From students' achievement to the development of teaching: requirements for feedback in comparative tests[1]

Andreas Büchter, Dortmund (Germany) &
Timo Leuders, Freiburg (Germany)

Abstract: In November 2004 Germany's largest federal state North-Rhine-Westphalia for the first time carried out central tests in the subjects German, English and Mathematics in grade 9 with about 210.000 students participating. One of the main goals in assessing students' performance was to improve of teaching. This imposed certain requirements on the construction of tasks and on the feedback of results. In this article we present – referring to the specific experience from the development of the mathematics test – concepts and requirements for comparative assessment that is intended to support desirable changes in teaching practice.

Kurzreferat: Im November 2004 wurden in Deutschlands bevölkerungsreichstem Bundesland, Nordrhein-Westfalen, erstmals Lernstandserhebungen in der Jahrgangstufe 9 in den Fächern Deutsch, Englisch und Mathematik durchgeführt. Bei diesem Projekt bearbeiteten 210.000 Schülerinnen und Schüler zeitgleich zentral gestellte Aufgaben. Neben vielen weiteren administrativ formulierten Zielen stand dabei vor allem die Unterrichtsentwicklung im Vordergrund. Dieses Ziel stellt besondere Anforderungen an die verwendeten Aufgaben und die Rückmeldung von Ergebnissen an die Schulen. In unserem Beitrag werden wir, ausgehend von der Entwicklungsarbeit und den Erfahrungen mit den Lernstandserhebungen sowie deren Analyse, konzeptionelle Überlegungen für und konkrete Anforderungen an vergleichende Leistungsmessung formulieren, die eine fachdidaktisch wünschenswerte Unterrichtsentwicklung unterstützen soll.

ZDM-Classification: B13, C73, D63, D73

1. Central comparative tests for grade 9 in Northrhine-Westphalia: Framework and goals of educational policy

At the moment the German educational system is in drastic change. The frequently used term "shift of paradigm" may be well chosen looking at the steering instruments of educational policy (see introduction by Büchter & Leuders in this issue).

1.1 Educational standards and core curricula: orientation on the outcome of learning.

In response to the mediocre results German students achieved in international large scale assessment studies government has decided to steer learning and teaching processes in schools no longer via detailed curricula which give precise prescriptions when to teach which content.

To replace them the government introduced *national educational standards* ("Nationale Bildungsstandards", see article by Blum et al. in this issue) and on the level of the federal state North-Rhine-Westphalia (henceforth abbreviated "NRW") so called *core curricula* ("Kernlehrpläne" see article by Leuders, Barzel & Hußmann in this issue). These new curricula focus on the outcome and are usually referred to as "standards". They describe which competencies are expected of students after a period of schooling (national standards at the end of grade 9 or 10, core curricula at the end of grade 6, 8 and 10).

1.2 Central comparative tests advance accountability, quality management – and development of teaching?

Defining standards implies that each school should strive that all their students acquire the described competencies on a level as high as possible. How this is achieved is left to the responsibility of every single school and their teachers. This change of perspective from steering by defining the input to steering by defining and testing the outcome is a realization of the demand to increase school's autonomy – a demand that has been expressed by school improvement research for a long time (cf. Rolff 1991, 1992; Fend 1986). Since schools are supported and equipped by society this freedom requires on the other hand a system of accountability for the effects of their work. This is not only necessary to legitimize the use of resources but also to ensure a standard of learning conditions and the equivalence of certificates (cf. Liket 1992).

Accountability and ensuring equivalence on the level of each subject will be organized by central comparative tests and central examinations. Some federal states (such as Bavaria and Baden-Württemberg) already have long-term experience in central examinations which were mainly used as an instrument to establish identical requirements for all students undergoing an examination. In NRW such central examinations at the end of lower and higher secondary education will be introduced in 2007.

Already in 2004 NRW introduced *central comparative tests* ("Lernstandserhebungen", henceforth abbreviated LSE) that do not relate to any grading or examinations but were designated to give informative feedback to schools and to society on the effectiveness of teaching in certain areas.

While central examinations – like any other summative evaluation – have no consequences for the tested students regarding their further learning, central comparative tests can be compared to a formative evaluation: their results can directly influence the quality of the learning processes of the tested students (cf. Rossi, Freeman & Lipsey 1993). Hence with central comparative tests like LSE school improvement, especially the improvement of teaching can be initiated by means of central tasks and feedback. To what extent schools and teachers can actually profit from tests and (comparative) feedback depends closely on the quality of the instrument and on the way results are used, that is on the question *who* is entitled to

[1] The analysis of this article is based on the authors' participation in the development of tasks and feedback for the mathematics part of the central comparative examinations in 2003/04 in Northrhine-Westphalia (NRW). Besides the scientific consultants the group consisted mainly of teachers of all school types and a representative of school inspectorate. The managing agency for the test development was the *Landesinstitut für Schule/Qualitätsagentur* in Soest, a subordinate institution of the ministry of education NRW (For administrative details cf. www.learn-line.nrw.de/angebote/lernstand9/).

use and interpret the data and what kind of consequences for schools and teachers are derived from the results.

1.3 Administrative goals for LSE in NRW

Announcing the introduction of the central test in grade 9 („LSE-9“), the government published an extensive list of intentions[2]:

(1) Assessing standards and assuring quality – informing schools about their effectiveness
(2) Recognizing students' necessities of support in the subjects German, English and Mathematics (related to the whole class and – if possible – to the individual)
(3) Increasing teachers' diagnostic competence by comparing their expectations to actual results of their students
(4) Giving orientation for individual grading by regarding objective results
(5) Improving classroom teaching as a consequence of recognizing the results
(6) Supporting the implementation of the new core curricula by closely relating the tasks to the standards
(7) Identifying possible less effective schools and preparing long-term support fort these schools
(8) Giving supplementary information for system monitoring (in addition to results from large scale assessments like PISA-E)

Since a central test in the large state of NRW implies the edition of approximately 210.000 test booklets which causes considerable costs, it is quite comprehensible that the education administration connects such an extensive list of goals with the project. Still the question remains whether these goals can be considered simultaneously and with equal weight.

