Olaf Bärenfänger

# Weibliche Sprachbegabung: Mythos oder Wirklichkeit?

„Schlaue Mädchen - dumme Jungen: Sieger und Verlierer in der Schule" - so lautet der provokante Titel des Leitartikels im „Spiegel" (21/2004). In eindrucksvoller Weise zeichnet der Beitrag massive geschlechtsspezifische Leistungsunterschiede im Bereich des primären und sekundären Bildungssektors nach, gar ist von einer „Jungenkatastrophe" in Deutschland die Rede.

Der Befund des „SpiegeF'-Artikels deckt sich insofern z. T. mit früheren Einschätzungen der Situation, als bei zahlreichen Pädagogen und in weiten Kreisen der Gesellschaft die Ansicht vorherrschte, das weibliche Geschlecht verfüge über eine besondere Sprachbegabung. Obschon die Geschlechterzusammensetzung in den Hörsälen deutscher Hochschulen diesen Eindruck auf den ersten Blick rein quantitativ zu bestätigen scheint, stellt sich dennoch die Forderung nach repräsentativen und zuverlässigen empirischen Daten. Die vorliegende Studie versucht daher mit einer großen Untersuchungsstichprobe die beiden Fragen zu beantworten:

1. Inwiefern lassen sich für den Bereich Deutsch als Fremdsprache Belege für eine größere Sprachbegabung von Mädchen bzw. Frauen finden?
2. Inwiefern manifestieren sich Unterschiede zwischen den Geschlechtern im Hinblick auf einzelne sprachliche Fertigkeiten?

Für die sachgemäße Beantwortung der beiden aufgeworfenen Fragen liegt die Wahl einer Forschungsmethodik nahe, die auf möglichst breiter Datenbasis einen ebenso zuverlässigen wie differenzierten Vergleich von weiblichem mit männlichem Geschlecht erlaubt.

Mit dem Test Deutsch als Fremdsprache (TestDaF) liegt ein nachgerade ideales Instrument für einen solchen Vergleich vor. Ein erster Vorteil des TestDaF liegt darin, dass er den Sprachstand der Testteilnehmer für die vier Fertigkeiten Lesen, Schreiben, Hören und Sprechen getrennt erfasst. Kalibrierte Itemschwierigkeiten[1] garantieren zudem die direkte Vergleichbarkeit sämtlicher Prüfungen. Durch die Einbeziehung von Daten aus mehreren Prüfungen lässt sich ferner eine hinreichend große und damit aussagekräftige Untersuchungsstichprobe erzeugen. Insofern Kandidaten aus der ganzen Welt an den Test-DaF-Prüfungen teilnehmen, ist schließlich von einer demographisch äußerst breit diversifizierten Stichprobe auszugehen, auf deren Grundlage sich verallgemeinerbare Schlussfolgerungen ziehen lassen.

## *1 Leistungsunterschiede zwischen den Geschlechtern im Licht der Forschung*

Bei einer Darstellung des Forschungsstands ist an erster Stelle die einflussreiche Pionierarbeit von Maccoby/Jacklin (1974) anzuführen. Die beiden Autoren kommen nach einer Sichtung der einschlägigen Literatur zu dem Ergebnis, Mädchen/Frauen verfügten generell über höhere sprachliche Fähigkeiten als Jungen/Männer, diese allerdings seien dem weiblichen Geschlecht im Bereich quantitativer und räumlicher Fertigkeiten überlegen. Diese Einschätzung geht in dieselbe Richtung wie Befunde aus dem Advanced Placement Program (AP), einem Wissenstest für den Zugang zu amerikanischen Colleges und Universitäten, bei dem Mädchen wiederum in den allgemeinsprachlichen Fertigkeiten besser abschneiden - allerdings nur in diesen (vgl. Buck/Kostin/Morgan 2002: 3; vgl. auch Clark/ Grandy 1984; Zeidner 1986). Eine weibliche Überlegenheit hinsichtlich der Schreibfertigkeitkonstatiert Cole (1997) für die zahlreichen vom amerikanischen Educational Testing System (ETS) durchgeführten Prüfungen. Auch beim Differential Aptitude Test (DAT) und dem Preliminary Scholastic Aptitude Test (PSAT) erzielen Mädchen in mehreren allgemeinsprachlichen Fertigkeitsbereichen bessere Resultate als Jungen, wobei die Unterschiede zwischen den Geschlechtern über die Jahre hinweg immer mehr abnehmen (vgl. Feingold

[1] Bei der Kalibrierung von Testitems wird im Rahmen von so genannten Erprobungsprüfungen der Schwierigkeitsgrad jedes Items ermittelt (vgl. Eckes 2003).

