Springer

Dear Author

Here are the proofs of your article.

- You can submit your corrections **online,** via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
  Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL: http://dx.doi.org/[DOI]. If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to: http://www.springerlink.com.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.

Springer

To: Email:- lucy.pradeepa@crest-premedia.in
Fax :- +91 (20) 30580715
Address :- 101/B, Delta 1, Giga space IT Park
S. No. 198/1B, Viman Nagar
Pune 411014
Maharashtra, India

Re: Zeitschrift für Erziehungswissenschaft (10.1007/s11618-010-0147-2)
Aspects of accountability and assessment in the Netherlands

Anton Béguin
Melanie Ehren

Permission to publish

I have checked the proofs of my article and

☐ I have **no corrections**. The article is ready to be published without changes.
☐ I have **a few corrections**. I am enclosing the following pages:
☐ I have made **many corrections**. Enclosed is the **complete article.**

Metadata of the article that will be visualized online

Please note: Images will appear in color online but will be printed in black and white.

ArticleTitle		Aspects of accountability and assessment in the Netherlands
Article CopyRight - Year		VS Verlag für Sozialwissenschaften 2010
Journal Name		Zeitschrift für Erziehungswissenschaft
Corresponding Author	Family Name	**Béguin**
	Particle	
	Given Name	**Anton**
	Suffix	
	Organization	CITO
	Address	PO Box 1034, 6801 MG Arnhem, The Netherlands
	Email	anton.beguin@cito.nl
Author	Family Name	**Ehren**
	Particle	
	Given Name	**Melanie**
	Suffix	
	Organization	University of Twente
	Address	PO Box 217, 7500 AE Enschede, The Netherlands
	Email	m.c.m.ehren@utwente.nl
Schedule	Received	
	Revised	
	Accepted	

Abstract	This article describes aspects of test-based accountability in the Netherlands. It provides a description of the design of the Educational system in the Netherlands, it gives a short introduction to the role of the Dutch Inspectorate of Education in the accountability of schools and describes different assessments that are used as sources of information in the accountability system. For each assessment, the primary function in education and its role in the accountability system are discussed. Finally, the factors that can potentially influence the validity of the accountability indicators and the strong and weak points of the current system are identified and some directions are presented of potential developments of this system.
Zusammenfassung	In diesem Artikel werden Aspekte der testbasierten Rechenschaftslegung in den Niederlanden präsentiert. Zunächst werden das niederländische Bildungssystem und die Rolle der Bildungsinspektion beschrieben. Sodann werden verschiedene Typen der Überprüfung vorgestellt, die die Informationsbasis für das System der Rechenschaftslegung bilden. Für jede Überprüfungsform werden ihre wesentlichen Funktionen und die Rolle, die sie im System der Rechenschaftslegung spielen, diskutiert. Abschließend werden Faktoren vorgestellt, die die Validität der Indikatoren beeinträchtigen können, die den Rechenschaftssystemen zugrundeliegen; es werden die Stärken und Schwachpunkte des gegenwärtig in den Niederlanden angewandten Systems präsentiert und einige Überlegungen zu seiner Weiterentwicklung vorgestellt.

Keywords(seperated by –)	Accountability – Validity – Educational assessment – High-stakes testing
Schlüsselwörter	Validität der Rechenschaftslegung – Bildungsevaluation – high-stakes testing

Z Erziehungswiss
DOI 10.1007/s11618-010-0147-2

ZfE

Aspects of accountability and assessment in the Netherlands

Anton Béguin · Melanie Ehren

Abstract: This article describes aspects of test-based accountability in the Netherlands. It provides a description of the design of the Educational system in the Netherlands, it gives a short introduction to the role of the Dutch Inspectorate of Education in the accountability of schools and describes different assessments that are used as sources of information in the accountability system. For each assessment, the primary function in education and its role in the accountability system are discussed. Finally, the factors that can potentially influence the validity of the accountability indicators and the strong and weak points of the current system are identified and some directions are presented of potential developments of this system.

