Bergh, H.H. van den & Thije, J.D. ten (2005). Beurteilung sprachlicher Kompetenz: Die Entwicklung von Beurteilungsverfahren für das Schulsystem und für individülle Schülerleistungen in den Niederlanden (Übersetzt von Guido Schnieders und Winfried Thielmann). In K. Ehlich (Ed.), *Anforderungen an Verfahren der regelmässigen Sprachstandsfeststellung als Grundlage für die frühe und individülle Sprachförderung von Kindern mit und ohne Migrantionshintergrund* (pp. 217-241). Bonn: Bundesministerium für Bildung und Forschung.

Projekt
"Anforderungen an Verfahren der regelmäßigen Sprachstandsfeststellung als Grundlage für die frühe und individuelle Sprachförderung von Kindern mit und ohne Migrationshintergrund"

Leitung: Prof. Dr. Dr. h.c. Konrad Ehlich
Institut für Deutsch als Fremdsprache / Transnationale Germanistik
LMU München

Huub van den Bergh & Jan D. ten Thije

Beurteilung sprachlicher Kompetenz: Die Entwicklung von Beurteilungsverfahren für das Schulsystem und für individuelle Schülerleistungen in den Niederlanden[1]

Übersetzt von Guido Schnieders und Winfried Thielmann

März 2004

[1]. Wir bedanken K. de Glopper (Universität Groningen), H. Kuhlemeier (Cito), und Dr. M. Zwarts (Nationale Schulinspektion), and H. Breevelt für ihre Kommentare.

Utrecht: Universität Utrecht

1. Einleitung

In jüngerer Zeit sind in Europa mehrere vergleichende Untersuchungen zur Sprachkompetenz von Kindern und Jugendlichen durchgeführt worden (wie z.B. die PISA-Studie). Diese Untersuchungen haben in etlichen europäischen Ländern eine leidenschaftliche gesellschaftliche und politische Diskussion über die Möglichkeit und Notwendigkeit von nationalen Standards zur Sprachstandsmessung ausgelöst. Dies war besonders in Deutschland der Fall, wo verschiedene Bundesländer in der PISA-Studie enttäuschend abgeschnitten hatten. Um in die gegenwärtige Debatte auch wissenschaftlich basierte Argumente einzubringen, hat Prof. Dr. Konrad Ehlich, der Leiter des Instituts für Deutsch als Fremdsprache/Transnationale Germanistik der Ludwig-Maximilians-Universität München, eine Pilotstudie zur Sprachstandsmessung initiiert. In diesem Projekt geht es um eine Evaluierung derzeit angewendeter Instrumente zur Sprachstandsmessung sowie um die Bedingungen, unter denen eine regelmäßige bundesweite Überprüfung der Sprachkompetenz von Kindern und Jugendlichen durchgeführt werden könnte. Dabei steht vor allem die Frage im Mittelpunkt, wie solche bundesweiten Tests der individuellen Sprachentwicklung von Muttersprachlern und Nichtmuttersprachlern (z.B. Kindern von Migranten) Rechnung tragen können. Deshalb wurden aus einigen Ländern, die auf dem Gebiet der Sprachstandsmessung schon eine längere Tradition aufweisen können (z.B. Schweden, Australien und die Niederlande), Expertisen angefordert.
Der vorliegende Bericht befasst sich mit den niederländischen Entwicklungen zur Sprachstandsmessung.
In den Niederlanden wurde vor etwa zwanzig Jahren eine Studie zur Praktikabilität nationaler Sprachstandsmessung durchgeführt. Die Studie brachte ans Licht, dass sieben Prozent der Schüler am Ende der Grundschulzeit ‚funktionale Analphabeten' waren (Wesdorp/Van den Bergh/Bos/Hoeksma/Oostdam/Scheerens/Triesscheijn 1986). Daraufhin brach eine leidenschaftliche Diskussion über den erforschten Leistungsstand der Schüler los, ohne dass dabei di Zwecke, den Aufgabentypen oder den angewendeten Normen der Studien in der Öffentlichkeit diskutiert wurden. Von da an wurden alle fünf Jahre Sprachstandsmessungen auf nationaler Ebene durchgeführt, also 1989, 1994 und 1999.
Um die Entwicklung und Praxis dieser nationalen Sprachstandsmessungen genauer charakterisieren zu können, setzen wir sie zunächst zu anderen Verfahren der Leistungsmessung in Beziehung, die in den Niederlanden eine längere Tradition haben. Während ihrer Pflichtschulzeit, also vom vierten bis zum sechzehnten (bzw. achtzehnten) Lebensjahr, werden niederländische Schüler zu verschiedenen Stadien ihrer Schullaufbahn mit vier verschiedenen Test- bzw. Examenstypen konfrontiert, die je spezifischen Zwecken genügen:

- Während der Grundschulzeit machen alle Schüler standardisierte Tests im Rahmen eines *Systems zur allgemeinen Leistungskontrolle* (Pupil Monitoring System).
- In der sechsten und in der achten Klasse der Grundschulzeit werden nach dem Zufallsprinzip ausgewählte Schüler alle fünf Jahre im Rahmen der *Nationalen Überprüfung der Fortschritte im Bildungswesen* (National Assessment) getestet.
- Am Ende der Grundschulzeit (8. Klasse) machen alle Schüler (dann ca. elfjährig) einen standardisierten Abschlusstest (oft den sog. Cito Grundschulabschlusstest), von dem weitgehend ihre Wahlmöglichkeiten für die weitere Schullaufbahn abhängen.
- Schließlich machen alle Schüler am Ende ihrer Schullaufbahn an einer weiterführenden Schule in allen Fächern Abschlussexamen, die aus zwei Teilen bestehen: zunächst ein Examen, das von der jeweiligen Schule gestellt wird, und dann ein landesweites, also zentral erstelltes, mit standardisierten Tests.

Um den deutschen Leser mit den niederländischen Bedingungen der Durchführung dieser verschiedenen Prüfungsarten vertraut zu machen, geben wir zunächst einige Hintergrundinformationen über das niederländische Schulsystem sowie die dortige Bildungspolitik, wobei wir auch auf die Situation von Migrantenkindern eingehen. Abschließend diskutieren wir kritisch die linguistischen, pädagogischen und bildungspolitischen Aspekte des niederländischen Prüfungssystems.

2. Grundstruktur des niederländischen Bildungswesens

2.1 Das niederländische Schulsystem

Als ein Hauptunterschied zum deutschen System fällt sofort ins Auge, dass niederländische Kinder im Alter von vier bis fünf Jahren eingeschult werden und ihre Schulpflicht mit dem vollendeten achtzehnten Lebensjahr endet. Während der letzten zwei Jahre ist zumindest teilzeitiger Schulbesuch Pflicht. Abb. 1 gibt eine Übersicht über das niederländische Schulsystem, wobei die Größe eines jeden Blocks den Schülerzahlen entspricht. Im folgenden versuchen wir, die niederländischen Schultypen auf das deutsche System zu beziehen, wobei hier nicht von exakten Entsprechungen auszugehen ist:

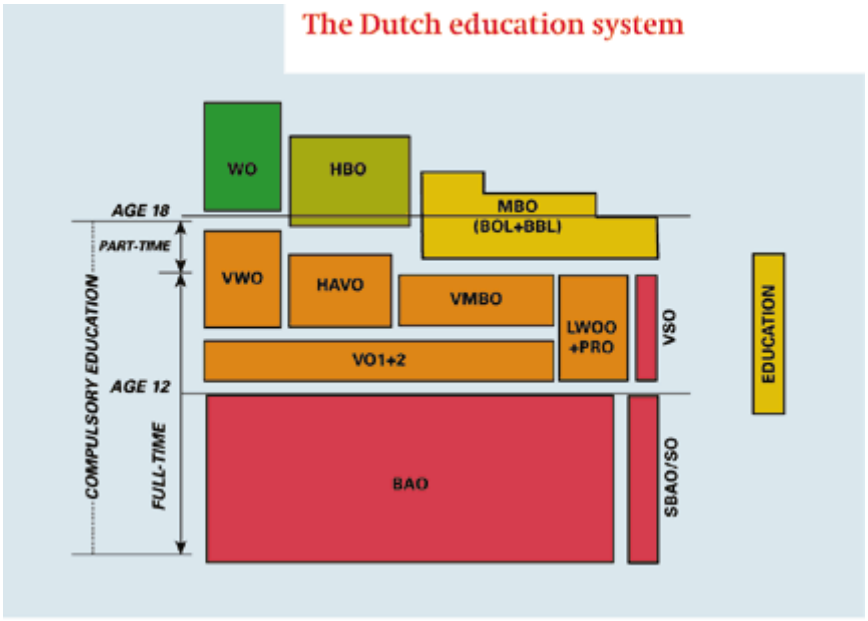

Abb. 1: Das niederländische Schulsystem **(Quelle: Minister of Education 2004)**

Nach der Grundschule (BAO) besteht grundsätzlich die Wahl zwischen verschiedenen weiterführenden Schultypen, also dem Gymnasium (VWO), der Fachoberschule (HAVO) und der berufsvorbereitenden Realschule (VMBO). Danach besteht die Wahl zwischen der Berufsfachschule (MBO) (**mit einer Teilzeit (BBL) und Vollzeitvariant (BOL)) oder dem Studium (HO) an einer Universität (WO) oder einer Fachhochschule (HBO). Für Schüler mit Lernschwächen gibt es zum einen den besonderen Zweig der (LWOO), ferner besondere Grundschule (SBAO und SO) , weiterführende Schulen SVO), sowie berufspraktische Programme (PRO).
Nach einem vom Niederländischen Institut für Soziale Planung durchgeführten Vergleich europäischer Schulsysteme (SCP 2000) müssen sich sowohl im deutschen als auch im niederländischen Schulsystem die Schüler relativ früh zwischen verschiedenen weiterführenden Schulen entscheiden. In den meisten anderen Ländern sind die

weiterführenden Schulen dagegen integriert und die Entscheidung steht erst im 15. oder 16. Lebensjahr an. Wenn auch die Möglichkeit ein- oder zweijähriger Orientierungsstufen (Abb. 1, VO1+2) es in den Niederlanden gestattet, die Entscheidung etwas hinauszuschieben, so ist doch zu bemerken, dass die Kluft zwischen Gymnasium (VWO) bzw. Fachoberschule (HAVO) auf der einen Seite und berufsbezogener Ausbildung auf der anderen (VMBO) immer tiefer und unüberbrückbarer wird – eine Situation, die weitgehend der deutschen in vielen Bundesländern entspricht. Damit haben die Niederlande eines der selektivsten Schulsysteme Europas.
In der Untersuchung SCP (2000) wird auch berichtet, dass das niederländische Bildungsniveau im europäischen Vergleich lange Zeit hoch war, aber in den letzten Jahren auf den europäischen Durchschnitt gefallen ist. Niederländische Kinder haben immer noch etwa 10% mehr Unterrichtsstunden als ihre europäischen Nachbarn, während die niederländischen Bildungsausgaben relativ niedrig sind und in den letzten Jahren noch weiterfallen. Gegenwärtig gibt es daher eine öffentliche Diskussion darüber, ob die Niederlande dabei sind, sich aus dem internationalen Wettbewerb zu verabschieden, besonders im Hinblick auf die Anforderungen, die an die Reproduktion einer modernen Wissensgesellschaft gestellt werden.