Some other aspects that are not explicitly mentioned in this list are of no less importance: the central tests, though practically suitable for it, are not to be used for school ranking, for sanctioning schools or teachers or for giving grades to students. Which of these intentions remain stable after a change of education policy under a new government is an open question.

The concept of comparative tests presented here is strongly influenced by similar instruments in Sweden (cf. the article by Kjellström & Pettersson in this issue). Since Sweden has already run through many years of data driven school improvement their instruments cannot be taken one-to-one but have to be adapted for the use in Germany. The construction of central tasks in Sweden puts less emphasis on objectivity and reliability of measurement but on usefulness as a diagnostic tool in the hands of the teacher. Nevertheless the performance results of each school in Sweden are publicly discussed without giving rise to inadequate comparison – which appears possible due to the long tradition of the long-term experience and the gradual development of such a system. Central tests in Sweden are supplemented by instruments for a long-term individual diagnosis. To exchange experiences with respective instruments the NRW *Landesinstitut für Schule/Qualitätsagentur* collaborated with the Swedish agency *Skolverket* from the beginning.

In the following sections we want to depict the phases of development and implementation of central comparative tests in mathematics. We put an emphasis on the development of adequate *tasks* (section 2) the construction of informative *feedback* (section 3) and the question, how these can influence the *development of teaching mathematics* in each classroom. Finally we give some concluding remarks as to the requirements for tasks and feedback that are optimized to reach this aim. We try to tackle the following questions that we consider decisive in this context (cf. Büchter & Leuders 2005b):

- What do tasks look like that assess whether a student has a true basic knowledge of a mathematical concept and that are not solvable by superficial application of a preformed technique?
- What should central tests look like which do not restrict themselves to expecting the application of algorithms?
- How can such tests create incentives for innovation and at the same tame assimilate traditions of the subject?
- What should feedback look like that is empirically valid and informative for everyday teaching?
- How can objective and reliable data on performance on a class level and at the same diagnostic information about the individual student be acquired?

2. Principles for developing tasks for a central comparative test

The potential quality of a central test is substantially dependent on the quality of its tasks. Blunt as this statement may seem at the first glance – it is an enormous challenge: The tasks have to fulfil requirements from mathematics teaching (didactics), from psychometrics and from school improvement at the same time. This proves a task of balancing multiple and partially conflicting quality criteria. To be more explicit, the three mentioned dimensions of criteria imply that

- the tasks chosen for a central test should not display a distorted picture of mathematics. Admittedly a paper-and-pencil-test that has to be solved individually can only focus on a certain sector of mathematical competencies anyway (aspects hardly considered are e.g.: working constructively with mistakes, presenting ideas to a public, reasoning discursively, constructing mathematical objects, exploring interactive computer environments, working together in solving a problem etc.) Still the activities induced by the tasks of the test can be constructed and selected by the criterion that they represent an authentic way of doing or using mathematics. In addition to that a central test should represent the subject in all its relevant aspects regarding content and complexity (*didactic requirement*).
- the performance of students in solving the tasks should be interpretable coherently within the chosen test model. For the LSE in NRW the model that was chosen in advance was the one dimensional, probabilistic model according to Rasch (cf. Rasch 1960; Fischer & Molenaar 1995; Rost 2004a), which was also used in many large scale assessment studies like PISA and TIMSS (*psychometric requirement*).

[2] cf. www.learn-line.nrw.de/angebote/lernstand9/ziele.html

- the tasks included in the test should be consistent with the "average" use of tasks by teachers and at the same time include distinct impulses for innovation (*school improvement requirements*).

The following sections will show that balancing these requirements in constructing adequate tasks is by no means a simple undertaking and that one has to consider many different levels of decision-making.

2.1 Curricular competence areas as a reference frame

The recent core curricula for NRW constitute the normative framework for LSE in grade 9. To be more precise the tasks largely relate to the competence expected at the end of grade 8. Several tasks also realize competence as described for the end of grade 6 to notify that these competencies are expected of students also two years after the end of grade 6.

Since the core curricula only became obligatory in the term 2005/06, a frequent argument against LSE was foreseeable: How could a central test assess competencies that were not yet "implemented" formally? This argument was easily rebutted since the new curricula did not present new content or new expectations but simply regarded the former syllabi from the point of view of expected competencies (cf. the article by Leuders, Barzel & Hußmann in this issue).

From the perspective of school improvement it even appeared mandatory to start the central evaluation from the outset of the introduction of new curricula and so to be able to measure annually palpable effects of the implementation.

The close interrelation between setting new standards in core curricula and assessing the standards in central tests has another important consequence that should not be underestimated: The introduction of standards as a new form of curriculum has evoked a strong necessity for orientation. Since only expected competencies are formulated and the path to reach them is left to choose for every single school and their teachers the question arises how these expectations – which seem abstract in comparison to concrete content lists – are to be interpreted and how students should acquire them effectively. New teaching materials (e.g. textbooks) are not yet issued and the majority of the old ones do not present the material from a "competence perspective" or make plausible suggestions. So teachers look for orientation and find it in the "official" tasks of central tests, since they give precise indication of competencies that are connected with each test item and thus fill the "implementation gap".

This connection seems helpful at the first glance but is dangerous at the second. Test items bear the problematic misinterpretation that they could be used for teaching instead for testing, e.g. by creating lessons that start from test items. Hence in lack of other support central tests could lead to an inadequate reductionist approach to learning mathematics, an approach that is extremely counterproductive regarding the acquisition of competencies. We give an example that illustrates this argument:

Choose a point C on the line g so that the triangle ABC has a right angle in C.

