NATIONAL EXAMINATIONS IN THE NETHERLANDS: STANDARD-SETTING PROCEDURES AND THE EFFECTS OF INNOVATIONS

Anton Béguin, Ed Kremers & René Alberts
Cito, National Institute for Educational Measurement

Paper presented at the IAEA Conference Cambridge, September 2008

(e-mail: anton.beguin@cito.nl; ed.kremers@cito.nl; rene.alberts@cito.nl)

Abstract
The Dutch secondary education system is divided into a number of pre-vocational education levels and two pre-university education levels. To complete one of these tracks successfully, pupils have to pass several examinations. These examinations are a mixture of national and school-based assessments in a number of subjects. This paper will address the specific goals and characteristics of the national examinations. Emphasis will be given to intra and inter-year comparability of examination results. Within a year, a national examination is comparable among all pupils, regardless of the school they attend. Inter-year comparability is obtained by using linking procedures. However, there are no common items among examinations of different years. For a large proportion of examinations, the linking is thus based on a statistical-result comparison among the population of two consecutive years, under the assumption of random equivalent groups. Under this assumption, the percentile ranks of cut-off scores on the examinations in previous years can be used to set the cut-off scores of the new examinations.

Alternatively, for some examinations a non-equivalent group-linking procedure is applied. A number of pilot studies have recently been set up to investigate possible changes to the national examination system aimed at more flexibility for schools and pupils. A key issue is that the intra and inter-year results comparability must remain intact. Changes to the national examination system may involve changes in time, place, content and type of examination. Recent developments mostly concern changes in time, some changes towards computer-based examinations and, to some extent, changes in content and standards. The details of these innovations and their impact on the appropriateness of traditional standard setting methods will also be discussed in this paper.

1. Introduction

In the Netherlands, secondary education traditionally ends with examinations. These examinations are a combination of national and school-based assessments on a number of subjects. This paper focuses on the national examinations. An important aspect of these examinations remains how to guarantee that intra and inter-year school results remain comparable. In order to ensure this, different standard setting methods are being used.

A number of pilot studies have recently been set up to investigate possible changes to the national examination system. These studies explore whether the examinations can be conducted in such a way that they are more flexible for both schools and pupils, e.g. more date options instead of the fixed dates currently set for all examinees. What impact will flexible examination dates have on, for example, the comparability of examination results? Other new developments in the examination system are the introduction of computer-based examinations and the construction of new examination formats, e.g. combined theoretical and practical examinations. Under these new developments, we must continue to ask ourselves if the 'traditional' standard setting methods are still applicable; or whether we need to develop new ones.

This paper starts with a short description of the Dutch education system, followed by an overview of the examination system and the difference between national and school-based examinations. We then discuss the construction and conduction of the national examinations. This is followed by a description of recent innovations in the Dutch examination system. Finally, we take a closer look at the current standard setting methods, before concluding the paper with an observation about whether these methods are also applicable to the innovations in the examination system.

2. The Dutch education system

The Dutch secondary education system is highly selective; it is a tracked system. After finishing primary school around age 12, pupils can choose between the following 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.

