The analysis of examination results as a prerequisite for improved teaching

36th IAEA Annual Conference
Bangkok, Thailand, August 22~27, 2010

Topic: Instrument and its Effectiveness
Key words: group reports, feedback, national examinations, benchmarks.

René Alberts (rene.alberts@cito.nl)
Peter Hermans (peter.hermans@cito.nl)

Cito, Arnhem, the Netherlands

peter.hermans@cito.nl

Abstract

Cito, the Dutch National Institute for Educational Measurement, processes test- and item analyses of the annual national examinations based on a sample of exam candidates. For these analyses, schools submit the scores at item level of the first five candidates of the alphabetically ordered list of their examinees through an online service called WOLF. The analyses are included in the information used by the National Council for Examinations (CvE) in the decision-making process of setting cut-off scores for each subject of the national examinations.
Since 2006, teachers willing to submit scores of all candidates in their classes can apply for customized reports of their students' results compared to the results of the national sample. This so called 'group reports' service, allows teachers to set up their own customized reports, based on a combination of subsets of questions (topics, domains, question types) and grouping of students (weak/ strong performers, male/ female etc.) Statistically, a group report is basic and limited in its ambitions. The group reports reinforce the intuitive knowledge of the teacher, because the results contribute to the assessment of the effectiveness of their teaching effort and of the quality of the curriculum.

Introduction

The Dutch national examinations are based on the examination syllabuses published by the Minister of Education. The responsibility for the examination papers and scoring guidelines lies with the Nationals Council for Examinations. A 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 those who were unable to sit the examination in May. The national examinations, which are the responsibility of the National Council for Examinations, are produced by Cito (Dutch National Institute for Educational Measurement). The examinations are marked by the pupils´ own teacher and a teacher from another school.

National examinations in the Netherlands are primarily developed as tools for the summative assessment of student achievement. In practice however, it turns out that examination results serve many evaluative purposes.
During the examination period, the national examinations draw a great deal of attention in and outside the media. While students and the public are merely interested in the fairness of the examinations, policymakers are focused on the possible effects of changes in educational policy, especially in times of major reforms in education. At a school level, national examinations have become one of the primary benchmarks, not only for certification purposes or determining achievement levels, but also for the school's overall performance. In the Netherlands examinations have been used as indicators for school quality for a long time, but from the moment that examination results were included in the annual 'quality charts' published by the schools inspectorate, examination data have become increasingly important.

A system of national examinations allows for inter-school comparisons and comparisons for the results of an individual school with national averages, even though the national examinations only cover a part of the syllabuses. One important question is how schools interpret anomalies and deviant results. More specific, schools are looking for answers on the following questions:

- At which point are deviations from the national average becoming something to worry about?
- Are national averages a valid benchmark for every school?
- In case of aberrations: which part of the examination, which questions or topics have caused these deviating results?
- Which instruments do teachers and schools have to improve their performance if necessary?

In 2006, Cito, the Dutch National Institute for Educational Measurement, introduced group reports to increase the usability of examination data for school self-assessment and quality management.
In a Cito group report, the scores of a group of students on subsets of items in a national written exam are compared to the scores of a nationally representative sample. These reports can help teachers to determine possible answers to the third question.

Data collection

Since 1976, student scores were collected using optical mark recognition technology, where teachers would mark a student's score for each question on a special form, which they then had to mail to Cito for further processing. For each subject, a school was obliged to submit the scores of the first five alphabetically listed students in taking the examination. These scores were used to produce test- and item analyses, which are used by the National Council for Examinations as a source of information for determining the cutt-off scores. Not only was this an expensive way of data collection, the number of forms rejected by the optical readers, was relatively high. In 2006 Cito introduced, after a series of tests, a computer program named WOLF and schools switched to submitting the item scores online. Due to the introduction of WOLF, the number of errors in submitting scores has minimized and at the same time, the speed of data collection and processing has increased drastic. Teachers can download

the data report models for each exam from a secured server. These data report models include al the necessary details of a particular exam, such as maximum item scores or the number of alternatives in a multiple-choice question. Thus, the program takes the teacher by the hand by giving online help and by supplying immediate feedback in case of errors. After the teacher has entered the data required, such as some student characteristics and the item scores and answers on multiple-choice questions in a digital form, these data can be uploaded to a central server at Cito.

