Reconciling the Pedagogical Goal and the Measurement Goal of Evaluation: The Perspectives of Teachers in the Context of National Standards

JACQUELIEN BULTERMAN-BOS
Vrije Universiteit

NICO VERLOOP
Leiden University

JAN TERWEL
WIM WARDEKKER
Vrije Universiteit

The intense conflict about whether standardized tests and national standards are appropriate is related to the goals one ascribes to student evaluation: a measurement goal for selection decisions or a pedagogical goal to support the learning process. This study focuses on how teachers balance these two goals. We interviewed 25 teachers who work in a context in which national standards in the final year are combined with classroom evaluation in the first years. The teachers’ perspectives show that both goals can coexist. The national standards cause classroom evaluation to have a measurement function, but many teachers use it to implement a pedagogical goal. Teachers adjust classroom evaluation to encourage better performance in the long run. Referring to the literature, we discuss first whether this practice is appropriate from a pedagogical perspective and then from a measurement perspective. We found that national standards may function as a common reference point when developing a qualitative framework necessary for the interpretive measurement of higher order skills. The side effect of discouraging low achievers is counterbalanced by the use of adjustment strategies to circumvent failing grades. At the same time, national standards counterbalance grade inflation and cause the function of evaluation in stimulating student effort to be retained. This shows that it is possible to reconcile the advantages of the measurement approach and the pedagogical approach to a large extent.
Current views of the evaluation of students require evaluation to comply with criteria of good measurement if it is used for selection decisions: Evaluation should be valid and reliable (Ebel, 1979; Phye, 1997). At the same time, the literature stresses the pedagogical goal of evaluation: Student evaluation should support and improve student learning (Crooks, 1988; Frederiksen & Collins, 1989; Phye, 1997; Verloop & Van der Schoot, 1995). Between the pedagogical goal to support the learning process and the measurement goal for selection decisions, a tension exists that has given rise to an intense conflict about which goal should be given priority. This often tends to be an either/or discussion. Some favor a measurement approach, including the use of standardized tests and external standards, believing that such measurement improves student learning. Others point out that objective measurement directs the learning process toward meaningless, trivial skills and that measurement, moreover, discourages low-achieving students. These people advocate authentic assessment. If no standards are available, however, authentic assessment may result in every kind of student performance being applauded without challenging students to achieve better. This is a weakness to which the advocates of standardized test refer.

Our study describes the perspectives of 25 Dutch secondary school teachers regarding classroom evaluation. In the Dutch system, decisions such as those concerning promotion to the next class (or to a lower or higher stream) are based on classroom evaluation (Verloop & Van der Schoot, 1995). Classroom evaluation is based on activities that students undertake as an integral part of the educational program in which they are enrolled (Crooks, 1988). It includes such tasks as teacher-made tests, curriculum-embedded tests, and oral questions. Evaluation refers to the process of judging the quality of the performance of the students on these tasks (Terwilliger, 1989). In the Netherlands, national standards on different levels exist for final examinations, which underline the necessity of proper selection during the years preceding the final year. This implies that classroom evaluation has a measurement goal during all years of secondary education.

Our study shows that, while using classroom evaluation, 23 out of the 25 teachers use adjustment strategies to prevent students from receiving failing grades. Some of these strategies entail actively encouraging students to perform better—they are attempts to drag students across the finish line, so to speak; other strategies entail avoiding discouragement. The use of adjustment strategies shows that teachers attach a pedagogical goal to evaluation that would otherwise primarily serve a measurement function. In other words, teachers try to reconcile aspects of the two goals of evaluation.

Using the literature and our own reasoning, we analyzed whether the use of adjustment strategies is appropriate from a pedagogical perspective.
Next, we analyzed whether the use of adjustment strategies is appropriate from a measurement perspective, particularly if measurement is seen from a developmental and interpretive view. We conclude by indicating the factors that enable teachers to reconcile the two approaches of evaluation (to some extent).

LITERATURE REVIEW

The literature ascribes two different goals to evaluation: a pedagogical goal to support student learning and a measurement goal for selection decisions. The older (psychometric) literature viewed evaluation primarily as a measurement activity, (implicitly) believing that good measurement reinforces a good pedagogical approach. Ebel (1979) disagreed with teachers who show an inclination to temper justice with mercy by yielding to the subtle pressures to give more high marks and fewer low ones. He considered the validity and reliability of evaluation to be central. Marks should indicate as accurately as possible the extent to which the student has achieved the objectives of instruction. Ebel described the goal of measurement as stimulating, directing, and rewarding student effort.

