996). Working for Washback: a review of the washback concept in language testing. *Language Testing, 13*, 257–279.

Baron, P., Curley, E. & Feigenbaum, M. (2000). *Indicators on SAT I DIF Analyses* . URL *Zugriff am 3.2.2010 unter http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/75/a8.pdf* . Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27)

Bausch, K.-R., Christ, H., Königs, F.G. & Krumm, H.J. (Hrsg.) (2003). *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion. Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts* . Tübingen: Narr.

Beck, B. & Klieme, E. (2007). *Sprachliche Kompetenzen : Konzepte und Messung - DESI-Studie (Deutsch Englisch Schülerleistungen International)* . Weinheim: Beltz.

Bialystok, E. (1981). The role of linguistic knowledge in second language use. *Studies in second language learning, 4 (1)*, 31–45.

Bialystok, E. (1986). Factors in the growth of linguistic awareness. *Child development, 57*, 498–510.

Bolton, S. (2000). *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar* . Köln: VUB Printmedia.

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

M., Vrignaud, P. & White, J. (2001). *The use of national reading tests for international comparisons: ways of overcoming cultural bias* . Paris: Ministère de l'éducation nationale, DPD Edition diffusion.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler. 6. Auflage* . Berlin: Springer.

Brown, G. (1996). Introduction. In Brown, G., Malkjoer, K. & Williams, J., *Performance and Competence in second Language Acquisition.* Cambridge University Press

Brown, J.D. & Hudson, T. (2002). *Criterion-Referenced Language Testing* . Cambridge, UK / New York: Cambridge University Press.

Brown, G., Malkjoer, K. & Williams, J. (Eds.) (1996). *Performance and Competence in second Language Acquisition* . Cambridge: Cambridge University Press.

Buck, G., Tatsuoka, K. & Kostin, I. (1997). The subskills of reading: rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47(3),* 423–466.

Camilli, G. (1993). The case against DIF techniques based on internal criteria: Do item bias procedures obscure test fairness?. In Holland, P.W. & Wainer, H. (Eds.), *Differential Item Functioning.* Lawrence Earlbaum Assiciates, Hillsdale, NJ

Camilli, G. & Shepard, L. (1994). *Methods for identifying test bias* . Thousand Oaks, CA: Sage.

Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1,* 1–47.

Carroll, J.B. (1961). *Fundamental considerations in testing for English language proficiency of foreign students* . Washington, D.C: Center for Applied Linguistics.

Chen, Z. & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2,* 155–163.

Chomsky, N. (1965). *Aspects of the Theory of Syntax* . Cambridge, Mass.: MIT Press.

Chomsky, N. (1980). *Rules and Representations* . Oxford: Blackwell.

Christ, H. (2003).Ohne Titel. In Bausch, K.-R., Christ, H., Königs, F.G. & Krumm, H.J. (Hrsg.), *Der Gemeinsame europäische Referenzrahmen für Sprachen in der Diskussion. Arbeitspapiere der 22. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Narr, Tübingen

CITO (2008). *Socrates Programme 113946-CP-1-1-2004-1-NL-Lingua-L2. Building a European Bank of Anchor Items for Foreign Language Skills* . Arnhem: CITO. Manuskript wurde an die Europäische Kommision übergeben; bisher unveröffentlicht, auf Anfrage bei CITO erhältlich

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences. 2nd edition* . Hillsdale, NJ: Lawrence Earlbaum Associates.

Cole, N. (1993). History and development in DIF. In Holland, P. W. & Wainer, H. (Eds.), *Differential Item Functioning.* Lawrence Erlbaum Associates, Hillsdale, NJ

De Jong, J.H.A.L. & Verhoeven, L. (1992). Modeling and assessing language proficiency. In Verhoeven, L. & De Jong, J.H.A.L. (Eds.), *The construct of language proficiency: applications of psychological models to language assessment*. John Benjamins Publishing Co, Amsterdem

De Saussure, F. (1916). *Cours de Linguistique Generale* . Paris: Payot.

DESI-Konsortium (Hrsg.) (2008). *Sprachliche Kompetenzen. Leistungsverteilungen und Bedingungsfaktoren. DESI-Ergebnisse Band 2* . Weinheim: Beltz.