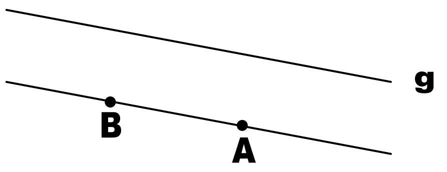
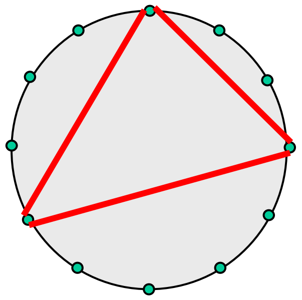


Fig. 1: "Theorem of Thales": Task for assessment

This is designed to test whether a student can apply the Thales theorem in a simple situation to create right angles – a competency that is required in the core curriculum in NRW. It would be quite a simplistic and hardly effective approach with regard to acquiring this competency to design a lesson based on this problem in which the theorem is developed and its application practised. Instead students should have the opportunity to explore situations in which they can find their individual ways to discover and maybe even to prove the theorem. Such learning environments should be rich, open and challenging and could look like this:



Try, by spanning the rubber band around the nails, to form as many different triangles as possible.

Copy and collect them, examine their angles and arrange them into groups.

Write down as many conjectures as you can make. Check them and give reasons for them if possible.

Fig. 2: "Theorem of Thales": Task for learning

This example illustrates how important it is – not only in mathematics – to clearly draw a line between tasks that serve the measurement of competencies (assessment tasks) and tasks that serve to initiate and promote the acquisition of competencies (learning tasks). While *learning tasks* should be part of or constitute open and rich learning environments that allow for differentiation, individual learning paths and mathematical invention, the items in central tests should be considered as *assessment tasks.* They are obliged to quality criteria of (psychometric) measurement performance. This different function implies in most cases a different form of the task (cf. Büchter & Leuders 2005a). This difference is not strict but it should be taken as a serious category for mathematics teaching in a context that defines curricula via outcome standards and therefore widely uses assessment tasks.

We shall now turn to potentially positive influences of the connection between core curricula and central tests that closely relate to them. Therefore we have to illustrate this connection by some concrete examples. A recent and in the German context rather innovative trait of the standards in the new core curriculum is the „competence area model“ which presents so called “process-related competencies” and “content-related competencies” on the same level.

process-related competencies	content-related competencies
reasoning and communicating	dealing with numbers and symbols
grasping, investigating and solving problems	describing and investigating dependence and change
finding and applying models	realising plane and spatial structures by measure and shape
using media and tools	working with data and randomness

Fig. 3: The competence areas of the NRW-core-curriculum

As to the content-related competence areas, the LSE takes into account all four areas. This is intended to lead to a consideration of aspects that have hitherto been neglected by some teachers – as is known for the area “working with data and randomness”. This content width could be realized in the two lessons that the test had at its disposal.

On the contrary the process-related competencies appeared comparatively unusual to (most of) teachers – and for the teachers developing the task as well. So it was decided to restrict to a single process-related competence area. Thus the test can contain sufficient items to illustrate this single process-related competence area extensively and to allow for a category specific one-dimensional Rasch modelling of the results. It was decided that the four process-related competence areas should rotate in an annual system (2004: modelling; 2005: problem solving, 2006: reasoning). About half of the test items in LSE 2004 illustrated different aspects of the competence area modelling, e.g. mathematizing, interpreting or validating. The other half of the test items concentrated on the other process-related areas or simply on content-related basic knowledge.

In the way depicted, also test items which are part of assessment tasks can indeed be useful for advancing teaching and learning - by giving useful and concrete interpretations of the competencies and by indicating important aspects of mathematical competence.

Candles

Two candles are lit simultaneously.

One has the initial length of 10 cm and burns down 1cm every hour.

The other has the initial length of 36 cm and burns down 3cm every hour.

At which point of time the two candles have the same length?

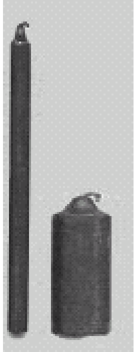


Fig. 4: Modelling task “Candles”

The example in fig. 4 from LSE illustrates how the often neglected aspect of validating a model can be transported by a task. The algebraic approach „linear equation“ leads to a system of two equations that has as solution the value 13. Students who write down “13 hours” as a solution have calculated correctly but if they had validated the model by calculating the length they would have remarked that “minus 3 cm“ indicates that the chosen model “linear equation” is problematic. Thus those students who chose this model and lose connection to the real situation can be identified. Those who solve the problem numerically and keep track of the interpretation realize that after 12 hours both candles have zero length.

On a much simpler level another test item also gives hints as to whether students interpret and reconsider their results in the real context:

School bus

1128 students must go from their school to a sports event. A single school bus can transport 36 students.

a) How many buses would be required to transport all students to the sports event?

b) 471 students go by bike. For the rest of the students 5 buses are available that – if necessary - can go several times. All buses should be equally used. How many times does every bus have to go?

Fig. 5: Task “School bus”

2.2 Item types

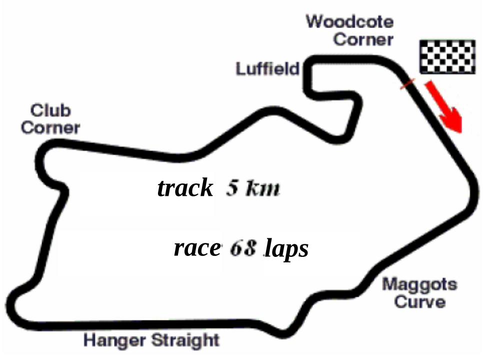
In LSE teachers evaluate the test results of their classes themselves. This is not a necessary trait of central tests as TIMSS or PISA show, where trained but not necessary didactically educated test assistants did the scoring. Whenever teachers evaluate their professional knowledge can be used for evaluation and a broader range of items types can be used in the test.

