Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs

Comparability of Students' Reading Literacy Performance Measured with Items Originating from Different Language Backgrounds

Cordula Artelt und Jürgen Baumert

Zusammenfassung: In internationalen Schulleistungsstudien stellt ein differenzieller Vorteil bei Leseaufgaben des eigenen Sprach- bzw. Kulturraums eine potenzielle Gefährdung der *fairness* des Tests dar. Durch die Analyse von IRT-basierten differenziellen Itemfunktionen (DIF) der PISA-Lesetestaufgaben wird geprüft, ob Schüler gleicher Fähigkeit, aber unterschiedlicher Sprachgruppen, systematische Vorteile bei Aufgaben haben, die ursprünglich aus ihren Ländern (Sprachgruppen) stammen. Besonders bei französischen und griechischen und z. T. auch bei deutschen Aufgaben lassen sich entsprechende Effekte nachweisen (d = .23). Aufgrund der geringen Anzahl wirkt sich dieser Vorteil jedoch kaum auf das mittlere Abschneiden der Länder aus. Auch die Vorteile englischsprachiger Schüler durch die Dominanz englischsprachiger Items im Lesetest lässt sich auf der Länderebene nicht zufallskritisch absichern. Die Ergebnisse machen insgesamt deutlich, dass in international vergleichenden Studien die sprachliche Herkunft der Aufgaben eine systematische Varianzquelle darstellt. Dem hieraus potenziell entstehenden *cultural bias* des Tests kann dabei – wie in PISA– durch eine möglichst multi-kulturelle Zusammensetzung von Testaufgaben begegnet werden.

***Schlüsselwörter*:** PISA, Lesekompetenz, Vergleichbarkeit, cultural bias, Sprache, DIF-Analyse

Summary: The fact that students are at an advantage when working on reading literacy items from their own cultural and linguistic background in an international large scale assessment can be seen as a threat to the fairness of a test. An IRT-based analysis of differential item functioning (DIF) in the PISA reading literacy items was performed to investigate whether students of equal ability but from different language groups have a systematic advantage when processing items originating from their own cultural and linguistic background. Such effects were discerned especially for French and Greek, but also for German items (d = 23). Because few items from these countries were contained in the PISA assessment, this advantage does not significantly affect the mean performance of these countries, as a re-analysis of student performance on a test without the biased items shows. Furthermore, the fact that most items stem from the Anglo-American background does not mean that students in the five English-speaking countries perform significantly better. The results presented confirm that the language source of the items in international student surveys can be regarded as a systematic source of variance. The potentially resulting cultural bias of the test might be addressed – as done in PISA 2000 – by administering a balanced, multi-cultural mix of test items.

Keywords: PISA, reading literacy, comparability, cultural bias, language, DIF-analysis

1 Einleitung

Die Aussagekraft international vergleichender Studien wie z. B. PISA, PIRLS und TIMSS ist u. a. davon abhängig, dass die in verschiedenen Ländern erhobenen Daten valide Rückschlüsse auf die zugrunde liegenden Kompetenzen erlauben und die erfassten Indikatoren in den Ländern dasselbe latente Fähigkeitskonstrukt abbilden. Zur Überprüfung der Gütekriterien der Instrumente und der Äquivalenz der Skalen werden im Rahmen von *large scale*-Untersuchungen zahlreiche Maßnahmen ergriffen, um diese notwendigen Voraussetzungen für die Interpretierbarkeit der Ergebnisse eines internationalen Vergleichs zu erfüllen (vgl. Adams & Wu, 2002; van de Vijver & Poortinga, 1997). Auch aufgrund des vermehrten Einsatzes von Tests für internationale Vergleichszwecke (Hambleton, 1994) sind die Standards in den Bereichen Stichprobenziehung, Messmethodik und Auswertung sehr hoch und entsprechen

Angenommenes Manuskript. Diese Fassung entspricht nicht exakt der in der
Zeitschrift für Pädagogische Psychologie, 18 (3/4), 2004, 171–185 (©Verlag: Hogrefe)
veröffentlichten Version und kann daher nicht als Quelle zitiert werden.
Der folgende Link führt zum Originalartikel: http://econtent.hogrefe.com/doi/abs/10.1024/1010-0652.18.34.171

in der Regel dem *state of the art* der Disziplin. Am Beispiel der PISA-Studie lässt sich dies anhand der Veröffentlichungen zur Methodik gut nachvollziehen (Adams & Wu, 2002; Baumert et al., 2001, 2003).

Nicht zuletzt im Zuge der Rezeption der Ergebnisse der PISA-Studie wurde jedoch die Frage nach der Vergleichbarkeit der in verschiedenen kulturellen Kontexten erhobenen Daten – und damit der *fairness* des Tests – erneut thematisiert und einer grundlegenden Kritik unterzogen (vgl. Bonnet, 2002). Auf Basis dieser Kritik wurde eine vom *European Network of Policy Makers for the Evaluation of Education Systems* unterstützte Pilotstudie zur Evaluation einer alternativen Rahmenkonzeption und Methodik der Erfassung von Lese- und Verstehensleistungen erarbeitet (Bonnet et al., 2003), bei der auf den Einsatz von für alle Länder verbindlichen Aufgaben fast gänzlich verzichtet wird. Statt dessen kommt vorrangig Testmaterial zum Einsatz, das einzig dem jeweiligen Sprach- und Kulturraum entnommen wurde. Dieses Verfahren wurde vorgeschlagen, um Probleme der Übertragung und Übersetzung von Texten auszuschließen. Die Realisierung der vorgeschlagenen Konzeption ist jedoch methodisch mit erheblichen Problemen verbunden und kann im jetzigen Stadium (vgl. Bonnet et al., 2003) keinesfalls als gelöst angesehen werden. So sind die von Bonnet et al. vorgeschlagenen Verfahren zur Sicherstellung der Vergleichbarkeit und der gemeinsamen Verankerung der Tests nicht ausreichend. Um sicherzustellen, dass die jeweils landesspezifisch ausgewählten Texte zur Messung der Schülerfähigkeit im Lesen tatsächlich eine als eindimensional aufgefasste Kompetenz abbilden, müssen die Tests miteinander verankert werden. Üblicherweise geschieht dies über eine in allen Ländern administrierte Aufgabenmenge. Wenn sowohl die gemeinsam administrierten Aufgaben (Ankeraufgaben) als auch die jeweils landesspezifisch administrierten Aufgaben eine als eindimensional aufzufassende Kompetenz abbilden, lässt sich eine gemeinsame Metrik zwischen den verschiedenen Tests in den einzelnen Ländern herstellen (s. a. Kolen & Brennan, 1995). Der Vorschlag der Projektgruppe, für die Verankerung die sprachlichen Skalen eines Intelligenztests (CFT) zu verwenden, führt im Endeffekt dazu, dass die jeweils landesspezifisch ausgewählten Texte eine latente Schülerfähigkeit messen, die mit der im CFT gemessenen Fähigkeit identisch sein muss. Das ist eine Annahme, die kaum vermittelbar ist und darüber hinaus auch nicht geprüft wurde.

Der Ansatzpunkt der Projektgruppe – die Bedenken bezüglich der Angemessenheit von übersetzten Tests für die Zwecke eines internationalen Vergleichs – ist vor dem Hintergrund bisheriger Forschungen zu diesem Thema jedoch durchaus begründet. So machen Gierl und Khaliq (2001) deutlich, dass der Anteil von Aufgaben mit differenziellen Itemfunktionen (DIF), d. h. Aufgaben, bei denen sich eine unbeabsichtigte Verschiebung der Schwierigkeit der Aufgaben durch die Übertragung in eine andere Sprache und/oder Kultur zeigt, in verschiedenen kanadischen Untersuchungen (Französisch/Englisch) sehr hoch sein kann. Ähnliches berichten Allalouf (2003) für den Vergleich von Testitems im Israelischen und im Russischen und Angoff und Cook (1988) für die Übersetzung des SAT vom Englischen ins Spanische. Einen Überblick über weitere DIF-Studien, die auch bei Instrumenten zur Persönlichkeits- und Einstellungsdiagnostik durchgeführt wurden, geben Budgell, Namburty und Douglas (1995). Unter anderem aufgrund dieser Ergebnisse wurden in den letzten Jahren mehrere Arbeiten publiziert, die sich mit den teilweise vermeidbaren Problemen der durch eine Übersetzung von Testmaterial entstehenden Verschiebung der Schwierigkeit der Aufgaben beschäftigen und Möglichkeiten der Vermeidung derartiger Probleme diskutieren (Allalouf, 2003; Allalouf, Hambleton & Sireci, 1999; Ellis, 1989; Gierl & Khaliq, 2001; Hambleton, 1994; Sireci, 1997; van de Vijver & Hambleton, 1996).

Neben technischen Fragen der Übersetzung werden in der Literatur auch immer wieder die spezifischen Merkmale einzelner Sprachen thematisiert (vgl. Adams & Wu, 2002; s. a. Gierl & Khaliq, 2001). So hat die zwischen Sprachen variierende Satz- bzw. Textlänge nachweislich einen Einfluss auf die Aufgabenschwierigkeit. Ähnliches gilt auch für spezifische grammatikalische und linguistische Besonderheiten einzelner Sprachen, die systematisch schwerer oder leichter zu verarbeiten sind. Systematische Vor- oder Nachteile, die bei *allen* Aufgaben gleich wirken, lassen sich jedoch mit den Daten der PISA-Studie, die die Datenbasis dieses Beitrags dargestellt, nicht von wahren Kompetenzunterschieden zwischen den Ländern trennen und sind daher nicht Gegenstand der empirischen Prüfung.