Keywords: Accountability · Validity · Educational assessment · High-stakes testing

Zusammenfassung: In diesem Artikel werden Aspekte der testbasierten Rechenschaftslegung in den Niederlanden präsentiert. Zunächst werden das niederländische Bildungssystem und die Rolle der Bildungsinspektion beschrieben. Sodann werden verschiedene Typen der Überprüfung vorgestellt, die die Informationsbasis für das System der Rechenschaftslegung bilden. Für jede Überprüfungsform werden ihre wesentlichen Funktionen und die Rolle, die sie im System der Rechenschaftslegung spielen, diskutiert. Abschließend werden Faktoren vorgestellt, die die Validität der Indikatoren beeinträchtigen können, die den Rechenschaftssystemen zugrundeliegen; es werden die Stärken und Schwachpunkte des gegenwärtig in den Niederlanden angewandten Systems präsentiert und einige Überlegungen zu seiner Weiterentwicklung vorgestellt.

Schlüsselwörter: Validität der Rechenschaftslegung · Bildungsevaluation · high-stakes testing

Dr. A. Béguin (✉)
CITO, PO Box 1034, 6801 MG Arnhem, The Netherlands
e-mail: anton.beguin@cito.nl

Dr. M. Ehren
University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands
e-mail: m.c.m.ehren@utwente.nl

Germany and many other nations are in the early stages of education reforms motivated in part by dissatisfaction with students' performance on PISA and TIMSS. These reforms vary, but many of them have as a central element the use of achievement tests to monitor the performance of the educational system and to hold educators accountable.

Reforms of this sort began decades ago in the United States, and test-based accountability has gradually become the cornerstone of U.S. education policy. The American experience holds lessons for the development of reform in other nations. This paper briefly describes the history of performance-based reform in the U.S., notes some key findings of research, and describes implications for the development of reforms in other nations.

1 Introduction

In recent years the results of tests and assessments receive increasing emphasis in accountability systems in some Western countries. This can partly be attributed to a relative low performance of their students on academic assessments when compared with students from certain Asian nations (Anderson 2005). Another cause is the notion that high-quality education is essential for economic development and innovation (Kok 2004). As a result, a number of countries have introduced strong accountability measures based on tests and assessments to try to ensure that public schools perform at the level necessary for economic supremacy. A well known example is the 'No Child Left Behind Act' in the United States of America that requires every state to develop standards, standardized tests and accountability systems. In the Netherlands, we also see a trend in which test results of schools and output indicators receive increasing attention by the Dutch Inspectorate of Education.

In this article a number of assessments and accountability indicators in primary and secondary education in the Netherlands will be described. The article starts with a description of the design of the educational system in the Netherlands and the use of tests as sources of information in the accountability system. For each assessment it is discussed what the function of this assessment is in education and how the results are applied in accountability. After a short general introduction to possible threads to validity of accountability systems, the strong and weak points with respect to the validity in the Dutch accountability system are identified.

2 The Dutch education system

In the Netherlands, students start primary education at the age of four. After finishing primary school around age twelve, students enter into secondary education. Secondary education in the Netherlands is highly selective; it is a tracked system in which students can choose between three school types:

- pre-vocational secondary education (VMBO): 4 year course;
- senior general secondary education (HAVO): 5 year course;
- pre-university education (VWO): 6 year course.