2.2 Bildungsfreiheit und Wandel in der Bildungspolitik

Die Entwicklungen innerhalb des niederländischen Bildungswesens können nur vor dem Hintergrund der verfassungsrechtlich garantierten Bildungsfreiheit verstanden werden, auf der das System seit 1917 beruht. Aufgrund der Bildungsfreiheit haben niederländische Schulen im europäischen Vergleich eine höhere Autonomie. Die Niederländische Verfassung gestattet es privaten Körperschaften nicht nur, eigene Schulen einzurichten, sondern gewährt ihnen auch eine weitgehend - aber nicht unbeschränkte - Freiheit bei der Bestimmung der Lehrinhalte und der Form ihrer Vermittlung. Allerdings hat der Staat mit der Zeit Instrumente der indirekten Einflussnahme entwickelt.
In diesem Zusammenhang ist die von Koole/ten Thije (1994) beschriebene „konstruktive Bildungspolitik“ der sechziger Jahre zu sehen, innerhalb derer zum ersten Mal der Staat als Initiator und Stimulator bildungspolitischer Innovation anerkannt wurde. „*Gesellschaftliche Notwendigkeiten wie rationalisierungsbedingte Änderung der Nachfrage auf dem Arbeitsmarkt und damit steigende Bedarf an ungelernten ausländischen Arbeitskräften waren die treibenden Kräfte dieses Wechsels in der Bildungspolitik.“* Aufgrund dieser neuen Bildungspolitik wurden viele eigenständige lokale, regionale und nationale Beratungsstellen für Schulen in einer großen Struktur zusammengefasst, die es der Regierung ermöglichte, auf dem so vereinheitlichten Beratungswege koordinierte Anreize zur systematischen Erneuerung von Strukturen und Lehrinhalten zu schaffen (Doornbos, 1986, 267). Dementsprechend wurden dann Beratungszentren auf nationaler Ebene gebildet, die mit spezifischen Aufgaben bezüglich der Lehrplanentwicklung (SLO), Forschung, Schulberatung sowie Test- und Examensentwicklung (Cito) betraut wurden. Die Entwicklung von Instrumenten der Leistungskontrolle und von Beurteilingsverfahren für das Schulsystem auf nationaler Ebene in den achtziger Jahren ist als direktes Resultat dieser bildungspolitischen Anreize zu sehen.
Im Lauf der letzten zehn Jahre hat sich der Staat jedoch sukzessive aus seiner innovativen Rolle zurückgezogen. Er tritt nicht mehr so oft als Initiator innovativer Projekte auf, sondern stattet Schulen eher direkt mit den Mitteln aus, die es ihnen ermöglichen, die Unterstützung lokaler oder staatlicher Beratungsstellen zu bezahlen – womit natürlich nun ein Wandel zum Effizienzdenken hin verbunden ist. Dieser bildungspolitische Wandel kann als weiteres Stadium der gesellschaftlichen Transformation des verfassungsrechtlichen Instituts der Bildungsfreiheit gesehen werden. Die nationale Bildungspolitik beschränkt sich nun auf ihre ‚Hauptaufgaben’, und der Bildungsminister konzentriert sich auf ‚input/output’-Vergleiche.

Im Rahmen dieses effizienzzentrierten Denkens kommt nun der nationalen Leistungskontrolle von Schulen eine neue Wichtigkeit zu.

2.3 Mehrsprachigkeit und Migrantenkinder

Als die ersten ausländischen Arbeitskräfte in den sechziger Jahren nach Nordeuropa gebracht wurden, wurde mit migranten Schülern nach der Maxime der ‚Integration unter Beibehaltung der kulturellen Identität' verfahren. Das Ziel dieser zweigleisigen Politik war es, einerseits die Kinder der Einwanderer in die Schulen und in die niederländische Gesellschaft zu integrieren, sie aber gleichzeitig auch für die Rückkehr in ihre Ursprungsländer vorzubereiten. Daher bekamen diese Kinder zusätzlichen Sprachunterricht – sowohl im Niederländischen als auch in ihrer jeweiligen Muttersprache.
In den achtziger Jahren stellte es sich dann heraus, dass es sich bei der Einwanderung um ein Dauerphänomen handelte und dass die Niederlande sich inzwischen zu einer multikulturellen Gesellschaft entwickelt hatten. Damit änderte sich auch die Bildungspolitik. Die Vermittlung von Niederländisch als Zweitsprache bekam einen höheren Stellenwert, und Kinder von Einwanderern erhielten dieselbe zusätzliche Förderung wie die Kinder niederländischer Arbeiterfamilien. Seit dieser Zeit erfolgt die staatliche Mittelzuweisung an Grundschulen nach einer dem Ausländeranteil proportionalen Multiplikatorskala von 1.0 über 1,25 bis 1.9, wobei bei der Ermittlung des Multiplikators der Geburtsort von ein oder beide Eltern, der Bildungsstand und und der Beruf beider Elternteile eine Rolle spielen. Die Skala entscheidet über zusätzliche Mittelzuweisungen für zusätzliche Lehrkräfte oder Unterrichtsmaterialien.
Heutzutage ist das Vermitteln von Sprachen ausländischer Minderheiten während der Unterrichtszeit verboten, und Unterricht im Niederländischen wird in allen Schulzweigen forciert. In einigen weiterführenden Schulen können Spanisch, Türkisch, Marokkanisch oder andere Minderheitensprachen als dritte Fremdsprache gewählt werden, aber diese Option wird gerade in den berufsvorbereitenden Schulzweigen, wo sich besonders viele Migrantenkinder befinden, selten angeboten. Im Rahmen der europäischen Vereinigung gibt es Tendenzen zum zweisprachigen Unterricht, der aber hauptsächlich auf Niederländisch-Englisch abgestellt ist. Neulich wird Deutsch und Fransosisch auch in der Sprachpolitik für die Grundschule beachtet.
Die gegenwärtige Schulsituation von Migrantenkindern ist in dem OESO-Bericht ‚Education at a glance' ('Bildung auf einen Blick') (Minister of Education 2003) zusammengefasst. Danach betrug in den Jahren 2001/2 der Anteil von Migrantenkindern in den Grundschulen insgesamt 15,3%, während er in den Sonderschulen bzw. -zweigen auf dieser Stufe 18,9% betrug. Innerhalb der letzten fünf Jahre ist der Anteil von Migrantenkindern in den weiterführenden Schulen auf 10% gestiegen, wobei auch hier der Anteil in den Sonderschulen (LWOO) mit über 33% am größten ist. Nach wie vor sind Migrantenkinder in den höheren Schulen (VWO und HAVO) mit nur 3,5% unterrepräsentiert.
Nach diesem kurzen Überblick über das niederländische Bildungssystem, die niederländische Bildungspolitik sowie die Situation von Migrantenkindern im niederländischen Bildungswesen befassen wir uns nun mit den verschiedenen Instrumenten der Leistungskontrolle bezüglich der Sprachkompetenz.

3. Beurteiloingsverfahren für das Schulsystem und für individuelle Schülerleistungen

Wir gehen nun ausführlicher auf die schon eingangs besprochenen Beurteilungsverfahren für das Schulsystem und für individuelle Schülerleistungen ein. Neben traditionellen, von Lehrern entworfenen Tests werden auch mehr standardisierte Tests durchgeführt, die nicht nur individuelle Lernfortschritte messen, sondern auch den Vergleich von Schulen untereinander gestatten und somit auch allgemein Fortschritte der gesamten Schülerpopulation über längere Zeitabschnitte erfassen können. In vielen Grundschulen

werden (standardisierte) Tests im Rahmen des *Systems zur allgemeinen Leistungskontrolle* (Pupil Monitoring System) durchgeführt. Der Hauptzweck dieses Systems ist die frühzeitige Identifizierung und sukzessive Kontrolle von leistungsschwachen Schülern.Wie bereits dargetan, müssen sich Schüler im Alter von etwa elf Jahren, also am Ende der Grundschulzeit (8. Klasse), für eine weiterführende Schule entscheiden. Diese Wahl basiert sowohl auf den Empfehlungen des Grundschulrektors als auch auf den Ergebnissen des standardisierten Tests. Obwohl im Prinzip für solche Zwecke jeder Schuleignungstest verwendet werden kann, solange er von einer unabhängigen Instanz durchgeführt wird, verwenden über sechzig Prozent aller Grundschulen den sog. Cito Grundschulabschlusstest. Ferner ist zu beobachten, dass trotz der ursprünglich gleichen Gewichtung von Rektorempfehlung und Testergebnis dem Testergebnis inzwischen die größere Rolle bei der Entscheidungsfindung zukommt. In einigen Städten nehmen manche weiterführende Schulen nur noch Schüler ab einem bestimmten Cito-Ergebnis auf, wodurch dieser Test immer wichtiger wird.
Während ihrer Grundschulzeit können Schüler an Tests im Rahmen der *Nationalen Überprüfung der Fortschritte im Bildungswesen* (National Assessment) teilnehmen, die in der sechsten und in der achten Klasse durchgeführt werden. Im Gegensatz zu den beiden bereits besprochenen Systemen der Qualitätskontrolle sind diese Tests jedoch stichprobenbasiert. Es nimmt also nur eine Auswahl von Schulen daran teil und innerhalb dieser Schulen wiederum nur eine Auswahl von Schülern. Auf die Vor- und Nachteile dieses Verfahrens gehen wir unten noch ausführlicher ein.
Der letzte zu besprechende Test- bzw. Examenstyp hat seinen traditionellen Ort am Ende der Schulzeit in einer weiterführenden Schule. Die Schüler legen in jedem Fach zwei Prüfungen ab – eine von der Schule gestellte und eine staatliche. Die schulischen Abschlussprüfungen und ihre Bewertungsmaßstäbe werden von Lehrern erstellt und finden in Abständen während des gesamten letzten Schuljahres statt. Die staatlichen Abschlussprüfungen bestehen aus einer Prüfung pro Fach am Ende des letzten Schuljahres. Diese Prüfungensaufgabe und ihre Bewertungsmaßstäbe wird zentral enwickelt. Die Testvoraussetzungen in den Schulen sind mehr oder weniger standardisiert. Die Endnote pro Fach errechnet sich aus dem arithmetischen Mittel der schulischen und staatlichen Abschlussprüfung. Auf der Basis von allen Endnoten wird die Entscheidung über das Bestehen der Abschlussprüfung getroffen.