Über die erwähnten Fehler bei der Übersetzung von Texten und Aufgaben hinaus kann jedoch der sprachliche Ursprung der Texte und Aufgaben noch eine weitere Varianzquelle beinhalten, die hier als kulturelle Färbung beschrieben wird. Die kulturelle Färbung von Texten und Aufgaben würde sich in einem Vorteil der Schüler aus dem kulturellen Hintergrund bemerkbar machen, aus dem die Texte und Aufgaben stammen. Für diesen Vorteil könnten dabei unterschiedliche Faktoren, wie z. B. textrelevantes Vorwissen, textrelevantes Alltagswissen sowie auch Wissen über typische Kommunikationsabsichten, Textgenres und Darstellungsformen verantwortlich sein, deren Ausprägungen aufgrund von schulischen und außerschulischen Lerngelegenheiten zwischen Kulturen variieren.

2 Fragestellung

Die von Bonnet (2002) formulierte Kritik an der Annahme der schwierigkeitsneutralen Übertragund Übersetzbarkeit der in international vergleichenden Tests verwendeten Texte und Aufgaben wird zum Anlass genommen, die Angemessenheit dieser Zweifel zu prüfen und damit die Frage der sprachlichen (und damit auch kulturellen) Nuan-cen und ihrer möglichen Auswirkungen auf die jeweils erfassten Schülerfähigkeiten für die PISA-2000-Untersuchung zur Lesekompetenz zu thematisieren.

Untersucht werden soll die Frage, ob die Messung von Lesekompetenz anhand von übersetzten Texten, die jeweils aus einem mehr oder weniger spezifischen kulturellen Kontext stammen, tatsächlich in jedem Land dieselbe Kompetenz abbildet, oder ob die Übersetzung und Übertragung von Texten und Aufgaben das Anforderungsniveau und das Anforderungsprofil dieser Aufgaben verändert. Sprach- und kulturbedingte Eigenschaften von Textaufgaben sollten sich – gleiches Fähigkeitsniveau vorausgesetzt – in höheren Lösungswahrscheinlichkeiten von Schülern beim Bearbeiten von Aufgaben und Texten aus ihrer Sprachgruppe nachweisen lassen. Schüler sollten also bei denjenigen Aufgaben und Texten einen Vorteil haben, die nicht übersetzt wurden und damit aus ihrem eigenen Sprachraum und/oder Land stammen. Unabhängig von der realen Zusammensetzung des PISA-Tests wird somit für die jeweils vertretenen Sprachgruppen analysiert, ob von einem potenziellen Vorteil auszugehen ist.

Neben der Frage, ob sich Effekte der Herkunftssprache der Aufgaben überhaupt bemerkbar machen und wie sich diese bei einem rein aus Texten und Aufgaben einer Landessprache bestehenden Test auswirken würde, ist es für die Beurteilung der Validität der Ergebnisse des internationalen Vergleichs zu PISA zentral, das Ausmaß eines solchen Effektes für den real eingesetzten PISA-2000-Test zu untersuchen. Mit Blick auf die Fairness des Tests wird daher untersucht, welchen Einfluss die erwarteten differenziellen Effekte tatsächlich auf die Ländermittelwerte im PISA-Lesetest haben. Zentral ist hier, ob die im PISA-Test zu findende Dominanz von Texten und Aufgaben aus dem Englischen dazu führt, dass englischsprachige Schüler allein hierdurch bessere Testergebnisse aufweisen und andere Sprachgruppen und Kulturen entsprechend benachteiligt werden (*cultural bias* des Tests)

In einem weiteren Schritt wird der Frage nachgegangen, ob der ggf. nachweisbare Vorteil bei Texten und Aufgaben des eigenen Sprachraums allein darauf zurückzuführen ist, dass das Aufgabenmaterial nicht übersetzt wurde. Hierzu wird geprüft, ob sich ein ggf. vorhandener Vorteil bei Texten und Items aus der Sprachgruppe der Schüler in allen Ländern, in denen diese Sprache gesprochen wird, im selben Umfang zeigt.

Annahmen über die erwarteten Effekte.

Eine Überlegenheit bei Aufgaben der eigenen Sprachgruppe kann dann vermutet werden, wenn die jeweils aus der Landessprache stammenden Texte und Aufgaben eine kulturelle Färbung aufweisen und die Schülerinnen und Schüler daher mehr explizite oder implizite Lerngelegenheiten hatten. Der PISA-Test setzt sich grundsätzlich aus Texten und Aufgabenmaterial zusammen, das von den Ländern als authentisches und für die Lebenswelt von Fünfzehnjährigen typisches Material eingereicht wurde. Berücksichtigt man ferner die Tatsache, dass bei der landesspezifischen Auswahl auch eine möglichst gute Passung der Inhalte der Rahmenkonzeption (vgl. OECD, 1999) für das jeweilige Land bzw. den jeweiligen Sprachraum maßgeblich war, ist anzunehmen, dass durch die Text- und Aufgabenauswahl auch tendenziell kulturspezifische Aspekte eingefangen werden, bei denen die Schüler des jeweiligen Landes bzw. Sprachraums im Vorteil sind. Der Effekt sollte aufgrund der anzunehmenden Homogenität in der Menge und Qualität der Lerngelegenheiten in einem Land besonders für die Sprachen nachweisbar sein, die lediglich in einem Land gesprochen werden.

Während der Nachweis eines differenziellen Vorteils für die Items aller Sprachgruppen angenommen wird, vermuten wir, dass sich dieser Vorteil aufgrund der z. T. geringen Anzahl entsprechender Items im Test in der Regel nicht auf die Ländermittelwerte auswirkt. Bezüglich der angloamerikanischen Texten und Aufgaben wird aufgrund der mehrere Länder umfassenden Gruppe der englischsprachigen Länder zwar ein geringerer Effekt vermutet, aufgrund der Dominanz dieser Texte und Aufgaben im Test kann jedoch angenommen werden, dass sich ein ggf. nachweisbarer Vorteil im Sinne eines *cultural bias* tatsächlich in veränderten Ländermittelwerten und damit auf der Ebene des Gesamttests bemerkbar macht.

Es wird weiterhin vermutet, dass ein eventuell vorhandener Vorteil bei Aufgaben aus dem eigenen Sprachraum nicht allein darauf zurückzuführen ist, dass diese Texte nicht übersetzt wurden. Differenzielle Itemfunktionen, die für einen Sprachraum insgesamt nachgewiesen wurden, sollten sich auch für die einzelnen Länder dieser Sprachgruppen belegen lassen. Entsprechend sollten sich auch zwischen den englischsprachigen Ländern unterschiedliche Vorteile und damit auch differenzielle Lösungswahrscheinlichkeiten bei im Original englischsprachigen Items nachweisen lassen.

3 Methode

3.1 Instrumente

In PISA werden entsprechend der Rahmenkonzeption Basisfähigkeiten gemessen, die für die Bewährung im späteren Leben, insbesondere beim Übergang ins Berufsleben, aussagekräftig sein sollen (Artelt, Stanat, Schneider & Schiefele, 2001; Baumert et al., 2001; OECD, 1999). Insgesamt besteht der Lesetest aus 129 Aufgaben, die sich auf 37 Texte beziehen. Die PISA-Tests wurden im *Multi-Matrix-Design* dargeboten, was bedeutet, dass nicht jeder Schüler alle dem jeweiligen Leistungstest zugrunde liegenden Aufgaben bearbeitet hat. Die Aufgaben verfolgen das Ziel, verschiedene lebenspraktisch relevante Leseanlässe, Inhalte und Leseanforderungen möglichst breit abzudecken. Die dem Test zugrunde liegende *literacy*-Konzeption zielt auf die Erfassung von funktionalen Basiskompetenzen, die als eine Voraussetzung für die Teilhabe am gesellschaftlichen und kulturellen Leben betrachtet werden (s. a. Artelt et al., 2001; Kirsch et al., 2003; OECD, 1999). Auf Basis der Ergebnisse eines im Jahr 1999 durchgeführten Feldtests wurden die Aufgaben bezüglich der Angemessenheit für die Zwecke eines internationalen Vergleichs überprüft. Neben den Hinweisen eines *Cultural Review Panels* und sehr detaillierten

Rückmeldungen der Länder zum gesamten Aufgabenmaterial wurden vor allem auch statistische Kriterien[1] bei der Auswertung der Feldtestdaten und der Daten der Hauptstudie berücksichtigt (Adams & Wu, 2002; Kirsch et al., 2003). Die Verfahren dienten dem Ziel, einen Test zu konstruieren, der in allen Ländern valide eine annäherungsweise als eindimensional zu konzipierende Lesefähigkeit abbildet. Als Testmodell wurde das einparametrische Rasch-Modell verwendet.

Der PISA-2000-Lesetest erfüllt aufgrund der nach dem Feldtest aus dem Jahre 1999 erfolgten Optimierung die vorab definierten Kriterien bezüglich der Angemessenheit für die Zwecke eines internationalen Vergleichs. Nicht enthalten bei dieser sehr aufwändigen Überprüfung waren jedoch DIF-Analysen in Abhängigkeit von der Sprachgruppe und der Ursprungssprache der jeweiligen Texte. Zwar wurde von Grisay (in Adams & Wu, 2002) analysiert, inwiefern die Länge der Texte in Abhängigkeit von der Sprache variiert, eine systematische Prüfung differenzieller Itemfunktionen blieb – nach Kenntnis der Autoren – jedoch aus. Eine erste Prüfung für einzelne Items des PISA-Tests, die im Original aus Frankreich stammen, findet sich bei Murat und Rocher (2003), die hierfür die Mantel-Haenszel-Statistik verwendet haben[2]

Mit Ausnahme der aus der *Adult Literacy Study* der *International Association for the Evaluation of Educational Achievement* (IEA) übernommenen Items wurden alle in PISA verwendeten Texte (z. T. auch Aufgaben) von den Teilnehmerländern zur Verfügung gestellt. Die Anzahl der Texte und Aufgaben, die von nicht-englischsprachigen Ländern eingereicht wurde und den Feldtest überstanden hat, ist dabei verhältnismäßig gering. Die insgesamt 129 Aufgaben des PISA-2000-Lesetests verteilen sich wie in Tabelle 1 dargestellt auf insgesamt 37 Texte und stammen aus acht unterschiedlichen Ursprungssprachen. Da der Test auf differenzielle Itemfunktionen einen Vergleich der Schwierigkeitsparameter der Aufgaben beinhaltet, beziehen sich die nachfolgenden Analysen auf die einzelnen Aufgaben und nicht auf die übergeordnete Einheit der Texte.

