Weeren, Jan van

# Wem nutzen Outputmessungen? Eine kritische Analyse ihrer Wirksamkeit und Nebeneffekte aus niederländischer Perspektive

*Die Deutsche Schule 99 (2007) 2, S. 210-223*

Jan van Weeren

# Wem nutzen Outputmessungen?

## Eine kritische Analyse ihrer Wirksamkeit und Nebeneffekte aus niederländischer Perspektive

Inwiefern können *Outputmessungen* zur *Unterrichtsqualität* beitragen? Um diese zentrale Frage der Schulentwicklung zu beantworten, ist es wichtig, ihre Typologie zu bestimmen. Wenn sie schulübergreifend, national oder überregional, in der Funktion eines Bildungsmonitorings also, eingesetzt werden, ist ihre Interpretation im Hinblick auf einzelschulische Schul- und Unterrichtsentwicklung problematisch und ihre Effektivität fragwürdig. Auf der anderen Seite treten nachweisbar *unbeabsichtigte und unerwünschte Rückwirkungen* auf, sobald die gewonnenen Daten landesweit auf der Schulebene aggregiert werden.

Outputmessungen, die schulintern durchgeführt werden und auf individuelle Unterstützung und Steuerung zunächst der Schülerinnen und Schüler, aber auch des Lehrpersonals ausgerichtet sind, dürften der Unterrichtsentwicklung und -verbesserung wohl am meisten nutzen. Ein *Erfahrungsbericht* aus den Niederlanden.

### 1. Qualitätssicherung und Outputmessung

Wesentliche Parameter und Zusammenhänge, die es bei der Qualitätssicherung und Outputmessung im Unterricht gibt, sind in Abbildung 1 dargestellt. Der eine oder andere Leser wird bereits die Aussage im oberen Kästchen in Frage stellen, da aus seiner Sicht Unterrichtsqualität und gemessener Output nur bedingt miteinander zusammen hängen. Zwar könne man Fachleistungen erfassen, doch weder *soziale* noch *kommunikative Kompetenzen* könnten als schulischer Lernertrag mit den üblichen Testverfahren geprüft werden.[1] Die Gefahr drohe, dass nicht gemessen wird, was wichtig ist, sondern dass *wichtig wird, was gemessen wird.* Pädagogische Prozesse seien nicht in der Weise ‚technologisierbar', dass ihre Wirksamkeit an normativen Vorgaben gemessen werden könnte.[2]

Solche und ähnliche Zitate sind symptomatisch und stellvertretend für Auffassungen von Personen, die den Ertrag von Outputmessungen zumindest re-

---

[1] So z.B. Eva-Maria Stange, ehemalige Vorsitzende der GEW, www.vds-bildungsmedien.de/pdf/forum/f_01/seite_32.pdf.

[2] Vgl. z.B. Walter Herzog, Vorstand des PH-Rates Bern, in einem Hauptreferat der Studientage der PH Bern ‚Bildungsstandards und Unterrichtsqualität', 18. Oktober 2006.

lativieren wollen und für einen *breiten Bildungsbegriff* plädieren. Dieses Meinungsbild lässt sich in *folgenden Thesen* zusammenfassen:

(1.) Nicht alles, was im Unterricht angestrebt wird, kann durch Outputmessungen getestet werden.
(2.) Outputmessungen reduzieren den Unterricht zu einem zielgerichteten Training der Fähigkeiten, die geprüft werden.[3]
(3.) Die Qualität unterrichtlicher Interaktionsprozesse lässt sich nicht durch Outputmessungen feststellen.

Zu These 1 kann Folgendes bemerkt werden. Eine Zielsetzung im Unterricht kann nur dann wichtig und erstrebenswert sein, wenn man Unterschiede zwischen denen beobachten kann, die das gesteckte Ziel erreicht haben, und solchen, die es noch nicht erreicht haben. Sind derartige *Unterschiede* beobachtbar, so bieten diese Beobachtungen *Ansatzpunkte zum Testentwurf.* Zielsetzungen können nur dann verwirklicht werden, wenn ihr Erreichen überprüfbar ist, d.h. diese Ziele bedürfen einer Präzisierung. Was man dagegen nicht genau beschreiben kann, kann man weder gut rechtfertigen, noch wirksam unterrichten. Ohne *operationalisierte*, d.h. *messbare* Zielsetzungen, können Lehr- und Lernerfolge nicht festgestellt werden. Wenngleich es Messverfahren gibt, die affektive und interaktive Personenmerkmale erfassen können, werden diese normalerweise nicht bei Outputmessungen eingesetzt, weil entsprechende Ziele nicht operational definiert sind.