Dogan, E., Guerrero, A. & Tatsuoka, T. (2005). *Using DIF to investigate strengths and weaknesses in Mathematic Achievement. Profiles of 10 Different Countries* . Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). Montreal, Canada, April 12-14

Dorans, N.J. & Holland, P.W. (1993). Description: Mantel-Haenszel and Standardization. In Holland, Paul W. & Wainer, Howard (Eds.), *Differential Item Functioning.* Lawrence Erlbaum Associates, Hillsdale, NJ

Dresselhaus, G. (1979). *Zur Klärung der Begrifffspaare bei Saussure und Chomsky, ihre Vorgeschichte und ihre Bedeutung für die moderne Linguistik* . Frankfurt am Main: Lang.

Ellis, R. (1990). Individual learning styles in classroom second language development. In de Jong, J.H.A.L. & Stevenson, D.K. (Eds.), *Individualizing the Assessment of Language Abilities.* Multilingual Matters, Clevedon

Ellis, R. (1990). A response to Gregg. *Applied Linguistics, 11 (4)*, 384–392.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* . Mahwah, NJ: Erlbaum.

Europäische Kommission (2002). *European Benchmarks in Education and Training. Follow-up to the Lisbon European Council* . URL *http://ec.europa.eu/education/policies/2010/doc/*

bench_ed_trai_en.pdf. Zugriff am 26.04.07

Europäische Kommission (2003). *Mitteilung der Kommission an den Rat, das Europäische Parlament, den Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen vom 24. Juli 2003 — Förderung des Sprachenlernens und der Sprachenvielfalt: Aktionsplan 2004-2006. Komm(2003)449 endgültig* . URL *http://europa.eu/scadplus/leg/de/cha/c11068.htm.* Zugriff am 20.03.07

Europäisches Parlament (2008). *Europäisches Parlament-Leitfaden Sprachenpolitik* . URL *http://www.europarl.europa.eu/facts/4_16_3_de.htm.* Zugriff am 29.01.08

Europarat (2001). *Gemeinsamer Europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen* . Berlin: Langenscheidt. Online-Version: www.goethe.de/referenzrahmen

Fandel, J.C., Gille, E., Hesse, H.G., Van Krieken, R., Lagergren, T., Tòth, T., Tovar, C., Troseille, B. (2007). *My B1 is not very likely your B1: A problem or a challenge?* . My B1 is not very likely your B1: A problem or a challenge? Discussion Group at the 8th Annual AEA-Europe Conference, November 8th-10th, 2007, Stockholm

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating Examinations to the Common European Framework: a manual. *Language Testing, 22 (3),* 257–261.

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

Fortus, R., Coriat, R. & Fund, S. (1998). Prediction of item difficulty in the English subtests of Israel´s Inter-University Psychometric Entrance Test. In Kunnan, A.J. (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach.* Earlbaum, Mahwah, NJ

Freedle, R. & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validation. *Language Testing, 10(2),* 133–170.

Geranpayeh, A. & Kunnan, A.J. (2007). Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination. *Language Assessment Quarterly, 4 (2),* 190–222.

Gesellschaft für angewandte Linguistik (2009). *Homepage* . URL *http://www.gal-ev.de/angewandte-linguistik.html.* Zugriff am 29.05.09

Gierl, M.J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit dgroup differences. *Educational Measurement: Issues and Practice, 24(1),* 3–14.

Gille, E. & Sluiter, S. (2005). *The European Anchor Item Bank* . Paper presented at the 2nd annual conference of the European Association for Language Testing and Assessment, Voss, Norway

Glaser, R. & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In Gagne, R.M. (Ed.), *Psychological principles in system development*. Holt, Rinehart & Winston, New York

Glaser, R., Lesgold, A. & Lajoie, S.P. (1987). Toward a cognitive theory for the measurement of achievement. In Ronning, R., Glover, J., Conoley, J.C. & Witt, J.C. (Eds.), *The influence of cognitive psychology on testing*. Erlbaum, Hillsdale, NJ

Gregg, K. R. (1990). The variable competence model of second language acquisition and why it isn't. *Applied Linguistics, 11 (4)*, 364-384.

Grotjahn, R. (2000). Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung des TESTDAF. In Bolton, S. (Hrsg., *TESTDAF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. VUB Printmedia, Köln

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* . Newbury Park, CA: Sage.

Hartig, J., Frey, A., Nold, G. & Klieme, E. (2008). *An Application of Explanatory Item Response Modelling For Model-based Proficiency Scaling* . Manuskript eingereicht

Helfrich, H. (1999). Beyond the Dilemma of Cross-Cultural Psychology: Resolving the Tension between Etic and Emic Approaches. *Culture & Psychology, 5(2)*, 131–153.