Rund 12 % der Aufgaben wurden aus der IEA *Adult Literacy Study* übernommen (s. a. Kirsch et al., 2003). Da diese bezüglich der jeweiligen Ursprungssprache nicht rückverfolgbar sind, konzentrieren sich die hier dargestellten Auswertungen nur auf die für PISA-Zwecke von den Ländern eingereichten Aufgaben und Texte. Die Aufzählung in Tabelle 1 verdeutlicht, dass nur zwei Länder bzw. Sprachgruppen jeweils mehr als 10 % der Aufgaben eingereicht haben. Neben Englisch als Ursprungssprache ist hier Französisch zu nennen. Aus Dänemark hat nur eine Aufgabe die Kriterien für die Aufnahme in den Haupttest erreicht.

Tabelle 1: Absoluter und relativer Anteil der Aufgaben aus dem PISA-Lesetest nach Ursprungssprachen

Ursprungssprache	Absoluter Anteil der Aufgaben und Texte[2]		Relativer Anteil der Aufgaben und Texte[2]	
Englisch	66	(17)	51 %	(46 %)
Finnisch	11	(3)	8 %	(8 %)
Französisch	18	(4)	14 %	(11 %)
Deutsch	5	(1)	4 %	(3 %)
Griechisch	3	(1)	2 %	(3 %)
Spanisch	4	(1)	3 %	(3 %)
Schwedisch	6	(1)	5 %	(3 %)
Dänisch	1	(1)	1 %	(3 %)
IEA-Items[1]	15	(8)	12 %	(22 %)
Total	129	(37)	100 %	(100 %[3])

[1] International Adult Literacy Study der IEA
[2] Texte in Klammern
[3] Abweichungen aufgrund von Rundungen

3.2 Stichprobe

Die PISA-Studie ist eine von der OECD-Staatengemeinschaft initiierte, zunächst für drei Erhebungszeitpunkte (2000, 2003, 2006) konzipierte Untersuchung der Leistungen von Schülerinnen und Schülern am Ende der regulären Schulzeit in den Bereichen Lesen, Mathematik und Naturwissenschaften. An dem ersten PISA-Zyklus, an dem sich ursprünglich 32 Staaten beteiligten, haben weltweit 174.896 Schülerinnen und Schüler teilgenommen (Baumert et al., 2001, 2003; OECD, 2001). Pro Land wurde eine (teils mehrfach-) stratifizierte Wahrscheinlichkeitsstichprobe gezogen (vgl. Adams & Wu, 2002). In Deutschland wurde dabei nach Bundesland und Schulform stratifiziert. Die Ziehungswahrscheinlichkeit einer Schule war proportional zu ihrer Größe (probability proportional to size, PPS). Innerhalb der Schulen wurden dann in der Regel (vgl. Baumert et al., 2003) 28 Fünfzehnjährige per Zufall gezogen. Zur korrekten Reproduktion der Populationszahlen ist es notwendig, bei den Datenanalysen die Ziehungswahrscheinlichkeit der Schüler in Form von Gewichten zu berücksichtigen. In die internationale Stichprobe, die auch diesen Analysen zugrunde liegt[3], gingen aus Deutschland die Daten von 5.073 Schülerinnen

[1] So wurden nur Aufgaben verwendet, die (1) nicht leichter oder schwieriger waren als im Durchschnitt, (2) keine positive Korrelation zwischen einem Distraktor und den Durchschnittsleistungen aufwiesen, (3) bei denen die Korrelation zwischen richtiger Antwort und der Durchschnittsleistung größer als .25 war, (4) Bei denen steps bei partial credits items geordnet waren und (5) bei denen der Itemfit größer als 1.20 oder kleiner als .80 war (vgl. Adams & Wu, 2002).

[2] Darüber hinaus präsentieren die Autoren eine erste grobe Analyse für alle Länder, bei der sie die Rangfolge der Itemschwierigkeiten im Sinne des Anteils korrekt gelöster Aufgaben verschiedener Länder clustern. Länder gleicher Sprachgruppen weisen nach diese Methode ähnliche Muster der Rangreihen der Items auf.

[3] Sämtliche Daten, die für die Analysen benötigt werden, stehen der interessierten Öffentlichkeit zur Verfügung. Dies gilt sowohl für die Rohdaten als auch die Itemparameter und die Information zur Ursprungssprache der Aufgaben, die jeweils auf der Homepage der OECD (www.pisa.oecd.org) veröffentlicht sind.

und Schülern aus 220 Schulen ein. Die Daten der PISA-Stichprobe erlauben es, Aussagen über den Leistungsstand der Population der fünfzehnjährigen Schülerinnen und Schülern in den untersuchten Ländern zu treffen.

3.3 Differenzielle Itemfunktionen (DIF) als Ausdruck systematischer Verzerrungen

Die in PISA verwendeten Modelle zur Bestimmung individueller Leistungswerte beruhen auf der Item-Response-Theorie (IRT; Rost, 2003) oder – wie man im Deutschen sagt – der probabilistischen Testtheorie. Die einparametrische Variante – das Rasch-Modell – wurde in PISA verwendet. IRT-Modelle bieten eine geeignete Möglichkeit, die auf den sprachlichen Ursprung der Texte und Aufgaben zurückzuführenden Verzerrungen zu untersuchen und zu quantifizieren. Skalenäquivalenz ist dann anzunehmen – und Verzerrungen damit auszuschließen – wenn die Beziehung zwischen den beobachteten Testwerten und der durch den Test gemessenen latenten Fähigkeit für unterschiedliche Subgruppen identisch ist (Drasgow, 1984). In IRT-Modellen wird die Beziehung zwischen der durch den Test gemessenen latenten Fähigkeit und der Antwort auf jede einzelne Aufgabe des Tests in Form einer Itemfunktion abgebildet, deren monoton ansteigende Funktion durch mathematische Modelle beschrieben werden kann (vgl. Rost, 2003). Wenn die Itemfunktionen für unterschiedliche Subgruppen bei einzelnen Items nicht gleich sind, drückt sich dies in differenziellen Itemfunktionen (DIF) aus, die – in Abhängigkeit von der Perspektive – als *bias*, als substanzielle Befunde über relative Stärken und Schwächen oder auch als nicht modellierte Mehrdimensionalität des Tests angesehen werden können (Baumert, Klieme & Watermann, 1998; Glöckner-Rist & Hojtink, 2003; Roussos & Stout, 1996).

Anders ausgedrückt: Ein perfekt eindimensionaler Test misst nach der dem Rasch-Modell zugrunde liegenden theoretischen Konzeption eine einzige homogene Fähigkeit, deren Ausprägung die Lösungswahrscheinlichkeit der Testaufgaben ausschließlich bestimmt. Demnach können alle Gruppenunterschiede in den Lösungswahrscheinlichkeiten von Testaufgaben – von zufällig verteilten Schätzfehlern abgesehen – vollständig auf unterschiedliche Fähigkeitsverteilungen zwischen den Gruppen (z. B. Länder) zurückgeführt werden. Wenn die Lösungswahrscheinlichkeit von Aufgaben zusätzlich und systematisch von weiteren Faktoren bestimmt wird, spiegelt sich dies in Abweichungen vom eindimensionalen Rasch-Modell wider. Diese Abweichungen lassen sich dann als DIF-Werte nachweisen, wenn sie für einzelne Aufgaben in einzelnen Schülergruppen (Sprachgruppen) bestehen. Im Sinne des Testmodells weisen diese Aufgaben somit eine kulturell bedingte Verzerrung (*cultural bias*) auf.

IRT-basierte Verfahren bieten einen optimalen Weg, differenzielle Itemfunktionen (DIF) zu entdecken und ihre Effektstärke zu bestimmen. Im Vergleich zu DIF-Prozeduren, die auf der klassischen Testtheorie aufbauen, sind IRT-basierte Verfahren nicht bzw. weniger durch mittlere Fähigkeitsunterschiede zwischen den Gruppen konfundiert, da die p-Werte bei der bedingten Wahrscheinlichkeit für Personen unterschiedlicher Gruppen, aber gleichen Fähigkeitsniveaus verglichen werden (Camilli & Shapard, 1994). Nur bei einer Übereinstimmung der gesamten latenten Fähigkeitsverteilung der Vergleichsgruppen sind Differenzen im Sinne des Anteils gelöster Aufgaben zwischen Gruppen verzerrungsfreie Schätzer differenzieller Itemfunktionen. Camilli und Shepard (1994) empfehlen deshalb zu Recht, diese Indizes trotz ihrer vordergründigen Plausibilität nicht zu verwenden.