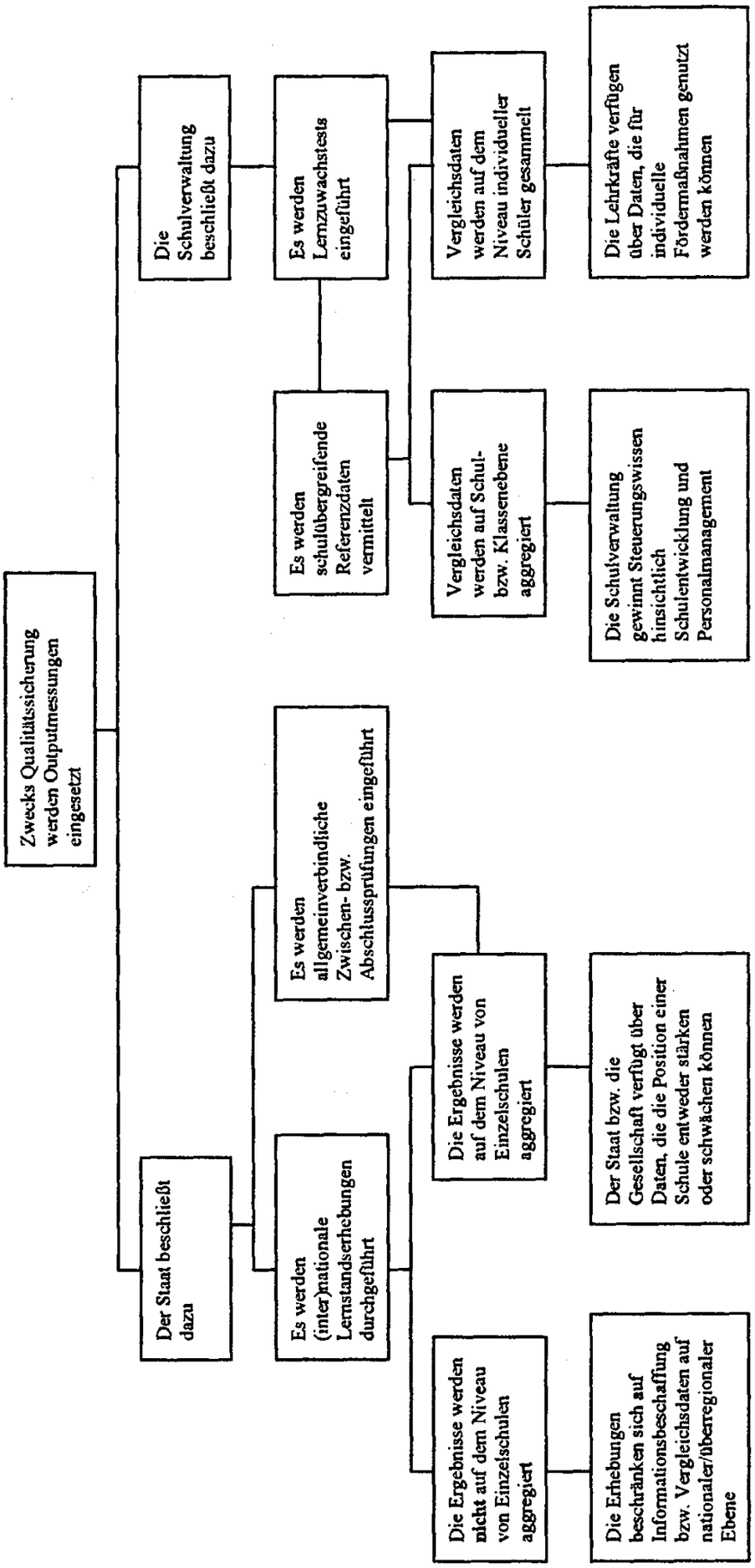
Der *backwash-Effekt*, der in These 2 angesprochen wird, kann sowohl negativ betrachtet als auch positiv genutzt werden. Wenn Outputmessungen die Unterrichtsziele – soweit sie *operational definiert* sind – nicht abdecken, kann dies unter Umständen zu einer Verarmung des Unterrichts führen, besonders dann, wenn die Testergebnisse Konsequenzen – in Form von positiven oder negativen Sanktionen – auf der Schul- oder der Personalebene nach sich ziehen können. Andererseits kann der Rückschlag-Effekt dahingehend ausgenutzt werden, dass als wesentlich erachtete und für unabdingbar gehaltene Lernziele im Unterricht auch tatsächlich verfolgt werden. Es gibt zahlreiche Fähigkeiten und Fertigkeiten, die sich ein Kind außerhalb der Schule aneignen kann, aber für Lesen, Schreiben und Rechnen ist es durchweg auf die Schule angewiesen, und diese hat demnach die Pflicht, die Schülerinnen und Schüler *auf jeden Fall* mit funktionsfähigem Grundwissen und Können auszustatten.

These 3 verwechselt Mittel und Zweck. Pädagogische Prozesse sind dazu da, Lernen auszulösen und Lernergebnisse zu ermöglichen. Sie sind *kein Ziel an sich.* Eine These zu einer rezenten Leidener Dissertation in der Erziehungswissenschaft lautet: ‚Eine affektive Beziehung zwischen Schüler und Lehrperson ist für den Schüler Grund, den normativen Erwartungen der Lehrperson entsprechen zu wollen.'[4] Die Prämisse ist hier wohl, dass man im Unterricht danach strebt, dass der Schüler den normativen Erwartungen – den Lernzielen – einer Lehrperson entspricht. Eine affektive Beziehung zwischen den beiden sei dazu ein geeignetes Mittel.

---

[3] Der so genannte ‚backwash-Effekt' oder ‚washback-Effekt' von Prüfungen und Tests.
[4] Ineke Henze, Science teachers' knowledge development in the context of educational innovation. 21. November 2006, meine Übersetzung.

Abbildung 1: Übersicht über den Einsatz von Outputmessungen



## 2. Die schulübergreifende Variante der Outputmessung

In Abbildung 1 sind drei Grundformen der Outputmessung dargestellt:

(1.) überregionale oder internationale Lernstandserhebungen,
(2.) allgemeinverbindliche Zwischen- bzw. Abschlussprüfungen,
(3.) schulspezifische oder schulübergreifende Lernzuwachstests.

Als Fallbeispiel der Outputmessung, bei der die Ergebnisse zunächst *bildungspolitisches Steuerungswissen* vermitteln sollen, werden hier die nationalen *Lernstandserhebungen* (Peilingsonderzoek PPON) in den Niederlanden beschrieben. Sie werden durch die vier äußerst linken Kästchen in Abbildung 1 charakterisiert. Die Erhebungen werden seit 1986 alljährlich von *CITO*, dem *Institut für Testentwicklung* (vgl. www.cito.nl/) im Auftrag des nationalen Bildungsministeriums durchgeführt. Ihre Legitimierung beruht darauf, dass dem Staat laut Verfassung ‚die Fürsorge für tauglichen Unterricht' obliegt. Aus diesem Grund werden jedes Jahr die Lernergebnisse in unterschiedlichen Domänen im Primarunterricht gemessen. Der Primarunterricht verfolgt als erste und einzige Schulstufe prinzipiell die gleichen Ziele für alle Schülerinnen und Schüler und bildet die Basis für den weiterführenden Unterricht, der in unterschiedlichen Schultypen aufgegliedert ist.

Abbildung 2: Ein mehrjähriger Terminkalender für die Lernstandserhebungen

| Lernbereiche | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
|---|---|---|---|---|---|---|
| Leseverstehen | Vorbereitung (V) | Messung (M) | Analyse (A) | Bericht (B) | | |
| Hörverstehen | | V | V | M | A | B |
| Schreibfertigkeit | B | | V | V | M | |
| Sprechfertigkeit | A | B | | | V | V |
| Rechnen/Mathematik | | V | M | A | B | |
| Weltkunde | | | V | V | V | M |
| Englisch | V | V | M | A | B | |
| Leibeserziehung | V | V | M | A | B | |
| Musik | | | V | V | M | A |
| Bildende Kunst | | | V | V | M | A |
| Handschrift | | | | V | M | A |
| Verkehrserziehung | | V | V | M | A | B |

Für den Primarunterricht gelten gesetzlich festgelegte Lernbereiche, deren Output beim Schulabschluss (im 12. Lebensjahr der Schüler) gemessen wird. Niederländisch und Rechnen/Mathematik werden als Hauptdomänen auch halbwegs (mit neun Jahren) getestet. Allerdings wäre es ein zu großer Aufwand, sämtliche Lernbereiche jedes Jahr zu testen. Zum einen würde der *Personalaufwand unerschwinglich* sein, zum anderen würde die *Belastung der Schulen und Schüler* zu groß werden. Aus diesem Grund wird entsprechend einem mehr-

jährigen Terminkalender für die Datensammlung, Datenanalyse und Berichterstattung vorgegangen.