3.1 Staatliche Prüfungen

Wie bereits ausgeführt, setzten sich die Abschlussexamen der weiterführenden Schulen aus einer von der Schule gestellten sowie einer staatlichen Komponente zusammen. Diese beiden Teile differieren hinsichtlich ihrer der in ihnen abgeprüften Lernziele. Lernziele, deren Erreichung komplexe Verfahren der Beurteilung erfordern, wie z.B. Ausdrucksfähigkeiten in einer Fremdsprache oder Literaturverständnis, werden in den schulischen Examen abgeprüft. Traditionelle Lernziele des Fremdsprachenunterrichts (Lese- und Schreibfähigkeiten) werden in den staatlichen Prüfungen getestet. Für die höheren Schulen (bzw. Gymnasium) besteht die Aufgabenstellung zumeist in einer Zusammenfassung eines fremd- oder muttersprachligen Textes; in den anderen Schulzweigen kommen hier Multiple-Choice-Tests zur Anwendung.
Die von den Schulen gestellten Abschlussprüfungen sind bezüglich ihre Inhälte und das Moment der Durchführung recht uneinheitlich. Normalerweise finden während des letzten Schuljahres in jedem Fach drei bis vier Prüfungen statt, deren Bedingungen von Schule zu Schule stark variieren können. Einige Schulen führen die Prüfungen unter sehr strikten Bedingungen durch, die denen der staatlichen Prüfungen ähneln. Andere sehen die Sache nicht so streng und lassen die Schüler die Prüfungen im eigenen Klassenzimmer unter Aufsicht ihres eigenen Lehrers schreiben. Auch die Bewertungsmaßstäbe können sich unterscheiden. Manchmal werden sie vorab festgelegt, manchmal erst nach der Prüfung, mitunter sogar gar nicht (Van den Bergh/Rohde/Zwarts 2003). Daher ist es insgesamt nicht

sinnvoll, die Examensergebnisse verschiedener Schulen untereinander zu vergleichen. Dies gilt auch für die Ergebnisse aus verschiedenen Jahren, da sich sowohl die Prüfungsinhalte als auch die Bewertungsmaßstäbe von Jahr zu Jahr ändern.
Die staatlichen Abschlussexamen sind demgegenüber sowohl hinsichtlich der Prüfungsbedingungen als auch der Inhalte stabiler. Sie differieren aber selbstverständlich hinsichtlich der spezifischen weiterführenden Schulzweige. Es ist nicht durchfürbar eine zentrale Prüfung für Schüler in verschiedenen Unterrichtstypen zu entwickeln, weil die Aufgabestellungen zu unterschiedlich sind. Eine standarisierte Prüfung für alle Schüler würde zu leicht für Gymnasiumschüler und viel zu kompliziert für Schüler in der berufsvorbereitenden Schuletypen sein. . Daher können nur Schulen ein und desselben Typs hinsichtlich ihrer Ergebnisse verglichen werden
Damit stellt sich die Frage, wie die Vergleichbarkeit schultypspezifischer Ergebnisse aus mehreren Jahren gewährleistet ist. Hier ist zunächst darauf hinzuweisen, dass Cito aufgrund der strengen Geheimhaltungsvorschriften keine Teile der staatlichen Prüfungen im Voraus testen darf. Dies hat zur Folge, dass die psychometrischen Eigenschaften der Abschlussexamen nur im Rückblick zu ermitteln sind. Trotz der Bemühungen von Cito um Konsistenz kann mithin die Schwierigkeit der Prüfungen über die Jahre hinweg schwanken. Deswegen kann die *Prüfungskommission für die Abschlussexamen weiterführender Schulen* (CEVO) die Bewertungsmaßstäbe und damit die Zuordnung von Testresultaten zu Abschlussnoten im Nachhinein ändern. Obwohl also Cito versucht, durch IRT-Modellierung[2] einen Schwierigkeitsstandard zu halten, ist weder garantiert, dass Prüfungen aus verschiedenen Jahren dieselben Fähigkeiten testen, noch, dass Schüler mit denselben Fähigkeiten zu verschiedenen Zeitpunkten dieselben Resultate erreichen.
Cito behauptet weiterhin, dass die Resultate staatlicher Examen aus verschiedenen Jahren vergleichbar sind, was aber in neueren Untersuchungen bezweifelt wird (z.B. Van den Bergh et al. 2003). In einer Analyse der Ergebnisse der schulischen und der staatlichen Abschlussprüfungen der letzten fünf Jahre erwies sich, dass die Durchschnittsergebnisse der schulischen Prüfungen weitaus weniger variierten als die der staatlichen. In Abb. 2 wird dieser Sachverhalt durch die nahezu horizontalen Linien für die einzelnen Schulfächer illustriert, was darauf schließen lässt, dass Lehrer ihre Bewertungsmaßstäbe von Jahr zu Jahr so anpassen, dass etwa dieselbe Durchschnittsnote erreicht wird.
Demgegenüber variieren die Ergebnisse der staatlichen Examen weitaus stärker. Verfolgt man die Linie für das Schulfach Französisch (FR), so waren die Abschlussergebnisse 1995 und 2000 sehr gut, in den dazwischenliegenden Jahren jedoch nur mittelmäßig. Es ist äußerst unwahrscheinlich, dass diese unterschiedlichen Resultate unterschiedlichen Fähigkeiten der Schülerpopulation korrespondieren. Außerdem tritt die Variation der Ergebnisse staatlicher Examen nicht nur in der graphischen Repräsentation der Durchschnittsergebnisse auf, sondern auch bei einer statistischen Analyse, in der die totale Abweichung der Ergebnisse nach Schülern, Schulen, Jahrgängen und Examen aufgeschlüsselt wurde.

[2] IRT-Modellierung (**I**tem-**R**esponse-**T**heory) ist eine statistische Methode, die eine populationsunabhängige Abschätzung der Schwierigkeit der Teilaufgaben eines Tests erlaubt. Hat man die Schwierigkeit der Aufgaben abgeschätzt, kann man die Fähigkeit der Aufgabenlöser abschätzen [bzw. das Abschneiden einzelner Individuen bei verschiedenen Aufgaben]. Die Fähigkeit der Schüler kann mithin als die Wahrscheinlichkeit einer korrekten Aufgabenlösung ausgedrückt werden. Sehr fähige Schüler bearbeiten die Aufgabe mit großer Wahrscheinlichkeit korrekt, weniger fähige mit weniger Wahrscheinlichkeit. Dies bedeutet natürlich nicht, dass alle guten Schüler die Aufgabe korrekt lösen. IRT-Modelle beruhen auf strikten Annahmen. Eine Grundannahme ist, dass es nur einen Grund für differente Testergebnisse geben kann, d.h. alle Aufgaben im psychologischen Sinne eindimensional sind. Nähme man an, dass sowohl der Schüler als auch die Schule an Unterschiede in den Testergebnissen Anteil haben, würde dies die IRT-Annahmen verletzen. Dies ist einer der Gründe, warum die Anzahl der Schüler pro Schule niedrig gehalten wird.

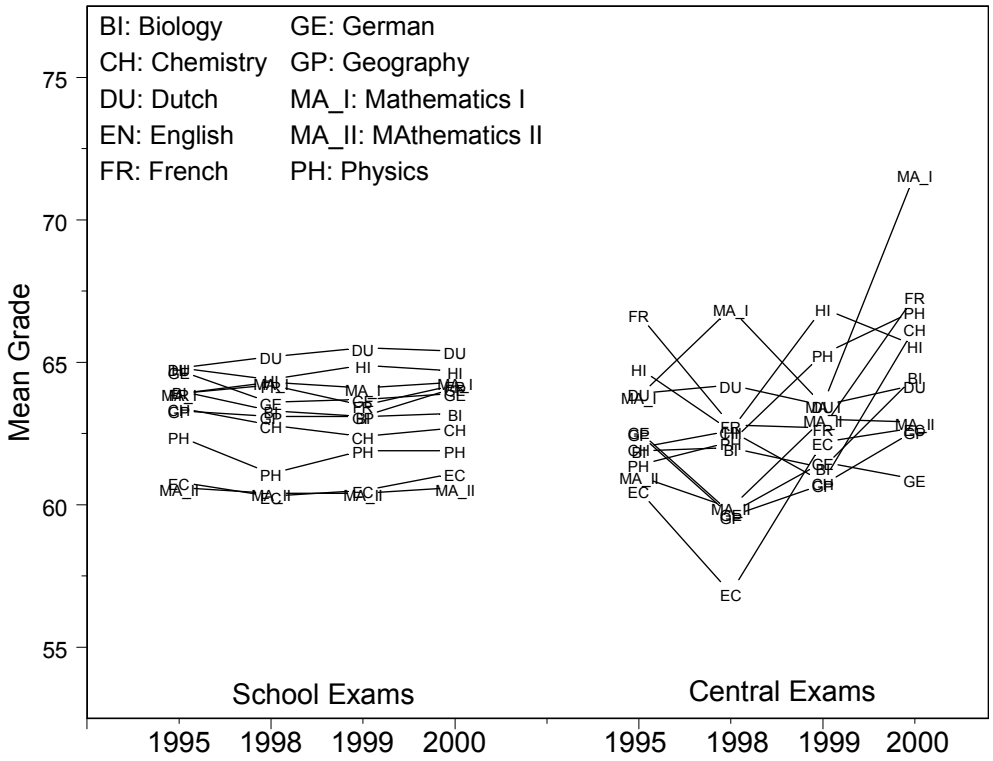

Abb. 2: Durchschnittsnoten (aus maximal 100 Punkten (10 bedeutet absolut unzureichend und 100 ausgezeichnet)) schulischer und staatlicher Abschlussprüfungen aus vier verschiedenen Jahrgängen

Für die meisten Schulfächer ist die Varianz zwischen den beiden Prüfungstypen erheblich – woraus sich schließen lässt, dass die bei der Erstellung der staatlichen Prüfungen angewendeten Prinzipien nicht hinreichen, Vergleichbarkeit der Ergebnisse über mehrere Jahre hinweg sicherzustellen (Van den Bergh et al, 2003). Es ist auch durchaus möglich, dass die Unterschiede zwischen den Examen nicht im Schwierigkeitsgrad bestehen, sondern darin, dass unter Umständen verschiedene kognitive Fähigkeiten getestet werden (De Glopper/Van Schooten, 2002). Da z.B. für jedes Sprachexamen andere Texte verwendet werden, werden natürlich auch je andere Fragen gestellt – so z.B. bei dem einen Text Fragen zu Beziehungen zwischen Absätzen, bei einem anderen hingegen Fragen, deren Beantwortung Schlussfolgerungen erfordert. Mithin können bei der Aufgabenstellung bereits kleine Unterschiede zwischen Texten in der Messung unterschiedlicher kognitiver Fähigkeiten resultieren.
Daher kann man mit ziemlicher Sicherheit annehmen, dass staatliche Abschlussprüfungen nur einen Vergleich zwischen Schulen desselben Typs – und auch dies nur für einen Jahrgang – zulassen. Was die Unterschiede der Ergebnisse aus verschiedenen Jahren betrifft, lässt sich – ohne sehr viel strengere Anforderungen hinsichtlich der getesteten Fähigkeiten sowie des Schwierigkeitsgrades – hingegen Vergleichbarkeit nicht herstellen.

3.2. Der Cito-Grundschulabschlusstest

Dieser am Ende der Grundschulzeit durchgeführte Test hat den in den sechziger Jahren entwickelten sogenannten *Amsterdamer Schultest* abgelöst, dessen Zweck in der

Erleichterung der Wahl zwischen den verschiedenen weiterführenden Schultypen bestand, für die damals im wesentlichen das Urteil des Rektors ausschlaggebend war.
Am Cito-Grundschulabschlusstest, einem Multiple-Choice-Test, nehmen jedes Jahr im Februar etwa 80% der achten Jahrgangsstufe teil. Getestet werden sprachliche und mathematische Fähigkeiten sowie die Fähigkeit des Umgangs mit verschiedenen Informationsmedien. Der Bereich „Weltwissen“ ist hingegen fakultativ. Es wird immer im Voraus angekündigt, wieviele Aufgaben pro Fach der Test enthält. Im Schnitt etwa 100 im Bereich Sprache (Rechtschreibung 20; Schreiben 30; Lesen 30; Wortschatz 20), 60 im Bereich Mathematik (Zahlen und Rechnen 25; Prozent- und Bruchrechnen 25; numerischer Umgang mit Zeit und Geld 15), 40 im Bereich des Umgangs mit Informationsmedien (Texte: 10; Branchenbücher und Nachschlagewerke: 10; Übersichten, Tabellen und Grafiken: 10; Landkarten: 10) und 60 aus dem Bereich des „Weltwissens“ (Geographie, Geschichte und Biologie je 20). Bei dem Test werden standardisierte Antwortblätter eingesetzt, die eine zentrale Korrektur der Arbeiten und eine Berichterstattung an den Schulen von rund 170 000 Schülern innerhalb von einigen Wochen gestatten.
Wie Tabelle 1 zeigt, werden die Resultate für jeden Schüler individuell nach der Anzahl der in jedem Fach korrekt beantworteten Aufgaben aufgeschlüsselt. Hieraus wird eine standardisierte Gesamtbewertung errechnet, die auf einer Skala von 501 bis 550 liegt. Außerdem wird für jede Schule auf der gesamten Schülerergebnisse eine Durchschnittsnote errechnet.