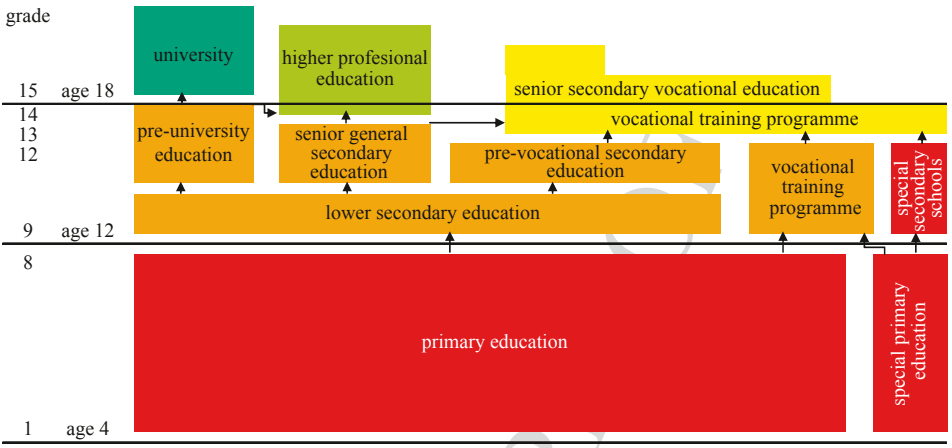

Fig. 1: The Dutch education system

After secondary education students go to a vocational training programme, higher professional education or university (Fig. 1).

At the different stages and different tracks of the educational system learning objectives are specified. The schools are autonomous in designing their own education as long as they focus on the objectives. National standardized examinations at the end of secondary education assess whether students meet the learning objects and qualify to graduate. At the end of primary education, schools are required to use relevant tests to advise students on the appropriate stream of secondary education. The examinations at the end of secondary education are compulsory, while placement tests and other tests are applied on a voluntary basis, although it is compulsory that some form of testing takes place.

3 Testing in the Netherlands

3.1 Tests in primary education

At the end of primary education schools are obliged to collect objective information about the most appropriate type of secondary education for the student. A vast majority of about 85% of the schools applies a test for this purpose called the '*Eindtoets Basisonderwijs*' (Van der Lubbe 2007; Resit 2009); other schools apply other tests and assessments. The *Eindtoets Basisonderwijs* contains a compulsory part of 200 multiple choice items on Dutch language, arithmetic and study skills and a voluntary part of 90 items on history, geography and science. Each year a new form of the test is constructed that contains all new items and of which the results are linked to the test of previous years. The *Eindtoets Basisonderwijs* and some of the other tests are based on the learning objectives that are specified for primary education. The primary function of the *Eindtoets Basisonderwijs* is to provide advice for the most suitable track of secondary education for a student. Aggregated over students the results can also be used for diagnostic purposes on school level.

In a similar way the Inspectorate of Education uses the results to construct an indicator of the effectiveness and for the accountability of the school. The average test result in the school is calculated and this average is corrected for social economic status and compared on a relative basis with the results of other similar schools. Schools that score below average for three consecutive years are identified as 'having a high risk of being weak'. These schools are scheduled for inspection visits in which the quality of educational processes is assessed.

A different, more formative way of assessment of progress of the student is provided by the '*Monitoring and Evaluation System*' of Resit (the Dutch Institute for Educational Measurement). This evaluation system is also applied by a large proportion of the schools. This system is a collection of formative tests on *language, arithmetic and mathematics* and *use of information* with two assessments in each grade. Results of students are evaluated on a vertical equated scale for each subject. In this way students' development can be monitored and based on the results and the diagnostic information it contains, the educational process can be adapted to suit these students' needs.

3.2 Examinations and certification in secondary education

At the end of secondary education students are to take a set of final examinations in a number of subjects within a profile that the student has chosen. The final examination is divided into two parts: a school examination and a national examination. Dutch language is a compulsory subject in the national examination in all types of secondary education. English language and some form of mathematics are compulsory elements in the national examination in pre-university and senior-general secondary education. Other compulsory subjects depend on the profiles (pre-university and senior-general secondary education) or type of vocational training the student has chosen. The elements to be tested in each examination are specified in the examination syllabus, approved by the Ministry of Education, Culture and Science. The syllabus also specifies the number and length of the tests that make up the national examination. After passing these examinations, students gain access to different forms of further education.

Schools are responsible for setting up the school examination. Every year schools are required to submit their own school examination syllabus to the Inspectorate showing which elements of the syllabus will be tested, when, and how marks are calculated, including the weight allocated to these tests and resit opportunities.