While multiple choice items grant a maximum of objectivity in LSE the type of fill-in items could be used extensively, since it is very similar to the familiar type of tasks that students and teachers use in most of the class tests. The benefits of fill-in over multiple-choice are evident: When choosing from a list the cognitive processes can be quite different from those that are necessary to reach a solution. Multiple-choice distractors – however intricately the may be constructed – still lead to a tendency of indirect choice by exclusion whenever the student is insecure about his or her solution. Moreover, when students do not find their individual solution in the list they may proceed by an unidentifiable strategy, e.g. that of choosing by maximum likeness.

These effects of guessing probability or systematic invalid interpretation can be minimized and may be irrelevant when the test results are used for a comparison of performance between national groups (cf. Lind 2004), but they are a serious impediment for using test items as a diagnostic tool. This can even be fatal with an underlying inadequate deterministic comprehension of diagnosis: „Choosing answer A has had the idea B“.

Items of the fill-in type reveal the general multitude of (wrong) solutions:

Racetrack



a) What is the total length of the race?

b) How many laps does a driver go in an hour if he can make 4 laps in 5 minutes?

c) How many minutes does he need for the total race?

Fig. 6: Task “Racetrack”

In part c) for a multiple-choice item one presumably would have chosen four distractors: the correct answer and three „typical wrong approaches“. In the pilot phase of LSE we used a fill-in format instead. From randomly chosen solutions of the pilot test of LSE we found considerably more sensibly interpretable wrong solutions:

85; 425; 135; 16,2; 78,2; 54,4; 272; 1700

Fig. 7: Students answers to the part c) of “Car racing”

Every single solution revealed after intensive analysis possible ideas of students – sometimes explicitly written down in the calculation part. These insights were concisely included in the feedback and can be used as a starting point for the teacher for an intensive and desirable analysis of students’ ideas. The openness of the item type is a necessary condition for this approach of using test items also as a diagnostic tool.

2.3 Tasks emphasizing understanding

What do tasks look like that assess whether a student has a true basic knowledge of a mathematical concept (*tasks emphasizing understanding*) and that are not solvable by superficial application of a preformed technique (*tasks emphasizing procedures*)? This question seems pivotal regarding the improvement of mathematics teaching and learning in German classrooms. Didactic expertise and later also interpretations of large scale assessments results as TIMSS II und III (see Baumert, Lehmann et al. 1997; Blum & Neubrand 1998; Henn 1999; Baumert, Bos & Lehmann 2000) testified a detrimental orientation on procedures (“Kalkülorientierung”). An example of an item that potentially circumvents the procedural approach and assesses true knowledge of a rather simple concept is part b) of the task shown in fig.8.

Throwing a ball

During a sports contest Sven achieved the following results in throwing ball:

Throw No.	*1*	*2*	*3*	*4*	*5*
Range in m	*48,5*	*51,5*	*31,0*	*49,0*	*40,0*

a) Calculate the average throwing range (arithmetic mean). Note down your calculation.

b) What range should Sven have thrown in his fifth throw, to reach an average range of 46,0m? Note down your calculation.

Fig. 8: Task “Throwing a ball”

In part b) of this task a competent use of a mathematical concept (“arithmetic mean”) is expected whereas part a) can be solved by applying a procedure (“assign all values and dividing”) irrespective of knowing their meaning.

Thus part b) can be seen as typical *task emphasizing understanding* which often can be constructed by simply inverting a given *task emphasizing procedure*. To determine the desired fifth value the student can e.g. use the fact that the sum of the five values must be 230,0m. Or, students that have determined the average 44,0m in part a) can reason that Sven has to throw 2,0m further “on average” and the sum thus needs a total increase of 10,0m. These ideas lead on different paths to a solution and they show a fundamental understanding of the underlying concepts.

This simple idea of creating tasks *emphasizing understanding* plays a central role in focusing on competencies rather than pieces of content knowledge. It should not be regarded as a universal recipe but as a general orientation for looking at students’ proficiency – in central tests and in class tests and examinations as well.

2.4 Assembling the test booklets

The results of the LSE-9 in NRW should allow an empirically valid and fair comparison of schools and classes. The attribute “fair” means that background variables, especially socio-economic factors are taken into account when comparing test results. This requirement makes an empirically solid approach imperative. The distribution of students on four school types in the secondary level (see fig. 1 in the article by Leuders, Barzel & Hußmann in this issue) and the large difference in performance of students of each of these types (cf. Deutsches PISA-Konsortium

2001; Prenzel et al. 2004), suggests that a distribution of items of different average difficulty over at least two test booklets could be sensible for two reasons: Firstly the difficulties could thus be better adapted to the respective populations and give more significant results. Secondly students (and teachers) were not used to writings tests that in part under- or overestimate the capabilities of the students. Very simple items that have to be included for some groups can actually irritate many students in others groups. So it had to be communicated in advance that a test booklet differs not only in item type but also in the difficulty distribution from neatly organized class tests.

The reliable estimated difficulties needed for this design can hardly be produced by a priori estimating by experts. They could be acquired in a broad pilot phase with carefully chosen representative sample of students. (In a pre-piloting phase tasks were examined concerning comprehensibility.) With the data from the piloting phase two test booklets could be assembled each consisting of 20 common items and 20 items of higher respectively lower difficulty. The coherence of the two test booklets was obtained by modelling the items in a common scaling procedure using the Rasch model.

This intended compatibility of the item set to a Rasch model brings up many questions that belong to the aforementioned problem of "balancing" the different requirements. This became especially manifest in the procedure of item selection. The psychometric approach that was suggested for the construction of tasks was that complex tasks should be broken into smaller ones with a hierarchical interdependence: students that solved a certain item should nearly necessarily solve an "easier" one. This procedure represented the *psychometric approach* and should a priori grant the one-dimensionality. The *didactic approach* was to create tasks that represented the width of performance throughout the whole curriculum irrespective of pre-planned interdependencies of tasks and look for an adequate model of students' performances afterwards. The *school improvement approach* mainly asked for the interpretability and usefulness of the model with which the results were to be reported to schools.