Die *Hauptfragen bei den Erhebungen* lauten nach wie vor: Was wird gelehrt? Was wird gelernt? Verschieben sich die Unterrichtsergebnisse mit der Zeit? In welchem Ausmaß werden die allgemeinverbindlichen Kernziele erreicht?

Diese Kernziele beziehen sich auf Fähigkeiten, die in den Schüler/innen in unterschiedlichem Maße ausgeprägt sind. Bei den Messungen wird in der Regel eine beträchtliche *Spannweite der Leistungsniveaus* aufgezeigt. Aber wie sollte man das Intervall ‚ausreichende Beherrschung' innerhalb der gemessenen Fähigkeitsbandbreite definieren? Als Norm gilt, dass die Leistungsstufe, die vorab als ‚ausreichende Beherrschung' definiert ist, von 70 bis 75% der Gesamtpopulation erreicht werden soll. Was vorab als ‚Mindestniveau' angesetzt ist, soll für 90 bis 95% der Probanden erreichbar sein. Schüler/innen, die über dem Niveau ‚ausreichend' abschneiden, gelten als ‚fortgeschritten'.

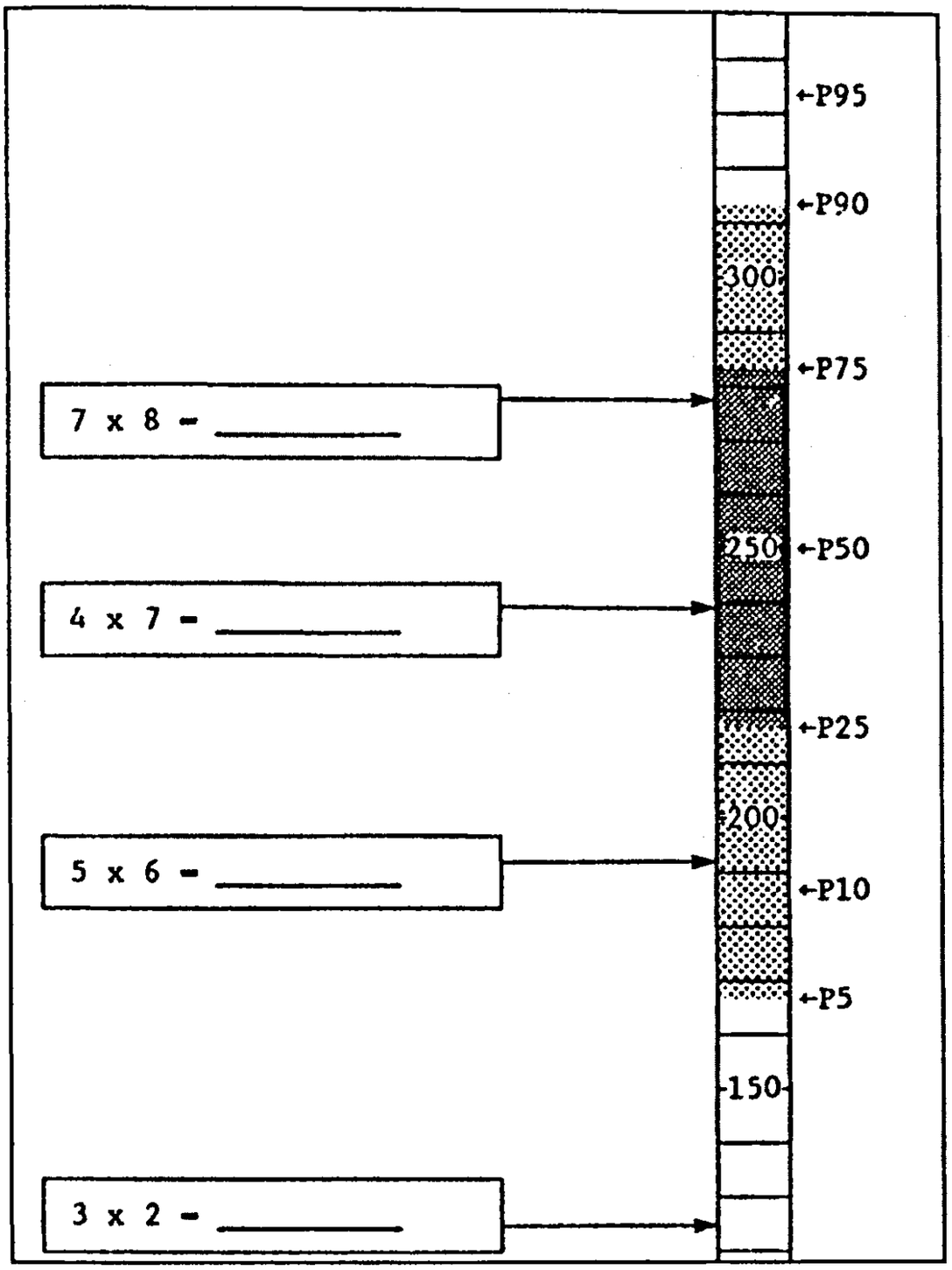
Für die Bestimmung der Leistungsniveaus werden Experten systematisch befragt. Es geht dabei um Ausschüsse von Primarschullehrern, PH-Dozenten und fachkundigen Schulberatern. Sie ordnen eine Vielzahl an Testitems den drei Niveaus – ausreichend, minimal bzw. fortgeschritten – zu. Die *Testitems sind vorerprobt und ausgewertet*, so dass den Experten auch die Beherrschungsperzentile bekannt sind: Sie wissen mithin, wie viel Prozent der Schüler/innen ein bestimmtes Item richtig lösen können.