1988). Auffälligerweise gewinnen Mädchen an amerikanischen High Schools überdies häufiger Preise für das Schreiben als Jungen, während Letztere vorwiegend in Bereichen wie Technik, Naturwissenschaften, Mathematik und Sport brillieren (vgl. Dwyer/Johnson 1997). Die Ergebnisse der US-amerikanischen Untersuchungen teilweise bestätigend lässt auch die weltweit mit 15-jährigen Teenagern durchgeführte Studie PISA 2000 auf eine wesentlich ausgeprägtere muttersprachliche Lesekompetenz von Mädchen schließen. Die weibliche Überlegenheit ist dabei besonders bei der Teilkompetenz „Reflektieren und Bewerten" gegeben und in etwas geringerem Umfang bei den übrigen allgemeinsprachlichen Fertigkeiten „textbezogenes Interpretieren" und „Informationen ermitteln". Bemerkenswerterweise liegt die Wahrscheinlichkeit, zu den leistungsschwächsten Schülern zu gehören, für die Jungen in allen untersuchten OECD-Ländern 1,3- bis 3,5-mal höher als für die Mädchen (vgl. OECD 2001).

Zur Erklärung der diagnostizierten Leistungsunterschiede drängen sich in erster Linie anders geartete Interessenlagen auf. So bekunden die in der PISA-Studie untersuchten Mädchen in der Regel ein stärkeres Interesse am Lesen als Jungen und konsumieren anspruchsvollere Literatur (vgl. OECD 2001). Ebenso legen US-amerikanische Schülerinnen ihre Schwerpunkte eher auf die Fächer Kunst, Englisch bzw. auf Fremdsprachen und Geisteswissenschaften, während Jungen tendenziell eher Vorlieben im Bereich der Naturwissenschaften oder der Informatik verfolgen. Auch an den Colleges streben Mädchen eher ein Studium der klassischen oder Fremdsprachen, der Literatur, der Künste oder der Hauswirtschaft an (vgl. College Board 2000).

Neben den Autoren, die dem weiblichen Geschlecht besondere Stärken hinsichtlich ihrer allgemeinsprachlichen Kompetenzen zuschreiben, sind auch solche zu nennen, die auf der Basis statistischer Metaanalysen keine differenziellen Effekte feststellen (vgl. Hyde/ Linn 1988) oder die sogar Jungen diesbezügliche Vorteile einräumen (vgl. Lynn/Mulhern 1991; Lynn/Dai 1993; Born/Lynn 1994).

Bezogen auf Tests zur Messung fremdsprachlicher Kompetenzen liegt lediglich eine Hand voll Studien vor. Bei der von ETS durchgeführten Englisch-Prüfung Test of English as a Foreign Language (TOEFL) werden keine diesbezüglichen Effekte für die Subtests „Listening Comprehension", „Structure and Written Expression" sowie „Vocabulary and Reading Comprehension" berichtet (vgl. Ryan/Bachman 1992; Wainer/Lukhele 1997), ebenso wenig für den Subtest „Reading Comprehension" des First Certificate of English (FCE) (vgl. Ryan/Bachman 1992). Auch eine Untersuchung zum Prüfungsgespräch im Rahmen des International English Language Testing System (IELTS) konnte keine geschlechtsspezifischen Unterschiede nachweisen (vgl. O'Loughlin 2002). Im Fall eines Englisch-Vokabeltests der Finnish Foreign Language Certification Examination hingegen beantworteten zwar Mädchen und Jungen einzelne Testitems unterschiedlich häufig korrekt. Gleichwohl schnitt kein Geschlecht insgesamt besser ab. Das voneinander abweichende fremdsprachliche passive Vokabular resultiert aller Wahrscheinlichkeit nach aus unterschiedlichen Geschlechterrollen (vgl. Takala/Kaftandijeva 2000).