Ergebnisse	Sprache	Mathe-matik	Umgang mit Informations-medien	Welt-wissen	Ge-samt	**Standard-Punktzahl**
Anzahl der Aufgaben	100	60	40	60	260	
Anzahl der korrekt bearbeiteten Aufgaben	73	48	23	52	195	**536**
prozentualer Anteil gleicher Ergebnisse	56	52	42	57	53	

Tabelle 1 Ein fiktives Cito-Abschlusstest-Zeugnis

Der fiktive Schüler in Tabelle 1 hat 73 von 100 Sprachaufgaben korrekt beantwortet, und 56% der Schüler haben dasselbe oder ein niedrigeres Ergebnis; die standardisierte Gesamtbewertung für diesen Schüler beträgt 536.
Für die Schulwahl dieses Schülers werden nun folgende Überlegungen durchgeführt: Nur zwei Prozent derjenigen Schüler, die sich für einen elementaren berufsvorbereitenden Schultyp entscheiden, haben ein höheres Ergebnis, wohingegen 54% derjenigen Schüler, die für eine weiterführenden Schultypen (Gymanium oder Fachoberschule) optieren, einen niedrigeres oder dasselbe Ergebnis haben und 46% haben eine höheres Ergebnis. Nur ein Prozent der Schüler, die sich für das Gymnasium entscheiden, haben dasselbe oder ein niedrigeres Ergebnis. Mithin wird dem Schüler empfohlen, sich für eine **Fachoberschule zu entscheiden.
Aufgrund der Bildungsfreiheit liegt aber die Entscheidung letztlich bei den Eltern. In einigen Städten haben allerdings die örtlichen Schulbehörden begonnen, die Zulassung zu bestimmten weiterführenden Schultypen von der Gesamtbewertung des Cito-Grundschulabschlusstests abhängig zu machen. So verlangt z.B. die Stadt Amsterdam für die Zulassung zum **Gymnasium (mit Latein und Griechisch) eine Gesamtergebnis von mindestens 545. Damit

ist der Grundschulabschlusstest faktisch nicht freiwillig, und in Amsterdam ist nicht nur Teilnahme Pflicht, sondern die Resultate werden auch veröffentlicht. Das Zurückhalten der Ergebnisse schlechterer Schüler zur Schönung der Schul-Abschlussstatistiken wird hierdurch unterbunden.
Der Cito-Grundschulabschlusstest ist jetzt eine Zulassungsprüfung – wobei er faktisch einer Eignungsprüfung gleichkommt, da er natürlich auch Rückschlüsse auf Intelligenz, Konzentrationsfähigkeit, Auffassungsgabe sowie auf das Durchhaltevermögen zulässt. In der Tat ist seine Vorhersagekraft so groß, dass noch bis zur elften Jahrgangsstufe Korrelationen zwischen Schulleistungen und Testergebnissen bestehen. Allerdings treffen diese Vorhersagen auf Migrantenkinder nur bedingt zu (Uiterwijk 1994).
Die Ergebnisse des Abschlusstests sind für die Beurteilung der Entwicklung der Qualität des Bildungswesens insgesamt relevant. Der Grund dafür ist, dass die Ergebnisse aus verschiedenen Jahren mit einem Gewichtungsverfahren jetzt miteinander verglichen werden können. Im Gegensatz zu den Aufgaben der in weiterführenden Schulen durchgeführten Abschlussprüfungen, die den bereits erwähnten Geheimhaltungsvorschriften unterliegen, können die psychometrischen Eigenschaften der Aufgaben des Cito-Tests in Vorversuchen ermittelt werden, wodurch die Vergleichbarkeit der Ergebnisse verschiedener Jahrgänge sowie verschiedener Schulen erleichtert wird.
Dennoch hat der Cito-Test einige gravierende Nachteile: Nicht alle Schüler machen ihn, da Schulen zwar zur Durchführung eines Eignungstests verpflichtet sind, aber unter verschiedenen Testtypen wählen dürfen. Daher kann nicht ausgeschlossen werden, dass die Entscheidung einer Schule für oder gegen diesen Test auf bestimmten pädagogischen Praktiken und Wertvorstellungen basiert. So behaupten einige Schulen, dass die Fähigkeiten ihrer Schüler nicht durch einen Multiple-Choice-Test zu beurteilen sind. Auch prüft dieser Test nur eine geringe Bandbreite von Fähigkeiten ab: Mit 100 Multiple-Choice-Aufgaben im Bereich Sprache lässt sich der hochkomplexe Gesamtbereich sprachlicher Fähigkeiten nicht ausmessen, da reales Schreibvermögen sowie wirkliche Sprech- und Hörfähigkeiten nicht geprüft werden können. Auch die Leseaufgaben sind so speziell, dass sie kaum Rückschlüsse auf die generelle Lesefähigkeit, also z.B. den Umgang mit verschiedenen Gattungen, erlauben.
In neueren Publikationen war davon die Rede, dass der Durchschnitt des Cito-Tests über mehrere Jahre hinweg gesunken ist. Genauere statistische Analyse hat allerdings ergeben, dass die Unterschiede zwischen den Testergebnissen aus mehreren Jahren nur sehr klein (Roeleveld 2002) sowie in den meisten Fällen nicht signifikant sind (Webbink 2002). Dasselbe Resultat ergibt sich auch, wenn man die wirtschaftliche Situation der Eltern oder den kulturellen Hintergrund der Schüler miteinbezieht. Wie bereits oben gesagt, prüft der Test nur Teilfähigkeiten ab, so dass sich die Ergebnisse auch fachspezifisch kaum generalisieren lassen. Mithin ist es keineswegs überraschend, dass andere Untersuchungen, wie z.B. die PRIMA-Studien, die die Leistungen kleinerer Schülergruppen über längere Zeiträume verfolgen, eine Verbesserung der Leistungen im Bereich Mathematik feststellen. Dies wird auch durch Abb. 3 illustriert (nach Roeleveld 2002), die die Entwicklung der Effektgrößen sowohl für den Cito-Test als auch für drei Indikatoren der PRIMA-Studien zeigt. Man sieht deutlich, dass sich die sprachlichen Fähigkeiten (P-LG) kaum verändern, während sich die mathematischen (P-AR) erheblich verbessern. Die PRIMA-Studien kommen also zu anderen Resultaten als der Cito-Test, wobei aber keineswegs klar ist, welche der beiden Messmethoden die exaktere ist. Hinzu kommt, dass in keinem der beiden Testverfahren Stichproben von Aufgaben/Testtypen durchgeführt wurden. Man kann also nur hoffen, dass die verwendeten Stichproben von Aufgaben für die jeweiligen Gesamtmengen der zum Einsatz kommenden Aufgaben repräsentativ sind. Solche Repräsentativität ist nur im Rahmen eines *National Assessment* (Nationale Überprüfung der Fortschritte im Bildungswesen) zu

erreichen, so dass z.B. Textauswahl und Fragen Rückschlüsse auf die tatsächliche Lesefähigkeit erlauben.

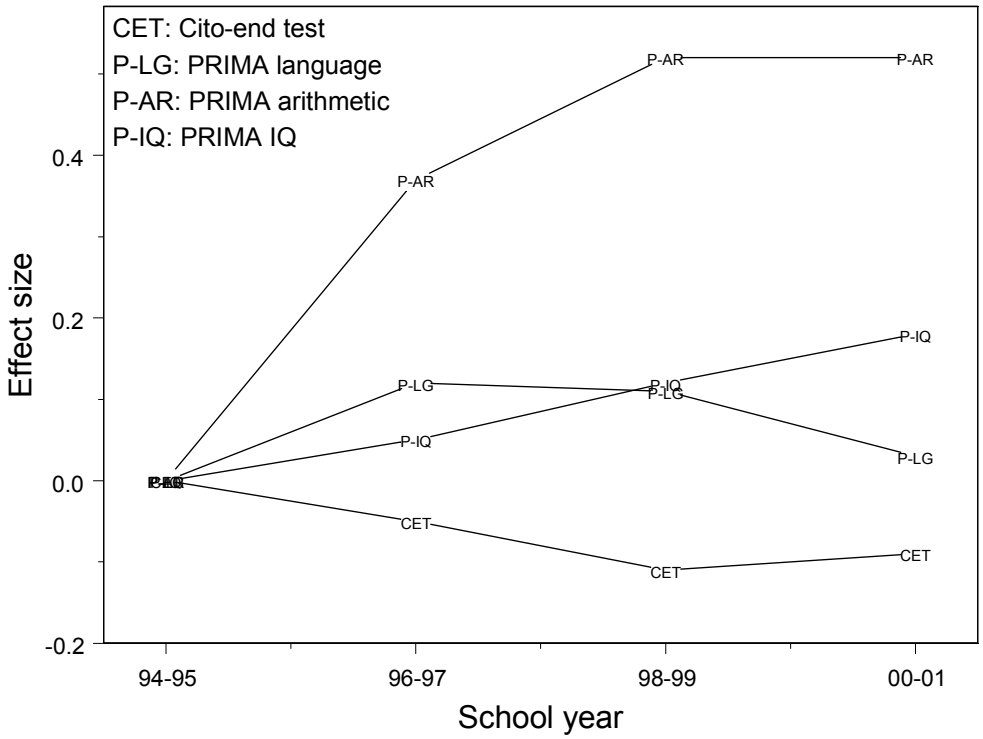

Abb. 3: Entwicklung der Effektgrößen sowohl für den Cito-Test (CET) als auch für drei Indikatoren der PRIMA-Studien

Ein weiterer Punkt, der zu berücksichtigen ist, sind die sog. *Rückkopplungseffekte* (washback effects), die dadurch entstehen, dass Schulen nur das vermitteln, was in einem Test abgefragt wird, oder ihre Schüler gleich mit früheren Testaufgaben ‚drillen'. Da der Cito-Test z.B. eigentliche Schreibfähigkeiten nicht abprüft, könnten solche Unterrichtsverfahren sich äußerst ungünstig auswirken. Zum Glück ist dies bisher nicht beobachtet worden, da sich Lehrer hinsichtlich ihrer Unterrichtsverfahren kaum von dem Cito-Test haben beeinflussen lassen (Wesdorp 1979). Auch eine neuere Untersuchung, die sich damit befasst, wieviel Unterrichtszeit Lehrer nicht Cito-relevanten Gegenständen widmen, hat diesbezüglich keine Änderungen feststellen können (Cito 2002). Mithin ist es nicht wahrscheinlich, dass der Verzicht auf die Messung eigentlicher Schülerfähigkeiten sich auf die Unterrichtspraxis ausgewirkt hat. Obwohl das Cito es entmutigt, ist das ‚Drillen' der Schüler mit früheren Tests allerdings gang und gäbe geworden – sowohl in der Schule als auch daheim. Dies macht die Schüler mit dem Format vertraut und kann die Resultate verbessern.