Generally speaking, a school examination consists of two or more tests per subject. These may be oral, practical or written. The school examinations are produced by the schools themselves or by test institutes. The school examinations are marked by the pupils' own teacher. There are also practical assignments for which no marks are given, only an acknowledgement that the examinee has completed them properly. The school examination must be completed and the results submitted to the Inspectorate before the national examinations start.

The national examination consists of tests with open or multiple-choice questions and, in some cases, a practical component. For some subjects, there is only a school examination. The national examination can be sat at three sessions during the school year—in May, June and August. All examinees sit the examination in May. The June and August

sessions are for pupils doing resits, or who were unable to sit the examination in May. The national examinations, which are the responsibility of the Dutch Ministry of Education, are produced by RESIT. The examinations are marked by the pupils´ own teacher and checked by a teacher from another school.

The head teacher is responsible for determining the examinees' final marks. The final mark in each subject is the average of the mark for the school examination and the mark for the national examination. To obtain a leaving certificate, an examinee must have scored passing marks in a specified number of subjects. For subjects with only a school examination, the mark obtained is the final mark (rounded off).

Marks are awarded on a scale ranging from 1 (very poor) to 10 (excellent). A six is a pass. It is clear that examinees with a final mark of six or higher for every subject have passed their school-leaving examination. However, even if they get a lower mark in some subjects, they can still be awarded an overall passing mark. Successful examinees receive a certificate and a transcript listing the marks scored in the school examination, the marks scored in the national examination, the final marks for each subject and the outcome of the school-leaving examination. Examinees who fail the examination after doing resits may decide to repeat the final year, go to an institute for adult secondary general education, or prepare for the state examination.

Next to the examinations a growing number of schools, currently approximately 40 percent apply a version of the CITO '*Monitoring and Evaluation System'* that is suitable for secondary education. There are tests in *Dutch* and *English Language*, *mathematics* and *use of information* and there are four test administrations divided over the first three years of secondary education. Schools are not obliged to use this system and can choose to participate with only a subset subjects and a reduced number of administrations. Again students are evaluated on a vertical equated scale for each subject. In this way students development can be monitored and the educational process can be adapted to students needs.

3.3 Entrance test in teacher training programs

Recently compulsory computer adaptive test were introduced to evaluate the proficiency level in arithmetic and spelling of students who wanted to enrol in a teacher training program. Although arithmetic and spelling are basis skills that are mostly taught in primary education, the impression existed that quite some students lost part of their proficiency during secondary education. For example the use of a calculator in secondary education prevents students to practice doing calculations by heart. The tests are presented in an adaptive form (see e.g. Van der Linden and Glas 2000). This means that harder items are selected if the student performed well on the previous items and that easier items are selected if the students made errors in the previous items. This procedure has the advantage that the proficiency of the student can be estimated efficiently and that the student gets items at the right level (challenging but not too hard). Students have to pass this test before the end of the first year otherwise they would be expelled from the training program. The standard of the tests were set at the proficiency level of a student at the 80th percentile rank at the end of primary education. Especially at the first administrations a very large proportion of the students failed this test. This was one of the causes to decide

to put more emphasis on maintaining basic skills in mathematics and Dutch language in secondary education.

3.4 National assessments of the educational system

Next to the test focussing on the evaluation of performance of the individual student, there are some studies to evaluate the educational system. The study Cool$^{5\text{-}18}$ (Driessen et al. 2009; Zijsling et al. 2009) collects longitudinal data of students in the age range between 5 and 18 years old. With three year cycles, background variables, educational position and both cognitive (mathematics, Dutch language, English language and citizenship) and non-cognitive skills (Big Five personality traits) are collected. Performance on the cognitive skills, mathematics and Dutch (and English) language is based on the results from the Cito '*Monitoring and Evaluation System'* for students in primary education (Rosier 2001; Sluijter and Rosier 2002). The tests that are used in the study in secondary education and vocational training are adapted versions of tests from the '*Monitoring and Evaluation System'* for secondary education. In pre-university and senior general education the final measurement of the mathematics and language skills is based on the examinations in secondary education.