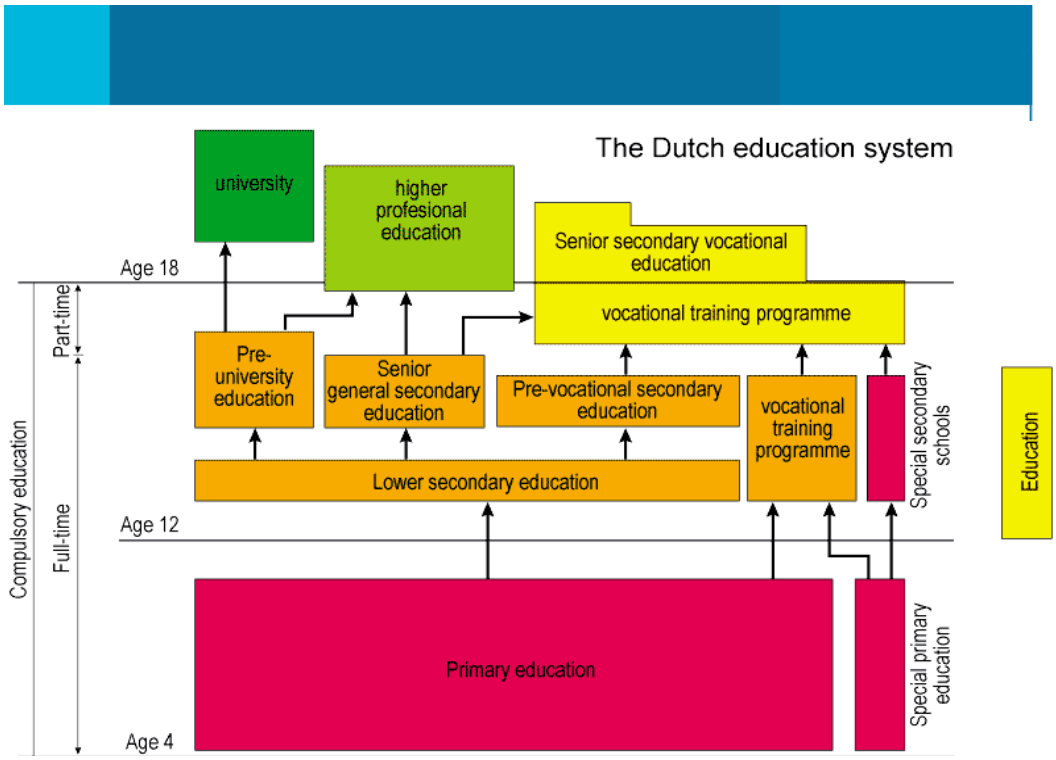


Figure 1. The Dutch education system

Interestingly, all proposals of the last forty years to make the system more comprehensive failed to meet sufficient approval. All secondary school studies end with examinations. After passing these examinations, pupils gain access to different forms of further education.

3. Overview of the examination system

3.1 Certification

The final examination is divided into two parts: a school examination and a national examination. 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.

The school examination
Schools are responsible for setting 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
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 Cito (Dutch National Institute for Educational Measurement). The examinations are marked by the pupils´ own teacher and checked by a teacher from another school.

School-leaving certificate
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.

Three national examinations
As we have seen, the Dutch secondary school system has three school types (see figure 1). Each school type ends with an examination that comprises a state and a school examination. We can distinguish examinations for pupils of the following ages:
- age 16: pre-vocational secondary education (VMBO: three sublevels);
- age 17: senior general secondary education (HAVO);
- age 18: pre-university education (VWO).

3.2 Test construction and validation

Participants
A number of bodies are involved in the examination process. First, there is **CEVO**, the Dutch National Examination Board, which represents the Ministry of Education. CEVO carries the overall responsibility for the examinations, including the standard-setting procedure. **Cito** acts as the examination contractor. The examination construction is done by a **Cito subject matter specialist** and a construction group composed of subject matter specialists (teachers). The validation of the examination is done by **validation groups** of subject specialists and higher education representatives (CEVO).

Construction and validation
Cito is responsible for constructing the examination papers. The Cito subject matter specialist is responsible for the construction process, i.e. producing questions and scoring schemes. Cito subject specialists are former teachers who have been trained in item and test construction and statistics. An important part of the job is done is by the construction group, a team of three teachers who write the examination questions. They still work as teachers, but spend on average one day a week on constructing questions. In our view, the involvement of active teachers is important, as they are familiar with the content of the subject and pupil ability levels. The construction group operates under the responsibility of the Cito subject matter specialist.

The CEVO validation group monitors the construction process. Once the group has accepted the proposed examination paper, it takes over responsibility for it.

3.3 Overview of the examination procedures

Below is a list of the eight phases that are normally followed for the conduction of examinations:

1. Examinations administered at schools (three-week period in May/June with resits in June and August).
2. School monitoring by the Inspectorate.
3. Score samples are sent to Cito.
4. Information about the quality reaches CEVO and Cito through teachers, pupils and schools.
5. The validation group, Cito and CEVO set the standards (cut-off scores).
6. The transformation tables are sent to the schools in mid June (CEVO, Cito).
7. The schools submit the results.
8. Evaluation (Cito, teachers, validation group, CEVO).