Evidence that the measurement of standards (or targets) directs and stimulates student effort appeared from various studies. In 1973, the international study of Kienitz stressed the importance of attainment targets, showing that a system of individual self-development perpetuated unequal chances, especially for students from lower social economic backgrounds (Kienitz, 1973; Vos, 1981, WRR-Rapport, 1986, p. 72). Natriello and Dornbusch (1984) found that higher standards generally lead to greater student effort and to students being more likely to attend class. Bishop (1995) concluded that curriculum-based external examinations (which are, of course, measured) are incentives for the learner to spend time studying and become actively engaged in learning. Assessment experts in the United States, such as Phye (1997), recommended the introduction of external standards to improve student learning. Thus, research shows that because selection takes place—requiring certain standards to be reached—students are challenged to do their best, which fosters the learning process.

The literature, however, also presents evidence of the negative aspects of measurement. Natriello and Dornbusch (1984) showed that students who perceived standards as unattainable were more likely to lose interest school. Paris, Lawton, Turner, and Roth (1991) warned that by adolescence low achievers have become suspicious and cynical about tests. Apparently, in their efforts to decrease personal anxiety and protect their own self-esteem, they relinquish both effort and appropriate strategies on standardized
achievement tests. They recommended balancing the psychometric perspective with a psychological perspective on educational assessment. Crooks (1988) made similar recommendations. He concluded that evaluation appeared to be one of the most potent forces in education—it guides the students’ judgment of what is important to learn; affects their motivation and perception of their level of competence; structures their approaches to, and timing of, personal study; consolidates learning; and affects the development of enduring learning strategies and skills. According to Crooks too much emphasis has been placed on the grading function of evaluation and too little on its role in assisting students to learn.

In time, the educational community observed the inclination of the educational system to modify itself in a direction that increased test scores: The practice of testing promoted the method of teaching for the test. Because standardized tests need to be scored unequivocally, the content of what is tested, and what is actually learned at school, is affected profoundly. Items such as “Write an essay about ...”, for instance, do not result in work that can be scored unequivocally; therefore, such items are unlikely to occur on standardized tests, and, as a result, the skills they require become less important. Items that do allow unequivocal answers and content that can be structured in such a way become more important. These tasks, however, hardly represent the kind of skills necessary for life outside of school. Newman (1997, p. 362) complained that “the kind of mastery required for students to earn school credits, grades and high scores on tests is often considered trivial, contrived, and meaningless, by both students and adults. This absence of meaning breeds low engagement in schoolwork.” The educational community began to realize that, if tests should include both useful and higher order skills, measuring students’ performance would be more complex than had been acknowledged by the psychometric movement (see also Cizek, 1997, p. 19; Good and Brophy, 1986). Frederiksen and Collins (1989) indicated that objective tests, which can be scored in an unequivocal way, lead to teaching strategies that emphasize the conveyance of information and to student learning strategies that emphasize the memorization of facts and procedures, rather than learning to generate solutions to problems. They recommend considering the advantages of subjective, direct assessment. With direct assessment, no procedures have been used to transform the objectives of education into questions that only allow unequivocal answers. A task such as “Write an essay about ...” directly addresses essential skills that students need outside of school and promotes the use of higher order skills. Because a completely unequivocal way of scoring of such test items is impossible, the term subjective testing has been introduced, which implies that a large degree of judgment is necessary to evaluate such tests. Concepts such as
direct testing, authentic assessment, and performance assessment refer to the focus on the essential or authentic skills that students need to acquire to be engaged in society, rather than on the kind of answer that can be scored unequivocally. Some authors, however, wonder whether such tests have anything to do with measurement. Shepard (1991) noted that authentic assessment has mockingly been called measurement-free assessment.

When summarizing the evidence for and against the measurement approach of evaluation versus the evidence for and against the pedagogical approach, we arrive at the following conclusion. Evaluation affects the pedagogical process in a way that guides the learning process in the direction of (more or less) relevant learning goals. The more complex the learning goals are and the closer they are to skills that students need in day-to-day life, the more difficult they generally are to stipulate and measure in objective, unequivocal terms. Judging the quality of student performance, however, is important because this stimulates students’ effort. The process of judging, on the other hand, not only affects the quality of students’ work but also affects the students’ perception of competence, possibly in a negative way. This implies that, depending on the situation, the pedagogical goal and the measurement goal of evaluation are intertwined in such a way that they sometimes work against each other and sometimes reinforce each other. It is likely that the tension between the two goals of evaluation becomes more visible when classes contain diverse learners, including those students who have troubles reaching the minimum standards. A separation of data for summative use (which reminds one of the measurement goal) and of formative use (which reminds one of the pedagogical goal), as suggested by the older literature (Bloom, Hastings, & Madaus, 1971), is impossible. Tests intended for summative use only inevitably have a formative impact on students (Paris et al., 1991). Stiggins, Frisbie, and Griswold (1989) observed, moreover, that in the day-to-day practice of teachers such a separation does not exist. If the two goals of evaluation are intertwined, it is likely that practitioners will look for a balance between the pros and cons of the two approaches.