Next to this longitudinal study there is a national assessment called *Periodieke Peiling van het OnderwijsNiveau* (PPON, eng= periodic assessment of educational level) in primary education that evaluates different subjects in detail but with irregular intervals (Van der Schoot 2008). Tailor made tests are used that evaluate relevant aspects of the subject. The test are assigned in an incomplete design, and for all the different aspects standards are set based on expert content judgements. For this judgments a bookmark procedure is applied. To provide an idea about the level of detail, the last assessment of mathematics was evaluated on 22 different aspects**.** Information from this study is used by content experts and decision makers to monitor the educational system and to evaluate if changes to the current curriculum are necessary.

4 Use of tests in accountability

In general, achievement tests can be important sources of information for evaluating school performance. In many accountability systems all over the world, schools or educators receive rewards, sanctions or both on the basis of students' test scores. Other output parameters may be the magnitude of student absenteeism, student drop-out rates etc. In most accountability systems worldwide, test scores are used to hold schools to account, but they can also be used to hold individual students, teachers or even districts or local communities to account (Ehren, submitted). Students may for example be confronted with sanctions such as having to repeat a class or having to follow compulsory summer classes when they fail to achieve nationally defined targets on test scores. Teachers may receive bonuses when meeting certain output targets, or be replaced or withheld from permanent contracts when failing to meet the targets. Districts or local communities may receive fines or bonuses or may loose government over their schools as a result of state

take-over when too many schools within the district function below some performance target.

In the Netherlands test scores are also an important part of educational accountability. The Inspectorate is legally complied to assess educational quality in schools (also including educational processes in the school such as a safe learning environment), but tests are expected to predict the quality of these educational processes in the school. The Inspectorate of Education uses test scores (next to annual reports and complaints and other signals on the functioning of schools) as part of an early warning system to detect schools with low educational quality. Schools that have declining or low test scores during a period of three years are considered to be failing or at risk of failing.

In primary education two different indicators of performance are used if there are test results available. The first indicator is based on the average test score in a school. This average is compared to the general mean score of a group of comparable schools. For this a total of 7 groups are defined based on the characteristics of student population and taking into account the level of education of the parents and the ethnic background of the students. The deviation of the average score to the relevant group mean is used to identify both over- and underperformance in a particular year. To evaluate performance of the school the inspectorate uses intervals of three years of test scores to increase the reliability of the results.

The use of tests by the Inspectorate of Education may have a number of difficulties. The first problem occurs when using test scores of schools with a limited number of students. In these schools the average score can be unstable over the years due to fluctuations in the performance level of the student population. Another complicating factor in this type of indicator is that a relative comparison is used. For smaller schools this leads to a higher probability of false positive and false negative evaluations (Verhelst et al. 2001). The second indicator corrects for this last problem and is based on a hierarchical regression model (Goldstein 2003) with a correction for the social economic composition of the student population in the school. The estimated school effects are used as indicator for over- and underperforming schools. Again reliability can be increased by evaluating over multiple years.

In secondary education an output indicator of a school is based on the average results on the national examination. The average is calculated over all students and all subjects. Next to the average result the difference between the national and the school based examinations is evaluated as a check on the standard used in the school based examinations. This last indicator was introduced after research showed that in some schools the average mark on the school based examinations were substantially higher than on the national examinations (De Lange and Dronkers 2007).

In both primary and secondary schools that are identified as failing, school inspectors will meet with the governmental boards of the schools to talk about causes of failure in schools; only when these meetings do not clarify the causes of failure in schools will they visit these schools to investigate the cause of risks. Schools that show sufficient test scores will not be visited, apart from a minor check up once every four years. This type of 'risk-based school inspections' was implemented in 2007/2008 to decrease the administrative load on schools and to enable a more efficient use of inspection capacity.

Compared to accountability systems in other countries and states the schools in the Netherlands are confronted with relatively low stakes in having to account for potentially low test scores. The Inspectorate only identifies these schools for increased monitoring and schools are obliged to write an improvement plan. There are no financial or legal sanctions, and schools do not have to meet specific targets related to test scores.