4. Recent innovations

4.1 The use of computers

National examinations can be classified along a continuum from completely paper-based examinations to completely computer-based examinations. Currently, most examinations are still paper-based, but the number of (partially) computer-based examinations is on the rise. First, there are the so-called **IMEX** examinations, which stand for **I**CT and **M**ultimedia in **EX**ams. Under this form of examination, pupils receive their questions on paper and the computer is used as a tool to answer some of the questions (video, subject-specific software). The answers themselves are written on paper.

Then there are fully computer based tests (**CBT**s) in which all questions are asked and answered using a computer, with nothing on paper. The ultimate aim is to convert the current IMEX examinations into full-fledged CBT examinations.

4.2 Combined theoretical and practical examinations

Another innovation involves the introduction of combined theoretical and practical examinations in pre-vocational secondary education (Van Hest & Zeelenberg, 2008). Although only recently introduced, this innovation has already proven a success. We would like to limit ourselves here to a short description of this form of examination. An important feature is the natural and integrated alternation of theoretical and practical tests, which highly motivates pupils. The total duration of the examinations can be up to 16 hours, spread over several days. While the theoretical examinations are all fully computer-based (CBT), the practical examinations can be either CBTs (e.g. using CAD-software: Computer Aided Design) or in the form of ‘real life’ practical skills (e.g. building a wall).

4.3 Flexible examinations

Recently, a number of pilot studies have been set up to investigate possible changes to the national examination system aimed at more flexibility for schools and pupils. Why flexible examinations? There are two important reasons. The first is to meet the growing demand for a more modern, pupil-tailored examination system.

With flexible examinations, it is possible to:
- link up with multimedia learning techniques;

- quickly organise resits for pupils who failed their examinations;
- facilitate early entry into higher education for high performing pupils.

The second reason is to reduce the organisational load for schools. Small groups of pupils can sit their examinations at different times instead of the whole-year group at the same time.

We distinguish three kinds of flexibility: flexibility in time, place and format (Kuhlemeier, Hermans & Kremers, 2004). We are currently experimenting with examinations that are flexible in time, and examinations that are flexible in time and place. We have not yet experimented with format flexibilisation.

Flexibility in time
A five-year written examination pilot project has been set up involving 12 schools, 800 pupils and 25 subjects, on all abovementioned school levels.

Some key-elements of this pilot project are:
- Three examination sessions in January, May and August;
- Flexible cycle runs from January to January;
- Two resits for each subject during each cycle;
- Administration and scoring by external, trained examiners.

The experiment started in 2007 will be thoroughly evaluated. Does flexibility produce the assumed benefits for schools and pupils? What are the costs?

Flexibility in time and place
A big CBT pilot project has been set up at pre-vocational level, involving 400 schools, 15.000 pupils and 12 subjects.

Some key-elements of this pilot project are:
- Schools get nine examinations of a similar difficulty level for each subject;
- Three out of nine examinations are earmarked for resits;
- Examinations are administered and marked by schools during a three-month period;
- Large number of automatically scored items;
- Schools organise resits.

This project started some years ago and the number of participating schools has been constantly increasing. In the past few years, the pilot has been extensively evaluated. The results of these evaluations are so encouraging that full implementation of this form of flexibilisation will take place within a year or two.

5. Standard-setting procedures

5.1 History

For many years, until 2000, the examination construction procedures and the definition of cut-off scores have remained largely unchanged.

During the 1990s, equivalence became a political issue. There had been some doubt in Parliament as to whether the rise in the number of pupils opting for higher forms of education may have been facilitated by a drop in standards. Doubts were also expressed in relation to the innovations introduced for lower secondary education, particularly regarding whether these innovations might lead to a performance level decrease.

In the 1990s, in response to these concerns over standards, studies were conducted which demonstrated the necessity and feasibility of using equating procedures. Acting on these outcomes, the State Secretary

of Education and Science provided funds to introduce and maintain equating as a standard procedure for a number of national examinations.