Our study investigates the perspectives of 25 Dutch secondary school teachers regarding evaluation. Due to the existence of national standards, (classroom) evaluation has a measurement function. We focused on whether teachers also acknowledge a pedagogical goal of evaluation. After describing the teachers’ perspectives, we discuss whether this practice is appropriate from a pedagogical perspective. Next, we analyze whether this practice is appropriate from a measurement perspective, particularly when viewing measurement from a developmental and interpretive perspective. In so doing, we show whether practitioners balance the advantages of the two approaches.
METHOD: COLLECTING AND ANALYZING NARRATIVES

The work reported here is part of a larger study about the perspectives of teachers regarding the teaching of diverse learners. We believe that to develop an ecologically valid theory the perspectives of practitioners should be taken into account (Brown & McIntyre, 1993; Cochran-Smith & Little, 1990; Elbaz, 1983; Feiman-Nemser & Floden, 1986; Fenstermacher, 1994; Hargreaves & Fullan, 1992; McDonald, 1988; Shulman, 1986; Verloop, 1992; Wardekker, 1989). To discover the teachers’ perspectives, we had conversations (Florio-Ruane, 1991) with them. Narrative is an important means of capturing the richness and indeterminacy of teachers’ experience and the complexities of their understanding. It offers ways to ascertain which questions teachers ask and the frameworks for understanding they use to improve their own classroom practice (Carter, 1995; Cochran-Smith & Little, 1990; Connelly & Clandinin, 1990; Doyle, 1997).

Noddings (1986) emphasized the importance of collegiality in research on teaching. However, in research–practitioner relationships, teachers have long been silenced by being used as objects for study. Therefore, we realized that giving teachers a voice is an active process to which the setup of our study had to contribute. We considered role taking as an essential criterion for the objectivity of qualitative research, both as a process of self-insertion in the other’s story to understand the other’s story and as a way of giving the other voice (Connelly & Clandinin, 1990; Elbow, 1986; Smaling, 1990). Our interviewer could relate to the teachers in various ways and had been trained extensively in interview and communication techniques developed in client-centered and behavioral therapy. Moreover, she was a teacher trainer with recent teaching experience. She was familiar with the literature on teaching diverse learners, as well as with the doubts of many teachers concerning formal decontextualized knowledge on this subject. Compared with the average age of the interviewed teachers, she was relatively young. In this way, she represented the beginning teacher who was expected to learn from the practical knowledge of the more experienced teachers. Thus, the interviewed teachers were pictured as more experienced and were also addressed as such in our letter, in which we invited the teachers to tell their story about teaching diverse learners to help us develop a course on this subject.

To offer the teachers the maximum opportunity to approach our subject from their own angle, we chose an interview format that was as open as possible. However, we felt that we needed to define beforehand the aspects of teaching that needed to be addressed. For this, we used a very commonly used planning model, consisting of entry situation, learning goals/content, learning processes, lesson format, media, and evaluation (the didactical analysis or DA model; De Corte, Geerligs, Lagerweij, Peters, &
Vandenberghe, 1981). While giving the teachers the opportunity to tell their own story about teaching diverse learners, the interviewer made sure that evaluation (and the other topics) were addressed in the language of practitioners (Clark & Yinger, 1978; Peterson, Marx, & Clark, 1978; Shulman, 1980; Yinger, 1977; Zahorik, 1975). If the conversation did not turn to evaluation spontaneously, she explicitly asked whether the diversity of learners within the class played a role in the way in which evaluation took place. For the rest, she primarily listened to the teachers' stories. Moreover, she invited the teachers to give concrete form to their experience and to comment on the experiences of other teachers. She tried to avoid suggesting that evaluation should have some pedagogical goal or that selection should take place in a proper way. This is why she did not bring up the subject of balancing the two goals directly but asked how individual differences played a role in evaluating students. Moreover, rather than steering the conversation in the direction of the opinions of the teachers, she tried to focus the conversation on their actions, their way of working.

We selected schools randomly from those to which our teacher education students are sent for practice training. The selected schools contacted us with teachers of English and biology. We did not set any teacher quality constraints. Apart from the fact that these qualities are hard to measure, we believe that educational theory should not only represent teachers' successes but also their problems. However, because we did not want our information to be overshadowed by the typical problems of beginning teachers, we only interviewed those who had more than 4 years' experience. For the rest, our only restriction was that the teachers taught classes in the first years of secondary school in which student ability is mixed to a certain extent. In all, 25 teachers participated in our project, 9 teachers of biology and 16 teachers of English. Schools that attracted many rural students, schools in medium-sized cities, and schools with inner city students were represented in our sample.