Compared to other accountability systems the Dutch system has a primary focus on the schools. Indicators describe output of schools and average performance targets for the entire pupil population in a school. Internationally also students, teachers, districts and local authorities are identified as potential stakeholders. For example in some countries the students have to meet (nationally defined) targets including minimum test scores with respect to math, reading and writing. In these countries sanctions and rewards may occur. Students for example may by central legislation be confronted with sanctions such as having to repeat a class or having to follow compulsory summer classes. Rewards may include scholarships for further education. Internationally also examples occur were teachers are assessed individually by school inspectors to determine whether they perform according to the indicators and performance targets in the accountability systems. In some countries, test scores of groups of students are used to assess performance of teachers. In the U.S., some states hold their districts accountable for the average performance of the schools within that district. Districts may receive fines or bonuses or may loose government over their schools as a result of state take-over when too many schools within the district function below some performance target. Accountability in the Netherlands does not have a direct focus on these other potential stakeholders. Students, teachers or local communities are not part of the educational accountability system, although the test results of a student will often have a direct impact on the education or educational career. It is also possible that the school government holds teachers to account for the results of the indicators produced by the inspectorate. By the same token, local politicians could focus on the results of schools in their jurisdiction for their political aims. In this way local initiatives could lead to local accountability systems that are derived from the school based accountability introduced by the Inspectorate. Figure 2 summarizes the Dutch educational accountability system.

To evaluate the validity of the Dutch accountability system, the aspects of accountability systems that are considered to be relevant for validity of the system are first identified. For this purpose, an overview of the literature on validity of accountability systems is presented. After this overview, the Dutch system is evaluated based on the relevant validity aspects.

4.1 Validity issues of accountability indicators

In general, validity addresses the concern of whether we are measuring what we intend to measure (Hill 2001). In the case of accountability systems, validity involves the inferences people are drawing from the results and whether these are consistent with actual results. When test scores are used to measure output of schools, validity refers to the degree to which evidence and theory support the interpretations of test scores for this purpose (Kane 2002; Sireci and Parker 2006). The test itself is not validated, and test

		Stakeholders			
		Students	Teachers	Schools	Districts/local communities
Standards	Output				
	Educational processes				
Measurement methods	Tests				
	Inspection visits				
	Results of internal evaluations				
	Desk research				
Targets	Targets on educational processes				
	Targets on output (e.g. cut-score)				
Stakes	Rewards				
	Sanctions				
	Interventions				

Fig. 2: Summary of Dutch educational accountability system

scores per se are not validated. It is the interpretation determined by the proposed use that is validated.

Several authors argue a more comprehensive view on validity when evaluating accountability systems. Marion and Gong (2003) for example state that the evaluation of accountability system validity must also specify how and why the system is intended to work in order to improve student learning and system capacity. Validation should include an evaluation of the consequences of uses and interpretations of the assessments, including both negative and positive consequences as well as the intended and probable unintended consequences (Lane et al. 1998). If, for example, tests are to help improve system performance, there should be information provided to document that test results are modifiable by quality instruction and student effort (Baker et al. 2002). Accountability systems are considered to be invalid when stakeholders in the system have no opportunity to control (some of) the components of the evaluation and when the consequences for stakeholders are not coordinated to support system goals or when stakes do not apply to adults and students symmetrically (Baker et al. 2002). In general incentives and sanctions that push in opposite directions for the professionals in education and for students can be counterproductive. They need to be consistent with each other and with the goals of the system. Probably, a situation with high to medium-high stakes for the individual student and not more than medium-high stakes for the aggregated indicator works best. In such a situation the test results service their primary purpose and provide valuable information at the individual level to both students and teachers, while the aggregated results can serve as unobtrusive indicators of outcome of education. But if the stakes on the aggregated indicator becomes too high due to pressure on the actors that are evaluated (e.g. the schools) this can ultimately lead to strategic behaviour on the test and consequently to invalidation of the indicator. Figlio and Getzler (2002) and Cullen and Reback (2006)