Die Differenz der Logit-Werte der Itemparameter (DIF-Werte der Aufgaben) für die in Frage kommenden Gruppen kann zugleich als direktes Maß der Effektstärke verwendet werden (Camilli & Shapard, 1994; Holland & Wainer, 1993). Die in Effektstärken ausgedrückten IRT-basierten differenziellen Itemfunktionen der Aufgaben des PISA-Lesetests für einzelne Sprachgruppen (inkl. ihrer Standardfehler) stellen daher die nachfolgend verwendete Methode der Überprüfung der Annahme von sprachlichen und/oder kulturellen Effekten bei den PISA-Leseaufgaben dar. Die Analysen wurden mit dem Computerprogramm ConQuest (Wu, Adams & Wilson, 1998) durchgeführt. Die Modellierung der differenziellen Itemfunktionen bzw. des Vorteils bei Items aus dem eigenen Sprachraum geschah dabei nach folgenden Prinzipien:

1) Die Analysen spezifischer Vorteile einzelner Ländergruppen erfolgen durch die Kontrastierung mit einer möglichst kultur-neutralen Referenzgruppe, was bedeutet, dass die Itemparameterschätzung über alle Sprachgruppen ermittelt und die DIF-Werte der Länder für einzelne Itemgruppen jeweils relativ hierzu interpretiert werden können (Ellis & Kimmel, 1992).
2) Basis der DIF-Analysen ist ein näherungsweise eindimensionaler Test, bei dem die Lösungswahrscheinlichkeit einzelner Aufgaben – von zufällig verteilten Fehlern abgesehen – lediglich auf die durch den Test gemessene Kompetenz zurückzuführen sein sollte. Abweichungen von dieser Annahme (Nachweis differenzieller Itemfunktionen) für Items unterschiedlicher Ursprungssprachen bei Schülern der jeweiligen Sprachgruppen werden als *cultural bias* bezeichnet, da sie Unterschiede abbilden, die nicht durch den Test intendiert waren (Roussos & Stout, 1996).
3) Um zu einer möglichst unverzerrten Schätzung differenzieller Itemfunktionen zu gelangen, wurde die Prüfung der erwarteten Vorteile der Schüler einer Sprachgruppe für Items dieser Sprachgruppe jeweils einzeln und im Verhältnis zu dem nur aus anglo-amerikanischen Items bestehenden Test durchgeführt. Bei Ländergruppen mit mehr als fünf Items aus dieser Ursprungssprache wurde die Prüfung auf eventuell vorhandene DIF-Werte bei diesen Items mit jeweils max. fünf Items im Vergleich zu den 66 englischsprachigen Items durchgeführt, was ausschließt, dass die Fähigkeitsschätzung der

Gruppen durch die vom *bias* betroffenen Items verfälscht ist (Roussos & Stout, 1996).

Aufgrund der klaren Dominanz anglo-amerikanischer Aufgaben im PISA-Lesetest wird die Prüfung der Effekte englischsprachiger Items nicht durch sukzessives Hinzunehmen der englischsprachigen Items zu einem ansonsten aus Items anderer Ursprungssprachen bestehenden Test modelliert, sondern durch eine Neuberechnung der Leistungswerte englischsprachiger Schüler auf Basis eines Tests, der a) keine im Original englischsprachigen Items enthält und b) lediglich aus englischsprachigen Items bsteht. Der Vergleich der Ländermittelwerte des Originaltests und der reduzierten Tests erlaubt eine Abschätzung des Vor- bzw. Nachteils englischsprachiger Schüler bei unterschiedlichen Zusammensetzungen des Tests.

4 Ergebnisse

Um die Eindeutigkeit der Interpretierbarkeit von differierenden Itemparametern zu erhöhen, wurde die DIF-Auswertung für näherungsweise eindimensionale Tests durchgeführt. Dies hat zur Folge, dass die Auswertung für die drei im PISA-Lesetest empirisch voneinander unterscheidbaren Kompetenzdimensionen separat durchgeführt wurde. Hierbei handelt es sich um die Unterscheidung in die Lesekompetenz bei (1) nicht-kontinuierlich geschriebenen Texten, (2) literarischen Texten und (3) anderen, kontinuierlich geschriebenen Texten (s. a. Artelt & Schlagmüller, 2004). Da die getrennte Darstellung der Befunde für die drei Kompetenzdimensionen jedoch aufgrund der kleinen Fallzahlen pro Sprachgruppe und Subdimension nicht für alle Sprachgruppen möglich ist, werden zunächst die Ergebnisse der getrennten DIF-Analysen der Logit-Parameter für alle Aufgaben zusammen besprochen, die für den Vergleich auf eine gemeinsame Metrik mit einem Mittelwert von 0 und einer Standardabweichung von 1 transformiert wurden.

Modelliert wurden die Effekte der Sprache, indem jeweils geprüft wurde, ob die Schüler unterschiedlicher Sprachen einen Vorteil bei Aufgaben haben, die aus ihrem eigenen Land bzw. Sprachraum stammen. Für die Sprachen Englisch, Französisch, Deutsch und Spanisch wurden hierbei jeweils Gruppen von Ländern gebildet, in denen die jeweilige Sprache gesprochen wird (vgl. Tabelle 2). Abbildung 1 gibt die Ergebnisse der DIF-Analysen wieder. Dargestellt sind Mittelwerte und Standardabweichungen der DIF-Parameter für die einzelnen Sprachgruppen. Negative Werte stehen dabei für einen differenziellen Vorteil der jeweils bezeichneten Sprachgruppe, positive Werte für einen differenziellen Nachteil. Der Mittelwert von –.21 für die französischsprachigen Schüler bedeutet dabei, dass die Leistungsmessung anhand der 18 Aufgaben, deren Ursprungssprache Französisch war, im Mittel zu einem Vorteil von d = .21 für französischsprachige Schüler führt. Da auch rund zwei Drittel der DIF-Werte (±1 Standardabweichung in Abbildung 1) der französischen Aufgaben ebenfalls deutlich im negativen Bereich liegen und damit auf einen Vorteil französischer Schüler bei Aufgaben aus ihrem Sprachraum hinweisen, kann dieser Effekt als substanziell und systematisch bezeichnet werden.

Tabelle 2: Gruppierung von Ländern zu Sprachgruppen

Ursprungssprache der Aufgaben	Länder aus der PISA-2000-Untersuchung
Englisch	Australien, Irland, Neuseeland, Vereinigtes Königreich, Vereinigte Staaten*
Französisch	Frankreich, französischsprachige Schweiz, Belgien (Wallonien)
Deutsch	Deutschland, Österreich, deutschsprachige Schweiz
Spanisch	Spanien, Mexiko

* Kanada musste aufgrund der fehlenden Information über die Regionalkennung (englisch- bzw. französischsprachig) ausgeschlossen werden.

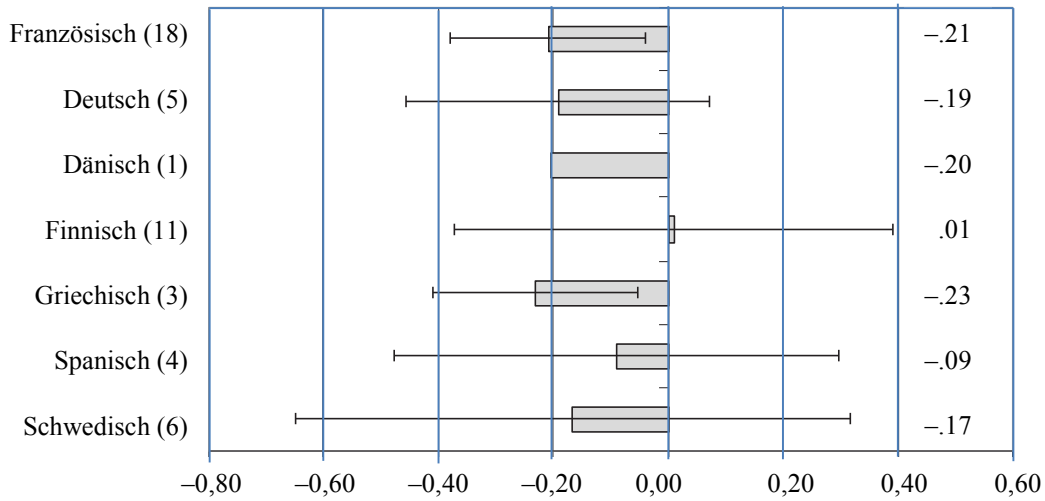
Auch für andere Sprachen – insbesondere Griechisch, z. T. auch Deutsch – ist ein deutlicher und für fast alle Items der eigenen Sprachgruppe nachweisbarer DIF-Wert zu Gunsten der Schüler dieser Sprachgruppen feststellbar. Kein Vorteil lässt sich hingegen bei finnischen Schülern nachweisen, wo die DIF-Werte zudem sehr stark streuen. Die ebenfalls in Abbildung 1 abgetragenen Standardabweichungen verdeutlichen auch für die schwedischen und spanischen Items eine erhebliche Streuung. Nicht für alle Items aus diesen Sprachen lassen sich Vorteile für Schüler dieser Sprachen feststellen.

Zur Veranschaulichung des Effektes bei französischen Aufgaben, die zahlenmäßig die größte Gruppe darstellen, sind die Unterschiede in den Itemparametern der Aufgaben auf der Logit-Skala in Abbildung 2 noch einmal gesondert aufgeführt. Jedes Item wird durch eines von vier verschiedenen Symbolen dargestellt. Die Symbole drücken die Zugehörigkeit der Items zu den vier verschiedenen Texten aus. Negative Werte stehen für leichte, positive für schwere Items. Würden keine Unterschiede zwischen den über alle Länder berechneten Item-Schwierigkeitsparametern und denen für Schüler aus französischsprachigen Ländern bestehen, lägen alle Werte auf der eingezeichneten Diagonalen. Abweichungen von der Winkelhalbierenden stellen hingegen differenzielle Itemfunktionen dar. Die Werte, die oberhalb der Diagonalen liegen, deuten an, dass französischsprachige Schüler hier einen relativen Vorteil haben; Werte unterhalb der Diagonalen hingegen stehen für einen relativen Nachteil. Die Abbildung verdeutlicht, dass sowohl im Durchschnitt wie auch fast bei allen Einzelaufgaben deutliche Vorteile französischsprachiger Schüler bei im Original französischsprachigen Aufgaben bestehen und dies allem Anschein nach bei den vier zugrunde liegenden Texten in vergleichbarem Maße der Fall ist.