WOLF was designed as a multipurpose, multifunctional tool and has many built-in extra features for teachers (automatic awarding of multiple choice questions, calculation of total scores and recalculation in case scoring rules are adjusted by the Nationals Council for Examinations). Marking schemes can be accessed directly from the program.
The idea of group reports came about when more and more uploaded forms showed that many teachers used Wolf to register the scores of all of their candidates and not just the first five as required by the Nationals Council for Examinations.

Group reports

The majority of the Dutch national examinations consist of a series of contexts (topics, cases or texts) each of these accompanied by a series of questions. The number of context varies from 5 to 10 and the average number of examination questions is 40.
Before the introduction of the group reports however, schools could only compare the average score of their students on an examination for a specific subject with the average of the national sample. This means teachers had no specific information about the origins of undesirable differences if these occurred, since an average test score does not tell much about which part of the examination, which group of questions caused these differences.

Group reports allow for multiple clustering of questions, these clusters are selected by Cito subject specialists. Normally questions are clustered by domains from the syllabus, by categories from the test blueprint, by examination context, by question type, computer use, etc. The types of clustering differ from examination to examination.
For each cluster of questions, a group report contains information about:

- the criteria for clustering questions;
- items in a cluster;
- item difficulty indexes for the group;
- item difficulty indexes for the national sample;
- the statistical significance of differences between the two difficulty

A prototype of the current group report was tested on a small sample of score sets submitted by teachers. A panel of 16 of the teachers in the sample was unanimously positive about the use of a more detailed picture of the achievement of their students on the national examinations for evaluative purposes.

The role of group reports in self-assessment

Self-assessment by teachers implies a customized approach and the use of instruments, chosen by the teacher, which fit the teachers' needs as well as the characteristics of the school. In other words, there is no such thing as a generic self-assessment.
External feedback can be advantageous for teachers and schools: on the one hand it will reduce the risk of blind spots, complacency and judgments guided by preconceptions and on the other hand, external feedback can open new perspectives. The initiative for self-assessments lies with the teacher or school and the same principle applies to requesting group reports. Whether or not to introduce and use this information in the overall evaluation of school quality and student performance is the teacher's /school's choice.

The notion that feedback can have a positive impact on a person's performance is widely accepted. It is also common belief that detailed feedback about student achievement and student performance is an important prerequisite for schools in order to maintain and raise the quality of education.

Research studies suggest a rather complex and ambiguous and relation between receiving feedback and improving the quality of education (Coe, 2002). Coe's findings are consistent with the meta analysis 131 research studies on the effects of feedback on performance by Kluger and DeNisi (1996).

It would therefore be premature to assume that group reports will have a positive impact on (future) examination results. One major conceptual issue is the idea of developing a comprehensive feedback system. Black and William (1998), identify four elements making up such a feedback system:

- data on the actual level of some measurable attribute;
- data on the reference level of that attribute;
- a mechanism for comparing the two levels, and generating information about the gap
- between the two levels;
- a mechanism by which the information can be used to alter the gap.

Kluger and DeNisi (1996) conclude that feedback interventions have a mild positive effect on performance. The variability in the reported effect sizes indicates that there are a number of factors which impact, either negatively or positively, on the effectiveness of feedback. Some of these factors refer to the content of the feedback as such, while others are more related to the nature of the task or the context in which feedback was given. Effective feedback should focus on progress, should be carefully presented and aimed at improving task oriented behaviour. Kluger & DeNisi (1996) found no empirical support for the popular belief that feedback based on social reference norms, where an individual's performance is related to the performance of the group, has a demoralizing, negative effect on performance.

Coe (1998) also emphasizes the importance of the way feedback is supplied. In addition to this he argues that feedback should be accurate and credible in order to generate a positive effect. Feedback should take such a shape that it is considered being information supportive to self-determination by the recipient. Feedback will have a positive effect when it enhances the feeling of being competent instead of enhancing complacency.
Coe (2002) concludes that, given the complexity of the many different types of feedback and the great number of different educational settings where feedback is given, and the many different forms of feedback it is just too complicated to propose generalised predictions about the effects of feedback.

It is important to understand the nature of the differences between feedback and its effects on performance. In order to increase the impact of feedback it is necessary to determine which variables can or need to be changed in order to generate a desired effect.
The merit of the work of Kluger and DeNisi lies in the attempt to integrate several perspectives on giving feedback into a comprehensive framework: Feedback Intervention Theory (FIT). The FIT is based on the notions that

- any reported discrepancy will evoke different reactions (denial, playing down or neglect;
- personal targets will prevail over task-related goals;
- the range of attention will be limited (only a limited number of discrepancies will be taken into consideration)
- discrepancies related to new and innovative tasks will result in more attention than discrepancies related to routine assignments
- feedback will change the scope of attention

Although there seems to be a consensus that giving feedback will be beneficial for performance, this general notion is not always supported by practical evidence. When applied to the potential effects of group reports, we must ask ourselves the question whether a positive impact of group reports on school performance is a realistic scenario.