for example describe how schools at risk of failing improve their state-assigned grade or classification by taking their poorest performing students out of the testing pool. This type of intended strategic behavior is usually referred to as 'reshaping the test pool'. Schools may do so by classifying (regular) students into the 'special education' or 'limited English proficient' categories that may be exempted from taking the test (Jacob 2005) or they may retain low-scoring students in grades below those in which the test is administered, allow an increase in absences on test days, or grant exemptions from testing by parents and increasing dropout rates. According to a study by Jacob and Levitt (2003) in 4–5% of the classrooms cheating occurs. Teachers may do so by prompting students with the right answer during a test, providing the actual test items in advance, providing hints during test administration, making changes to answer sheets before scoring or leaving pertinent materials in view during the testing session.

4.2 Evaluating the validity of the use of tests in the Dutch accountability system

The above aspects are also relevant for the evaluation of the validity of the Dutch accountability system. The first question is: Does the system actually give an indication of educational quality, or of the potential risk of a lack of quality? Based on their primary function as summative or formative assessment, it can be assumed that the tests and assessments will validly measure the proficiency of the individual student. This will not necessarily be the case for the aggregated results that are used as indicator for the quality of education at the school level. Two aspects are important. Firstly, the indicator can be a misrepresentation of educational quality if parts of the curriculum are not represented in the tests at hand. For example the tests at the end of primary education do not contain active writing skills. Although one can argue that not all relevant aspects are represented, a relatively low score on the tests could still validly be interpreted as a potential lack of quality. Secondly, one can argue that valid measurement of the individual ask for different characteristics and content of the tests than measurement of schools. Potentially some items in the tests are not valid for the assessment of quality of the school. For example, this will be the case if items to some extent measure aptitude instead of attainment. In that case the performance on these items is not so much influenced by the school, but more by external factors like intelligence or SES (Popham 1999). In the Netherlands this is at least to some extent the case in the *Eindtoets Basisonderwijs*, the test at the end of primary education. This test aims at predicting success in secondary education. It is known that the domain of reading comprehension in this test is less influenced by the school than most of the other content domains. Based on this one could argue that an indicator of school quality could be constructed in which the reading comprehension items are removed.

Another issue in the validity of the Dutch accountability system is the use of relative standards with respect to outcome indicators used to assess schools. For example, the average results in a school are compared with results from different schools that have a similar population. In this manner a dependency in the evaluation exists between the results of an individual school and all of the other schools in this comparison. The relative nature of this comparison can lead to a thread of validity if the idea takes post that this comparison is unfair. This could be the case if differences in the situation of the school

(e.g. a more difficult student population) do not occur in the other schools with which the comparison is made.

Finally, the condition in which the tests are administered is a crucial factor to the validity of the accountability indicators. As described above the school based indicators in the Netherlands are based on tests and assessments that serve multiple purposes. In such occasions the validity of the system of indicators at a school or system level are both influenced by the stakes of the test results and by the stakes of the indicators based on these test results. The test results do often have an effect on the educational career of the students and as such, especially the summative assessments like the examinations in secondary education, are high stakes to the students. Looking at the indicators based on the test results, the stakes in the Dutch accountability systems mainly affect the schools, while other actors, such as teachers are not held directly accountable for the performance of the school and the average test scores obtained by the school. There may be concerns about the degree to which they will contribute to making sure that students achieve high test scores. In practise this will often be no serious issue, since the primary function of the test should provide an incentive for the student to perform well. From this perspective the total system seems to be relatively balanced.

References

Anderson, J. A. (2005). *Accountability in Education.* Paris: Unesco.

Baker, E. L., Linn, R. L., Herman, J. L., & Koretz, D. M. (2002). Standards for educational accountability systems. *The CRESST Line, Winter,* 1–4.

AQ4 Cito (2009). *Handleiding Eindtoets basisonderwijs* (Manual). Arnhem: Cito.