Der Cito-Grundschulabschlusstest ist also der bei weitem einflussreichste Test in der Primarstufe. Dies gilt sowohl für die individuellen Schüler, denen je nach Ergebnis eine entsprechende weiterführende Schule empfohlen wird, als auch für die Schulen selbst, die ihre erzielten Durchschnittsergebnisse in Jahresberichten veröffentlichen müssen. Da diese Durchschnittsergebnisse auch Rückschlüsse auf die Schulqualität zulassen, wird im Moment sogar diskutiert, ob man nicht wenigstens einen Teil der zugewiesenen Mittel von den

erzielten Durchschnittsergebnissen abhängig machen soll. Da natürlich die von einer Schule erzielten Durchschnittsergebnisse auch von den Grundfähigkeiten der Schüler abhängen, wird im Moment versucht, einen diesbezüglichen Test in Kindergärten einzuführen, so dass die Mittelzuweisung von den Unterschieden in den Ergebnissen beider Tests abhängig gemacht werden kann. Zum Glück ist all dies noch in der Planungsphase, wenn auch die Politiker diese Maßnahmen zu begrüßen scheinen.

3.3 Pupil-Monitoring-Systeme – Systeme zur allgemeinen Leistungskontrolle

Es gibt eine ganze Reihe von Pupil Monitoring Systeme, die die Gewinnung verlässlicher Informationen über die Fortschritte einzelner Schüler gestatten. So werden – mehr oder weniger regelmäßig – standardisierte (und manchmal genormte) Tests von Grundfähigkeiten durchgeführt. Die Ergebnisse werden mithilfe spezieller Computerprograme verarbeitet, die für jeden Schüler und jede Schulklasse einen Bericht über Lernfortschritte erstellen. Wenn also ein Schüler z.B. sechs Tests gemacht hat, lässt sich zum einen sagen, ob er sein Niveau gehalten hat, zum anderen lässt sich seine Position innerhalb der Klasse sowie – bei normierten Tests – innerhalb der Gesamtpopulation ermitteln. Dieses Verfahren ist in Abbildung 4 dargestellt, die für drei fiktive Schüler die Ergebnisse von sechs Lesetests aus der dritten, vierten und fünften Jahrgangsstufe im Vergleich zu einer Vergleichsgruppe zeigt.

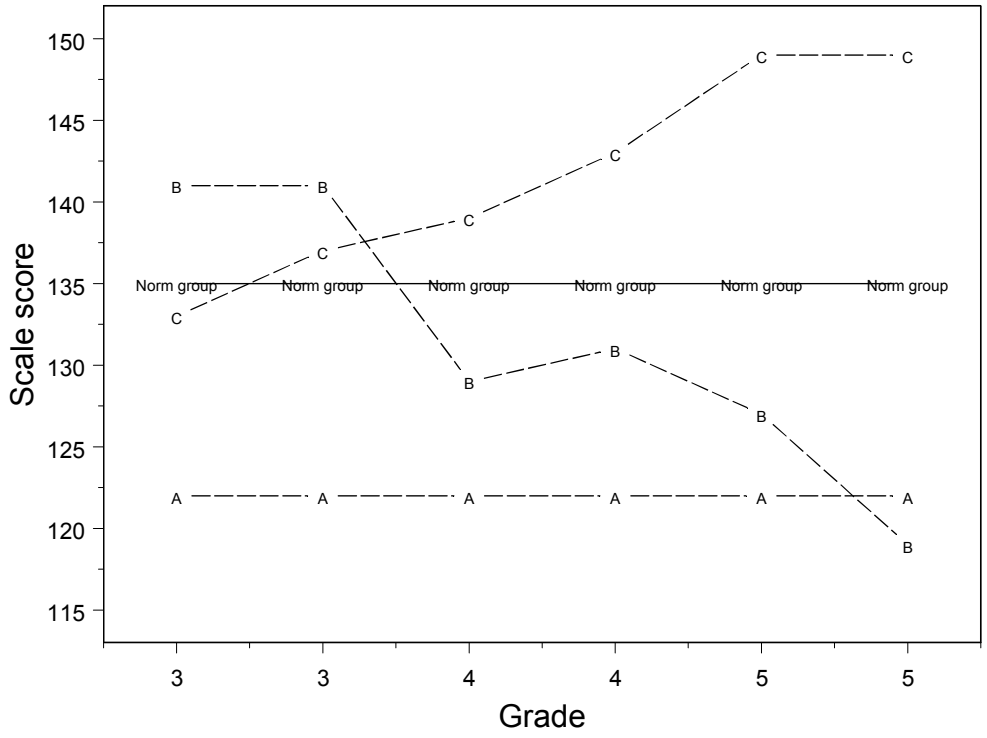

Abb. 4: Ergebnisse eines Tests innerhalb des Pupil Monitoring Systems (Grade = Jahrgangsstufe; Scale Score = Leistungsmaßstab; Norm Group = Vergleichsgruppe)

Schüler A ist ein schlechter Leser, der über drei Jahre hinweg konstante Fortschritte macht, die 13 Punkte unter dem Durchschnitt liegen. Die Leistungen von Schüler B fallen hingegen kontinuierlich ab, seine Fortschritte werden immer geringer und er bedarf besonderer Zuwendung durch den Lehrer, was auch daran deutlich wird, dass der Schüler im Vergleich

zu der Vergleichsgruppe zunächst überdurchschnittliche Leistungen erbrachte. Der dritte Schüler hat zwar mit unterdurchschnittlichen Leistungen begonnen, die Entwicklung seiner Lesefähigkeiten liegt aber dann weit über dem Durchschnitt – um ihn muss man sich keine Sorgen machen.
Solche Beobachtungsverfahren sind für Lehrer sehr nützlich, da sie zum einen Aufschluss darüber geben, welche Schüler sich immer schwerer tun (oder gefährdet sind), und zum anderen es dem Lehrer ermöglichen, seinen eigenen Unterricht zu evaluieren. Wenn zu viele Schüler bei bestimmten Themenbereichen scheitern, wäre eine Wiederholung dieser Thematik wünschenswert. Aber die Daten könnten natürlich auch zur Evaluierung des Bildungssystems selbst herangezogen werden, wenn sie systematischer erhoben würden und Vergleichbarkeit der verschiedenen Verfahren sichergestellt wäre. Im Moment erscheint es jedoch als unmöglich, aus diesen Daten ein zuverlässiges Gesamtbild des landesweiten Bildungsniveaus erhalten zu können. Einerseits werden zu viele unterschiedliche Verfahren neben ein ander verwendet, anderzeits ist die Art und Weise, wie die Daten gespreichert und präsentiert werden, nicht standardisiert.

3.4 National Assessment – nationale Überprüfung der Fortschritte im Bildungswesen

Der Zweck des *National Assessment* ist ein zwiefacher: Zum einen gestattet es eine Momentaufnahme des Leistungsniveaus zu einem bestimmten Zeitpunkt. Zum anderen können die Ergebnisse mit denen aus früheren Jahren verglichen werden, so dass festgestellt werden kann, ob Schüler generell in einem bestimmten Fach besser oder schlechter geworden sind. Damit sind National Assessments sozusagen ein Thermometer, das den Gesundheitszustand des Bildungswesens anzeigt – und die Bildungsdiskussion auf solide Grundlagen stellt. Wie allgemein bekannt, weiß das Gemecker immer schon, dass früher auch die Bildung besser war – weswegen es bildungspolitische Diskussionen nicht beeinflussen sollte. Schon 1892 klagte ein Professor, der mit der Bewertung von Abituraufgaben befasst war, dass diese ‚unter alle Kritik' seien.
Das National Assessement ist ergebnisorientiert, obwohl das grundsätzlich nicht die einzige Möglichkeit seiner Ausrichtung sein müsste. Bildungsqualität könnte z.B. auch nach den faktischen Lernprozessen beurteilt werden: Werden die Schüler genügend gefordert? Sind die Aufgabenstellungen angemessen und machbar? Bekommen die Schüler vernünftiges Feedback? Fühlen sie sich im Klassenzimmer sicher und gut aufgehoben? – Diese Aspekte werden zwar in den National Assessment inventarisiert, sind jedoch nicht dessen Hauptgegenstand. Eine dritte Definition von Bildungsqualität könnte im Sinne der angebotene Lern- und Entwicklungsmöglichkeiten formuliert werden. Damit würde die Möglichkeit einer Beurteilung des Bildungswesens nach der Qualität der angebotenen Qualifikationsstrukturen fokusiert. Das National Assessment enthält jedoch keine Aspekte, die mit solcher Definition verbunden sind. Wenn man die drei mögliche Definition betrachtet, fokussiert das National Assessement die Schulleistungen und einige Aspekte des realisierten Lehrangebots.
Das erste National Assessment in den Niederlanden wurde 1984 durchgeführt. Bei dieser Erhebung ging es um die Frage der Aussagekraft und Durchführbarkeit von Sprachtest-Verfahren (Wesdorp et al. 1986), d.h. um die Frage, ob es machbar und sinnvoll ist, eine relativ große Zahl von Schülern Sprachtests machen zu lassen und die Ergebnisse im Nachhinein zu bewerten.
Ein traditioneller Lesetest besteht ja aus einem Text mit Fragen, so dass zwei Lesetests – aufgrund der Verschiedenheit von Texten und Fragen – nie exakt dieselben Fähigkeiten messen können. Es ist auch bekannt, dass jeder Text einen Einfluss darauf hat, welche Einzelfähigkeiten getestet werden (z.B. Van den Bergh 1990). Selbst das Thema kann entscheidend sein: Kennen die Leser das Thema schon, werden durch den Text andere Fähigkeiten abgeprüft als wenn das Thema den Lesern unbekannt ist. Mithin weichen die

Ergebnisse verschiedener Lesetests immer voneinander ab – was übrigens auch für Tests von Schreib-, Hör-, und Sprechfähigkeiten gilt (Kuhlemeier/Van den Bergh 1998). Dies ist der Grund, warum scheinbar unwichtige Einzelentscheidungen im Rahmen der Entwicklung von Messinstrumenten erhebliche Auswirkungen haben können. Um eine präzise und zuverlässige Einschätzung der Schreibfähigkeit eines Schülers zu gewinnen, bedarf es bis zu zwanzig verschiedener Schreibtests (s. Van den Bergh/de Glopper/Schoonen 1987). Die Zufallsabweichung verschiedener Sprachaufgaben ist erfahrungsgemäß sehr groß – dieser Sachverhalt ist vor allem für Schreibtests sehr gut dokumentiert. Daher ist es – um Aufschluss über die tatsächlichen Fähigkeiten von Schülern zu einem bestimmten Zeitpunkt ihrer Schullaufbahn zu gewinnen – nicht ausreichend, sie je einem einzigen, Lese-, Hör- oder Sprechtest zu unterziehen, da ein Einzeltest keine Rückschlüsse auf die tatsächlichen Fähigkeiten erlaubt. Daher müssen sozusagen aus dem Universum möglicher Tests repräsentative Stichproben entnommen und den Schülern vorgelegt werden, was also bedeutet, dass ein und dieselbe Fähigkeit durch etliche verschiedene Tests abzuprüfen ist. Begrifflich gesprochen: Die Varianz der Tests muss als Zufallsfaktor in das Assessment-Design integriert werden (s. Clark 1973). Dies war der Grund warum man für das erste National Assessment 17 verschiedene Lesetests entwickelte, von denen jeder zwischen sechs und sechzehn Aufgaben hatte.
Im Gegensatz zu schulischen Examen oder Qualifikationstests geht es beim National Assessment nicht um die Leistungen individueller Schüler, sondern um Stichproben: Stichproben aus der Schul- sowie der Schülerpopulation. Da es auch nicht um die genaue Feststellung z.B. der Lesefähigkeit eines einzelnen Schülers geht, ist es auch nicht erforderlich, dass alle Schüler einer Stichprobe alle Lesetests bearbeitet haben. Tests können den Schülern über eine Matrix zugewiesen werden, solange sichergestellt ist, dass die Ergebnisse zu spezifischen Aufgaben über ein statistisches Modell miteinander verbunden werden können. Dann können nämlich die Ergebnisse zu allen Aufgaben auf ein- und derselben Skala repräsentiert werden, was Rückschlüsse über die individuellen Testresultate hinaus gestattet. Deswegen ist das National Assessment der Niederlande in hohem Maße der Item-Response-Theory verpflichtet, die es gestattet, die Fähigkeit der Schüler nach dem Schwierigkeitsgrad der einzelnen Aufgaben zu beurteilen. Hierbei ist zu berücksichtigen, dass Aufgaben, die verschiedene Fähigkeiten messen, nie in derselben Dimension dargestellt werden können, weswegen unter den verschiedenen Aspekten einer Fähigkeit zu unterscheiden ist. So müssen bei einem Lesetest am Ende der Primarstufe Unterschiede gemacht werden zwischen der Lesefähigkeit bezüglich der verschiedenen Gattungen – Fachtext, Bericht, reflektierender Text, Anweisung, argumentierender Text, fiktionaler Text, Nachschlagewerk sowie Tabelle, Grafik und Landkarte. Für Sprechen und Schreiben ist nach sprachlichen Handlungen zu unterscheiden, was in sieben funktionalen Texttypen wie Beschreibung, Frage oder Überredung resultiert. Für jeden funktionalen Texttyp werden drei verschiedene Aufgaben entwickelt. In Tabelle 2 finden sich einige Beispiele für Aufgaben aus dem National Assessment.