Another unpublished study of the Inspectorate revealed huge differences over the years between the results of school examinations and national examinations. Therefore, strict guidelines were prepared for the standard setting decision-making procedure.

5.2 Framework

We can roughly identify three steps in the standard-setting decision-making process. The first step involves the recommendation given by Cito subject matter experts based on an analysis of the examination data. A Cito subject matter expert submits a recommendation discussed during a meeting of CEVO's standard-setting group.

The second step consists of the recommendation submitted by the standard-setting group to CEVO's executive committee. After a content-based assessment of the examination and the reactions and comments that have been received, CEVO's subject section decides if Cito's recommendations need to be modified. The standard-setting recommendation is then submitted to CEVO's executive committee. In the third step, CEVO's executive committee surveys and assesses all recommendations, taking into account, among other things, the level of consistency among subjects and school types before setting the definitive standard.

The advice given by a Cito subject matter expert is mainly based on an interpretation of the analyses of the examination data. This is primarily technical advice, in the sense that, based on the collected examination data, a type of validation is being proposed that can be used to set similar requirements to examinees over years and time spans. The standard used as the starting point, is set by CEVO. For most subjects, CEVO has done this by selecting a reference examination for each subject using a reference standard. For new examinations, a Cito examination expert generally recommends a standard that is equivalent to the reference standard of the reference examination.

The data collection and subsequent standard-setting procedures are not identical for all subjects. For some examinations, more and different data are available to support the standard-setting process. For examinations where data collection is less complete, a number of assumptions have to be made. The weight of Cito s recommendation depends on the correctness of these assumptions. However, given the point of departure, the style of argumentation is transparent. Cito's recommendations have no 'absolute' pretensions; but if the assumptions made explicit are correct, then the result should also be correct. We will further elaborate on this below.

5.3 Procedures

The standard-setting procedure can be roughly divided into the following four variants:

1. Standard-setting based on systematically collected qualitative expert opinions. This method is used for examinations with a limited number of examinees with potentially strong performance level variations from year to year.
2. The standard data collection procedure in place for (almost) all examinations. This involves the collection of partial scores of each question of around 2000 random examinees;
3. For some examinations the standard procedure is complemented with additional data collection of the examination questions, combined with reference items.
4. Finally, there are "experimental" examinations, generally taken by a relatively small group of examinees that tend to overlap with regular examinations, so that the standards of both examinations are related to each other. Based on the overlap between both examinations, additional analyses are conducted.

Note 1) Standard setting procedures based on assessments by experts can be divided into two situations: with or without a reference examination and reference standard. In both situations, the experts need to have a good understanding of the final examination programme. They must be up to date with what is reasonably obtainable within the allotted study load.

If a reference examination is available, the experts will then evaluate the difficulty level of each single item of a collection of items. This collection also includes a large number of reference items. The preparation of the recommendation for the new examination is based on the estimated difference in difficulty.

If no reference examination is available, they will follow the Angoff procedure. Under this procedure, the experts will set, for each item, the score obtained by a hypothetical borderline examinee. The cut-off score is the sum of these scores.

Note 2) The standard procedure is used for the majority of the examinations. The recommendations are based on the data produced by random scores of candidates after a first correction round.