Damit keine Lernziele übersehen werden, ist die *Aufgabensammlung* breit angelegt. Sie basiert auf einer Analyse von Lehrwerken und auf Expertenbefragungen. Sie enthält nicht nur schriftliche, sondern (selbstverständlich) auch praktische Aufgaben, z.B. in den Bereichen Sprechfähigkeit, Leibeserziehung, Kunstunterricht, Naturkunde und Informationssuche mit IuK-Technologie. Mit Hilfe dieser ‚flächendeckenden' Aufgabensammlung wird bei der Erhebung die *Ist-Situation* ermittelt. Bei den Lehrpersonen werden durch Fragebögen zusätzliche Daten gesammelt, u.a. über das konkrete Lernstoffangebot und den damit verbundenen Zeitaufwand, und über ihr pädagogisch-didaktisches Vorgehen, z.B. im Bereich der Binnendifferenzierung und beim ‚remedial teaching' (Förderunterricht). Damit die Datensammlung planmäßig abläuft, werden die Tests mit Hilfe angeleiteter Testassistenten durchgeführt.

Da bei den Erhebungen manchmal über 500 Aufgaben bearbeitet werden müssen, findet die Testdurchführung im *Matrixdesign* statt: Stichproben von 200 bis 300 Schüler/innen aus einer Gesamtpopulation von 4.000 bis 6.000 Probanden bearbeiten jeweils einen Teil der Aufgaben; durch Überlappungen der Tests können die Ergebnisse dennoch zu einem Gesamtbild vereint werden. Bei binären Lösungsmöglichkeiten, bei denen die Antworten entweder richtig oder falsch sind, wird für die Testanalyse die *Item-Response-Theorie* angewandt. Der Vorteil dieses Analyseverfahrens ist, dass die *Schwierigkeit* der Items und die *Fähigkeit* der Schüler/innen auf derselben Skala abgebildet werden können.

In Abbildung 3 zeigen wir eine Ergebnisskala der Mathematikmessung zur Mitte der Primarschule (9-jährige Schüler/innen), und zwar zur Elementarfähigkeit ‚Multiplizieren'. Alle Skalen haben 250 als Landesdurchschnitt und eine Standardabweichung von 50. Die Items sind so skaliert, dass Schüler/in-

Abbildung 3: Ergebnisskala für 'Multiplizieren'




nen mit einem entsprechenden Fähigkeitsscore eine Richtig-Antwort-Rate von 80% erreichen. Mit anderen Worten: Im Durchschnitt sind 8 aus 10 Schülern mit diesem Score fähig, die Aufgabe richtig zu lösen. Ein durchschnittlicher neunjähriger Schüler (Perzentilscore P50 auf der Skala) hat also eine Chance von mehr als 80%, die Aufgabe ‚4 x 7' richtig zu lösen. Praktisch alle Neunjährigen können die Aufgabe ‚3 x 2' richtig lösen. Ein wenig mehr als ein Viertel dieser Schüler/innen bewältigt die Aufgabe ‚7 x 8'.

Bei der Datenauswertung wird eine Reihe von Durchschnittswerten berechnet, z.B. für *Schulen* auf der Basis der Zusammensetzung der Schülerschaft (in Bezug auf ethnische bzw. soziale Herkunft) und der benutzten Unterrichtsmaterialien, und für *Schülergruppen* auf der Basis von Geschlecht, etwaiger Verzögerung der Schulkarriere und soziodemographischer Merkmale, z.B. Bildungsnähe der Elternhäuser, Migrationsstatus und Familiensprache.

Die *Berichterstattung* über die Ergebnisse erfolgt jeweils in dreifacher Ausfertigung:

(1.) Technische Berichte für forschungsorientierte Interessenten,
(2.) Berichte für Sachkundige (Institute für Bildungsforschung, Schulaufsicht, Schulberatungsstellen, Lehrerausbildungen) und
(3.) allgemeinverständliche Broschüren mit den Hauptergebnissen. Diese werden an alle Schulen und ihre Aufsichtsbehörden verschickt.

*Evaluation*

Wenngleich die Berichte generell günstig aufgenommen werden – wie durch Fragebögen festgestellt wurde –, haben die Resultate *keine nachweisbaren bildungspolitischen Folgen.* Durch die Berichterstattung wird zwar versucht, sowohl auf der Makro- als auch auf Mesoebene eine Diskussion anzustoßen, doch diese ergibt sich nicht zwangsläufig aus der Rückmeldung der Resultate. So wurde z.B. bei der Lernstandserhebung Rechnen/Mathematik im Jahr 2002 festgestellt, dass die Vorgabe (70 bis 75% der Schüler/innen erreichen die Leistungsstufe ‚ausreichend') in 13 der 22 Teilbereichen nur von 50% erreicht wird. Wenngleich diese und andere Ergebnisse Fragen aufkommen lassen bzw. bestimmte bildungspolitische Maßnahmen nahezuliegen scheinen, gibt es bislang *keine aussagekräftigen Studien*, in denen den Rückmeldungen aus den Lernstandserhebungen bestimmte Effekte zugesprochen werden. Es bleibt demnach bei der Aussage im unteren linken Kästchen der Abbildung 1: Es werden lediglich Lernstandsinformationen und Vergleichsdaten ermittelt.

## 3. Die externe schulspezifische Variante der Outputmessung

Wenden wir uns jetzt der Situation zu, dass der Staat *allgemeinverbindliche Abschlussprüfungen* im Primarbereich beschließt, d.h. in der gemeinsamen Unterrichtsphase, die im niederländischen Schulsystem bis zum 12. Lebensjahr dauert. Dazu gibt es in den Niederlanden gute Voraussetzungen. Beim Abschluss des Primarunterrichts wählen die Eltern für ihr Kind eine passende Schulform. Der Direktor der Grundschule berät sie bei der Wahl: Schließlich geht es nicht allein um den Wunsch der Eltern, sondern auch um die Kapazitäten des Kindes. Um diese genauer bestimmen zu können, ist der Einsatz eines objektiven Testverfahrens in den Niederlanden obligatorisch. Dabei ist es den Schulen jedoch freigestellt, welchen Test sie einsetzen, wenngleich etwas mehr als 80% der Schulen den Test des zentralen Testinstituts CITO einsetzen. Hat man sein Kind für eine bestimmte Sekundarschulform angemeldet, wird *vor der Zulassung das Testergebnis überprüft.* Im Prinzip kann die Sekundarschule die Annahme des Kindes wegen eines zu niedrigen Testwertes verweigern.