Tabelle 2: Zwei Beispiele für Aufgabentypen aus dem National Assessment für den Bereich Sprache (Cito 2002)

Lesen von Graphiken und Tabellen

Entfernungen in Kilometern				
	Pelo	Puki	Suka	Tremp
Alomi	30	100	50	90
Septono	70	20	315	10
Manuk	54	210	38	20

Kura	165	85	340	310
Leksa	40	90	115	50

Was ist die Entfernung zwischen Kura und Suka?
_____Kilometer

Informationen einholen (Schreiben)

Lies das zuerst!

Stell dir vor...
Du isst regelmäßig SMUCO Kartoffelchips. Auf der Packung steht Folgendes:

AKTION

Unglaublich, aber wahr: mit den neuen SMUCO Kartoffelchips könnt ihr Kopfhörer bekommen, die ihr in euren Walkman oder in eure Stereoanlage einstecken könnt.

Dazu müsst ihr Folgendes tun:
Auf jeder SMUCO-Tüte ist ein SMUCO-Sammelpunkt. Sammelt drei Punkte und schickt sie in einem frankierten Umschlag an SMUCO. Dann kriegt ihr einen Gutschein, für den ihr bei Dixy-Radio einen brandneuen Kopfhörer UMSONST bekommt! UMSONST!

Dixy-Radio-Läden gibt es überall in den Niederlanden. Ihr könnt euch den Kopfhörer auch schicken lassen, aber das kostet EUR 3,-, die ihr mit Scheck bezahlen könnt.

Den Umschlag mit den Sammelpunkten schickt ihr an

SMUCO Sammelaktion
Postfach 3333
1200 AD Hilversum

Es ist wichtig, dass wir euren Namen und eure genaue Adresse bekommen – Postleitzahl nicht vergessen! Und sagt uns in dem Brief, ob ihr einen Gutschein wollt, oder ob wir euch den Kopfhörer direkt schicken sollen.

Aufgabe
Du hast drei Sammelpunkte und möchtest einen Gutschein für einen Kopfhörer. Du möchtest auch wissen, wo der nächste Dixy-Radioladen ist.
Schreib einen Brief an SMUCO.
Schreib dann, was du auf den Umschlag schreiben würdest.

Sprechen

Der Schüler wird gebeten, auf die Anzeige zu reagieren und eine Geschichte für die Radiosendung einzuschicken

Stell dir vor... In der Zeitung siehst du folgende Anzeige:

RADIO VERY YOUNG
Das jüngste Radio in den Niederlandsen

Ein neuer Sommer – ein neues Programm! Jeden Abend zwischen 7 und 8

ALLES IST WIRKLICH PASSIERT, ALLES IST WAHR.

Schick uns eine Audio-Cassette mit deiner Geschichte über

BÖSE LEUTE

Wer weiß, vielleicht hörst du sie schon bald im Radio! Schick deine Cassette an: RADIO VERY YOUNG Postfach 12345 1014 RP Jonge Tonge
Du findest die Anzeige ziemlich gut. Du kannst sicher auch eine Geschichte über böse Leute erzählen, denn du hast es sicher schon einmal mit einem richtig bösen Menschen zu tun gehabt, mit einem, der dich unfair behandelt hat dich im Stich gelassen hat dich geärgert hat über dich geklatscht hat dich lächerlich gemacht hat über dich gelacht hat sein Versprechen nicht gehalten hat dich ausgestoßen hat sich vorgedrängelt hat dich .....

Der Zweck des National Assessment ist es, Aufschlüsse über den Leistungsstand von Schülern einer bestimmten Altersgruppe in verschiedenen Schulfächern zu geben. Wenn dieser Zweck ernst genommen wird, ergeben sich weitreichende Konsequenzen für die Art der eingesetzten Tests, die Testsituationen und die Bewertung.

Natürlich kann man die ausgewählten Schüler nicht alle 21 Schreib- und Leseaufgaben bearbeiten lassen, da hierfür zuviel Unterrichtszeit verbraucht würde und Schulen mithin weniger bereit wären, am National Assessment teilzunehmen. Außerdem ist es keineswegs nötig, dass alle Schüler einer Schule alle Aufgaben bearbeiten, da es ja nicht um die Beurteilung von Schulen, sondern von Schülern geht. Es reicht also aus, wenn eine Stichprobe von Schülern eine Stichprobe von Aufgaben bearbeitet – in der Regel reichen hierfür schon drei Schüler pro Schule, die je dieselben Aufgaben bearbeiten.

Hierbei stellt sich natürlich folgendes Problem: Eine Stichprobe von Schülern ist gleichbedeutend mit einer Stichprobe von Schulen, der eine Stichprobe von Schülern entnommen wird. Die Entnahme der Stichproben hat also zweischrittig zu erfolgen: Zunächst wird eine Stichprobe von Schulen durchgeführt, aus der dann eine Stichprobe von Schülern entnommen wird. Dieses zweischrittige Verfahren ist grundsätzlich weniger präzise als ein einschrittiges (bei gleicher Anzahl der Elemente), da die Leistungen zweier zufällig ausgewählter Schüler derselben Schule mit größerer Wahrscheinlichkeit ähnlich sind als die Leistungen zweier zufällig ausgewählter Schüler. Daher wäre es präziser, aus jeder Schule nur einen Schüler auszuwählen. Das würde aber die Kosten vergrößern, da für jeden Schüler eine andere Schule besucht werden müsste. Deshalb machen in der Regel drei Schüler pro Schule denselben Test – obwohl aus Kostengründen oft auch mehr Schüler pro Schule verschiedene Tests bearbeiten, so dass einer drei Schreibtests, der nächste drei Lesetests und der dritte drei Sprechtests macht.

Der zentrale Punkt bei allen Stichprobenuntersuchungen besteht natürlich in der gewählten Genauigkeit des Durchschnitts, also dem erwartbaren Verhältnis zwischen empirischem Mittel der Stichprobe und dem Mittel der Gesamtpopulation. Wenn es darum geht, sowohl den Status quo zu bestimmen als auch Veränderungen zwischen verschiedenen Assessments erfassen zu können, muss der Genauigkeitsgrad so hoch wie möglich gewählt werden. Normalerweise entscheidet man sich für eine Genauigkeit von 95% (Wesdorp et al. 1986), so

dass der faktische Durchschnitt nicht weit von dem geschätzten entfernt ist. Dasselbe gilt für Teilpopulationen, obwohl hier die Stichproben natürlich weniger repräsentativ sind.
Teilpopulationen entstehen zum einen durch Differenzierung zwischen den Geschlechtern, zum anderen durch Gewichtung von Schülern, da diese für die Mittelzuweisung an Schulen eine große Rolle spielt: Die Gesamtpopulation der Schulen wird auf Basis dieser Gewichtung in drei Schichten eingeteilt: In der ersten befinden sich Schulen, in deren Einzugsgebiet Eltern mit Abschlüssen weiterführender Schulen vorherrschen, in der zweiten Schulen, die vorwiegend von Arbeiterkindern und einigen Migrantenkindern besucht werden, und in der dritten Schulen, in denen Arbeiter und Migrantenkinder dominieren.
Für jede Schicht wird eine Stichprobe entnommen, um eine Gesamtgenauigkeit des Mittels von etwa 95% zu erreichen. Die Genauigkeit pro Stichprobe ist dann etwas geringer (s. Cito 2002). Diese Zahlen mögen zunächst attraktiv erscheinen, aber es ist zu berücksichtigen, dass die Schulen am National Assessment auf freiwilliger Basis teilnehmen, so dass es zwischen Teilnehmern und Nicht-Teilnehmern Unterschiede geben könnte, die die Repräsentativität der Stichproben beeinflussen. Dieser Bias ist in dem ersten National Assessement von Wesdorp et al. (1986) quantifiziert worden (Tabelle 3).

Angenommenes Ergebnis aller Nichtteilnehmer (auf einer Skala von 0-6)	Prozentualer Anteil der Teilnehmer mit einem Ergebnis von 0-6	Bias bei der Schätzung des Durchschnitts X
0	2.8	2.00
1	4.5	1.52
2	8.1	1.04
3	12.3	0.56
4	21.6	0.08
5	31.9	-0.40
6	18.9	-0.88

Tabelle 3: Bias der Genauigkeit des Populationsdurchschnittes durch Nicht-Teilnahme bei einem Lesetest (X = 4,17)

Wenn also z.B. das Ergebnis aller Nichtteilnehmer bei einem Test 0 ist, beträgt der Bias der Genauigkeit des Populationsdurchschnittes 2 Punkte. So ein großer Unterschied ist allerdings unwahrscheinlich, da nur 2,8% der Teilnehmer dieses Ergebnis erzielten. Wenn aber alle Nichtteilnehmer die Höchstpunktzahl erreichen würden, wäre der Schätzwert für den Populationsdurchschnitt 0,88 unten vom wahren Durchschnitt entfernt. Selbst wenn der durchschnittliche Unterschied zwischen Teilnehmern und Nicht-Teilnehmern nur in einer einzigen korrekt bearbeiteten Aufgabe bestehen würde, wäre der Bias des geschätzten Durchschnitts erheblich, da er etwa das Sechsfache der Standardabweichung betragen würde. Dieses Beispiel zeigt, dass das Problem der Nicht-Teilnahme zu erheblichen Abweichungen führen kann, vor allem, wenn man berücksichtigt, dass die Gesamtbeteiligung am National Assessment unter 50% sowie in der untersten Schulschicht (mit vorwiegend Arbeiter- und Migrantenkindern) nur bei etwa 30% liegt.
Die Resultate der National Assessments in den Niederlanden werden auf eine Skala aufgetragen. So haben beim Lesen argumentativer Texte, **die 18 Aufgaben enthält, die

10% schwächsten Schüler eine geringere Chance von 50%, überhaupt eine Aufgabe korrekt zu bearbeiten. Der durchschnittliche Schüler beantwortet sechs Aufgaben korrekt, sechs weitere teilweise korrekt und besteht zwölf Aufgaben nicht ((** 6+6+12= 24??)) Die sehr gute Schüler (**prozentual-90) beantworten elf Aufgaben korrekt, sieben teilweise korrekt und sechs unzureichend. (Cito, 61). Neben sölche allgemeine Darstellungen werden die Ergebnisse aber auch nach Geschlecht, zu Hause verwendeter Sprache, Alter und Schüler-Gewichtung aufgeschlüsselt. Normalerweise erreichen Mädchen etwas höhere Punktzahlen als Jungen (nur bei reflektiven Texten sind Jungen den Mädchen überlegen). Außerdem sind die Ergebnisse von Schülern, die die niederländische Standardsprache sprechen, etwas höher als die von Dialektsprechern oder Schülern aus mehrsprachigen Haushalten. Schüler, die zu Hause kein Niederländisch sprechen, sondern z.B. ausschließlich die Muttersprache ihrer eingewanderten Eltern, haben die schlechtesten Ergebnisse in Sprachtests. Es kommt heraus, dass die Gewichtung von Schülern ein wichtiger Indikator für das sprachliche Leistungsvermögen: Je höher die Gewichtung, desto geringer das Leistungen.