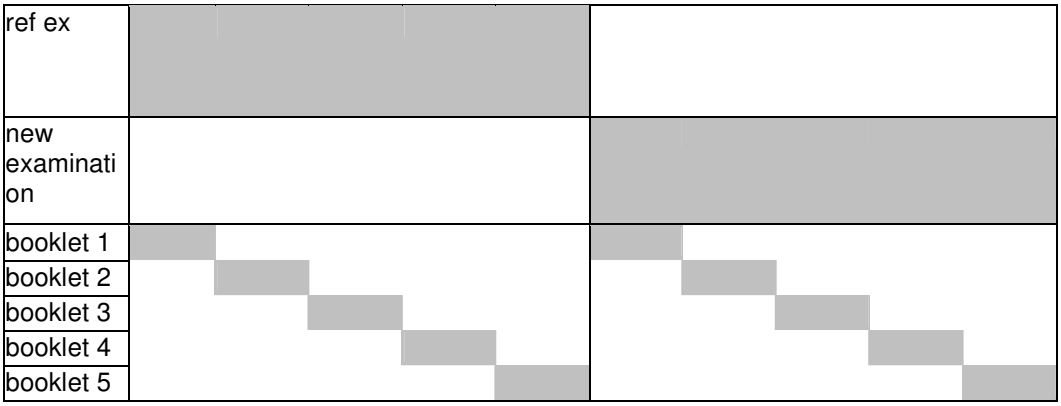
Cito's recommendations assume that populations have similar performance levels from year to year. Differences between the average score and the reference examination are thus considered, because of a difference between the difficulty level of the reference examination. Under this assumption, the percentile ranks of cut-off scores of the reference examination of a previous year can be used to set the cut-off score of the new examination.

The data design:

	Items 1 to 50 of the reference exam	Items 51 to 100 of the new exam
Sample students of the reference population N = 2000		
Sample students of the new population N = 2000		

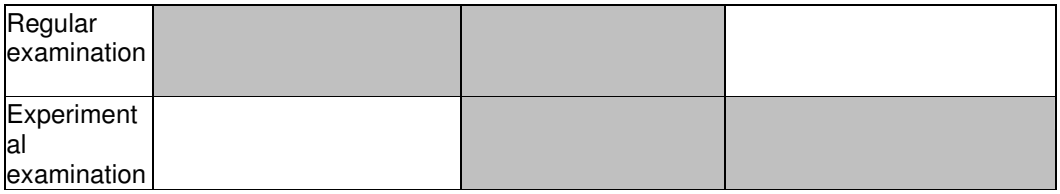
Note 3) For examinations where, in addition to the standard procedure, supplemental data is used we make a distinction between post-test and pre-test procedures. The post-test procedure works as follows: when composing and determining the examinations, the aim is to create the same difficulty level as that of a preselected reference examination. After the examinations have been conducted, the actual difficulty level of the new examination is compared with that of the reference examination. This is done by offering items of the new examination combined with items of the reference examination.

For the purpose of this additional research, the examinations are split up in sections, with booklets comprised of one section of each examination. The booklets are handed out to pupils of another level. Under this method, it must be assumed as a possibility that the ability levels of the two groups differ: they cannot be considered as a random sample from the same ability group. If the results of a comparison between the standards do not unequivocally show that the new population has a higher performance level than the reference population, this will also lead to higher scoring in the sense of lower fail rates and higher average marks.



Under the pre-test procedure, the results of the pre-test have already been processed before drawing up the definitive examinations and formulating the correction instructions.

Note 4) With experimental examinations, the overlap with regular examinations plays an important role. Based on the overlap, an attempt is made to set the standard of the examinations in such a way that both groups of examinees face similar requirements. An important starting point of the definitive setting of the standard is that examinees who take the experimental examination are not let down by its experimental character.



5.4 Appropriateness for innovations

Examinations - even if they are constructed by experienced test constructers – will differ in difficulty. Procedures for standard setting including equating methods are necessary to compensate for these differences and to underpin the selection of cut off scores on each individual exam.
Critical audits of the inspectorate stimulated a strict use of the results of test analyses in standard setting. However, as a consequence of recent innovation in which some schools are offered more examination opportunities the number of students for each sit will dramatically decrease. We will be less sure about the performance level of these small populations. Standard setting procedures based on teachers estimations of the degree of difficulty will become more important. Different research projects are started that evaluate the effectiveness of these kind of standard setting procedures. The results of one of these studies are presented below.