Bei der bisherigen Darstellung wurden die getrennt für die drei Teilkompetenzen im Lesen berechneten

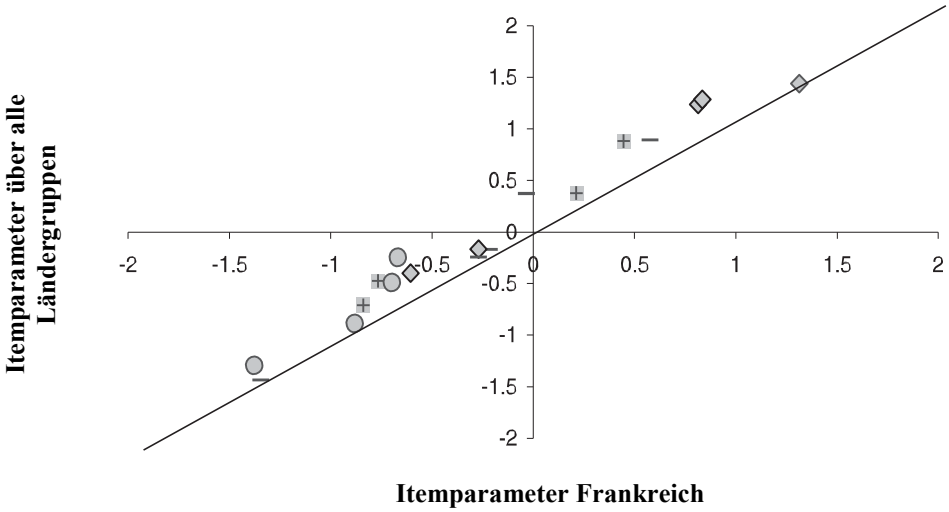
DIF-Parameter für Schüler der jeweiligen Sprachgruppe gemeinsam berichtet. Da sich die Teilskalen im Lesen jedoch auch hinsichtlich hier relevanter Merkmale unterscheiden, werden die entsprechenden Ergebnisse – trotz des Verlustes ganzer Sprachgruppen – unter Berücksichtigung der teilweise sehr geringen Fallzahlen noch einmal getrennt berichtet.

Auch die getrennte Betrachtung der drei Subskalen im Lesen führt zu strukturell ähnlichen Ergebnissen. Schülerinnen und Schüler – mit Ausnahme der finnischen Schüler – haben bei Aufgaben, die ihrem Sprach- und Kulturraum entstammen, unabhängig von der betrachteten Kompetenzdimension einen Vorteil (s. Abbildung 3). Der Effekt ist jedoch – besonders im Fall der spanischen und schwedischen Items – nicht durchgängig bei allen Aufgaben nachweisbar. Auf den beiden Subskalen im Lesen, in denen französischsprachige Items vorkommen, zeigt sich ein Vorteil der französischsprachigen Schüler. Mit einem Mittelwert, der im Falle der kontinuierlich geschriebenen Texte bei $M = -.20$ ($SD = .17$) und im Fall der nicht-kontinuierlichen Texte bei $M = -.27$ ($SD = .19$) liegt, wird dies unabhängig von der Subskala im Lesen deutlich.



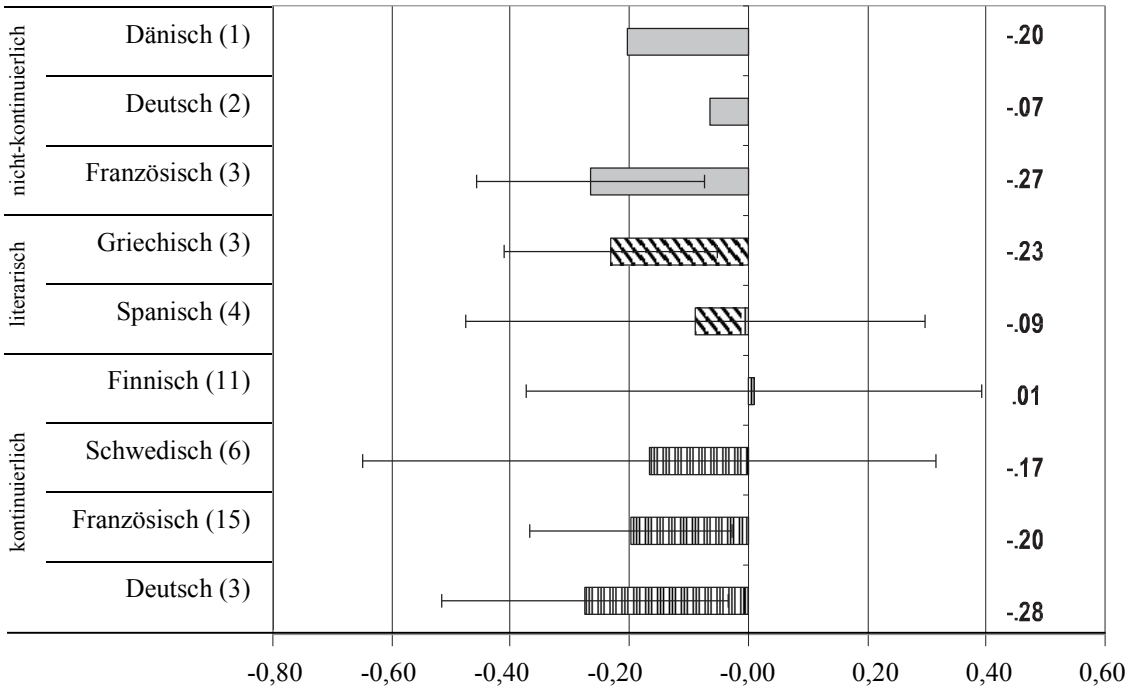
Anmerkungen:
Fehlerindikator: Standardabweichung (nur bei 3 und mehr Items abgetragen)
Negative Werte: Vorteil für Schüler derselben Sprachgruppe wie Ursprungssprache der Items
Positive Werte: Nachteil für Schüler derselben Sprachgruppe wie Ursprungssprache der Items

Abbildung 1: Differenzielle Itemfunktionen bei Aufgaben im Lesen.
Vor- bzw. Nachteil für Schüler aus Sprachgruppen der Ursprungssprache der Aufgaben.



Anmerkung: Die Symbole der Items repräsentieren die vier unterschiedlichen zugrunde liegenden Texte.

Abbildung 2: Scatterplot der Itemparameter bei Aufgaben, die im Original französischsprachig waren ($N = 18$), für französischsprachige und nicht-französischsprachige Schüler.



Anmerkungen:
Fehlerindikator: Standardabweichung (nur bei 3 und mehr Items abgetragen)
Negative Werte: Vorteil für Schüler derselben Sprachgruppe wie Ursprungssprache der Items
Positive Werte: Nachteil für Schü- ler derselben Sprachgruppe wie Ursprungssprache der Items

Abbildung 3: Differenzielle Itemfunktionen bei Aufgaben der drei Subskalen im Lesen. Vor- bzw. Nachteil für Schüler aus Sprachgruppen der Ursprungssprache der Aufgaben.

Eine Deutung differenzieller Befunde in Abhängigkeit von der Teilkompetenz im Lesen ist auch für die anderen Sprachgruppen aufgrund der z. T. geringen Itemzahlen und der Beschränkung einiger Sprachen auf eine einzelne Teildimension kaum möglich. Es lässt sich jedoch vermuten, dass der Effekt der Ursprungssprache in Deutschland bei nicht-kontinuierlichen Texten geringer ist. Für Frankreich bestätigt sich das Bild eines geringeren Effektes bei nicht-kontinuierlich geschriebenen Texten jedoch nicht. Bei literarischen Texten zeigt sich erwartungsgemäß dasselbe Bild. Auch hier finden sich – besonders bei griechischen Aufgaben – Vorteile für Schüler beim Bearbeiten von Items aus dem eigenen Land und der eigenen Sprache.

4.1 Wie stark macht sich der nachgewiesene DIF-Wert in den Leistungen der Schüler dieser Länder bemerkbar?

Das Ausmaß des auf die Ursprungssprache zurückführbaren Effektes lässt sich auf zwei verschiedene Arten abschätzen. Eine Möglichkeit besteht darin, die Veränderungen im mittleren Abschneiden der Länder für einen jeweils nur aus Aufgaben der Landessprache bestehenden Test «hochzurechnen». Wenn man von einem hypothetischen Test ausgeht, der ausschließlich aus Aufgaben der jeweiligen Landessprache besteht und der im Mittel denselben DIF-Wert wie in den hier vorgestellten Analysen aufweisen würde, lässt sich der mittlere Leistungszugewinn für Schüler derselben Landessprache direkt aus den in Abbildungen 1 und 3 dargestellten Werten ablesen. Bei einem mittleren DIF-Wert von –.21, der sich bei französischen Aufgaben zeigt, würden französischsprachige Schüler in einem solchen Test .21 Einheiten der Standardabweichung (Effektstärke d = .21) besser abschneiden. Auf der internationalen PISA-Metrik, die einen Mittelwert von 500 und eine Standardabweichung von 100 aufweist, entspräche dies 21 Punkten (Artelt et al., 2001). Für die anderen Ländergruppen lässt sich der Effekt analog ablesen. Schweden würde bei einem hypothetischen PISA-Test mit rein schwedischen Ursprungsitems 17 Punkte besser abschneiden und der deutsche Mittelwert würde 19 Punkte höher liegen. Es sei allerdings angemerkt, dass bei diesem Gedankenspiel die Streuung der DIF-Parameter nicht berücksichtigt wird. Diese Berechnungen dienen zur Verdeutlichung des potenziellen und aufgrund der vorgestellten Ergebnisse plausiblen Effektes, sie sagen jedoch keinesfalls aus, dass ein Test, der tatsächlich nur aus Aufgaben der jeweiligen Sprachgruppe bestünde, notwendigerweise dieselben Effekte erzielen würde.