These different requirements lead to irresolvable discrepancies: Some items that represented didactically important concepts and that promised to give helpful insights to teachers (as e.g. the task "candles", fig. 4 on p. 327 regarding the "validation" aspect). Unfortunately they proved incompatible with a one-dimensional Rasch scaling and thus "internal" model validity and "external" didactic validity lead to different and irreconcilable views of item quality.

The pragmatic approach was that relevant tasks could remain in the test irrespective of their model fitting while a sufficient quantity of model conform tasks had to be included to enable Rasch scaling.

Another compromise in view of LSE 2005 is also important with respect to the quality of the feedback: The selection of tasks regarding a one dimensional Rasch model should be restricted only to the central process-related competence area (2004: modelling; 2005: problem solving). Thus only half of the items are subjected to the scalability criterion.

Aside from the question of fitting a model all items are tested and will be tested with respect to reliability and objectivity: Do independent raters of the testbooks reach the same results? This is indeed an important criterion for a test that is to be used for comparing schools and classes. Still there are many tasks, especially the open tasks that are excluded by this criterion although they produce valid results when interpreted by the teacher who knows the students whose solutions he or she assesses.

3. Principles for designing feedback from central comparative tests

In a review article on feedback of performance data Bonsen & von der Gathen pose a central question of data driven school improvement:

> „One of the problems concerning feedback that has to be solved in the future is the question how data can be processed in a way that teachers can actually derive relevant 'messages' concerning the design of instructional processes. When this process of ‚feedback deciphering' becomes too complex or time-consuming the teachers interest dwindles abruptly." (Bonsen & von der Gathen 2004, S. 238)

In the first LSE 2004 this problem was tackled by three complementary types of feedback:

- a detailed commentary for every task, containing hints to its diagnostic potential and suggestions for a possible follow-up work in the classroom (section 3.1),
- the solution frequencies for every item for the class, the school and an average over all schools of comparable types (section 3.2), and
- a distribution of the competence level concerning the competence area ‚modelling' (section 3.3).

These three types of feedback and their respective benefits and drawbacks will be discussed in more detail in the following.

3.1 Commentary on the tasks

The annotation that was added to every task comprised three main aspects:

First of all, the competencies from the curriculum that every item could be an indicator for, are represented in a tabular form. Here the core curriculum and LSE can mutually support their recognition among teachers as was mentioned in section 2.1.

The emphasis in the interpretation of the respective competencies is on a probabilistic interpretation: students who possessed a certain competence were expected to solve the task rather easily. Conversely, if a student failed a task this could be a hint for a probable lack of at least one of the competencies in question. On the other hand the correct solution should never be seen as the proof for a competence. In fact this approach could lead to a deficit-oriented interpretation of a single student's performance. From the point of view of evaluating the whole class it seems sensible to ascribe a competence only if a majority of students solves an item rather confidently. Quite on the contrary in class tests teachers tend to give partial credits to students for every solved part of a task – mostly for pedagogical considerations. This can indeed lead to an “additive” assessment policy, summing up the result of many simple activities to a false picture of proficiency and obstruct the view on the true competence spectrum of a class. To reduce the effect of deficit-oriented assessment in the interpretation of single solutions there will be a partial credit model with respect to partial competencies in future tests.

Furthermore a detailed perception of students’ performance can reverse the negative thinking into seeing the deeper meaning of students ideas. This already refers to the second aspect of the commentary: the detailed description of solutions and especially of individual difficulties. To this end the students’ solutions from the piloting phase have been analysed using qualitative research methods. This analysis resulted in a collection of students’ workings and their typical ideas. With these examples teachers could sharpen their view on students’ solutions and problems by these examples. Further hints were given how to face such problems in classroom, e.g. by conceiving adequate practising tasks which focus on fundamental notions and connections.

Two examples should illustrate the quality of the commentary – that almost never exceeded one page!
The analysis of the students’ solution of the task in fig. 9 in the pilot phase produced a multitude of individual (right and wrong) solutions and several underlying ideas could be identified by the students’ writings.

Wheel of fortune

Peter and Tanja turn the wheel. They agree on the following rules:

Peter wins, when a number appears that is bigger than 6.

Tanja wins, when a number appears that is less than 6.

Whose chances to win are better? Give a reason.

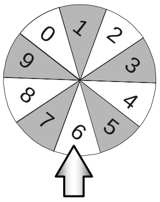


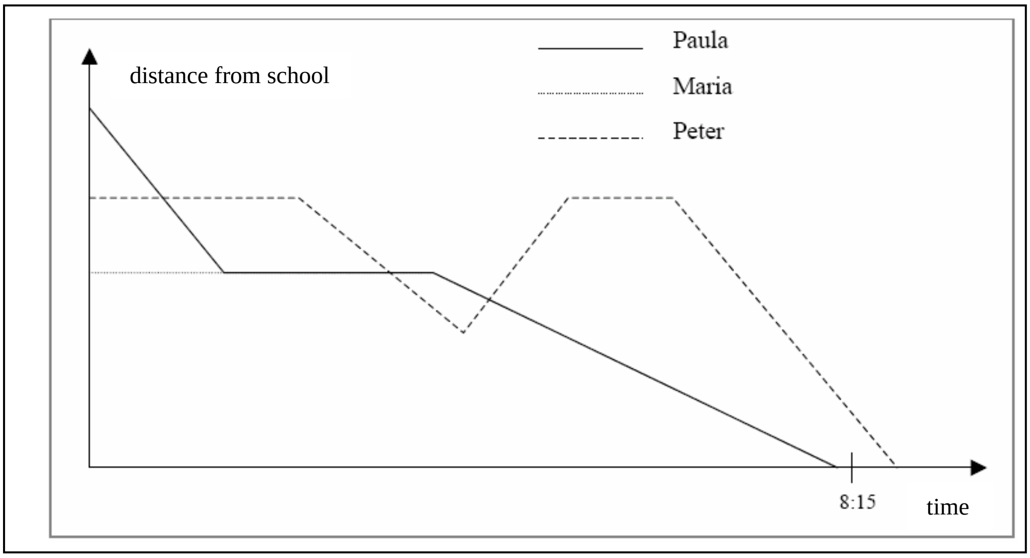
Fig. 9: “Wheel of fortune”

These ideas leading to a wrong solution could almost entirely be assigned to one of the following types:

- The students did not find an adequate mathematical model for the rule of the game, e.g. he or she assigned the “6“ to Peter or to Tanja. Some students missed to take the „0“ into account.
- Some students had individual conceptions about the turning of the wheel that did not reflect the intended (standard) interpretation as a random experiment. Some students regarded the illustration not as a metaphor for an experimental situation but as the depiction of a concrete outcome and thus assigned the highest winning probability to the number “6”.