Before meeting the teachers personally, we sent each a letter in which we explained that because practical knowledge tends to be tacit (Feiman-Nemser & Floden, 1986; Schön, 1983) it would be valuable if they used a notebook to record some events and reflect on them before the interviews took place. After the interviews, some of the interviewed teachers reacted by saying they were not aware that they knew so much. This showed that the conversation had raised their consciousness and had helped them to make their tacit knowledge more explicit.

All conversations were recorded on audiotape and transcribed. They offered students, teachers, and teacher trainers the opportunity to relate their personal experiences to those of the interviewed teachers. Our colleagues who participated indirectly in our research commented positively on the warm atmosphere and the collegial tone of the
conversations. Thus the criterion that a narrative if it is a good narrative constitutes an invitation to participate and can be read and lived vicariously by others, seemed to be fulfilled (Connelly, 1978; Connelly & Clandinin, 1990; Guba & Lincoln, 1989).

Stories are multidimensional. Our a priori formulated categories were based on the planning model, out of which we focus on evaluation in this study. Even if a remark of a teacher concerned evaluation, however, it often at the same time referred to other aspects of teaching. To be able to grasp the interrelatedness of the various aspects of teaching, we created an outline map of each interview, in which all expressions concerning the main categories, as well as important characteristics of the concrete situation and relations with other aspects of teaching were summarized. This made detailed comparison (of excerpts about evaluation) possible without losing the context of each fragment.

The excerpts about evaluation showed that for many teachers evaluation also had a pedagogical goal. By reading the fragments repeatedly (in their context), we detected that patterns could be formulated about the way in which teachers had acknowledged this goal. Formulating patterns, however, depersonalizes the material. It removes the individual idiosyncrasies that make an appealing story. On the other hand, the formulation of patterns from a whole body of conversations increases its generative power (Wardekker, 2000) or transferability (Guba & Lincoln, 1989) for a larger group of people. (Both concepts are introduced to avoid the concept of generalizability.) Formulating patterns is a creative process, in which the material as well as the inventiveness of the formulator plays a role. It can be compared to focusing a camera lens: One should not get too close to the subject (otherwise only a few details are captured), nor too far from the subject (otherwise any detail is lost; the summary becomes meaningless). Therefore, the photographer (i.e., the person who formulates the patterns) should have a clear idea of the subject of the photo (the processes that the patterns are supposed to summarize). The formulator of patterns should also be familiar with educational theory because a pattern is an element of (prospective) educational theory.

In our study, the patterns presented were formulated by using the outline map, by knowing the context in which a remark belonged, by being familiar with the literature on evaluation, and by checking later on whether the patterns really fitted the original stories.

DUTCH BACKGROUND

The Dutch system of secondary education is based on streaming, meaning that students are assigned to a stream in which all subjects are given on a
designated level. There are four levels. Students leave primary education, which is neither streamed nor tracked, with a recommendation for placement in an appropriate follow-up stream. During the first year(s) of secondary education, it often occurs that at least two streams are merged into one mixed class that splits at the end of the year(s). Completely heterogeneous classes, consisting of four streams, also exist. In the classes of the teachers interviewed in the present study, the second lowest stream is always represented.

Deciding to which stream a student will be assigned begins with a recommendation given by his or her primary school (based on a standardized test, the opinion of the principal of the primary school, who is often the teacher of the final year, or both). The final decision depends on the way in which the student performs during the first years of secondary education. Performance is measured by classroom evaluation (Verloop & Van der Schoot, 1995). Within a stream, the student’s report grades must (on average for all subjects) be satisfactory to be promoted to the next year. Grading usually occurs on a scale from 1–10, with the cut-off score at 5.5/6. During the first years, students with an average of very high grades are usually promoted to a higher level, and students with an average of too low grades repeat the class or move to a lower level. The students’ parents are generally consulted about the placement of students in the right stream.

Each stream has its own final examinations. Exams consist of two parts: The external part of the exam is constructed by the National Institution for Educational Measurement in close cooperation with a team of practicing teachers. Essay components are marked by the students’ own teacher and by another teacher with the aid of a marking scheme supplied by the National Institute for Educational Measurement (Bishop, 1995). The school-leaving exam also consists of an internal exam. Although the external part is the same nationally, the content of the internal examination may vary from school to school and sometimes from student to student. Nevertheless, the dual exams are recognized to represent the same level nationally. After graduation from one level, students have access to the stream at the next level.

Classroom evaluation, partly consisting of authentic evaluation (such as essays, presentations, or reports of experiments), has always been used for decisions such as promotion to the next class and for the internal school-based part of the final examinations. Nevertheless, psychometric voices saying that classroom evaluation should be unequivocal have affected the evaluation practices of Dutch schools. Standardized noncurriculum-based tests, however, are not used in secondary education.