Die National Assessments der letzten Jahre haben ergeben, dass sich die meisten sprachliche Leistungen zwischen 1993 und 1998 kaum geändert hat: Schüler in 1998 können ebenso gut lesen und schreiben wie die Schüler in 1993. Dasselbe gilt für Teilfähigkeiten wie Grammatik, lexikalisches Wissen, Orthographie usw.. Lediglich was das Hören berichtender und fiktionaler Texte betrifft, scheint das Niveau zwischen 1988 und 1998 leicht gesunken zu sein.

Es ist wichtig einzusehen, dass die Leistungsskala ihre eigene Interpretation nicht mitliefert. Was bedeutet es zum Beispiel, dass 50% der Schüler sechs Aufgaben korrekt beantwortet haben? Ist dies ausreichend oder ungenügend? Es müssen also den Zahlen Bewertungsmaßstäbe zugeordnet werden. Cito verwendet hierzu eine Methode, bei der kompetente Personen entscheiden, wieviele Aufgaben Schüler korrekt bearbeiten müssen, um hinsichtlich ihres Leistungen als minimal, ausreichend oder fortgeschritten bewertet zu werden. Diese Methode kann zwar für alle Schulfächer eingesetzt werden, produziert aber nicht immer relevante Ergebnisse. So wurden z.B. Politiker, die mit der Einführung solcher Normen betraut sind, befragt, wieviele Rechtschreibfehler ihrer Ansicht nach Schüler im Durchschnitt machen. Ohne Aussnahme schätzten sie das diesbezügliche Leistungen der Schüler zu gering ein. Daher könnte eine andere Methode der Festsetzung von Bewertungsstandards vorzuziehen sein, so z.B. bei Schreibaufgaben die Schlüsselaufgabenmethode ('core-item method'). Diese Methode besteht darin, dass dasjenige Element einer Aufgabe identifiziert wird, das eine Schüler korrekt bearbeiten müssen, um das fiktive Gesamtziel der Aufgabenstellung überhaupt erreichen zu können. Wenn ein Schüler bei der Kartoffelchips-Aufgabe (Tabelle 2) die Absenderadresse vergisst, hat er Kommunikation unmöglich gemacht, so dass diesem Element größere Wichtigkeit zukommt als anderen (vgl. Kuhlemeier/Van den Bergh 1990).

Schülerleistungen hängen, zumindest zum Teil, von der Art des Unterrichts ab. Daher werden beim National Assessment auch Informationen zu Lehrmethoden und fachspezifischen Unterrichtszeiten gesammelt, um den Status quo sowie Trends ermitteln zu können. So sind z.B. heutzutage ergebnisorientierte Methoden wesentlich populärer als strategische oder eklektische. Und im Schnitt werden in den letzten drei Jahren der Primarstufe fast 5 Stunden wöchentlich auf den Bereich Sprache verwendet, von denen sich durchschnittlich 2 Stunden und 18 Minuten auf Leseaktivitäten beziehen.

Trotz einer so detaillierten Aufschlüsselung der Ergebnisse des National Assessment nach Methoden etc., lassen sich hier keine Input-Output-Analysen durchführen, da die Anzahl der Schüler pro Schule viel zu gering ist, um präzise Schätzungen zu gestatten. Die gesamte Auslegung des National Assessment sowie die Item-Response-Methode läuft solchen Interpretationen zuwider, so dass diese Verfahren lediglich für eine effiziente Schätzung des Leistungsstandes der Schülerpopulation in Anspruch genommen werden sollten.

4. Diskussion

4.1 Der hinter Sprachtests stehende Sprachbegriff

Nach dem bisher Entwickelten konzentrieren wir uns nun auf die Frage, welcher Sprachbegriff den Sprachaufgaben des National Assessment zugrundeliegt. Traditionelle Sprachtests sind überwiegend auf der Unterscheidung zwischen rezeptiven und produktiven Fähigkeiten basiert und häufig auf Lese- und Höraufgaben beschränkt (Van Berkel 2002). Aber durch das Abtesten von Teilfähigkeiten kann nur ein fragmentarisiertes Bild sprachlicher Kompetenz entstehen.[3]
Dieser fragmentarisierte Sprachbegriff wird von Befürwortern mit dem Hinweis auf die Notwendigkeit der Validität von Tests verteidigt, da nur das Abtesten von Teilfähigkeiten zuverlässige Vergleiche der Testergebnisse aus mehreren Jahren und von mehreren Individuen **gestattet. Interessanterweise enthält jedoch der Abschlussbericht des National Assessment (Berkel et al. 2002) den Vorschlag, die fragmentarisierten Teilfähigkeiten auf Basis eines interaktionalen Sprachkonzepts zu integrieren (Berends 2002, 12). Dieser Vorschlag hat jedoch bei den zuständigen Institutionen keine Reaktionen ausgelöst. Stattdessen stellt Cito sicher, dass die Aufgaben des National Assessment sogenannten ‚Lehrplanelementen', die in den Schulen faktisch zum Einsatz kommen, entsprechen. Es ist auch fraglich, ob Berkels Vorschlag einer Integration der Teilfähigkeiten mit den Zielen des National Assessment vereinbar ist, die ja in der Evaluierung des Bildungswesens selbst bestehen, weswegen Vergleichbarkeit der Ergebnisse allgemein sowie zwischen Teilpopulationen höchste Priorität hat. Daher müssen die durchgeführten Tests statistischen und psychometrischen Grundanforderungen genügen. Dabei gestattet die Praxis der matrixgesteuerten Stichprobenerhebung, bei der jeder Schüler nur einen kleinen Teil der Tests absolviert, keine tiefere Analyse der Beziehungen zwischen den Ergebnissen verschiedener Tests. Für eine Analyse der Relationen zwischen sprachlichen Teilfähigkeiten müsste man daher ein weniger effizientes, d.h. im statistischen und psychometrischen Sinne weniger strenges Testdesign verwenden (s. Kuhlemeier/Van den Bergh 1998).
Verglichen mit dem Cito-Abschlusstest decken die Sprachtests des National Assessment einen relativ großen Bereich ab – indem alle Unterbereiche und Begleitfähigkeiten durch insgesamt 166 verschiedene Tests abgeprüft werden –, während der Cito-Test nur 100 Sprachaufgaben beinhaltet. Dies ist umso erstaunlicher, als der Cito-Test einen wesentlich größeren Einfluss auf individuelle Schullaufbahnen hat und in der Öffentlichkeit auf weitaus stärkere Resonanz stößt: Jedes Jahr ist er ein Medienereignis, das die öffentliche Bildungsdebatte befeuert.

4.2 National Assessment, individuelle Sprachaneignung und Qualitätskontrolle

[3] Die traditionellen Bereiche des Sprachunterrichts, die auch in dem National Assessment verwendet werden, werden in Unterbereiche unterteilt. Was das Lesen betrifft, unterscheidet man zwischen fiktionalen Texten und Sachtexten, und für das Hören zwischen den Untertypen Argumentation, Reportage, Kommentar und fiktionaler Text. Darüber hinaus werden die folgenden sogenannten ‚Begleitfähigkeiten' (supportive skills) getestet: (1) Wortbedeutungen, (2) semantische Relationsbeziehungen wie Ober- und Unterbegriffe, Gegensätze und Bedeutungsüberlappungen, (3) Morphologie nominaler Ausdrücke: Plural, Genus, Nominalkomposita, Diminutive, Steigerungsstufen der Adjektive; (4) Zerlegung von Wörtern in Silben; (5) Funktionswörter, (6) die Konstruktion komplexer Sätze aus einfachen Hauptsätzen, die Umformung deklarativer Sätze in Fragesätze sowie Änderungen der Wortstellung ohne Änderung des propositionalen Gehalts, (7) Orthographie, (8) Interpunktion und, schließlich, (9) das Alphabet. Alle diese Elemente des Niederländischen werden separat getestet, wobei die Frage, in welchem Verhältnis Beherrschung dieser Teilfähigkeiten zur faktischen Sprachbeherrschung steht, ignoriert wird.

Aus pädagogischer sowie curricularer Perspektive ist es eine wichtige Frage, in welchem Maße sich verschiedene Beurteilungsinstrumente zur Beobachtung des Sprachaneignungsprozesses individueller Schüler eignen. Schulische Examen sowie der Cito-Abschlusstest leiten sich von den Normen und Zwecken der niederländischen Qualifikationsstruktur her und sind mithin nicht auf die Dokumentation der Entwicklung einzelner Schüler abgestellt. Das National Assessment ist mit der Gesamtentwicklung des Bildungsstandards befasst. Beide Verfahren tragen dem individuellen Schüler nur wenig Rechnung.
Die Fortschritte individueller Schüler können nur durch das *Pupil-Monitoring-System* sichtbar gemacht werden, das regelmäßig die Einzelergebnisse auf eine standardisierte Norm bezieht. Obwohl bei solchen Verfahren nur einige Teilfähigkeiten abgeprüft werden, hat das Pupil-Monitoring-System eine diagnostische Funktion, da es den Lehrer auf unzureichende Fortschritte aufmerksam macht. Es ist dann die Aufgabe des Lehrers, die Testergebnisse auf Basis seiner eigenen Einschätzung der Fähigkeiten des Schülers zu interpretieren und zu entscheiden, ob besondere Förderungsmaßnahmen nötig sind.
Dem Pupil-Monitoring-System liegt eine stillschweigende Annahme zugrunde, die für das niederländische Bildungswesen keineswegs selbstverständlich ist, nämlich dass alle Schulen nach demselben (Sprach-)Lehrplan vorgehen. Denn die Tests des Pupil-Monitoring-Systems sind nur dann sinnvoll, wenn eine Schule nach dem diesem System korrespondierenden Lehrplan vorgeht. Daher ist es mitunter keineswegs klar, ob ein Schüler einen Test wegen mangelnder Fähigkeiten nicht bestanden hat, oder einfach nur deswegen, weil die abgeprüften Fähigkeiten und Inhalte gar nicht Unterrichtsgegenstand waren. Da in den Niederlanden Bildungsfreiheit sehr wichtig ist, haben Schulen das Recht, ihre eigenen Lehrpläne zu erstellen – und von diesem Recht machen sie auch Gebrauch. Demzufolge werden in der Primarstufe mehrere Monitoring Systeme parallel eingesetzt, was aber kein eigentliches Problem darstellt, da hinsichtlich der in der Primarstufe zu vermittelnden Grundfähigkeiten weitgehend Konsens besteht. In den weiterführenden Schulen sind hingegen die Unterschiede der Sprachlehrpläne so gravierend, dass alle Versuche, hierfür ein übergreifendes Monitoring System zu entwickeln, gescheitert sind.
In diesem Kontext ist eine im niederländischen Schulwesen traditionsreiche Institution zu erwähnen, über die bisher nichts gesagt worden ist: die *Nationale Schulinspektion*, deren Aufgabe es ist, auf nationaler Ebene sowie auf Gemeindeebene die Einhaltung von Minimalstandards zu überwachen. Die Schulinspektion hat das Recht und die Pflicht, in Schullehrpläne und Unterrichtsstrukturen einzugreifen, wenn es Probleme gibt, so z.B. bei Beschwerden der Eltern. Alle Schulen sind verpflichtet, mit der Schulinspektion zu kooperieren und ihren Weisungen Folge zu leisten, und die Schulvorstände sind hierfür verantwortlich. So wurde kürzlich die Stadt Amsterdam als Schulvorstand für Amsterdams öffentliche Schulen dazu verpflichtet, zusätzlichen Sprachunterricht zu bezahlen, nachdem sich Eltern über den schlechten Sprachunterricht an einer Schule beschwert hatten, die wiederholt die Weisungen der Schulinspektion ignoriert hatte. Die Nationale Schulinspektion hält den Bildungsminister über wichtige Themen im Bildungssektor durch regelmäßige Berichte auf dem Laufenden und berät in Fragen der Entwicklung der in unserer Untersuchung dargestellten Beurteilungsinstrumente. Es ist nicht immer ganz offensichtlich, welchen Einfluss die Schulinspektion tatsächlich besitzt; es kommt aber vor, dass sie eine politische Debatte über Bildungsstandards oder -qualität initiiert.