Heyworth, F. (2004). Why the CEF is important. In Morrow, K. (Ed.), *Insights from the Common European Framework*. Oxford University Press

Holland, W. P. & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H. & Braun, H.I. (Eds.), *Test validity*. Lawrence Earlbaum Associates, Hillsdale, NJ

Holland, P. W. & Wainer, H. (Eds.) (1993). *Differential Item Functioning* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Hymes, D. (1971). Competence and Performance in Linguistic Theory. In Huxley, R. & Ingram, E. (Eds.), *Language Acquisition, Models and Methods*. Academic Press, New York

Janssen, R., Schepers, J., Peres, D. (2004). Models with item and item group predictors. In De Boeck, P., Wilson, M. (Eds.), *Explanatory item response models: A generalized linear and*

nonlinear approach. Springer, New York

Jones, N. (2002). Relating the ALTE Framework to the Common European Framework. In Alderson, J. C. (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies.*. Council of Europe, Strasbourg

Jöreskog, K.G. & Moustaki, I. (2001). Factor Analysis of ordinal Variables: A comparison of three approaches. *Multivariate Behavioural Research, 36*, 347-387.

Jöreskog, K.G. & Sörbom, D. (1979). *Advances in factor analysis and structural equation models* . New York: University press of America.

Jude, N. (2008). *Zur Struktur von Sprachkompetenz* . Frankfurt am Main: www.dissonline.de.

Kaftandjieva, F. & Takala, S. (2002). Council of Europe Scales of language proficiency: a validation study. In Alderson, J. C. (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies.* Council of Europe, Strasbourg

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M (2000). What do Test Scores in Texas Tell Us?. *Education Policy Analysis Archives, 8(49).*URL *http://epaa.asu.edu/epaa/v8n49/*. Zugriff am 10.10.2008

Klieme, E. (1989). *Mathematisches Problemlösen als Testleistung* . Frankfurt am Main: Lang.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16*, 385–402.

Klieme, E. & Bos, W. (2000). Mathematikleistung und mathematischer Unterricht in Deutschland und Japan. *Zeitschrift für Erziehungswissenschaften, 3*, 359–379.

Klieme, E., Eichler, W. Helmke, A., Lehmann, R.H., Nold, G., Hans-Günter Rolff, H.-G., Schröder, K., Thomé, G. & Willenberg, H. (2006). *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Zentrale Befunde der Studie Deutsch-Englisch-Schülerleistungen-International (DESI)* . Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung.

Kline, R.B. (2005). *Principles and Practice of Structural Equation Modelling* The Guilford Press.

Komorowska, H. (2002). The Common European Framework in Poland. In Alderson, J. C. (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies.*. Council of Europe, Strasbourg

Koretz, D. (2005). Alignment, High Stakes and the Inflation of Test Scores. *Yearbook of the National Society for the Study of Education, 104 (2),* 99–118.

Koretz, D.M. & Barron, S.I. (1998). *The Validity Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)* . Santa Monica, CA: Rand Education.

Lewis, E.G. & Massad, C. (1975). *The Teaching of English as a Foreign Language in Ten Countries* . New York: John Wiley & Sons.

Li, Y., Cohen, A.S. & Ibarra, R.A. (2004). Characteristics of Mathematics Items Associated With Gender DIF. *International Journal of Testing, 4 (2),* 115–136.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems* . Hillsdale, NJ: Lawrence Earlbaum Associates.

Loveday, L.J. (1982). *The sociolinguistics of Learning and Using a Non-Native Language* . Oxford: Pergamon.

Maris, G., Bechger, T. & Veldhuijzen, N. (2006). *EBAFLS: Preliminary DIF Analysis* . Arnhem: CITO: Unveröffentlichtes Manuskript.

McNamara T. (1996). *Measuring second language performance* . Essex, UK: Addison Wesley Longman Ltd..

McNamara, T. (2006). Validity in Language Testing: The Challenge of Sam Messick´s Legacy. *Language Assessment Quarterly, 3 (1,* 31–51.

Messick, S. (1989). Validity. In Linn, R.L. (Ed.), *Educational measurement (3rd ed.).* American Council on Education and Macmillan, New York

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences From Persons´ Responses and Performances as Scientific Inquiry Into Score Meaning. *American Psychologist, 50(9),* 741–749.