Way to school

Peter, Paula and Maria are in the same class and live on the same street. There school is at the end of the street. Every morning they walk to school, lessons begin at 8:15. The illustration shows where they have been yesterday at different points of time.



a) Please look at the illustration. Which statements could be true?

	yes	*no*
Peter lives farthest from school.	□	□
Maria and Paula walk faster when the walk together than when walking alone	□	□
Maria is not ready to leave when Paula passes her home.	□	□

b) Write a story about Peter’s way to school.

Fig. 10: Task “Way to school”

- Some students went even further and – possibly referring to experiences with real wheels of a similar kind – concluded that numbers that follow closely after the “6” had a lesser probability when the wheel is turned again.

These results from the pilot phase could be used to compile a manual to support a diagnostic approach that contained suggestions for possible interpretations and some original solutions that could be used as indicators. Some of the solutions may e.g. suggest that the students’ everyday experience still interferes with the idea of a mathematical random experiment – a possible reason for students difficulties in reasoning with probability.

This analysis was possible due to the “give a reason” part of the task. These answers of the students were not evaluated. The request “give a reason” was frequently added to the items although it was not regularly evaluated. Still the teacher can draw important diagnostic information from the students’ productions.

This idea of using the test items apart from the central evaluation for individual diagnosis is best illustrated in the following example. The part b) of the task „way to school“ (fig. 10) was originally intended for the test but could not be included because it could not be evaluated objectively.

Analysing the number of correct solutions to item a) gives an estimate whether the class can profit interpret such graphs. This knowledge can be important for further lesson planning. Nevertheless the individual patterns of right or wrong responses give almost no clues to the individual ideas or problems of students.

Hence, for working with this task after completing the test, we recommended to do part b) in class and to read and/or discuss the students’ texts. Such extensive student productions are hardly evaluable centrally and objectively. For the single teacher who has background knowledge about his or her students and who can pose questions if necessary, such production give valuable diagnostic information as the following two authentic texts may prove:

- *Peter was standing in front of his house when he got the feeling that he had forgotten something. That way he kept standing for a few minutes. Then he started walking. He had made half way to school when he realized* that *he didn’t wear his panties. He was ashamed and ran home. He met two girls, who laughed at him. For a long time he couldn’t find his trousers. That’s why he came late for school...*
- *Peter had* not *packed his books and had not had breakfast either. So when he left his house it was very late. So he hurried a little and because the shortest way was a dead end he turned one road before. That’s why he moved away from school a little because the road went the opposite direction. Now he has reached the level of his house and…*

3.2 Feedback of solution frequencies and comparison of classes

The natural interest of teachers (and students) in the results of the test refer to the numerical proportions of correct solutions. To construct an adequate numerical feedback one has to consider several questions:

- How far should the data be processed and aggregated? Should there be individual results? Over which groups should an average be calculated? Should there be a summative calculation of performance over the whole test or only per competence area, per task or even per item?
- Which comparisons between group results should be made? Haw can significance of differences be intelligibly displayed?
- How should the information be put into graphs?

The number of possible feedback formats is almost boundless. In LSE 2004 decided for a “profile” format (fig. 11).

For every competence area teachers could regard the item profile of his class. Immediately after entering the data of his class into the system teachers can compare the profiles of their classes and identify items with striking discrepancies. Significance of differences is illustrated by indicator bars. In a second step when the results of all schools are available (e.g. 2 weeks after the test) the comparison to classes of comparable schools can be made.

In contrast to a summative score that would presumably lead to an inadequate summative attempt of explaining differences (“teaching quality”, “class composition” etc.)

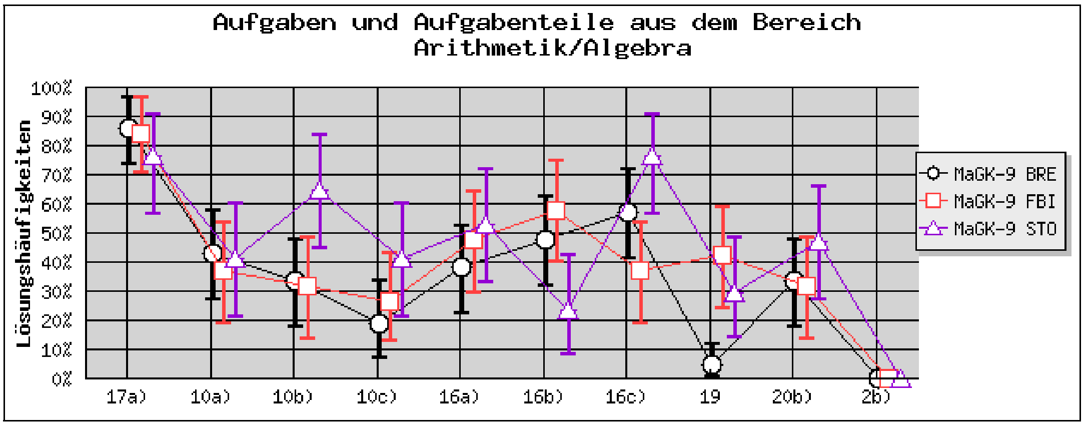


Fig. 11: Feedback of class performance as a “profile”

the profile approach encourages an analysis that begins with students performance related to concrete tasks. To promote this the online display of results is immediately linked to the commentary described in sec. 3.1.