Eine zweite Art der Betrachtung der Effekte, die in Bezug auf die Frage der *fairness* des Tests bzw. des Ausmaßes der Verzerrung aufgrund des *cultural bias*

entscheidender ist, geht von der tatsächlichen Zusammensetzung des PISA-Tests aus. Aufgrund der z. T. nur geringen Repräsentanz von Aufgaben aus unterschiedlichen Ursprungssprachen ist der Effekt bezüglich der Gesamtleistungder Schülerinnen und Schüler der betrachteten Länder nur sehr gering und in keinem Fall zufallskritisch absicherbar: Durch eine erneute Skalierung der Leistungsdaten der Schüler wurden hierzu die Ländermittelwerte ermittelt, die sich für einen Test ohne die jeweils aus der Sprachgruppe der Schüler stammenden Aufgaben ergeben würden. Durch die Nicht-Berücksichtigung dieser Items, die insbesondere im Fall von Frankreich deutliche Effekte zeigen, wird also deutlich, wie die Länder abgeschnitten hätten, wenn aus ihrer Sprachgruppe überhaupt keine Aufgaben enthalten gewesen wären.

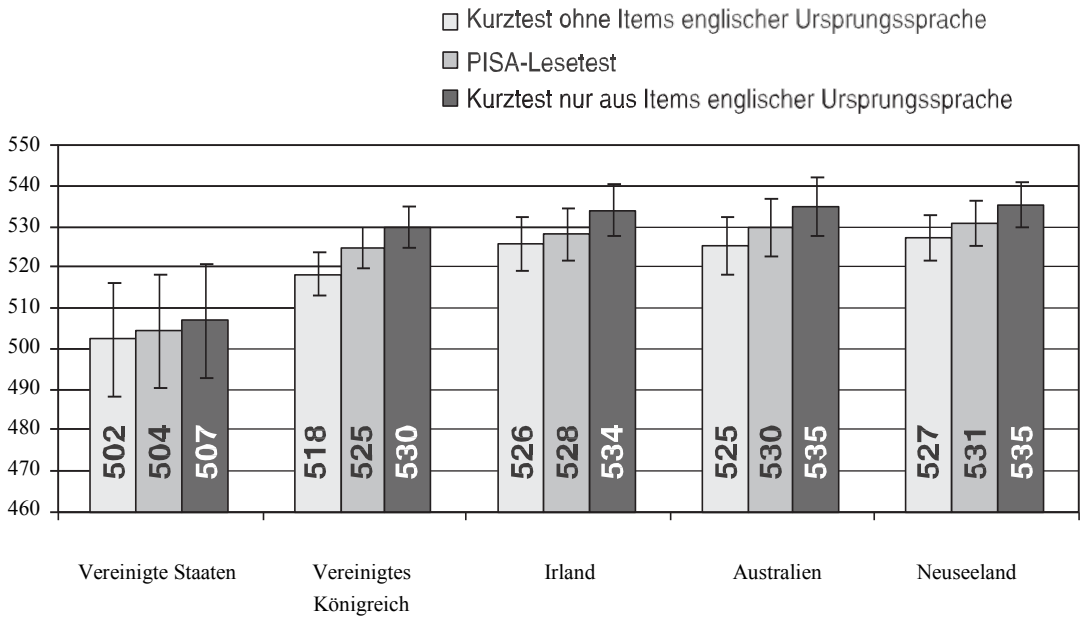
Wie aufgrund der DIF-Analysen zu erwarten war, zeigt sich für die Länder jeweils ein numerisch schlechteres Ergebnis bei der Skalierung eines Tests, der keine Aufgaben enthält, die ursprünglich aus der Sprachgruppe des untersuchten Landes stammen. Der Effekt ist jedoch insgesamt sehr gering und macht auch im Fall der französischsprachigen Schüler (N = 18 Aufgaben) lediglich eine Differenz von 2.6 Punkten auf der PISA-Skala (M = 500, SD = 100) aus. Vor dem Hintergrund der ebenfalls in der Größenordnung dieses Unterschieds liegenden Standardfehler der Mittelwerte (für Frankreich SE = 2.7) wird deutlich, dass diese Differenz nicht zufallskritisch abgesichert werden kann. Für die Schüler aus Deutschland führt die Nicht-Berücksichtigung der ursprünglich deutschsprachigen Items zu einer «Verschlechterung» von 1.7 Punkten; ein Effekt der ebenfalls nicht statistisch nachweisbar ist.

4.2 Führt die Dominanz anglo-amerikanischer Aufgaben im PISA-Lesetest zu Vorteilen für englischsprachige Schüler?

Rund die Hälfte der Aufgaben aus dem PISA-Lesetest stammen ursprünglich aus dem anglo-amerikanischen Sprachraum. Die deutliche Dominanz wird durch die aus der IEA-Studie übernommenen Items noch verstärkt, die aller Wahrscheinlichkeit nach auch aus dem englischen Sprachraum stammen. Die Frage, ob Schüler aus englischsprachigen Ländern bei diesen Aufgaben einen Vorteil haben, lässt sich aufgrund der deutlichen Dominanz englischsprachiger Items im Test nicht optimal mit DIF-Analyse nach dem o. g. Muster beantworten. Da die durch den Gesamttest festgestellten Fähigkeitsniveaus der Gruppen für die Berechnung von DIF-Parametern benötigt werden, würde eine Dominanz von evtl. mit bias behafteten englischsprachigen Items zu einer Verzerrung der DIF-Schätzungen dieser Itemgruppe führen (Roussos & Stout, 1996). Ein sukzessives Hinzufügen englischsprachiger Items zu einem rein aus Items der anderen Sprachen bestehenden Test hätte eine Alternative dargestellt. Hier wurde jedoch statt dessen das anschauliche Maß einer Neuskalierung ohne englischsprachige Items bzw. allein auf Basis englischsprachiger Items gewählt. Analysiert wird also, wie die Schüler aus englischsprachigen Ländern bei zwei Kurzformen des PISA-Tests im Vergleich zum Originaltest abschneiden. Die Kurzformen bestehen entweder nur aus Items, die ihrer Ursprungssprache nach aus dem Englischen stammen oder lediglich aus Items, die alle ihrem Ursprung nach aus anderen Sprachen stammen. Hierzu die Leistungswerte für englischsprachige Schüler bei diesen Testkurzformen getrennt skaliert und mit den Mittelwerten der Originalskalierung verglichen.

Der Richtung nach verdeutlichen die in Abbildung 4 dargestellten Mittelwerte der verschiedenen Testfassungen einen Vorteil englischsprachiger Schüler bei englischsprachigen Items. Die Unterschiede, die sich zu dem Mittelwert der Originalskalierung ergeben, sind jedoch in keinem der Länder statistisch nachweisbar. Vergleicht man zunächst über alle englischsprachigen Länder hinweg die Differenz zwischen dem Mittelwert des Gesamttests mit dem derjenigen Kurzform, die keine Items englischer Ursprungssprache enthält, ergibt sich ein «Nachteil» von 3.8 Punkten. Der Vorteil, der sich – wiederum über alle Länder betrachtet – im Vergleich zum Originaltest bei dem lediglich aus englischsprachigen Items bestehenden Kurztest ergibt liegt bei 4.6 Punkten. Beide Differenzen sind statistisch nicht signifikant. Auch bei einer getrennten Betrachtung der einzelnen Länder sind die Unterschiede zwischen den Kurzformen und dem Gesamttest nicht signifikant. Von den insgesamt fünf englischsprachigen Ländern der Studie ist der Effekt der sprachlichen Herkunft der Items im Vereinigten Königreich und in Irland am stärksten. Aber auch hier ist nachweisbare «Verschlechterung» (Effektstärke d = .05 Vereinigtes Königreich; d = .06 Irland) bzw. «Verbesserung» (Effektstärke d = .06 Vereinigtes Königreich; d = .03 Irland) im mittleren Leistungsniveau zwischen dem jeweiligen Kurztest und dem Originaltest statistisch nicht signifikant, was sich auch an den überlappenden Vertrauensintervallen um die zwei Mittelwerte pro Land aus Abbildung 4 entnehmen lässt.

Betrachtet man, wie sich die Ergebnisse der englischsprachigen Länder im Verhältnis zu denen der anderen PISA-Länder verändern würden, wenn die englischsprachigen Schüler überhaupt keine englischsprachigen Aufgaben bearbeitet hätten, so ändert sich auch die Rangfolge der nach dem mittleren Abschneiden geordneten PISA-Teilnehmerstaaten (s. a. Artelt et al., 2001) kaum: Drei der fünf Länder verändern ihre Rangposition um jeweils einen Platz. So rückt Australien von Platz 4 auf Platz 5, während Irland vom fünften auf den vierten Platz aufsteigt. Insgesamt beträgt der Unterschied zwischen diesen beiden Ländern jedoch nur 0.5 Punkte. Das Vereinigte Königreich würde zudem vom ehemals siebten auf den achten Rangplatz rücken. Für keines der fünf Länder würden sich darüber hinaus die vorhandenen Unterschiede zum OECD-Durchschnitt verändern. Mit Ausnahme der Vereinigten Staaten liegen auch bei einer Skalierung ohne englischsprachige Items alle englischsprachigen Länder oberhalb des Durchschnitts der OECD-Länder, während der Wert der USA nicht vom OECD-Mittelwert verschieden ist.



Anmerkung: Fehlerindikator: Standardfehler (± 2 SE).

Abbildung 4: Mittelwerte von fünf englischsprachigen Ländern im Gesamttest Lesen und in zwei Kurzformen, die entweder nur aus engischsprachigen Items bestehen oder überhaupt keine englischsprachigen Items enthalten.

Der Vorteil, der sich aus der Neuskalierung für die englischsprachigen Länder ableiten lässt, ist vor dem Hintergrund der Verschiebung der Rangplätze als auch der statistisch nachweisbaren Veränderung als relativ gering zu bezeichnen. Deutlich wird jedoch, dass das Vorhandensein von englischsprachigen Items im Test zu einer numerischen Verbesserung führt. Im Vereinigten Königreich, wo die sprachlichen Effekte insgesamt relativ stark ausgeprägt sind, ist der Unterschied zwischen dem aus englischen Items bestehenden Test und dem einem Test, der keine englischsprachigen Items enthält, dabei signifikant.