Alles in allem mag die geschilderte Befundlage widersprüchlich erscheinen. Manche der Inkongruenzen relativieren sich jedoch bei einer genaueren Betrachtung der verwendeten Forschungsdesigns. So sind beispielsweise die zitierten Studien, die eine männliche Überlegenheit im Bereich sprachlicher Fertigkeiten behaupten, alle mit einem klinischen Forschungsinstrument durchgeführt worden, nämlich dem Test Wechsler Intelligence Scale for Children - Revised (WISC-R) bzw. dem Test Wechsler Adult Intelligence Scale - Revised (WAIS-R). Fraglich ist, ob hier ein für Geschlechtervergleiche valides Untersuchungsinstrument verwendet wurde. Desgleichen erhebt sich die Frage nach der Verallgemeinerbarkeit der Ergebnisse bei der Studie zum mündlichen Subtest von IELTS, die mit nur 16 Probanden durchgeführt wurde. Und bei denjenigen Studien, die keine Unterschiede zwischen den Geschlechtern feststellen, sollte zumindest folgende Möglichkeit bedacht werden: Eventuell sind diese Sprachtests so beschaffen, dass sie Mädchen bzw. Frauen leicht benachteiligen; diese könnten ihren Leistungsvorsprung dann nicht unter Beweis stellen. Eine solche Interpretation ist insofern plausibel, als sowohl die Studien zum TOEFL als auch zum FCE auf den Testteilen mit Multiple-Choice-Aufgaben beruhen; bei diesem Aufgabenformat schneidet üblicher-

weise das männliche Geschlecht besser ab (vgl. OECD 2001).

## 2 *Forschungsdesign*

Die vorliegende Studie basiert auf den Ergebnissen von 6.552 Teilnehmern des TestDaF. Der TestDaF ist eine Sprachprüfung, mit dem ausländische Studienbewerber hinreichende Sprachkenntnisse für ein Studium an deutschen Hochschulen nachweisen sollen. Dieser Zielsetzung gemäß prüft der TestDaF schwerpunktmäßig den Kompetenzstand in der akademischen Domäne der Sprachverwendung, und zwar getrennt für die vier Fertigkeiten Leseverstehen, Hörverstehen, schriftlicher Ausdruck und mündlicher Ausdruck.

Die Teilnehmer verteilen sich auf sechs unterschiedliche Prüfungsereignisse, T004 (10. 09. 2002), T005 (14.11. 2002), T006 (4. 02. 2003), C001 (5. 04. 2003), T007 (10. 04. 2003) und T008 (26.06.2003). Von den Teilnehmern waren 3.562 weiblich und 2.990 männlich. Mit Ausnahme von C001 (nur in China) wurden alle Prüfungen weltweit durchgeführt.

Um die Teilnehmerleistungen aus den verschiedenen Prüfungsereignissen direkt miteinander vergleichbar zu machen, wurde mithilfe der Software Winsteps (Version 3.37; vgl. Linacre 2002) bei den Subtests Leseverstehen und Hörverstehen für jeden Prüfungskandidaten im Rahmen eines einparametrischen Raschmodells für dichotome Daten der individuelle Personenfähigkeitsparameter 8 berechnet (auf eine Darstellung der komplexen mathematischen Grundlagen wird an dieser Stelle verzichtet; verwiesen sei diesbezüglich auf Bond/Fox 2001 und McNamara 1996). 6 wird in Logits ausgedrückt und stellt im gegebenen Fall ein objektives Maß für die individuelle Lese- bzw. Hörverstehensfähigkeit dar. Analog wurde für den schriftlichen Ausdruck und den mündlichen Ausdruck mithilfe der Software Facets (Version 3.40; vgl. Linacre 1999) eine Multifacetten-Rasch-Analyse durchgeführt (vgl. zu diesem Verfahren Linacre 1989 sowie Linacre/Wright 2002), um für jeden Prüfungsteilnehmer den Personenfähigkeitsparameter 8 zu ermitteln. Im Anschluss daran wurden für jeden der vier Subtests auf der Basis der individuellen Personenfähigkeitsparameter 8 Gruppenmittelwerte für das männliche und für das weibliche Geschlecht berechnet und mithilfe von T-Tests für unabhängige Stichproben auf statistische Signifikanz überprüft.