Schools are responsible for assigning students to the right streams. An incorrect assignment to a stream becomes apparent when student grades appear to be exceptionally low (or high). The National Inspectorate
monitors the numbers of students who repeat classes and who continue their study in a higher or lower stream. It also monitors the difference between the average grades for internal and external examinations. If an unusually high number of students fail, the Inspectorate takes action. Recently, the Inspectorate judged that selection in the first years of secondary education takes place in a proper way (Van den Bergh, Peters-Sips, & Zwarts, 1999).

In the Dutch situation, teachers or teams of teachers have a lot of freedom in how they use classroom evaluation. In regular practice, however, tests that are of much importance for the final grade (based on the weighted grade-point average) are generally designed by the subject department. These tests are given to all students and judged with the same norms. This limits the freedom of individual teachers to implement classroom evaluation in their own way. More freedom is allowed for low-stakes evaluations. Report grades are a weighed average of both high- and low-stakes evaluations.

During the conversations, we invited teachers to define the term *diversity* in their own way. Although various teachers referred to issues such as social background or ethnic identity, all teachers primarily elaborated on diversity in terms of difference in achievement. This underlines that in a system in which all students have to meet the same external requirements, cultural differences are framed in terms of whether students have reached the standards. In the present study, when we speak about heterogeneity or about diversity in learning, we thus primarily refer to differences in performance.

**RESULTS**

Our results show that 23 of the 25 teachers interviewed are aware that grades used for selection also have a psychological and pedagogical impact. They acknowledge this pedagogical impact by implementing adjustment strategies to avoid failing judgments. Two teachers, however, did not report using evaluation as a pedagogical instrument. They view evaluation primarily as a measurement activity, concluding the learning process. Three kinds of adjustment strategies to avoid failing judgments (grades below the cut-off score) were mentioned: (1) adjusting instruction to improve student grades; (2) adjusting the way information is gathered to improve student grades; (3) adjusting the way information is judged to improve student grades. Many teachers reported using more than one of the above strategies. Failing grades are the starting point for using evaluation as a pedagogical instrument. For students whose achievement is above the cut-off score, a similar kind of adjustment was hardly mentioned.
ADJUSTING INSTRUCTION TO AVOID FAILING GRADES

Adjusting Instruction (17 Teachers)

If student results are below the cut-off score, or if it is feared they will fall below the cut-off score, some teachers try to help students by giving extra instruction. This means making a second attempt to explain the subject, using a different method to explain the subject, or giving the students another opportunity to work with the subject matter. Feedback on earlier work may be given as well. In six cases, extra practice for students who needed it was embedded in the system, which was based on independent work: After each course, the textbook offered a diagnostic test, pointing out which parts of the curriculum had to be practiced. In addition to this system, four of the six teachers also took personal action. This implies that at least 15 teachers reported taking personal action.

Often, individual instruction takes place while other students carry on independently:

Mrs. Akkermans:

Because of your experience and what you’ve agreed upon with your colleagues about tests you’ve designed together, you know the requirements a student has to meet. If a child is weak, he can be stimulated in a number of different ways. One of my students was frequently absent and regularly missed tests and quizzes. When he had to retake them, he would get low grades. I had his mother tell me about the home situation on the phone. In subsequent classes, I tried to make better contact with him and succeeded. I then spent half of a lesson working with him (the rest of the class had to work on their own and they did!). We agreed that he would work at home to make up for lost time. He does that homework diligently; his grades have improved considerably. The more vulnerable, insecure students need extra positive attention: more turns speaking, compliments, clear agreements on what they are supposed to do, and second chances.

Mrs. Akkermans shows that not only the method of instruction but also the kind of relationship that is built up with students plays a role. The time available for individual instruction during lessons, however, is limited. Therefore, extra instruction sometimes takes place after school.

Mrs. Koning:

I have 29 students in my class. That’s a big class and there are a few weaker students in the group. I make appointments for after school; I get together a
small group. If nothing else, the attention is something they seem to need. There are only three of us in the room, and I’ve noticed that then they have the courage to ask questions, something they might not have the courage to do in a full classroom. I always ask them to describe what they are having trouble with; that in itself is a step towards understanding. Then we use that to get to work or I talk about an exercise; often, that moves too quickly in class for these kids. Yes, I find this to be very worthwhile.

Some teachers use oral turns for individual instruction:

Mr. Bogaard:

If a student gets a failing grade, you have to be able to provide some support. “There’s going to be an oral test soon, try to make up for it there.” You have to let the kids know you’re there for them. And if you give them a turn in class, in a sense, you can explain things while you’re quizzing them.

Teachers have thus found various ways to adapt instruction. The quotes show that teachers view the personal element (building up contact, paying attention to individuals, letting kids know you’re there for them) as important. Their attempts to circumvent failing grades, however, are not always successful, as is illustrated by Mr. Heerma, who is very committed to heterogeneous classes in general and to addressing the diversity of his students in particular, and does not hesitate to invest his personal free time in remedial instruction. Nevertheless, he makes the following remark:

Mr. Heerma:

One of the central ideas at our school is that there are communal goals that all of the students should reach. It’s our goal: we reach a lot of students, but there are always students who just don’t understand. I think you have to take that as a given. I used to have to work extremely hard at math, but despite the effort, I just did not understand it.