Der *CITO-Test* (Eindtoets Basisonderwijs) existiert seit 1968. Er nimmt momentan drei Testtage in Anspruch und enthält insgesamt 200 Aufgaben in den Bereichen Rechnen, Muttersprache, Informationssuche und Weltkunde. Der Zyklus der Aufgabenentwicklung, Aufgabenerprobung und Testzusammenstellung findet jedes Jahr statt. Damit die Tests jährlich *angeglichen* werden können, werden zwischen den neu entwickelten Aufgaben so genannte *Ankeritems* aus früheren Tests eingefügt. Im Jahre 1985 hat man 535 als durch-

schnittlichen Standardscore festgestellt. Die möglichen Standardscores der Schülerinnen und Schüler rangieren zwischen 500 und 550; für die Zulassung zum Gymnasium wird normalerweise der Standardscore 545 verlangt. Der Schule wird der Durchschnitt der Schülerleistungen in den unterschiedlichen Lernbereichen rückgemeldet. Die Rückmeldung ermöglicht einen *Vergleich mit dem Landesdurchschnitt sowie mit dem vorjährigen Resultat.*

Bis vor einem Jahrzehnt waren die *Schulscores* vertraulich. Sie wurden nicht vom CITO freigegeben. Der Schulvorstand konnte sich entscheiden, ob und wem er die Resultate bekanntgeben möchte. Seit 1997 werden die Schulergebnisse jedoch veröffentlicht. Die Schulaufsicht wurde dazu durch die Klage einer Zeitung unter Berufung auf das *Gesetz über die Verwaltungsöffentlichkeit* gezwungen. Die Ergebnisse werden seitdem von der Schulaufsicht publiziert. Zwar werden die Daten für die Zusammensetzung der Schülerschaft korrigiert, damit bei einem Vergleich der jeweilige *Standorttyp* der Schule berücksichtigt werden kann, doch es werden öfters leistungsschwache Schüler/innen von der Testteilnahme ausgeschlossen. Manche Schulen versuchen, auf diese Weise ihre Ergebnisse zu ‚frisieren'. Da die Schulwahl frei ist, und die Finanzierung der Schulen an die Zahl der eingeschriebenen Schüler/innen gekoppelt wird, empfiehlt es sich durchaus, mit relativ guten Leistungen aufzuwarten.[5] Die Schulinspektion stellt deshalb eine *verstärkte Kontrolle der Testdurchführung* in Aussicht, um solche Effekte zu vermeiden. Die Bildungsministerin hat zudem in jüngster Zeit eine verbindliche Teilnahme für alle Schüler/innen in den Bereichen Sprache und Rechnen angekündigt, zumal in einem Expertisebericht festgestellt wurde, dass sich der CITO-Test wegen seines Messbereichs und seiner Messqualität durchaus als zentraler Test eignet (Bosker und Heeringa 2006).

### *Evaluation*

Damit hat sich ein *Beratungstest* innerhalb von 40 Jahren nahezu schleichend zu einer *regelrechten zentralen Abschlussprüfung* weiterentwickelt. Einheitliche Durchführungsbestimmungen werden dazu führen, dass eindeutige Leistungsvergleiche zwischen Schulen im Bereich der Kernfähigkeiten gemacht werden können. Dieses können dazu führen – entsprechend dem zweiten Kästchen von links in Abbildung 1 –, dass die Position der Schulen entweder gestärkt wird, z.B. wenn sie bei unterschwelligen Leistungen irgendwie unterstützt werden, oder aber geschwächt, z.B. wenn bei freier Schulwahl die Marktkräfte prävalieren und/oder die Schulen wegen schlechter Leistung in ihrem Budget gekürzt werden.

## 4. Die interne schulspezifische Variante der Outputmessung

Wenn Outputmessungen von der Schulverwaltung zur *Leistungskontrolle* eingeführt oder genehmigt werden – die äußerst rechte Spalte in Abbildung 1 –, handelt es sich um *Lernzuwachstests*. Damit sind standardisierte normbezo-

---

[5] So wurden im Schuljahr 2005/2006 im Durchschnitt 5,2% der Schüler von der Teilnahme ausgeschlossen, während sich etwa 3% von ihnen durchaus an dem Test hätten beteiligen können. Von zwei Drittel der Großstadt-Schulen wurden die Testscores der leistungsschwächsten regulären Schülerkategorie bei der Berichterstattung ausgeklammert. (Inspectie van het Onderwijs 2006)

gene Tests gemeint. Die *Standardisierung* bezieht sich einerseits auf die Testdurchführung, andererseits auf die Testauswertung und -interpretation. Die *Normorientierung* ermöglicht es, die Position eines Individuums relativ zu einer vergleichbaren Gruppe zu bestimmen.[6] Die Tests ergänzen und objektivieren die Progressionsinformation, die durch Klassenarbeiten gewonnen wird. Ein Problem bei solchen Tests ist allerdings, dass sie in der Regel in relativ großen zeitlichen Abständen abgehalten werden. Etwaige Lernrückfälle werden dann zu spät aufgezeichnet, um das Lehrpersonal zu effektiven Fördermaßnahmen zu befähigen.

Als Alternative bieten sich die *Schülerbegleitsysteme* (*pupil monitoring systems,* niederländisch: *leerlingvolgsystemen*) an. Dabei handelt es sich um Informationssysteme, welche die Lehrpersonen und die Schulleitung zu einer systematischen Überwachung des Lernfortschritts und der Unterrichtsqualität befähigen.

Abbildung 4: In den Jahresgruppen getestete Lernbereiche

| | Gruppen im Primarunterricht (4. – 12. Lebensjahr) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Ordnen | * | * | | | | | | |
| Sprache | * | * | | | | | | |
| Räumliche/zeitliche Orientierung | * | * | | | | | | |
| Technisches Lesen | | | * | * | * | * | * | * |
| Verstehendes Lesen | | | * | * | * | * | * | * |
| Hörverstehen | | | * | * | * | * | * | * |
| Wortschatz | | | * | * | * | * | * | * |
| Rechtschreibung | | | * | * | * | * | * | * |
| Ausdrucksfähigkeit | | | | * | * | * | * | * |
| Schreibfertigkeit | | | | | | * | * | * |
| Rechnen/Mathematik | | | * | * | * | * | * | * |
| Gesellschaftskunde | | | | | | * | * | * |
| Studierfähigkeit | | | | | | * | * | * |

Gruppen 1-2 : Anfangsstufe
Gruppen 3-4-5: Unterstufe
Gruppen 6-7-8: Oberstufe

Wir stellen hier ein System (Leerlingvolgsysteem LVS) vor, bei dem eine ständige (halbjährliche) Lernstandsmessung *zwecks frühzeitiger individueller Unterstützung* und Ausgleichens von Defiziten im niederländischen Primarunterricht vorgesehen ist. Es sammelt individuelle Vergleichsdaten in unterschiedlichen Lernbereichen über die Schuljahre hinweg, vermittelt nationale Referenzwerte und enthält zusätzliche *Förderungshinweise und Übungsmaterialien,* die zur Lernentwicklung eingesetzt werden können. Die Daten können auch auf Klassen- und Schulebene aggregiert werden. Obige Abbildung gibt die Lernbereiche wieder, die in den unterschiedlichen Stufen des Primarunterrichts geprüft werden.