## 3 *Ergebnisse*

### 3.1 *Leseverstehen*

Im Fall des Subtests Leseverstehen lassen sich nur für T006 ($p < 0,01$), C001 ($p < 0,05$) und T008 ($p < 0,01$) signifikante Mittelwertsunterschiede zwischen den Geschlechtern feststellen. Bei diesen drei Testereignissen erzielen Frauen minimal bessere Ergebnisse als Männer, wobei die Effektstärken mit $e = 0,29$ (T006), $e = 0,24$ (C001) bzw. $e = 0,13$ (T008) als gering anzusehen sind (nach Bortz 1999: 140 liegt mit $e = 0,20$ ein schwacher und mit $e = 0,50$ erst ein mittlerer Effekt vor; die Effektstärke e berechnet sich für unabhängige Stichproben aus der Differenz der beiden Mittelwerte dividiert durch die Standardabweichung der Gesamtstichprobe).

### 3.2 *Hörverstehen*

Anders als beim Subtest Leseverstehen zeichnen sich beim Hörverstehen systematische Effekte ab: Bei allen berücksichtigten Testereignissen erreichen Frauen im Durchschnitt signifikant höhere Logit-Werte als Männer. Tab. 1 veranschaulicht diesen Sachverhalt: Spalte zwei gibt für die untersuchten Prüfungsereignisse die durchschnittlichen Logit-Werte der Männer mit den zugehörigen Standardabweichungen an, Spalte drei die durchschnittlichen Logit-Werte der Frauen mit den zugehörigen Standardabweichungen, Spalte vier die eigentlich aussagekräftigen Mittelwertsdifferenzen zwischen den Geschlechtern (wiederum ausgedrückt in Logits) und Spalte fünf die jeweilige Effektstärke e.

Auch wenn der Durchschnitt der Frauen durchweg besser abschneidet als der der Männer, so bewegen sich die Effektstärken beim Hörverstehen doch lediglich zwischen $e = 0,27$ und $e = 0,40$, also im Bereich schwacher bis mittlerer Effekte.

Angesichts der unterschiedlichen Performanz der Geschlechter erhebt sich die legitime Frage, ob die diagnostizierten Mittelwertsdifferenzen tatsächlich auf Kompetenzunterschiede zwischen den Geschlechtern zurückgehen oder womöglich auf Besonderheiten des Testverfahrens (Problematik von Messartefakten). Um hierauf eine Antwort zu finden, wurde für sämtliche eingesetzten Hörverstehensitems eine so genannte Bias-Analyse durchgeführt. Bei dieser wird die statistische Hypothese geprüft, dass die Wahrscheinlich-

| Testereignis | Mittlerer Personenfähigkeitsparameter 0 männlich (Logits) | Mittlerer Personenfähigkeitsparameter 0 weiblich (Logits) | Mittelwertsunterschied (Logits) | Effektstärke s |
|---|---|---|---|---|
| T004 | 0,39 *(std.* 1,15) | 0,73 *(std.* 1,28) | 0,35 | 0,28 |
| T005 | 0,41 *(std.* 1,28) | 0,78 *(std.* 1,31) | 0,38 | 0,29 |
| T006 | 0,15 *(std.* 1,18) | 0,62 *(std.* 1,12) | 0,47 | 0,40 |
| C001 | 0,52 *(std.* 1,23) | 0,90 *(std.* 1,41) | 0,38 | 0,29 |
| T007 | 0,51 *(std.* 1,25) | 0,85 *(std.* 1,33) | 0,35 | 0,27 |
| T008 | 0,38 *(std.* 1,17) | 0,75 *(std.* 1,17) | 0,37 | 0,31 |

Alle Mittelwertsunterschiede sind auf einem Niveau von $p < 0,05$ signifikant.

*Tab. 1:* Geschlechtsspezifische Unterschiede im Subtest Hörverstehen

keit, ein Testitem korrekt zu beantworten, mit der Zugehörigkeit zu einer bestimmten Gruppe (z. B. Alter, Geschlecht, Nationalität) zusammenhängt (zu diesem Verfahren vgl. Cole/Moss 1993). Als Ergebnis dieser gleichfalls mit Winsteps berechneten Analyse ist festzuhalten, dass zwar bei allen untersuchten Prüfungen die Wahrscheinlichkeit einer korrekten Beantwortung einzelner Testitems für die Angehörigen eines Geschlechts signifikant höher ist als für das andere Geschlecht, dass aber insgesamt kein Geschlecht durch den Test benachteiligt wird. Als Beispiel hierfür mag T005 dienen, wo sechs von 25 Testitems einen Geschlechterbias aufweisen: Bei fünf Items ist die Wahrscheinlichkeit einer korrekten Beantwortung für Männer signifikant höher als für Frauen, bei einem Item für Frauen höher. Gleichwohl lässt die Gesamtwahrscheinlichkeit für den Test als Ganzes keine Benachteiligung eines Geschlechts erkennen. Hierfür spricht auch, dass die Angehörigen des weiblichen Geschlechts ja letzten Endes im Schnitt bessere Ergebnisse als die Männer vorweisen können.