Although the solution frequency is a value that can be easily interpreted the feedback per item may be too detailed information to draw more general conclusions. The per-item analysis could prevent a discussion and explanation of substantial problems. This could be remedied by agglomerated and interpretable values such as the performance in a certain competence area (see section 3.3)

Another way of supporting teachers in evaluating the data could be a “guided exploration” that displays only significant differences and groups them according to preselected relevant aspects, e.g. students’ performance in dealing with probability. This exploration could be realised in an online platform.

3.3 Feedback concerning a competence area (‘modelling’)

When a certain competence can be identified as a latent dimension underlying the students’ performance this can be an appropriate basis of meaningful feedback. An instrument to identify such latent dimension can be drawn from probabilistic test theory (IRT, item response theory, cf. Rost 2004a). The model that is most widely used at the moment is the one dimensional Rasch model. When items are selected in conformity to such a model a one dimensional scale is achieved that relates students’ competence and task difficulties on a common scale. This scale can be interpreted as a latent variable and be used to explain the variation in performance.

Such a model can be used to give substantial feedback about students’ performance. In the LSE a Rasch scaling was used to describe the students’ competence in the area of ‘modelling’. A unified scale for ‘general mathematics competence’ was deliberately avoided since this would have meant to describe the students’ performance in mathematics with a single variable – an interpretation which seems inadequate from the didactic perspective and which could lead to an oversimplification in the interpretation by teachers.

To supplement the numerical performance data of the Rasch scale with interpretable categories competence levels were introduced. This was done by grouping the items on the performance scale by the quality of their respective demands. To every group a competence level was assigned and described in terms of students cognitive activities when solving tasks from this level. The result was a competence scale for ‘modelling’ that is similar to the representations of competence scales known by PISA (cf. Deutsches PISA-Konsortium 2001; Prenzel et al. 2004) and TIMSS (cf. Baumert, Lehmann et al. 1997; Baumert, Bos & Lehmann 2000) and can be used for giving feedback to schools about the competence distribution of students (see fig.12).

Communicating the notions that are necessary to understand such a probabilistic competence model is somewhat difficult. To explain the underlying concepts several simplifications had to be applied:

> “Depending on the demands of the tasks they can be collected in groups of comparable level of demand. These levels are defined so that tasks of a higher level include demands of those of a lower level and pose additional demands. Depending on the proficiency of students they can solve tasks from lower or higher levels with a certain grade of secureness (i.e. they solve the least difficult tasks of the level with a probability of 62%) When they are rather secure with tasks of a specified level then their competence corresponds to the demands of that level. Hence they have reached a certain competence level in the test 2004. Students of a higher competence level perform better than those of a lower level.”
> (www.learn-line.nrw.de/angebote/lernstand9/)

One of the main problems in communicating an adequate interpretation of a Rasch scale appears to be the probabilistic design: Competence levels cannot be ascribed to single students but are a trait of a hypothetic average student achieving a certain score. The complicated explanation in the quotation was designed to avoid that the notion of ‘competence level’ is interpreted as a quality of a single student that the test only detects.

Consequently a distribution of students over competence levels should only be seen as information about the average performance of a group.

So in spite of the simple appearance of such a feedback relying on competence scaling some serious drawbacks should be taken into account:

- The problem of generating valid competence levels is still in discussion (cf. Rost 2004b; Kleine 2004, 2005). In the LSE we solved this task by cumulating tasks with similar demands. On a basis of 20 items this is still

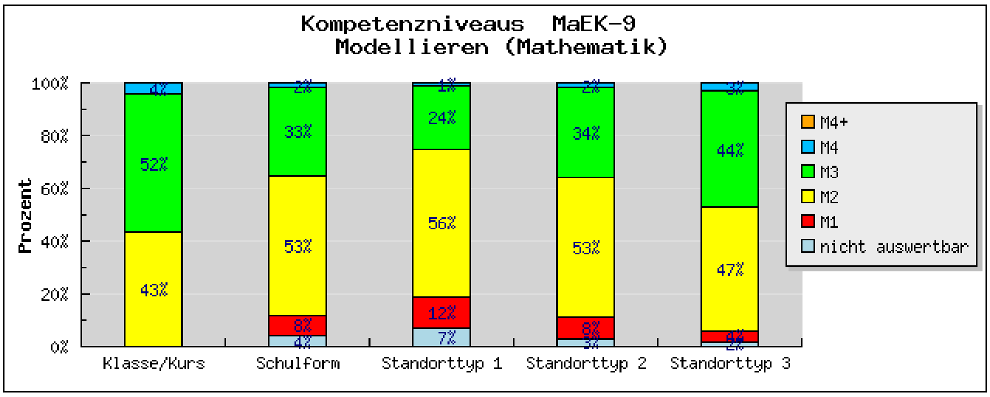


Fig. 12: Feedback of the competence distribution per class, per school type and in the state average. The categories M1 to M4+ relate to increasing competence levels with

a rather subjective solution and does not give rise to a psychometric universal model. The achieved scale and its interpretation must be seen as a result of the specific test.

- The requirement that tasks have to fit into the one-dimensional model seems a too restrictive one. From the perspective of didactics it would be desirable to explore other models of mathematical competence before postulating one-dimensionality.
- The interpretation of competence levels can be misunderstood as objective traits of pupils and lead to a labelling. Traditional grades could be substituted by pseudo-objective competence levels. The interpretation of scores of individual students is known to be inadequate (cf. Lind 2002) but frequently encountered under the term of “individual diagnosis”.
- The technical complexity of a procedure resulting in psychometric scales is rather high and requires an extensive piloting phase. Due to this fact the feedback can be delayed for several weeks or months and thus the impact and acceptance by the teachers can be considerably diminished. This can be remedied by a system of item banking.
- Competence levels of a cross sectional test can give a picture of the variation of performance in a cohort of students. There is no reason that they give an adequate orientation for organizing a learning process for students over a period of many years. To this end more sophisticated models describing learning development (including e.g. conceptual changes) would be required to.