---

[6] vgl. Leutner 2001.

Der Einsatz des Systems erfolgt im Rahmen eines *Deming-Zyklus*[7]. Es fängt mit einer *check-Phase* an: Zur Datenerhebung werden Tests durchgeführt. In der *Planungsphase* werden Zusatzinformationen über die Schüler/innen gesammelt, gesichtet und gewertet: Was weiß die Lehrperson über sie? Wie verhält sich der Schüler in der Klasse? Wie ist die Lage zu Hause? Dies alles führt zu einer *Interventionsplanung*. In der *do-Phase* werden die geplanten Maßnahmen in die Tat umgesetzt. Sodann erfolgt eine neue check-Phase, indem die Maßnahmen durch neue Tests auf ihre Zielwirksamkeit überprüft werden.

Theoretisch könnte man den Erfolg der Eingriffe an den Rangplätzen der Schüler/innen in der Testgruppe abmessen.[8] Der Leistungsstand eines Schülers hängt in diesem Fall von dem seiner Mitschüler/innen ab. Leider kann dann weder sein Leistungsfortschritt festgestellt werden, noch die Frage beantwortet, inwieweit seine Entwicklung normal verläuft. Damit die *Lernprogression* eindeutig festgestellt werden kann, werden die Testitems Rasch-skaliert. Dieses Verfahren macht es möglich, den Schwierigkeitsgrad der Testitems und die für die Lösung der Testitems erforderliche Fähigkeit der Schüler/innen auf einer und derselben Skala abzubilden (siehe Abbildung 3 für ein einfaches Beispiel.) Die Aufgabe ‚5 x 6' ist (erfahrungsgemäß) schwerer zu lösen als die Aufgabe ‚3 x 2', und bedarf also einer größeren Fähigkeit im Multiplizieren. *Itemschwierigkeit und Lösungsfähigkeit* werden in derselben runden Zahl ausgedrückt (siehe Abbildung 5). Ein Schüler, der bei einem Test beispielsweise die Fähigkeit 88 erreicht hat, wird generell in der Lage sein, Items bis Schwierigkeitsindex 88 zu lösen. Schafft er bei einem späteren Test die Fähigkeit 93, dann hat er deutlich dazugelernt. Wie bereits gesagt, findet im vorgestellten System die Leistungsmessung alle halben Jahren statt, hier z.B. zur Mitte der 3. Klasse (M3), am Ende der 3. Klasse (E3), halbwegs der 4. Klasse (M4), und beim Abschluss der 4. Klasse (E4).

Peters Lernzuwachs ist unverkennbar, aber reicht er aus? Bleibt er nicht zurück im Vergleich zu anderen Schülern? Um das zu erfahren, müssen wir über *Referenzdaten* verfügen können. Es werden dazu *landesweit* Referenzwerte gesammelt, die wie folgt eingestuft sind:

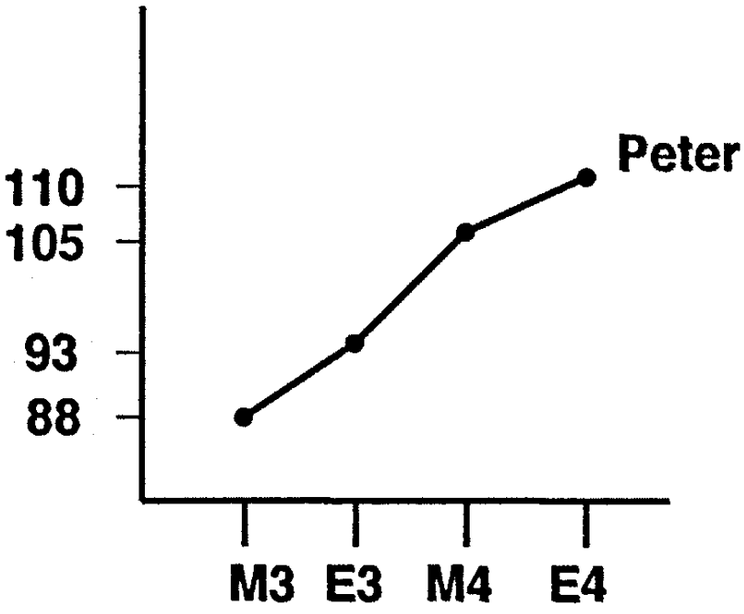
(1.) Leistungsstufe A: diese Resultate werden von den oberen 25% der Schüler/innen erzielt.
(2.) Leistungsstufe B: diese Resultate liegen zwischen dem Landesdurchschnitt und der Leistungsstufe A.
(3.) Leistungsstufe C: diese Resultate werden von dem Viertel der Schüler/innen gleich unter dem Landesdurchschnitt erreicht.
(4.) Leistungsstufe D: diese Resultate liegen weit unter dem Landesdurchschnitt und werden von 15% der Schüler/innen erreicht.
(5.) Leistungsstufe E: die Resultate der unteren 10% der Schüler/innen.

---

[7] Als Deming-Zyklus wird eine Folge von wiederholten Arbeitsschritten bezeichnet, die zu kontinuierlichen Verbesserungen führen.