### *3.3 Schriftlicher Ausdruck*

Tab. 2 enthält die Ergebnisse der Frauen- und Männerteilstichprobe für den Subtest schriftlicher Ausdruck.

Ähnlich wie beim Hörverstehen erzielen Frauen im Subtest schriftlicher Ausdruck im Durchschnitt signifikant bessere Ergebnisse als Männer. In einem vergleichbaren Bereich wie beim Hörverstehen bewegen sich auch die Effektstärken, bei denen wiederum ein schwacher bis mittlerer Effekt vorliegt.

Um sicherzustellen, dass die Mittelwertunterschiede nicht auf Korrektoreneffekte zurückzuführen sind, oder - mit anderen Worten - um bei der Bewertung von Prüfungsleistungen einen Zusammenhang zwischen dem Geschlecht der Beurteiler und dem der Testkandidaten ausschließen zu können, wurde für jedes Testereignis eine zweifaktorielle Varianzanalyse (ANOVA) mit den Faktoren „Geschlecht des Testkandidaten" und „Geschlecht des Beurteilers" berechnet. Als Resultat ließ sich für kein Prüfungsereignis ein signifikanter Interaktionseffekt, also eine systematische Benachteiligung von Testkandidaten des einen oder des anderen Geschlechts durch die Beurteiler, feststellen.

### *3.4 Mündlicher Ausdruck*

Tab. 3 veranschaulicht die Ergebnisse für den Subtest mündlicher Ausdruck.

Wie Tab. 3 nahe legt, sind die mittleren Leistungsunterschiede zwischen Männern und Frauen im Subtest mündlicher Ausdruck am ausgeprägtesten, wobei besonders bei T006 und T007 Frauen bessere Testresultate erbringen. Allerdings bewegen sich die Effektgrößen e auch hier lediglich im Bereich schwacher bis mittlerer Effekte.

Die gleichfalls für den Subtest mündlicher Ausdruck durchgeführte zweifaktorielle Varianzanalyse mit den Faktoren „Geschlecht des

| Testereignis | Mittlerer Personen-fähigkeitsparameter 0 männlich (Logjts) | Mittlerer Personen-fähigkeitsparameter 0 weiblich (Logits) | Mittelwerts-unterschied (Logits) | Effektstärke s |
|---|---|---|---|---|
| T004 | -0,53 *(std.* 2,86) | 0,39 *(std.* 2,83) | 0,92 | 0,32 |
| T005 | -0,15 *(std.* 3,05) | 0,96 *(std.* 2,95) | 1,12 | 0,37 |
| T006 | -0,15 *(std.* 4,78) | 1,40 *(std.* 4,22) | 1,55 | 0,34 |
| C001 | -0,48 *(std.* 5,09) | 0,64 *(std.* 4,37) | 1,12 | 0,24 |
| T007 | 0,40 *(std.* 4,43) | 1,70 *(std.* 4,23) | 1,30 | 0,32 |
| T008 | -0,29 *(std.* 4,62) | 0,99 *(std.* 4,62) | 1,28 | 0,27 |

Alle Mittelwertsunterschiede sind mit Ausnahme von C001 auf einem Niveau von $p < 0,05$ signifikant (bei C001 $p = 0,05$).

*Tab. 2:* Geschlechtsspezifische Unterschiede im Subtest schriftlicher Ausdruck

| Testereignis | Mittlerer Personen-fähigkeitsparameter 0 männlich (Logits) | Mittlerer Personen-fähigkeitsparameter 0 weiblich (Logits) | Mittelwerts-unterschied (Logits) | Effektstärke **g** |
|---|---|---|---|---|
| T004 | 2,32 *(std.* 2,98) | 2,99 *(std.* 2,64) | 0,67 | 0,24 |
| T005 | 2,43 *(std.* 3,13) | 3,45 *(std.* 2,92) | 1,02 | 0,33 |
| T006 | 3,08 *(std.* 3,61) | 4,76 *(std.* 3,28) | 1,68 | 0,48 |
| C001 | 0,35 *(std.* 3,97) | 1,90 *(std.* 4,14) | 1,55 | 0,38 |
| T007 | 2,80 *(std.* 3,83) | 4,59 *(std.* 3,57) | 1,80 | 0,47 |
| T008 | 3,13 *(std.* 3,43) | 4,27 *(std.* 3,08) | 1,14 | 0,35 |

Alle Mittelwertsunterschiede sind auf einem Niveau von $p < 0,05$ signifikant.