Messick, S. (1996). *Validity and Washback in Language Testing. Technical Report* Educational Testing Service.

Morrow, K. (1981). Principles of Communicative Methodology. In Johnson, K. & Morrow, K. (Eds.), *Communication in the Classroom.* Longman

Morrow, K.(Ed.) (2004). *Insights from the common European Framework* . Oxford: Oxford University Press.

Muthén, B.O. (1988). Some uses of structural equation modelling in validity studies: Extending

IRT to external variables. In Wainer, H., Braun, H. (Eds.), *Test Validity*. Lawrence Earlbaum Associates, Hillsdale, N.J.

Muthén, B.O. & Lehman, J. (1985). Multiple IRT Modelling: Applications to Item Bias Analysis. *Journal of Educational Statistics, 10*, 133–142.

Noijons, J. & Kuijper, H. (2006). *Report of a research project commissioned by the Dutch Ministry of Education, Culture and Science* . Arnhem: CITO.

North, B. (2000). *The development of a common framework scale of language proficiency* . New York: Peter Lang.

North, B. (2002). A CEF-based self assessment tool for university entrance. In Alderson, J. C. (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies.* Council of Europe, Strasbourg

Oller, J.W. (1976). Evidence for a General Proficiency Factor. In Oller, J.W. (Ed.), *Issues inLanguage Testing Research.* Newbury House, Rowley

Perkins, K. & Linville, S.E. (1987). A construct definition study of a standardized ESL vocabulary test. *Language Testing, 4*, 125–41.

Prenzel,M., Artelt,C., Baumert,J., Blum,W., Hammann,M., Klieme, E. & Pekrun,R. (Hrsg.) (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* . Münster: Waxmann.

Projektgruppe TestDaF (2000). TestDaf: Konzeption, Stand der Entwicklung, Perspektiven. *Zeitschrift für Fremdsprachenforschung, 11(1)*, 63–82.

Rauch, D.P. & Hartig, J. (in Vorbereitung). A Differential Analysis of Abilities Related to Answering Open-Ended Test Items vs. Multiple Choice Test Items in Reading Comprehension Assessment. Ohne Verleger.

Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion. 2.überarbeitete Auflage* . Bern: Huber.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF Analysis paradigm. *Applied psychological measurement, 20*, 355–371.

Roussos, L., & Stout, W. (2004). Differential Item Functioning Analysis: Detecting Dif items and testing DIF hypotheses. In Kaplan, D. (Ed.), *Sage Handbook of quantitative methodology for the social sciences.* Sage, Thousand Oaks

Rumelhart, D.E., McClelland, J.L. & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I, Foundations;*

Volume II, Psychological and Biological Models . Cambridge, MA: MI Press.

Ryan, K.E. & Chiu, S. (1997). *An Examination of Item Context Effects, DIF, and Gender-DIF* . Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, March 24-28

Sasaki, M. (1991). A comparison of thwo methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8(2)*, 95–111.

Scheuneman, J.D. & Gerritz, K. (1990). Using Differential Item Functioning Procedures to Explore Sources of Item Difficulty and Group Performance Characteristics. *Journal of Educational Measurement, 27(2)*, 109–131.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., Wiley, D.E. (1997). *Many Visions, Many Aims. Volume 1: A Cross-National Investigation of Curricular Intentions in School Mathematics* . Dordrecht: Kluwer Academics Publisher.

Schneider, G. & North, B. (2000). *Fremdsprachen können-was heißt das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit* . Zürich: Rüegger.

Shealy, R., & Stout, W.F. (1993). A model-based standardization approach that seperates true bias/DIF and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Skinner, B. F. (1957). *Verbal behaviour* . Acton, MA: Copley Publishing.

Smith, R.L., Ager, J.W. & Williams, D.L. (1992). Suppressor Variables in Multiple Regression / Correlation. *Educational and Psychological Measurement, 52*, 17–29.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2*, 31–40.

SPSS Inc. (2009). SPSS Statistics 18. Ohne Verleger.URL *http://www.spss.com/de/*. Zugriff am 2.3.2010

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361˘-370.

Tarone, E. (1985). Variablilty in interlanguage use. A study of style shifting in morphology and syntax. *Language Learning, 35*, 373–403.

Thiessen, D., Steinberg, L. & Wainer, H. (1993). Item Functioning Using the Parameters of Item Response Models. In Holland, P.W. & Wainer, H., *Differential Item Functioning.* Lawrelnce Earlbaum Associates, Hillsdale, NJ

Thurstone, L.L. (1938). *Primary Mental Abilities* . Chicago: University of Chicago Press.