Aus Abbildung 4 wird ebenfalls deutlich, dass der Vorteil für englischsprachige Schüler bei im Original aus dem Englischen stammenden Items nicht in allen englischsprachigen Ländern gleich stark ausgeprägt ist. Mit 5 bzw. 6 Punkten Unterschied gegenüber dem Originaltest ist der Vorteil in Irland und im Vereinigten Königreich am größten, in den Vereinigten Staaten hingegen mit 2.0 Punkten verhältnismäßig klein. Die Unterschiede verdeutlichen indirekt, dass es auch innerhalb der englischsprachigen Länder Unterschiede gibt, der Vorteil bei Texten und Aufgaben aus der eigenen Sprache also nicht allein darauf zurückzuführen ist, dass diese Texte nicht übersetzt wurden. Die Unterschiede machen deutlich, dass auch innerhalb eines Sprachraums mit Abweichungen in der Lösungswahrscheinlichkeit der Aufgaben zu rechnen ist. Diese Abweichungen sind nicht auf Übersetzungen zurückzuführen, sondern kennzeichnen Variationen, die als kulturelle bedingte potenzielle Verzerrungen auf der Itemebene (*cultural bias*) zu interpretieren sind.

5 Diskussion

Es wurde vermutet, dass sich die kulturelle und sprachliche Herkunft der Aufgaben des PISA-Tests systematisch in den Schülerleistungen nachweisen lässt. Eine kulturelle Färbung der Aufgaben wurde u. a. auch deshalb als sehr wahrscheinlich angesehen, weil die Texte und Aufgaben durch die in der PISA-Studie angestrebte Authentizität der Texte von den Teilnehmerländern selbst eingereicht werden konnten. Die Ergebnisse zu den gefundenen Effekten der sprachlichen Herkunft der Aufgaben lassen sich auf zwei unterschiedliche Arten diskutieren: Einerseits bezüglich des prinzipiellen Nachweises eines auf die Ursprungssprache der Items zurückführbaren Effektes. Andererseits bezüglich des *cultural bias*, d. h. des Ausmaßes an Verzerrung, dass durch die differenziellen Vorteile für den PISA-Gesamttest nachweisbar ist. Beide Betrachtungsarten sollen hier noch einmal aufgegriffen werden.

Die Tatsache, dass sich der Ursprung der Sprache prinzipiell nachweisen lässt – und im Fall eines rein aus den Items einer Ursprungssprache bestehenden Tests auch Veränderungen der Ländermittelwerte von 1/5 Standardabweichung nach sich ziehen könnte – steht dabei im Einklang mit den bisher publizierten Befunden zum DIF beim Übertragen von Texten in andere Sprachen und Kulturen (vgl. Allalouf, 2003; Allalouf et al., 1999; Budgell et al., 1995; Gierl & Khaliq, 2001; Murat & Rocher, 2003). Die Vorteile bei Aufgaben, die im Original aus der eigenen Sprachgruppe stammen, lassen darauf schließen, dass Faktoren wie etwa die Vertrautheit mit den Inhalten und der Form des einge-

setzten Stimulusmaterials die Lösungswahrscheinlichkeit der Aufgaben der Schüler mitbestimmen. Die Analysen zeigen, dass ein hypothetischer Test, der lediglich Aufgaben enthält, die aus einem Sprachraum stammen, im Mittel zu einen 17 Punkte höheren Mittelwert auf der 500/100er-PISA-Skala für die Schülerinnen und Schüler dieser Sprachgruppe führen würde. Die Annahme, dass die Effekte der Ursprungssprache dann am stärksten sind, wenn die Sprache nur in einem Land gesprochen (und getestet) wurde, ließ sich nur z. T. bestätigten. Besonders die Ergebnisse zu den finnischen Items, bei denen kein Vorteil finnischer Schüler nachweisbar war, sind hiermit nicht vereinbar (s. u.).

In Bezug auf den von internationalen Vergleichsstudien anvisierten Kompetenzbereich stellen die hier vorgestellten Sprach- und Kultureffekte einen potenziellen *cultural bias* dar, da es sich um unbeabsichtigte Variationen in der Lösungswahrscheinlichkeit von Aufgaben handelt, die die Validität des Tests beeinträchtigen können. Die Ergebnisse sprechen jedoch nicht dafür, dass sich auf der Ebene des Tests insgesamt ein *cultural bias* nachweisen lässt, der die *fairness* des Tests beeinträchtigen würde. Unter Beachtung der tatsächlichen Zusammensetzung des Tests lassen sich für die getesteten Sprachgruppen *keine* zufallskritisch nachweisbaren Veränderungen der Gesamtmittelwerte nachweisen. Auch wenn deutlich wird, dass die Herkunftssprache der Aufgaben im PISA-Test für einige Länder hinsichtlich der Fähigkeitsmessung einen Unterschied machen kann (und bei einem ausschließlich aus diesen Aufgaben bestehenden Test auch machen würde), verdeutlicht das Ausmaß des Effektes bei Berücksichtigung der realen Zusammensetzung des Tests, dass die gezeigten Vorteile von Schülern beim Bearbeiten von Aufgaben aus dem eigenen Land keinen Anlass bieten, die Länderergebnisse von PISA anzuzweifeln. Auch für die dominante Gruppe der englischsprachigen Items ließ sich auf der Länderebene kein signifikanter Vorteil für englischsprachige Schüler nachweisen.

Der Vorteil bei englischsprachigen Items für die Schüler aus den fünf teilnehmenden englischsprachigen Ländern ist insgesamt weniger deutlich, als man aufgrund der starken Dominanz im Test erwarten könnte. Zudem zeigen die nach Ländern getrennten Ergebnisse, dass es nicht allein die nicht erfolgte Übersetzung sein kann, die für den Vorteil beim Bearbeiten von Aufgaben aus der eigenen Sprachgruppe verantwortlich ist. Auch zwischen den englischsprachigen Ländern variiert der Effekt und zeigt sich am deutlichsten für das Vereinigte Königreich und Irland, am geringsten hingegen für die Vereinigten Staaten. Auch die Rangfolge der Länderergebnisse zu PISA 2000 würde sich kaum verändern, wenn die englischsprachigen Schüler lediglich einen Test ohne englischsprachige Items bearbeitet hätten.

Die hier verwendeten IRT-basierten DIF-Analysen der Items sind im Vergleich zu Verfahren, die auf einem Vergleich der Schüleranteile mit korrekten Antworten pro Item beruhen, deutlich weniger fehlerbehaftet (Baumert et al., 1998; Camilli & Shapard, 1994). Gleichzeitig führt die Verwendung eines omnikulturellen Referenzpunktes, der durch die Berücksichtigung mehrerer Länder kulturell neutral wird, zur Möglichkeit, die kulturellen und/oder sprachlichen Spezifika einzelner Länder und Sprachgruppen klar abzugrenzen (Ellis & Kimmel, 1992). Eine noch adäquatere Modellierung der DIF-Effekte würde sich allerdings durch das von Roussos und Stout (1996) vorgeschlagene Verfahren des multidimensionalitätsbasierten DIF-Paradigmas ergeben, bei dem – wie hier auch verwendet – die Effekte von Bündeln von Items konfirmatorisch getestet werden und – wie hier nicht beachtet – die Schätzung der Itemparameter der Referenzgruppe jeweils ganz ohne die «verzerrte» Subgruppe geschieht. Dieses Verfahren ließ sich aufgrund des Multi-Matrix-Designs der PISA-Daten nicht anwenden, wurde jedoch durch die jeweils sukzessive Hinzunahme der verzerrten Items näherungsweise auch erreicht.

Die hier dargestellten Analysen belegen einen z. T. systematischen differenziellen Vorteil von Schülerinnen und Schülern bei der Bearbeitung von Aufgaben aus ihrem Sprach- und Kulturraum. Sie sind allerdings nicht in der Lage, Aussagen über die differenzielle Schwierigkeit einzelner Sprachen zu treffen. Systematische, d. h. bei allen Aufgaben gleich wirkende und schwierigkeitsbestimmende Effekte von Sprache wie die Satzlänge und die grammatikalische Struktur können nicht modelliert werden, da sie sich nicht von allgemeinen Kompetenzunterschieden trennen lassen (vgl. Sireci, 1997). Die Analysen von Grisay (in Adams & Wu, 2002) deuten jedoch darauf hin, dass dieser systematische Effekt der Satzlänge einzelner Sprache vermutlich als relativ gering veranschlagt werden kann, was sich auch an der Konsistenz der Befunde in den drei in PISA untersuchten Leistungsbereichen Mathematik, Lesen und Naturwissenschaften bestätigt (s. a. Stanat & Lüdtke, in Vorb.).

Unabhängig von der Frage, wie die Fairness des PISA-2000 Lesetests durch die nachgewiesenen Effekte beeinflusst wird, muss der auf der Ebene von Itembündeln gefundene substanzielle Effekt der Ursprungssprache als systematische Varianzquelle gesehen werden, die es in konzeptionellen Arbeiten genauer zu untersuchen gilt. Dies gilt insbesondere auch für die erwartungswidrigen Effekte der finnischen Items bzw. für die Erklärung der unterschiedlichen Effekte in den englischsprachigen Ländern.