Tab. 3: Geschlechtsspezifische Unterschiede im Subtest mündlicher Ausdruck

Testkandidaten" und „Geschlecht des Beurteilers" konnte auch hier keine Abhängigkeit zwischen dem Geschlecht des Beurteilers und dem der Testkandidaten ermitteln.

## *4 Diskussion und Ausblick*

Der Darstellung der Ergebnisse folgend sollen diese abschließend im Hinblick auf die beiden eingangs aufgeworfenen Problemkreise diskutiert werden.

Ob sich nun im Bereich Deutsch als Fremdsprache Belege für eine größere Sprachbegabung von Mädchen bzw. Frauen finden lassen, diese Frage lässt sich auf dem Hintergrund der vorliegenden Studie klar bejahen: Weibliche Sprachbegabung erscheint keineswegs als Mythos, denn im Mittel zeigen Frauen tatsächlich bei allen Prüfungsereignissen und fast allen Subtests einen Leistungsvorsprung gegenüber den M ännern. Allerdings relativiert sich diese Erkenntnis insofern, als die beobachteten Unterschiede zwischen den Geschlechtern nur vergleichsweise geringe Ausmaße annehmen.

Im Hinblick auf die vier sprachlichen Fertigkeiten ergeben sich erkennbare geschlechtsspezifische Unterschiede. Am geringsten ausgeprägt ist die weibliche Überlegenheit im Subtest Leseverstehen und zudem nur in drei Fällen signifikant. Vor dem Hintergrund der PISA-Studie (vgl. OECD 2001) wäre hier vielleicht ein deutlicherer

Vorsprung der Frauen zu erwarten gewesen. Möglicherweise wäre der Leistungsunterschied deutlicher ausgefallen, wenn beim Subtest Leseverstehen nicht bei zwei der drei Aufgabenblöcke trichotome Aufgaben eingesetzt worden wären, bei denen das weibliche Geschlecht meist eine etwas schlechtere Performanz erbringt als das männliche. Demgegenüber nimmt bei den Subtests Hörverstehen und schriftlicher Ausdruck die Größe des Effekts ähnliche Ausmaße an. Die besten Ergebnisse scheinen Mädchen bzw. Frauen im Subtest mündlicher Ausdruck erzielen zu können. Offenbar sind Frauen also vor allem bei den produktiven Fertigkeiten im Vorteil. Die analysierten Daten stellen somit weitere empirische Evidenz für die weit verbreitete Behauptung dar, Frauen seien kommunikativer als Männer.

Den Ergebnissen der vorgelegten Studie kommt nicht allein wegen der Größe der Stichprobe ein gewisses Gewicht zu, insofern auch mittels statistischer Verfahren die Hypothese wenigstens teilweise entkräftet werden konnte, die Leistungsunterschiede gingen auf die Benachteiligung eines Geschlechts durch Besonderheiten der Aufgabenstellung oder auf Korrektoreneffekte zurück (ein Umstand, der zugleich für die Qualität des TestDaF als Sprachprüfung spricht).

Gleichwohl stehen die gemachten Beobachtungen unter einem gewissen Vorbehalt, denn aufgrund der Domänenspezifik des TestDaF - die Konzentration auf Sprache in Hochschulkontexten - konnte nur ein vergleichsweise kleiner Ausschnitt der Sprachverwendung erfasst werden. Denkbar wäre es, dass Männer in anderen Zusammenhängen der Sprachverwendung bessere Resultate erzielen. Zugleich erlaubt das gewählte Forschungsdesign für die aufgeworfenen Forschungsfragen Schlussfolgerungen nur für die Gruppe der ausländischen Studienbewerber, also lediglich eine (wenn auch beachtliche) Subgruppe der Sprachlerner.

Wünschenswert wäre daher besonders die Durchführung von Geschlechtervergleichen mit anders stratifizierten Untersuchungsstichproben und im Hinblick auf andere Bereiche der Sprachverwendung. Von großem Interesse sind weiterhin Untersuchungen, ob sich die beobachteten Leistungsunterschiede an Charakteristika der Testverfahren festmachen lassen, also etwa an besonderen Itemtypen, die von Mädchen bzw. Frauen tendenziell besser bearbeitet werden als von Jungen bzw. Männern, oder an spezifischer Lexik.