TNS opinion & social (2005). *Special Eurobarometer 63.4. Europeans and their languages* . URL *http://ec.europa.eu/public_opinion/archives/ebs/ebs_237.en.pdf* . Zugriff am 20.03.07

Trautner, H.M. (1997). *Lehrbuch der Entwicklungspsychologie Bd. 2 Theorien und Befunde* . Göttingen: Hogrefe.

Trim, J. (2001). The Work of the Council of Europe in the field of Modern Languages, 1957-2001. Ohne Verleger. Paper given at a Symposium to mark the European Day of Languages 26 September 2001 at the European Centre for Modern Languages, Graz.

Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing, 22*, 211–234.

Van den Noortgate, W. & de Boeck, P. (2005). Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models. *Journal of Educational and Behavioral Statistics, 30*, 443-464.

Van de Vijver, F.J.R. & Leung, K. (1997). *Methods & Data Analysis for Cross-Cultural Research* . Thousand Oakes: Sage.

Van de Vijver, F.J.R. & Poortinga, Y. H. (1990). A taxonomy of cultural differences. In Van de Vijver, F.J.R & Hutschemaekers, G.J.M., *The investigation of culture. Current Issues in Cultural Psychology*. Tilburg University Press

Van de Vijver, F. & Tanzer, N.K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Revue européenne de psychologie appliqué, 54*, 119–135.

Van Ek, J.A. (1986). *Objectives for Foreign Language Teaching. Volume I: Scope* . Strasbourg: Council of Europe.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1995). *One Parameter Logistic Model* . Arnhem: Cito.

Weir, C.J. (2005). *Limitations of the Common European Framework for developing comparable examinations and tests* , Seiten 281–300.

Wertenschlag, L., Müller, M. & Schmitz, H. (2002). The Common European Framework and the European Level Descriptions for German as a Foreign Language. In Alderson, J. C. (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case Studies.*. Council of Europe, Strasbourg

Widdowson, H.G. (1983). *Learning Purpose and Language Use* . Oxford: University Press.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach* . Mahwah, NJ: Erlbaum.

Wilson, M. & deBoeck, P. (Eds.) (2004). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach* . Berlin: Springer.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modelling software manual* . Hawthorne, Australia: ACER Press.

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

Für ihre Unterstützung bei meiner Doktorarbeit schulde ich vielen Menschen ganz besonderen Dank. Zunächst möchte mich beim Deutschen Institut für Internationale pädagogische Forschung für die Möglichkeit bedanken, dort zu arbeiten und zu einem solch interessanten Thema zu promovieren. Des Weiteren danke ich ganz herzlich meinem Bertreuer Prof. Dr. Eckhard Klieme, der mir mit wertvollen Anregungen während der Erstellung der Arbeit stets zur Seite gestanden hat. Er hat mir dabei geholfen, die Thematik immer wieder neu und kritisch zu durchdenken. Auch hat er mir ermöglicht, die Arbeit am Dipf fertigzustellen, wofür ihm mein ganz besonderer Dank gilt. Ich danke auch den Mitgliedern der internationalen EBAFLS-Projektgruppe, ohne deren Daten ich das Thema der Dissertation nicht hätte bearbeiten können. Ferner danke ich meinen Kolleginnen und Kollegen am DIPF, mit denen ich immer wieder wertvolle wissenschaftliche und methodische Diskussionen zu dem Thema führen konnte. In dieser Zeit habe ich in sehr angenehmer Atmosphäre arbeiten, wertvolle Freundschaften schließen und interessante Kontakte knüpfen können. Auch danke ich meinen Eltern und dem Rest meiner Familie nicht zuletzt für ihre fortwährende geistig-moralische Unterstützung während meiner Promotionszeit. Mein ganz besonderer Dank gilt aber meinem Mann Oliver, ohne dessen fortwährende und uneingeschränkte Unterstützung und Toleranz vor allem in arbeitsintensiven Phasen diese Arbeit sicherlich nicht fertiggestellt worden wäre.

Dieses Dokument wurde mit dem freien Schriftsatzprogramm Lout erstellt (Download unter http://sourceforge.net/projects/lout/. Die erstellte Postscript-Datei wurde unter Zuhilfenahme von AFPL *Ghostscript* (http://pages.cs.wisc.edu/ ghost/) nach PDF konvertiert.