Summary of adjusting instruction. Adjusting instruction, thus, takes place while other students carry on independently, after school or during individual turns. The teachers who apply this strategy, although they believe that their efforts can be successful, observe at the same time that it does not bridge all the differences in achievement.

Adjusting Information Gathering (9 Teachers)

The strategy of adjusting information gathering refers to both the kind of information teachers gather (what) and the method of gathering (how). Adapting information gathering occurs particularly in the case of low stakes
evaluations. Stipulating the factors that determine the results of a test is complicated. One such factor is the actual level of difficulty of a test: Even when remaining within the official goals, any one test can be more difficult than another. Often, precise prescriptions about the difficulties of the questions do not exist. This gap is used by teachers.

Mr. Morssink:

You get quite a few students who have trouble keeping up. The ones who get such a kick out of getting a high grade that they just blossom. This really has an effect in the long term. It gives them new drive. I consciously give really easy quizzes every now and again. In that way, those weaker students also get a high grade once in a while. The others get high grades too, but that’s beside the point. The kids (who need the boost) go home with a high grade. They always get bad grades and now, this is great. You hear parents tell you this. The children get a kick out of it and I can imagine they do.

Adjustment of level of difficulty of the teacher-designed test occurs for the whole class. Only for very low-stakes evaluations, such as oral turns that are individual anyway, did we find examples of adjustment of difficulty on an individual level.

Mrs. Tulp:

If I ask them questions orally about their homework, then it is clear that I’m going to ask one student tougher questions than another. If a given student can answer an easier question, I’ll ask a tougher one. But if I were to ask that student a tough question right off, then there is a chance that the student will cave in while I believe that those students should also be given the chance to show what they know.

The level of difficulty is also determined by the amount of content that a student has to master for one test. Frequent testing of small amounts is a way to generate higher grades.

Mrs. Akkermans:

Every child that does its best deserves a pat on the back. If you do your best and it doesn’t work, that’s discouraging. You can’t keep that up. A child that does its best should not be plagued by a constant stream of failing grades. I try to do something about that. There are moments when a good grade can be scored by just studying a little piece of content. If you try to give a child some extra encouragement, “really try to do your best on this one,” then there’s a good chance of the child getting a better grade at that point. There are so many test moments, times that you give small quizzes, that everyone can get a decent grade at those points.
Some teachers vary the timing of performances, giving low achievers the opportunity to see the performances of high achievers and learn from them.

Mrs. Vogel:

If a lesson has to be read and you’ve got five sentences, then I have good students repeat the first two. After that, the weaker students get turns, once they’ve had some examples. Students also have to take turns telling a little bit about themselves, a kind of in-class talk, that they start doing in the third or fourth weeks. I never let the weakest students lead this off; they don’t start giving their talks until well into the year.

By gathering information on achievement related behavior, some teachers foster an attitude of working neatly. This contributes to better motivation.

Mr. Morssink:

Once they’ve finished a given module, they have to hand in their workbook and it’s graded. I consciously do it this way because I know that colleagues have complained that it becomes such a mess, not just among the weaker students but all of the students. This led to the decision that these things be handed in and then graded, even though the grade counts for very little, in order to ensure that students do the work neatly. It’s not much of a measure, but it works. I think they need that; after all, I know I wouldn’t like it if I put a lot of work into something and nobody seemed to appreciate it. That’s not very motivating.

By making the effort component visible in the grades she assigns, Mrs. Vink uses evaluation to improve student learning. She believes that for English vocabulary, it is mostly effort that is required, while for English grammar aptitude might be involved. She splits up the grades for vocabulary and grammar with the intention of gathering information on whether the students have done their best.

Mrs. Vink:

I always split up grammar and memorization. It’s easy for a child to say, I don’t understand it. If the score says grammar 7, memorization 5, you can show this to the parents.

Other teachers apply the notion of multiple intelligences (without mentioning this term). They realize that although the educational system generally takes traditional forms of intelligence into account, some aspects of their subjects (e.g., drawing of microscopic images) require a different kind of intelligence. By attending to more kinds of intelligence, they give a broader range of students a chance to score well.
Mr. Morssink:

Students are also given grades for drawings they do of microscopic images. This has a moderating effect. You are not just testing their intellectual level, but also their observation skills and their ability to commit these observations to paper. Often, an average student will be quite good at this. When I have to advise on their promotion to the next grade, however, I often look more closely at how they did on the tests. I don’t just look at their final grade but primarily—and this is something that becomes quite clear on a test—at their ability to deal with large quantities of information, how they have incorporated this into their own body of knowledge. I believe this to be a good criterion for making such a decision. Not just the average, which is often inflated by grades on drawings and the easier quizzes which are given so that weaker students can get a good grade.