Eine intensivere Erforschung dieses Felds erscheint erforderlich, um aussagekräftige Untersuchungsinstrumente für Geschlechtervergleiche zu schaffen. Deren Durchführung ist auch nach Jahrzehnten von „gender studies" noch immer geboten, denn eine Frage ist nach wie vor ungelöst: das Problem der Gleichbehandlung der Geschlechter in den Bildungssystemen.

## Literatur


Bond, Trevor G. / Fox, Christine M. (2001): Applying the Rasch model. Fundamental measurement in the human sciences. Mahwah, N.J.

Born, M. / Lynn, Richard (1994): Sex differences on the Dutch WISC-R. A comparison with the USA and Scotland. Educational Psychology 2, 249-255.

Bortz, Jürgen (1999): Statistik für Sozialwissenschaftler. 5., vollst. Überarb. Aufl. Heidelberg/ New York.

Bück, Gary/Kostin, Irene / Morgan, Rick (2002): Examining the relationship of content to gender-based Performance differences in advanced placement exams. New York.

Clark, Mary J. / Grandy, Jerilee (1984): Sex differences in the academic Performance of scholastic aptitude tests takers. Princeton.

Cole, Nancy S. (1997): The ETS gender study: how females and males perform in educational settings. Princeton.

Cole, Nancy S. / Moss, Pamela (1993): Bias in test use. In: R. L. Linn (Hg.), Educational measurement. O. O., 201-219.

College Board (2000): College Bound Seniors 2000. New York.

Dwyer, Carol A. / Johnson, Linda M. (1997): Grades, accomplishments, and correlates. In: W. Willingham/N. Cole (Hg.), Gender and fair assessment. Hülsdale, N.J., 127-156.

Eckes, Thomas (2003): Qualitätssicherung beim TestDaF. Konzepte, Methoden, Ergebnisse. In: Fremdsprachen und Hochschule 69, 43-68.

Feingold, Alan (1988): Cognitive gender differences are disappearing. In: American Psychologist 2, 95-103.

Hyde, Janet S. / Linn, Marcia C. (1988): Gender dif-

ferences in verbal ability. A metaanalysis. In: Psychological Bulletin 104, 53-69.
Linacre, J. Michael (1989): Many-facet Rasch measurement. Chicago.
Linacre, J. Michael (1999): A user's guide to Facets. Rasch measurement Computer program. Chicago.
Linacre, J. Michael (2002): A user's guide to WINSTEPS-MINISTEP. Rasch-model Computer programs. Chicago.
Linacre, J. Michael / Wright, Benjamin D. (2002): Construction of measures from many-faceted data. In: Journal of Applied Measurement 3, 484-509.
Lynn, Richard / Dai, Xiao Y. (1993): Sex differences on the Chinese standardization sample of the WAIS-R. In: Journal of Genetic Psychology 4, 459-464.
Lynn, Richard/ Mulhern, Gerry (1991): A comparison of sex differences on the Scottish and American standardisation samples of the WISC-R. In: Personality and Individual Differences 12, 1179-1182.
Maccoby, Eleanor E. / Jacklin, Carol N. (1974): The psychology of sex differences. Stanford, Calif.
McNamara, Tim (1996): Measuring second language Performance. London.
OECD (2001): Lernen für das Leben. Erste Ergebnisse der Internationalen Schulleistungsstudie PISA 2000. Verfügbar im Internet unter www.pisa. oecd. org.
O'Loughlin, Kieran (2002): The impact of gender in oral proficiency testing. In: Language Testing 19, 169-192.
Ryan, Katherine E. / Bachman, Lyle F. (1992): Differential item functioning on two tests of EFL proficiency. In: Language Testing 9, 12-29.
Takala, Sauli / Kaftandijeva, Felianka (2000): Test fairness. A DIF analysis of an L2 vocabulary test. In: Language Testing 17, 323-340.
Wainer, Howard / Lukhele, Robert (1997): How reliable are TOEFL scores? In: Educational and Psychological Measurement 5, 741-759.
Zeidner, Moshe (1986): Sex differences in scholastic ability in Jewish and Arab College students in Israel. In: Journal of Social Psychology 126, 801-803