Thus a competence model derived from a performance study like LSE is still far away from the expectation stated in Klieme et al (2003, p.18):

> Competency models should also convey information about the contexts and age groups within which the individual competency domains develop and about the influences operative on this development. Only under these circumstances can schools be expected to contribute appropriate measures to the systematic development of competencies and to cumulative learning.

Concluding one can say that the question whether a feedback referring to competence levels is an adequate and sustainable way for improving teaching is not yet resolved.

4. Conclusion: Requirements for feedback in comparative assessment

The main goal of feedback for comparative tests undoubtedly is in general the support of school development and in particular improvement of teaching. Therefore adequate feedback should meet three main challenges:

- *relevance*: teachers should promptly receive relevant data that can be a starting point for developing their everyday work.
- *intelligibility*: the feedback should use concepts and terminology that are consistent with the customary professional view of teachers
- *acceptance*: the results should be constructive and avoid accusations of established practice
- *innovation*: the feedback (and the collection of tasks) should transport innovative ideas and thus exploit a “positive washback effect” whereas undesired side effects (“negative washback”) like teaching to the test should be avoided by communicating the difference between acquisition and assessment of competencies.

References

Baumert, J., Lehmann, R. et al. (1997). *TIMSS. Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde.* Opladen: Leske+Budrich.

Baumert, J., Bos, W. & Lehmann, R. (Eds.) (2000). TIMSS/III. *Dritte internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Grundbildung am Ende der Schullaufbahn.* 2 Bände. Opladen: Leske+Budrich.

Blum, W. & Neubrand, M. (Eds.) (1998). *TIMSS und der Mathematikunterricht: Informationen, Analysen, Konsequenzen.* Hannover: Schroedel.

Bonsen, M & von der Gathen, J. (2004). Schulentwicklung und Testdaten. In H.G. Holtappels, K. Klemm, H. Pfeiffer, H.-G. Rolff & R. Schulz-Zander (Eds.), *Jahrbuch der Schulentwicklung. Band 13. Daten, Beispiele und Perspektiven* (p. 225-252). Weinheim & München: Juventa.

Büchter, A. & Leuders, T. (2005a). *Mathematikaufgaben selbst entwickeln. Lernen fördern – Leistung überprüfen.*Berlin : Cornelsen Scriptor.

Büchter, A. & Leuders, T. (2005b). Zentrale Tests und Unterrichtsentwicklung ... ... bei guten Aufaben und gehaltvollen Rückmeldungen kein Widerspruch. *PÄDAGOGIK, 57 (5),* p. 14-18.

Deutsches PISA-Konsortium (Ed.) (2001). *PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske & Budrich.

Fend (1986). „Gute Schulen – Schlechte Schulen“. Die einzelne Schule als pädagogische Handlungseinheit. *Die Deutsche Schule, 78 (3),* p. 275-293.

Fischer G. H. & Molenaar, I. W. (Eds.) (1995) *Rasch models. Foundations, recent developments, and applications.* New York: Springer-Verlag.

Henn, H.-W. (Ed.) (1999). *Mathematikunterricht im Aufbruch.* Hannover: Schroedel.

Kleine, M. (2004). *Quantitative Erfassung von mathematischen Leistungsverläufen in der Sekundarstufe I.* Hildesheim: Franzbecker.

Kleine, M. (2005). Latent-Class-Analyse. Ein Bindeglied zwischen Empirie und Theorie zur quantitativen Erfassung mathematischer Leistungen. *Journal für Mathematikdidaktik, 26 (2) , p.* 97-113.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E. & Vollmer, H. (2004) *The Development of National Educational Standards. An Expertise.* Bonn: Bundesministerium für Bildung und Forschung. (http://www.bmbf.de/publikationen/2713.php)

Liket, T. M. E. (1992). *Vrijheid en Rekenschap. Zelfevaluatie en externe evaluatie in hetvoorgezet onderwijs.* Amsterdam: Meulenhoff.

Lind, Detlef (2004). Welches Raten ist unerwünscht? Eine Erwiderung. *Journal für Mathematik-Didaktik, 25(1),* p. 70-74.

Prenzel, M., Baumert, J. , Blum, W. , Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J.; Schiefele, U. (Eds.) (2004). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland - Ergebnisse des zweiten internationalen Vergleiches* . Münster: Waxmann.

Rasch, G. (1960). *Probabilistic models for some intelligence an attainment tests.* Copenhagen: Nielsen & Lydicke.

Rolff, H.-G. (1991). Schulentwicklung als Entwicklung von Einzelschulen. *Zeitschrift für Pädagogik, 37 (6),* p. 865-886.

Rolff, H.-G. (1993). *Wandel durch Selbstorganisation. Theoretische Grundlagen und praktische Hinweise für eine bessere Schule. Weinheim & M*ünchen: Juventa.

Rossi, P. H., Freemann H. E. & Lipsey M. W. (1993). *Evaluation. A systematic approach.* London: Sage Publications.

Rost, Jürgen (2004a). Lehrbuch Testtheorie – Testkonstruktion. Bern: Huber

Rost, Jürgen (2004b). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik 50 (5),* p. 662-678.

Authors

Büchter, Andreas, Institut für Entwicklung und Erforschung des Mathematikunterrichts, Universität Dortmund, D-44221 Dortmund.
E-Mail: andreas@buechter.net

Leuders, Timo, Prof. Dr., Institut für Mathematik und Informatik und ihre Didaktiken, Pädagogische Hochschule Freiburg, Kunzenweg 21, D-79117 Freiburg.
E-Mail: leuders@ph-freiburg.de