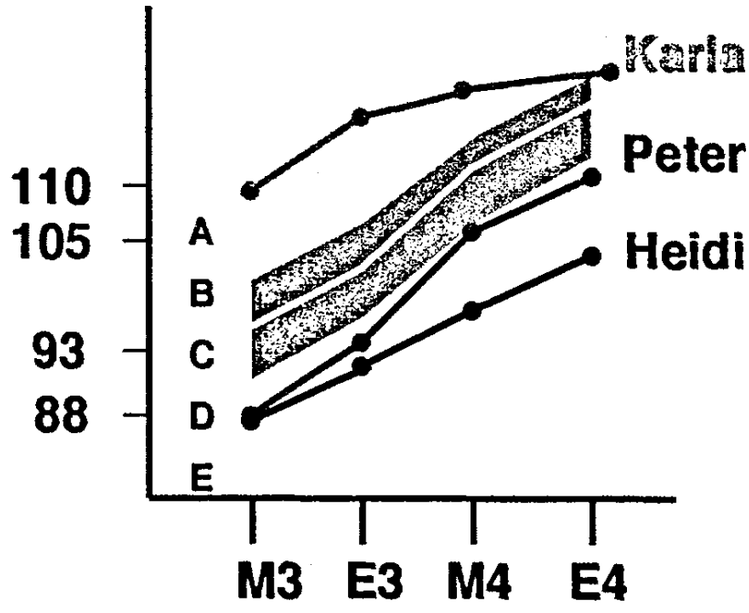
[8] Beispielsweise durch den Vergleich von Perzentilrängen oder durch die Berechnung von Z-Scores.

Abbildung 5: Verzeichneter Lernzuwachs




Wird der gemessene Lernzuwachs entsprechend dieser Einteilung abgebildet, so leuchtet die Antwort auf die oben gestellten Fragen nach dem Fortschritt sofort ein.

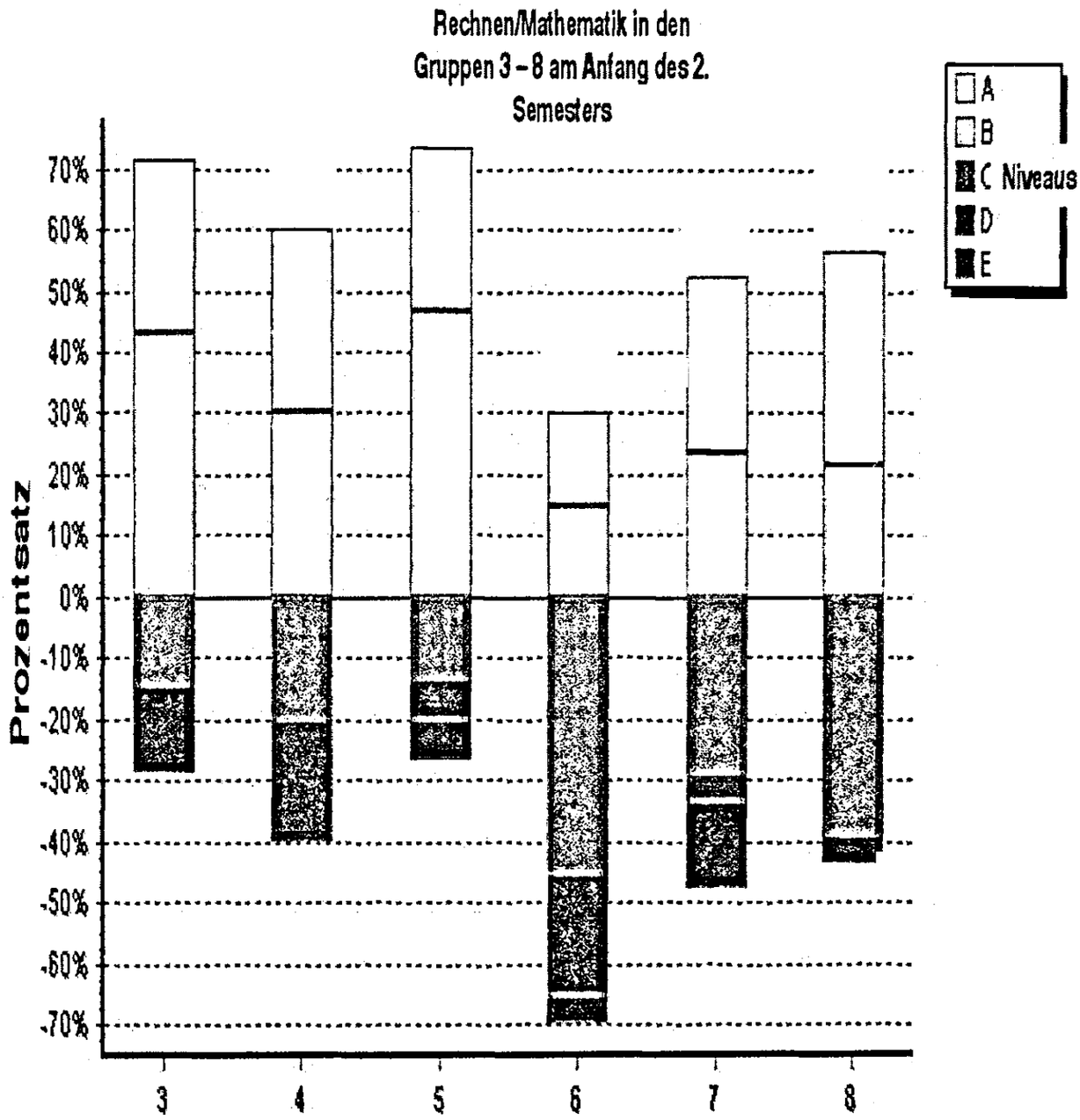
Abbildung 6: Lernzuwachs und Referenzdaten




Aus Abbildung 6 wird ersichtlich, dass sich Peter entsprechend seiner Leistungsgruppe entwickelt. Heidi fängt mit Peter auf gleicher Ebene an, bleibt dann aber zurück. Vielleicht wäre Heidis Testergebnis am Ende der 3. Klasse bereits Grund zur Aufmerksamkeit; das Resultat halbwegs der 4. Klasse hätte sicherlich zu Begleitmaßnahmen veranlassen sollen. Karla ist halbwegs der 3. Klasse eine überaus leistungsstarke Schülerin, doch zeigt sich später ein Leistungsabfall. Womöglich fühlt sie sich unterfordert und langweilt sich in der Schule. *Ohne zusätzliche Daten bleiben diese und andere Erklärungen jedoch spekulativ.* Ein Schülerbegleitsystem lässt sich mit einem Fieberthermometer vergleichen: Es misst zwar Fieber (Leistungsrückstand), bietet aber an und für sich keine Diagnose.