By collecting information twice or in other words, giving second chances, some teachers try to avoid failing grades.

Mr. Heerma:

If they get a failing grade, they haven’t met the basic goals, then they get a chance to redo the test.

Summary of adjusting information gathering. As we see, teachers have found various ways of adjusting information gathering to help low achievers get better grades: putting forth easy questions, frequently testing small amounts of content, grading effort-related behavior or making the effort component more visible, grading diverse aspects of the content to give a broader range of students a chance to achieve good grades, giving students second chances to write tests, and scheduling the test in such a way that low achievers may benefit from the example of high achievers. In general, they remain fair to the whole group by applying the same procedures to everybody. Also, second chances are given to all students who have not reached the minimum goals. As for individual turns, no agreements about which questions should be asked exist: Each turn is different anyway. Teachers make use of the absence of such agreements to give pedagogics a chance.

Adjusting Judgment (16 Teachers)

To reduce the amount of failing grades, some teachers adjust the way in which student performance is judged. The most obvious way of adjusting judging procedures is simply liberalizing the norms—for example, counting three rather than two spelling mistakes as one failing point.
Mr. Visser:

You want to regularly give children the impression that they can do better than a failing grade. So you give them a decent grade. Of course, this means that if you want to remain fair to the group as a whole, the other students are going to wind up with even higher grades. I have to keep this in mind later in the school year. You have to keep in mind that you don’t promote children into a grade or level they cannot yet cope with scholastically. So you are constantly correcting your grades. At least I do; I don’t know how my colleagues do this. I can’t recall that we’ve discussed this, but I think that, basically, that’s how it works.

You grade in a range from four to nine, something like that, or if it’s a written quiz, five to eight. Or sometimes in a range from the passing grade to a nine or a ten. It depends on whether I feel it is appropriate, whether it serves a real purpose.

Although adjusting judgment might be intended to help a few students pass, teachers believe they should remain fair toward the rest of the group. Therefore, the new norms are generally applied for all students. We rarely found examples of teachers who just gave good grades to just some individuals (which does not imply that it never occurs).

Adjusting judgment also occurs at the school level. In connection with the merging of streams, teachers and school boards have found a way to bring more students of different abilities above the cut-off score. Therefore, the traditional scale of 1–10 has sometimes been stretched.

Mrs. Vink:

We use a 1 to 14 grading system. From 10 to 14 is defined as the highest level, 8 and 9 as HAVO the higher intermediate level, and below that are the MAVO the lower intermediate grades. In other words, you can get a twelve on your report card. Grandparents find this extremely confusing. I have to admit, it took a bit of getting used to on my part that students could score a 14. But we started using the system when the MAVO was added.

Another (similar) policy, which has also been implemented at the school level, entails judging on both basic and extra levels: Students who are expected to qualify for higher streams are subject to more rigorous norms or are often also given more difficult parts of the test, or both. Then two grades are assigned: one at the basic level and one at the extra level. Low achievers are told to look at the basic grade and ignore the insufficient grade at the extra level. Sometimes on a school level, it is decided that the extra grade will no longer be given to certain (groups of) students.
Mr. Winter:

In principle, the marking of the tests starts from the highest level. By means of a conversion table, the mark can be converted into a (higher) mark, belonging to a lower level. If pupils consistently have low scores, we decide, at the school report meetings or during class consultations, henceforth only to give (higher) marks, belonging to a lower level.

The policy of grading on different levels, however, is denoted as cosmetic by some teachers. They indicate that students are aware that their higher grades are artificial and do not believe that such marks really encourage students.

Although some tested aspects of content are obviously correct or incorrect, other aspects of content require judgment. English spelling mistakes, for instance, can be counted, but no procedures are available to quantify the quality of a student’s report about a novel that he or she has read. Some teachers make use of the absence of procedures indicating what is right or wrong and fill this gap with pedagogical goals. Mr. Langen refers to this practice as subjective grading:

Mr. Langen:

What should be prevented at all cost, especially in the first year, is children failing lots of tests. It is my contention that everybody who does their best, more or less, should score a pass. If you want to keep children working, you should not constantly give them unsatisfactory marks, because they are bound to give up sooner or later. There’s a couple of things you can do to prevent this. Marks can easily be manipulated. You can, for example, ask them to retell stories. I never do that in front of a class; I have a few children come to me. If you can see that a child has done its best, you can simply give a pass.