Die Psychologie des Textverstehens bietet durch differenzierte hypothesengeleitete Studien sowohl begriffliche als auch methodische Werkzeuge, um den hier auf globaler Ebene festgestellten *cultural bias*, im Hinblick auf zugrunde liegenden Mechanismen weiter aufzuklären. Methodischistdabei besonders die systematische experimentelle Variation von Aufgabenmerkmalen, kombiniert mit einem quasiexperimentellen Vergleich von Schülerinnen und Schülern aus verschiedenen Ländern viel versprechend. Über Gründe für systematische Unterschiede lassen sich – teils aus der Psychologie des Textverstehens, teils aus der Kulturanthropologie – zahlreiche Hypothesen ableiten. So wäre ein Land, in dem literarische Geschichten weder in der Schule noch im Rahmen der Familie gelesen und rezipiert würden, durch die Aufnahme eines literarischen Textes benachteiligt. Gleiches gilt für die Aufnahme

von Texten über Inhaltsbereiche, die in bestimmten Ländern eher relevant sind als in anderen.

Sprache an sich ist ein konstituierendes Merkmal von Kultur (s. a. Ellis, 1989). Auch sozialisationstheoretische Ansätze gehen oft davon aus, dass Lesesozialisation zu großen Teilen mit dem Hineinwachsen in eine Kultur (Enkulturation) gleichgesetzt werden kann (s. a. Oerter, 1999). Nicht nur die Inhalte, sondern auch die Vorlieben, die Gewohnheiten und die Art der Kommunikation über Gelesenes können sich zwischen Kulturen erheblich unterscheiden. Darüber hinaus hängt die Art der Inferenzen beim natürlichen Lesen auch von den Zielen und Erwartungen der Lesenden ab (vgl. Graesser, Singer & Trabasso, 1994; Kintsch, 1998; s. a. Schnotz, 1994). Auch die Kenntnis des jeweiligen Genres oder der Kommunikationsabsicht führt zur Bildung von Erwartungen und Zielen beim Lesen und zur Auswahl bestimmter Strategien. So konnte Zwaan (1994) zeigen, dass allein die Ankündigung eines Textes als literarischer Erzähltext oder als Zeitungsartikel zu unterschiedlichen mentalen Repräsentationen mit variierendem Informationsgehalt führt.

Für eventuelle kulturelle Unterschiede ist auch die Unterscheidung zwischen bereichsspezifischem und bereichsunspezifischem Vorwissen relevant. Ausgehend von der Annahme, dass sich kulturelle Unterschiede sehr stark beim Alltagswissen zeigen (Olson, 1980; Olson & Pelletier, 2003), lässt sich ein differenzieller Effekt besonders bei narrativen Texten vermuten, bei denen Kohärenzlücken im Text über implizites Wissen des Lesers geschlossen werden müssen. Beim Lesen von narrativen Texten, die sehr nahe an kulturell geprägten Alltagserfahrungen sind, wird also in erster Linie allgemeines Weltwissen benötigt, das vorrangig aus Skripten und Schemata besteht und hochgradig überlernt und automatisiert ist.

Grundsätzlich unterstreichen die hier vorgestellten Analysen die Notwendigkeit, die Herkunftssprache der Items als systematische Varianzquelle zu berücksichtigen. In Tests, wie dem PISA-Test, die sich aus unterschiedlichen Herkunftssprachen zusammensetzen, mitteln sich die begünstigenden Effekte der jeweiligen Sprachgruppe jedoch tendenziell aus. Zumindest kann dies für die analysierten Sprachgruppen gesagt werden. Für diejenigen Länder, die keine Items und Texte in den PISA-Lesetest eingebracht haben (z. B. Japan, Korea) kann dabei zunächst keine Aussage gemacht werden. Es lässt sich jedoch vermuten, dass sie durch die Nicht-Berücksichtigung von Texten und Aufgaben aus ihrem Sprachraum potenziell benachteiligt werden.

Das Verfahren zum Nachweis von Vorteilen bei Items aus dem eigenen Sprachraum ist vor dem Hintergrund der im PISA-Test abgedeckten Breite der Aufgabenanforderungen sehr grob. Aber gerade der Nachweis eines Effektes für einzelne (homogene) Sprach- und Ländergruppen unter Missachtung der hierzu quer liegenden anderen Aufgabenanforderungen deutet darauf hin, dass es sich um einen deutlichen Effekt handelt, den es in zukünftigen international vergleichenden Untersuchungen zu kontrollieren und in konzeptionellen Arbeiten genauer zu untersuchen gilt.

Literatur

Adams, R. & Wu, M. (2002). *PISA 2000 technical report.* Paris: OECD.

Allalouf, A. (2003). Revising translated differential item functioning items as a tool for improving cross-lingual assessment. *Applied Measurement in Education, 16*, 55–73.

Allalouf, A., Hambleton, R. K. & Sireci, S. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*, 185–198.

Angoff, W. H. & Cook, L. K. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test (College board report No. 88–2)*. New York: College Entrance Examination Board.

Artelt, C. & Schlagmüller, M. (2004). Der Umgang mit literarischen Texten als Teilkompetenz im Lesen. Dimensionsanalysen und Ländervergleiche. In U. Schiefele, C. Artelt, W. Schneider & P. Stanat (Hrsg.), *Struktur, Entwicklung und Förderung von Lesekompetenz: Vertiefende Analysen im Rahmen von PISA-2000* (S. 169–196). Wiesbaden: VS Verlag für Sozialwissenschaften.

Artelt, C., Stanat, P., Schneider, W. & Schiefele, U. (2001). Lesekompetenz: Testkonzeption und Ergebnisse. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann & M. Weiß (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (S. 69–137). Opladen: Leske + Budrich.

Baumert, J., Artelt, C., Klieme, E., Neubrand, J., Prenzel, M., Schiefele, U., Schneider, W., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2003). *PISA 2000: Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland im Vergleich*. Opladen: Leske + Budrich.

Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J. & Weiß, M. (Hrsg.). (2001). *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich*. Opladen: Leske + Budrich.

Baumert, J., Klieme, E. & Watermann, R. (1998). Jenseits von Gesamttest- und Untertestwerten: Analyse differentieller Itemfunktionen am Beispiel des mathematischen Grundbildungstests der Dritten Internationalen Mathematik- und Naturwissenschaftsstudie der IEA (TIMSS). In H.-J. Herber & F. Hofmann (Hrsg.), *Schulpädagogik und Lehrerbildung. Festschrift zum60. Geburts tag von Josef Thonhauser* (S. 301–324). Innsbruck: Studien Verlag.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education, 9*, 387–399.

Bonnet, G., Daems, F., Clopper, C. D., Horner, S., Lappalainen, H.-P., Nardi, E., Remond, M.,

Robin, I., Rosen, M., Solheim, R. G., Ronnessen, F.-E., Vertecchi, B., Vrignaud, P., Wagner, A. K. H. & White, J. (2003). *Culturally balanced assessment of reading [c-bar]*. Retrieved 2003, from http://cisad.adc.education.fr/reva/pdf/cbarfinalreport.pdf.

Budgell, G. R., Namburty, S. R. & Douglas, A. Q. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19*, 309–321.

Camilli, G. & Shapard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). London: Sage.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95,* 134–135.

Ellis, B. B. (1989). Differential item functioning: Implications for test translations. *Journal of Applied Psychology, 74*, 912–921.

Ellis, B. B. & Kimmel, H. D. (1992). Identification of unique cultural responses patterns by means of item response theory. *Journal of Applied Psychology, 77*, 177– 184.

Gierl, M. J. & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*, 164–187.

Glöckner-Rist, A. & Hojtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling, 10*, 544–565.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229–244.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kirsch, I., de Jong, J., Lafontaine, D., Mc Queen, J., Mendelovits, J. & Monseur, C. (2003). *Reading for change. Performance and engagement across countries. Results from PISA 2000*. Paris: OECD.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating. Methods and practices*. New York: Springer.

Murat, F. & Rocher, R. (2003). *On the methods used for international assessments of educational competences* [La méthodologie des évaluations internationales]. Paris: OECD.

OECD (1999). *Measuring student knowledge and skills. A new framework for assessment.* Paris: OECD. [In deutscher Sprache: Deutsches PISA-Konsortium. (2000). *Schülerleistungen im internationalen Vergleich: Eine neue Rahmenkonzeption für die Erfassung von Wissen und Fähigkeiten.* Berlin: Max-Planck-Institut für Bildungsforschung.]

OECD (2001). *Knowledge and skills for life – First results from PISA 2000*. Paris: OECD. [In deutscher Sprache: OECD (2001). *Lernen für das Leben: Erste Ergebnisse der internationalen Schulleistungsstudie PISA 2000.* Paris: OECD].

Oerter, R. (1999). Theorien der Lesesozialisation – Zur Ontogenese des Lesens. In N. Groeben (Hrsg.), *Internationales Archiv für Sozialgeschichte der deutschen Literatur. 10. Sonderheft: Lesesozialisation in der Mediengesellschaft* (S. 27–55). Tübingen: Niemeyer.

Olson, D. R. (1980). Social aspects of meaning in oral and written language. In D. R. Olson (Ed.), *The social foundations of language and thought. Essays in honor of Jerome S. Bruner* (pp. 90–108). New York: Norton & Company.

Olson, D. R. & Pelletier, J. (2003). Schooling and the development of literacy. In J. Valsiner & K. Conolly (Eds.), *Handbook of developmental psychology* (pp. 358–369). London: Sage.

Rost, J. (2003). *Lehrbuch Testtheorie, Testkonstruktion* (2. Aufl.). Bern: Huber.

Roussos, L. A. & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355–371.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen: Untersuchungen zur Kohärenzbildung bei Wissenserwerb mit Texten* (Bd. 20). Weinheim: Beltz.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16*, 12–19.

Stanat, P. & Lüdtke, O. (in Vorb.). Internationale Schulleistungsvergleiche. In *Enzyklopädie der Psychologie: Kulturvergleichende Psychologie*. Göttingen: Hogrefe.

van de Vijver, F. J. R. & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89–99.

van de Vijver, F. J. R. & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29–37.

Wu, M., Adams, R. & Wilson, M. (1998). *ACER ConQuest: Generalised item response modeling software manual*. Camberwell: ACER Press.

Zwaan, R. A. (1994). Effect of genre expectation on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 920–933.