Die Ergebnisse des *CITO-Schülerbegleitsystems LVS* lassen sich auf Klassenebene aggregieren Hier werden die Testergebnisse unterschiedlicher Jahresgruppen nach den genannten fünf Leistungsniveaus gestaffelt. Der Null-Linie entspricht der Landesdurchschnitt. Oberhalb dieser Linie hat ein Viertel der Schüler/innen der Referenzgruppe ein B-Niveau und ein weiteres Viertel ein A-Niveau. Unterhalb der Null-Linie hat die andere Hälfte der Schüler/innen ein C, D oder E-Niveau, in der Verteilung 25%, 15% bzw. 10%.

Abbildung 7: Klassenvergleich und Landesdurchschnitt




Aus der Abbildung wird ersichtlich, dass die Klasse 7 der nationalen Staffelung entspricht, obgleich die E-Kategorie relativ stark vertreten ist. Die Klassen 3, 4 und 5 haben im Vergleich zum Landesdurchschnitt viele Schüler/innen der A- und B-Kategorie. Die Klasse 6 weicht stark ab: Nur 30% der Schüler/innen befinden sich über dem Landesdurchschnitt, 70% befinden sich darunter. Es kann sein, dass diese Klasse *rein zufällig* Schüler/innen einer niedrigen Leistungsstufe enthält; diese Hypothese kann man überprüfen, indem man die Ergebnisse vorheriger Jahre zu Rate zieht. Es kann auch sein, dass der getestete Lernbereich in der Periode unterbetont ist, und zwar zugunsten an-

derer Bereiche. Auch dies ist anhand von Testdaten zu überprüfen. Der Lernrückstand kann aber auch mit Erkrankung der Lehrperson oder mit derer pädagogisch-didaktischen Qualitäten zu tun haben. Einmal mehr gilt, *dass das System nur misst, aber nicht deutet.*

### *Evaluation*

Aus einer Untersuchung (Roeleveld u.a. 2002) zum positiven Effekt von Schülerbegleitsystemen gehen zwei wichtige Punkte hervor. Erstens, ein Schüler ist keine *tabula rasa* mehr bei einem Klassenübergang oder einem Lehrerwechsel. Resultate werden fast ausnahmslos weitergegeben. Durchaus bejaht wird auch die These, dass Schülerbegleitsysteme eine frühzeitige Diagnostik von Lernproblemen ermöglichen.

Unverkennbar ist auch, dass die Schulleitung durch Aggregation von Schülerdaten über solide Informationen verfügen kann, die beim Personalmanagement eingesetzt werden können. Eine Lehrperson, deren Klasse schlecht abschneidet, ohne dass ein Alibi vorhanden ist, bedarf der Professionalisierung. Bislang liegen keine Daten vor, inwieweit diese Informationen tatsächlich zu entsprechenden Zwecken genützt werden. Damit hat jedes der beiden rechten Kästchen in Abbildung 1 seine eigene Bewandtnis.

## 5. Fazit

Die direkte Auswirkung von Bildungsmonitoring auf nationaler/überregionaler Ebene ist unbekannt. *In den Niederlanden gibt es dazu keine Effektstudien.* Ein genereller Leistungsrückfall bzw. Leistungsanstieg ist kaum festzustellen, da wichtige Lerngebiete wie Muttersprache und Rechnen aus mehreren, z.T. stark verschiedenen Teilbereichen bestehen. Ein Fortschritt bzw. Rückgang in einem Teilbereich muss daher immer im Zusammenhang mit anderen Teilbereichen betrachtet werden (Onderwijsraad 2006, S. 33). Hinzu kommt, dass bestimmte Domänen mit der Zeit einen anderen Stellenwert bekommen können. So dürfte beispielsweise das schriftliche Rechnen infolge der Abundanz an Taschenrechnern in der Praxis weniger geübt werden, während die Lösung von Sachaufgaben an Bedeutung gewonnen hat.

Sobald jedoch in einer Lernstandserhebung die Leistung einer ‚organischen' Gruppe von Einzelschülern, lies: einer Klasse bzw. einer Schule, zum Bewertungspunkt auch für Lehrer oder Schulleiter wird, treten Mechanismen auf, die darauf ausgerichtet sind, die wirklichen Ergebnisse zu vertuschen bzw. zu verschönern. Es werden daraus durchaus *unerwünschte Rückwirkungen* entstehen. Van Ackeren (2005) beschreibt die taktischen Mechanismen zur Verbesserung erreichter Test- und Examensresultate im englischen Bildungssystem.

Alles in allem erscheint die *graphische* rechte Spur in Abbildung 1 – allerdings unter Hinzuziehung schulübergreifender Referenzwerte – auch als die *metaphorische* rechte Spur. Nur in dieser Weise können diejenigen, um die es geht – die Schüler/innen – unmittelbar Nutzen ziehen aus den Leistungsmessungen, zu denen sie ja selber am meisten beitragen.

## Literatur

Ackeren, Isabell van 2005: Vom Daten- zum Informationsreichtum? Erfahrungen mit standardisierten Vergleichstests in ausgewählten Nachbarländern. Pädagogik 57(5), S. 24-28

Bosker, R. J.; J. de Jong-Heeringa 2006: Leeropbrengsten van scholen. Groningen: GION, Instituut voor Onderzoek van het Onderwijs, Rijksuniversiteit Groningen

Inspectie van het Onderwijs 2006: Eindtoets in het basisonderwijs – Een onderzoek naar leerlingen die niet meedoen en/of niet meetellen. Zwolle: IvhO

Leutner, Detlev 2001: Pädagogisch-psychologische Diagnostik. In Detlef H. Rost: Handwörterbuch Pädagogische Psychologie, Weinheim: PVU, S. 521-529

Onderwijsraad 2006: Versteviging van kennis in het onderwijs. Den Haag: Onderwijsraad

Roeleveld, J.; M.E. Otter, H Blok. 2002: Leerlingvolgsystemen in de jaren 90: Secundaire analyses op gegevens uit PRIMA en IST. Amsterdam: SCO-Kohnstamm Instituut

*Jan van Weeren,* geb. 1946, Dr. phil.; wissenschaftlicher Mitarbeiter für Germanistik und angewandte Linguistik an der Universität Leiden; Leiter der Sprachenabteilung und der Abteilung Sekundarstufe II am CITO, dem niederländischen Institut für Testentwicklung; z.Zt. mit Innovationsprojekten für den höheren und berufsbildenden Unterricht beauftragt;
Anschrift: CITO, Postfach 1034, NL – 6801 MG Arnhem;
Email: jan.vanweeren@cito.nl