5.4.1 Standard setting with the examinations in the pilot study flexibility in time.
In the pilot study flexibility in time, 12 schools were allowed to administer examinations in January, May, and August. Owing to the small number of schools, the number of candidates was extremely limited, no more than 15 candidates in most subjects and no examinations with more than 100 candidates. The examinations will probably more often than normally be used by either less proficient students who need a resit or more proficient students that want to finish their education earlier. Consequently, it is not reasonable to estimate the difficulty level of the examinations with linking procedures based on the assumption of random equivalent groups. To be able to set the standard, a mixture of the following sources of information on the difficulty level of the examinations is used:

- Results on the examinations in January;
- Historical difficulty level of the examinations;
- Two forms of standard setting by different groups of stakeholders;

These different sources provide some information on the difficulty level of the examinations but none of them is accurate enough or influential enough to independently determine the standard. In the final standard setting, these sources are combined by a standard-setting panel. The resulting standards based on these different sources of information are graphically presented together with a confidence interval that represents the accuracy of the information.

In this study the focus is on the effectiveness of the standard-setting procedures that is carried out by the test construction team. The objective of this procedure is to transpose the existing standard of an old examination to a new examination. A set of items is then rated that contain items from both the old and the new examination. The old items function as an anchor for the standard on the old examination. The raters in the construction team are teachers which are also experienced item writers. The raters *r* are asked to estimate $\hat{p}_{i,r}$, the probability of a correct answer on item *i* in the total population of candidates[1]. The severity or lenience, s_r of rater *r* is defined by $s_r = \sum_{i=1}^{k} (\hat{p}_{i,r} - p_{i,r}^{obs}) / k$, where the sum is over all *k* anchor items and $p_{i,r}^{obs}$ is the observed probability correct in the population of candidates. Subsequently, for the new items $p_{i,r}^{c}$, the p-value corrected for severity, is determined by $p_{i,r}^{c} = \hat{p}_{i,r} - s_r$. Taking the average over items $\overline{p}_{.,r}^{c}$ and over raters $\overline{p}_{.,.}^{c}$, the expected difficulty level of the new examination is determined. A confidence interval is based on the variance of $\overline{p}_{.,r}^{c}$ over raters.

With the combination of observed and estimated data, it is also possible to evaluate the effectiveness of the procedure and the quality of the raters. An indicator of the effectiveness of the procedure is the reduction of variance between the corrected and uncorrected ratings. In an example, the effectiveness is investigated using results from an Angoff procedure with 20 raters for which also observed data are available. By using half of the examination to estimate s_r and the other half to investigate the variance between raters, the reduction in variance can be estimated. Table 1 shows that after correction the standard deviation between the raters increased from 4.67 to 5.13

Table 1

	SD	Number of raters removed
Uncorrected	4.67	
Corrected	5.13	
Rho 0.9	4.40	10
Rho 0.85	4.72	4
Gower 0.8	2.72	14
Gower 0.75	4.12	6

Taking into account the quality of the raters we tried to reduce the variance between the ratings. The quality of the judgements of a rater is assessed by the correlation between p^{obs} and $\hat{p}$. The higher the correlation, the better the judgement is in line with the actual data. Results are presented in row 4 and 5 of Table 1. It can be seen that a slight improvement in the variance occurred if we only took into account the raters with a correlation of 0.9 or higher. It should be noted that in this case we only evaluated the standard deviation between the 10 raters that met our criterion and that 10 other raters were left out of our

[1] Theoretically, it is also possible to ask to estimate $\hat{p}_{i,r}$ in a population of *'just sufficiently proficient'* candidates (Angoff, 1971), but the construction teams were more familiar with the population of candidates and therefore preferred this population.

evaluation. A somewhat better reduction of the variance could be obtained if we only considered those raters that with at least a Gower distance of 0.8 or 0.75 to the actual data. But again a substantial number of raters is removed from the analysis.
Although the above procedures are easily applied and can potentially improve the effectiveness of standard-setting procedures, we were not able to show that this procedure is effective with the current data.

In general, research into the standard setting procedures based on teachers estimations of the degree of difficulty will be important to be able to adapt the standard setting to be suitable for the innovative forms of central examinations of the future.

References

Kuhlemeier, H., Hermans, P., & Kremers, E. (2004). Wordt het een 'flexamen'? .
Examens, 2, 5-9.
Van Hest, E., Zeelenberg, R. & Dietvorst, P. (2008). Het centraal schriftelijk en praktisch examen- Een geschikte examenvorm voor het vmbo. *Examens, 1.*