Cullen, J. B., & Reback, R. (2006). *Tinkering toward accolades: school gaming under a performance accountability system.* Cambridge: National Bureau of Economic Research. Working paper 12286. http://www.nber.org/papers/w12286.

De Lange, M., & Dronkers, J. (2007). *Hoe gelijkwaardig blijft het examen tussen scholen? Discrepantie tussen de cijfers voor het schoolonderzoek en het centraal examen in het voortgezet onderwijs tussen 1998 en 2005.* [The equivalence between schools of the Dutch secondary final examination. Discrepancies between the grading of the central and school part of the final examinations of secondary education between 1998 and 2005]. European University Institute working paper EUI SPS 2007/3. Florence: EUI.

Driessen, G., Mulder, L., Ledoux, G., Roeleveld, J., Van der Veen, I. (2009). *Cohortonderzoek Cool5-18: technisch rapport basisonderwijs, eerste meting 2007/08.* Longitudinal study Cool5-18: technical report, first measurement primary education 2007/2008. Den Haag: NWO.

AQ5 Ehren, M. C. M. (submitted). Effectiveness of strong accountability systems in education. *Educational Policy.*

Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: gaming the system.* (Working Paper). Cambridge: National Bureau of Economic Research.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.

Hill, R. (2001). *Issues related to the reliability of school accountability scores.* Report on the reliability lecture from the 2000 Annual Edward F. Reidy Interactive Lecture Series Dover: National Center for the Improvement of Educational Assessment, Inc.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*(5–6), 761–796.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, (August),* 843–877.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kok, W. (2004). *Facing the Challenge. The Lisbon strategy for growth and employment.* Report form the High Level Group chaired by Wim Kok. Office for Official Publications of the European Communities.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 24–28.

Marion, S., & Gong, B. (2003). Evaluating the validity of state accountability systems. The 2003 Reidy Interactive Lecture Series. http://www.nciea.org/publications/RILS2003_BGSM03.pdf. Zugegriffen: 18 June 2009

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice, 18*(3): 13–17.

Rosier, W. (2001). *Computerprogramma leerlingvolgsysteem. Versie 3.0.* Computer program Pupil Monitoring System. Arnhem: Citogroep.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, *25*(3), 27–34.

Sluijter, C., & Rosier, W. (2002). Goede toetsen maken is een kunst. Toetsen en leerling volgsystemen. The art of making good tests. *Jeugd in School en Wereld, 87*(4), 12–14.

Van der Linden, W. J., & Glas, C. A. W. (Eds). (2000). *Computerized adaptive testing: Theory and practice*. Boston: Kluwer-Nijhoff Publishing.

van der Lubbe, M. (2007). *The end of primary school test (better known as Citotest)*. Paper presented at the 33 annual conference of the International Association for Educational Assessment. September 16–21, Baku, Azerbijan.

van der Schoot, F. (2008). *Onderwijs op peil? Een samenvattend overzicht van 20 jaar PPON*. Education up to the standard? A summary of 20 years of national assessment. Arnhem: Cito.

Verhelst, N., Staphorsius, G., & Kleinjes, F. (2001). Scholen langs de meetlat. Measurement of schools. *De psycholoog, 36*(12), 658–664.

Zijsling, D., Keuning, J., Kuyper, H., Hemker, B., & Van Batenburg, T. (2009). *Cohortonderzoek Cool5-18: Technisch rapport voortgezet onderwijs, eerste meting 2007/08.* Longitudinal study Cool5-18: technical report, first measurement secondary education 2007/08. Den Haag: NWO.

Questions to the Author(s)

AQ1. Please provide a German version of the article title.

AQ2. Figure 1 was not cited in the text. Please check if the citation has been inserted at the correct place.

AQ3. “Resit 2009” is cited in the text but is not given in the reference list. Please provide a full reference or delete the citation.

AQ4. “Cito (2009)” is not cited in the text. Please provide the citation or delete the entry from the reference list.

AQ5. Please update the reference “Ehren (Submitted)”.