Methodological implications

Moelands (2006) distinguishes types of information about results that are important to schools.

1. information about the school population based on descriptive statistical data, enabling schools to determine to what extent a school's population differs from that of other schools;
2. detailed information about student characteristics in order to track the progress of different groups over time;
3. a model based estimate of the added value of a school.

If we compare the group reports to these characteristics, we must conclude that group reports do not provide this type of information.
In essence a group report makes a comparison of a specific group of students with a nationwide sample and thus conclusions can only refer to the question whether or not the group of students is representative for the entire population of examinees.

Reliability
Estimates of reliability (Cronbach's Alpha/ GLB) are at the heart of the quality control process of the Dutch examination system. For the majority of the examinations these indices tend to be somewhere in the .75 - .85 range. Clustering questions will normally result in subscales with a lower reliability estimate than the examination itself, because of the smaller number of questions included in the subset. On the other hand, the average score of a group of students is less sensitive to errors of measurement related to the individual and, more important, these subscales are used for program evaluation only and not for assessment at the individual student level as is the case with the examination itself.

Equating and standard setting
Equating of examinations is based on test scores and other test related indices. Differences in difficulty can be compensated for on an annual basis, thus securing a stable achievement standard over the years. This is not the case with the subscales resulting from clustering questions. As a result the representation of a specific domain of the syllabus in a national examination can differ from year to year.
A .25 difference in the average item difficulty index from year to year might well be the result of a different difficulty level of the questions and not be related to variations in achievement level at all. In the case of group reports, the results of a group are reviewed in a referential context because of the comparison with a national sample. Group reports should therefore be considered as a form social referencing

Conceptual interpretation
The conceptual interpretation of data is a special problem in the context of program evaluation because teachers tend to have different views on the educational implications of the underlying syllabus. The big question is whether or not teachers accept the group report as a valid representation of the conceptual structure of the program they teach. In theory, information from the group report could lead to alterations in the sequence of domains taught, or to changes in time allocated to certain domains, or even to changes in teaching methods, media and assignments.

But teachers normally do not have this kind of mechanistic view of their teaching. Therefore, in terms of Black and William, there is no single alternative measure that will lead to a reduction of unwanted discrepancies. Educational reality is even more complex, because more and more students are given the opportunity to be responsible for their own learning process, and not all discrepancies will be considered as being 'undesirable' by teachers.
Ideally, teachers are able to predict discrepancies on the basis of the priorities in their teaching programme. Teachers who devote a relatively large portion of teaching time to a specific domain in the syllabus in order to raise the average level of attainment, can use the group report to check whether they have succeeded or not. Group reports should not result in teachers constantly adjusting their teaching in an attempt to match the national profile. The primary function of the group reports is to

alert teachers and schools if certain areas of the syllabus structurally (and inadvertently) result in test scores below the national average.

References

Black, P., & William, D. (1998). Assessment and classroom learning. Assessment in Education, 5, 1, 7-74.

Coe, R. (1998) Can feedback improve teaching? A review of the social science literature with a view to identifying conditions under which giving feedback to teachers will result in improved performance. Research Papers in Education, 13 (1), 43-46

Coe, R., 2002). Evidence on the role and impact of performance feedback in schools; In: Visscher, A. J. & Coe, R. (Ed.), School improvement through performance feedback, Swetz & Zeitlinger, Lisse, The Netherlands.

Coe, R, & Fitz–Gibbon, C.T. (1998) School Effectiveness Research: criticism and recommendations. Oxford Review of Education, 24 (4), 421-438.

Jong deOzn, (1981), Examennota: Examens in het voortgezet onderwijs, SDU, Den Haag,

Kluger, A.N., & DeNisi, A. (1996). The effects of feedback Interventions on performance: a historical review, a meta-analysis, and a preliminary Feedback Intervention Theory. Psychological Bulletin, 1, 19, 2, 254-284.

Koning de, P., (1979), Afsluitingen, doelen en functies, Pedagogische Studiën

Moelands, H.A. (2006) Schoolzelfevaluatie, Het evalueren van en door scholen, Cito, Arnhem