4.3 Beurteilung sprachlicher Fähigkeiten von Migrantenkindern

Nach diesem umfassenden Überblick über das niederländische Bildungswesen kann nun die Frage erörtert werden, wie die vorhandenen Beurteilungsinstrumente den Bedürfnissen von Migrantenkindern Rechnung tragen. Wie bereits gesagt, ist – trotz aller Bemühungen – die

Position dieser Kinder im niederländischen Schulsystem immer noch problematisch. Im allgemeinen erzielen sie bei den Tests schlechte Resultate und sind daher in den Sonderzweigen der Primar- und Sekundarstufe sowie in den elementar berufsvorbereitenden Schulen überrepräsentiert. Im Schulwesen ist eine immer stärkere ethnische Segregation zu beobachten, die sich durch Akzeptanz zu verfestigen droht (Karsten et al. 2002). Die Sprachtestresultate von Migrantenkindern haben zwar einen großen Einfluss auf deren Schullaufbahn, aber sind nicht der einzige Grund für die Probleme, da die Schulwahl der Eltern und u.U. ihre Präferenz für kulturell homogenere Schulen auch eine große Rolle spielt. Alltäglicher und institutioneller Rassismus **räumt mit dem Mythos der sprichwörtlichen niederländischen Toleranz auf (Koole/ten Thije 1994).
Was die Angemessenheit der Beurteilungsinstrumente für den Sprachstand von Migrantenkindern betrifft, sind drei Punkte besonders hervorzuheben.
Zunächst ist auf den Ethnozentrismus der Sprachtests hinzuweisen. Etliche Untersuchungen haben erwiesen, dass die Aufgaben des Cito-Abschlusstests einen kulturellen Bias besitzen (z.B. Kok 1988, Uiterwijk 1994). Allerdings hat dieser Bias auf die Testresultate höchstens minimalen Einfluss: die Cito-standarisierte Gesamtbewertung wird kaum durch einen kulturellen oder geschlechtsspefischen Bias bestimmt Das heisst, diese Bias bestehen und auch significant sind, sich aber nur auf sehr weniger Aufgabestellungen besiehen (Uiterwijk 1994).Für die staatlichen Prüfungen oder die Tests des Pupil Monitoring Systems liegen diesbezüglich noch keine Daten vor.
Eine weitere Frage ist der Einfluss der Sprachtests auf die Chancen von Migrantenkindern. Sie werden durch das Pupil Monitoring System als ‚Risikogruppe' identifiziert, und der Cito-Abschlusstest verhindert, dass sie sich für eine Schulart entscheiden, die ihren Fähigkeiten nicht angemessen ist. Hierdurch werden auch die u.U. zu hohen Erwartungen der Eltern gedämpft. Als sich herausstellte, dass der Cito-Test für Migrantenkinder unpräzise Resultate lieferte, etablierte sich vorübergehend die Praxis, die Empfehlungen für die weitere Schullaufbahn nach oben hin zu korrigieren. Es stellte sich dann aber heraus, dass die Migrantenkinder nicht die erforderlichen Leistungen erbrachten und die Schule ohne Abschlusszeugnis verließen (Van Veen/Berdowski 2000; Vedder/Kloprogge 2001). Der umgekehrte Fall wird durch die Situation in Amsterdam illustriert: Seit der Cito-Abschlusstest dort für alle Schüler Pflicht ist und die Ergebnisse die Schulwahl beschränken, hat sich im allgemeinen die Schulwahl auch von Migrantenkindern geändert, und es ist zu beobachten, dass sie seltener die Schule ohne Abschlusszeugnis verlassen (Minister of Education 1999).
Abschließend ist anzumerken, dass wir uns nur auf Durchschnittsergebnisse und deren Veränderung über mehrere Jahre hinweg beschränkt haben – so beziehen sich unsere Aussagen über Migrantenkinder auf solche Durchschnittsergebnisse. Auf Unterschiede zwischen Teilpopulationen innerhalb der verschiedenen Migrantengruppen wurde nicht eingegangen, obwohl diese de facto so groß sind, dass sie kaum Verallgemeinerungen zulassen. Auch die Variation innerhalb der Teilpopulationen wurde nicht erörtert, obwohl sie gerade bei Migrantenkindern erheblich ist.

4.4 Die Niederlande: Europas Testmeister?

Die besprochenen Testtypen – staatliche Examen, Cito-Abschlusstest und Pupil-Monitoring-Systeme – messen den Leistungsstand an bestimmten Punkten der Schullaufbahn. Jedes dieser Verfahren hat seine eigenen Stärken und Schwächen. Die staatlichen Examen, die nur zwischen bestanden/nicht bestanden differenzieren, dienen dem jährlichen Vergleich von Schulen desselben Typs. Sie gestatten weder den Vergleich verschiedener Schultypen noch den Vergleich der Ergebnisse aus verschiedenen Jahren. Ihr größter Nachteil ist jedoch, dass sie aufgrund ihres kleinen Umfangs, der kaum Rückschlüsse auf die faktischen Fähigkeiten der Schüler erlaubt, nicht als Indikator für Bildungsstandards dienen können.

Seit der institutionellen Einführung des National Assessment vor fünfzehn Jahren muss sich das System aufgrund auch der involvierten wirtschaftlichen Interessen ständig selbst rechtfertigen. Bisher ist aber nur wenig über seinen lokalen und landesweiten bildungspolitischen Einfluss bekannt geworden. Die erste, 1985 durchgeführte Untersuchung brach eine heftige Debatte über ‚funktionalen Analphabetismus' vom Zaun, während die drei folgenden Untersuchungen relativ stabile Lese- und Hörfähigkeiten belegten, wenn sie auch auf die strukturellen Probleme von Migrantenkindern aufmerksam machten. Aber solange diese Untersuchungen nicht dazu dienen, politischen Handlungsbedarf sichtbar zu machen, legitimieren sie die gegenwärtige Sprach- und Bildungspolitik.

Literatur

Berends, R. (2002) In Nederland missen we een goed concept voor taalonderwijs. In: Berkel, S. van et al. (red.), 12.

Bergh, H. van den (1990) On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement*, *14*, 1-12.

Bergh, H. van den & Kuhlemeier, H. (1990) De haalbaarheid van eindtermen voor de basisvorming. *Pedagogische Studiën*, *67*, 1-15.

Bergh , H. van den, Glopper, K. de & Schoonen, R. (1987) Directe metingen van schrijfvaardigheid: Validiteit en taakeffecten. In: F.H. van Eemeren & R. Grootendorst (Red.), *Taalbeheersing in ontwikkeling*. Dordrecht: Foris Publications.

Bergh, H. van den, Rohde, E, & Zwarts, M. (2003) Is het ene examen het andere? Over de stabiliteit van schoolonderzoek en centraal examen. *Pedagogische Studiën*, *80*, 176-191.

Berkel, S. van, Schoot, F. van der, Engelen, R. & Maris, G. (red.) (2002) *Balans van het taalonderwijs halverwege de basisschool 3. Uitkomsten van de derde peiling 1999.* Arnhem: Cito.

Cito (2002) *Balans van het taalonderwijs aan het einde van de basisschool 3*. Arnhem: Cito.

Clark, H.H. (1973) The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335-359.

Doornbos, K. (1986) De verzorgingsstructuur van het onderwijs. In: Kemenade, J.A. et al. (red.) *Onderwijs: Bestel en beleid I: Onderwijs in hoofdlijnen*. Groningen: Wolters, 244-270.

Glopper, K. de & Schooten E. van (2002) Dalende leerlingprestaties op de centraal schriftelijke examens Duits, Engels en Frans in mavo, havo en vwo? *Pedagogische Studiën*, *79*, 5-17.

Karsten, S., Roeleveld, J., Ledoux, G., Felix, C. & Elshof, D. (2002) Schoolkeuze en etnische segregatie in het basisonderwijs. *Pedagogische Studiën* 79/5, 359-376.

Kok, F. (1988) *Vraagpartijdigheid*. Amsterdam: Universiteit van Amsterdam.

Koole, T., Thije, Jan D. ten (1994) *The Construction of Intercultural Discourse. Team discussions of educational advisers* (Utrecht: diss.) Amsterdam / Atlanta: RODOPI.

Kuhlemeier, H. & Bergh, H. van den (1998) Relationships between language skills and task effects. *Perceptual and Motor Skills*, *86*, 443-463.

Minister of Education (1999) *Plan van Aanpak Voortijdig Schoolverlaten*. Den Haag: SDU.

Minister of Education (2003) *Onderwijsprofiel van Nederland. Samenvatting van de belangrijkste beelden van 'Education at a Glance'. Het onderwijs indicatoren rapport van OESO.* Website: http://www.minocw.nl/brief2k/2003 /doc/44136g. PDF, 15 January 2004.

Minister of Education (2004) *Fact and figure*. Website: http://www.minocw.nl/english/ figures2003/008.html, 15 januari 2004.

Roeleveld, J. (2002) De kwaliteit van het basisonderwijs: dalen de Cito-scores? *Pedagogische Studiën*, *79*, 389-403.

Sociaal en Cultureel Rapport (2000*) Nederland in Europa*, Sociaal en Cultureel Planbureau, Den Haag.

Uiterwijk, H. (1994) *Eindtoets basisonderwijs: De bruikbaarheid van de eindtoets basisonderwijs voor allochtone leerlingen*. Arnhem: Cito.

Vedder, P., Kloprogge, J. (2001) *Onderwijskansen op tafel: het bestrijden en voorkomen van onderwijsachterstand*. Den Haag: Management Landelijke Activiteiten Onderwijskansen PMPO.

Veen, D. van, Berdowski, Z. (2000) *Preventie van schoolverruim en zorg voor risicoleerlingen*. Leuven: Garant.

Webbink, D. (2002) Moeten we ons zorgen maken over dalende scores op de Eindtoets basisonderwijs? *Pedagogische Studiën*, *79*, 184-191.

Wesdorp, H. (1979) *Studietoetsen en hun effect op het onderwijs*. Staatsuitgeverij: Den Haag.

Wesdorp, H., Bergh, H. van den, Bos, D.J., Hoeksma, J.B., Oostdam, R.J., Scheerens, J. & Triesscheijn, B. (1986) *De haalbaarheid van periodiek peilingsonderzoek*. Lisse: Swets & Zeitlinger.