To circumvent a negative judgment, some schools have replaced the 1–10 scale with symbols. Two teachers (working in complete heterogeneous groups) judge effort rather than achievement. Their approach is unusual in Dutch schools because it is an attempt to downplay the selective function of evaluation. Although most teachers work with minimum standards that are (indirectly) related to the external standards, the grades assigned by these two teachers are not related to any objective achievement. This kind of grading only occurs during the first year of a 2-year period in which classes are completely heterogeneous (the four streams are merged). During the first year, placement decisions need not to be taken. Grading student efforts, however is not without problems.
Mrs. Wolf:

After a test has been completed and graded, it is discussed with the whole class. Then they are told which answers were possible. At the bottom of the test, they are given a grade: good, pass, or fail. That is aimed at each student individually. It could be that a student comes to me, and at the start that sometimes happens, and says, I counted and I see that I made more or fewer mistakes than so-and-so and I got a Pass while he got a Good. And I tell them that they can do better, so I think they did not perform as well as the other student.

Teachers who grade effort, however, are not always adequately informed about the amount of effort the students have exerted. Mrs. Wolf continues,

And then I say, “Yes, I think you can do better; I think you did worse, given your potential, than that other student.” If that is really the case, then the student usually doesn’t say that much, or just says, “Yeah, you’re right.” If I’m among them and I notice that a student perceives it as being extremely unfair, I respond by saying, “If you find this unfair, then what you are saying is that you probably can’t do much better than this. It could be that I was wrong but I am only trying to get the best out of you.”

Mrs. Wolf shows that students who compare their work to other students’ are able to see through the mark and still become informed about the objective value of their results. While trying to downplay the selective function of evaluation, she cannot avoid that a positive judgment on a student’s work can be interpreted as a negative judgment on the student’s ability, which implies a selective function. Grades retain a selective function that directly or indirectly refer to an objective judgment (for a certain stream). The selective function remains so when teachers use symbols instead of grades or grade effort. This illustrates that the selective function of evaluation is unavoidable. Even when heterogeneous classes have been formed, the differences in performance cause the borders of the old streams with accompanying achievement levels to remain visible. Selection does not disappear with the integration of streams. Students select themselves by their level of performance, as De Koning (1988) noticed.

Summary of adapting judgment. Teachers encourage students by adjusting judgment. They (temporarily) liberalize the norms, use a stretched scale, grade on different levels, grade effort instead of achievement, exploit the absence of standards for authentic tasks, and use subjective grading (which should take place anyway) to support students. At the same time, they
remain fair to the group by applying the same (adjusted) procedures of judgment to everybody. Only if procedures are not available (which is, e.g., the case with grading novel reports), they fill this measurement gap with pedagogics.

**An Example of the Combination of Adjustment Strategies**

Teachers often apply more than one strategy. Eight teachers mentioned only one strategy; 17 teachers mentioned two or three strategies. Mrs. Bouw illustrated how different strategies support each other.

**Mrs. Bouw:**

> The basic program at our school is such that all of the students should be able to pass. If it looks like they are going to fail—which happens sometimes, right, that students can’t even meet the basic demands—then you have to do extra things with those students: give them more opportunities, extra chances to get good grades; you have to set really minimal standards. Not too low, of course, because then they might get lazy. I try to make sure that students like this get passing grades. I’ll give them a quiz that is quite easy, and they get better grades. I’ll write something on the paper like “well done” or “keep up the good work,” some kind of encouragement. Sometimes, I go to one of these students, because they often think after they’ve tried a number of these tasks that they can’t do it, and I tell them, “You do it like this, then it will work. You can do it. I am sure you can do it.” That gives them confidence and then they succeed. If you make sure that they get a few good grades, that spurs them on. If students consistently get bad grades, well, those grades just keep getting worse, so you have to make sure that they are somehow stimulated. An easy quiz is something I give to the whole class, otherwise you just wind up stigmatizing people. A test on the basic material is equally easy for the whole class. Everyone can get a high grade and the extra material is more difficult. The weaker students won’t get a good grade on that part of the test, but that’s not crucial, as long as they pass the basic tests.

> I sometimes tell one of the weaker students who has failed a number of tests to really take a good look at the homework assignment, because I could ask him or her to work out a problem on the board in class. That gives them a hint and they tend to then study like crazy. I make them work the problem out on the board and then sometimes they get 100% on the quiz. They get a kick out of that. I help them in that way, once in a while.

This teacher thus adapts instruction, way of information gathering, and way of judging evaluation.
REFLECTION ON THE ADJUSTMENT STRATEGIES FROM A PEDAGOGICAL PERSPECTIVE

We have shown that teachers generally apply three kinds of adjustment strategies to reduce the number of failing grades: adjusting instruction, adjusting information gathering, and adjusting judgment. The reported purpose of all these strategies is pedagogical: Teachers want to keep up student motivation, to encourage students to achieve better, to avoid disillusionment, and to build self-esteem.

In various cases, the adjustment strategies represent ways to challenge students in an active way. By using the strategies, teachers try to drag students across the finish line, so to speak. In this respect, teachers not only acknowledge (see, e.g., Phye, 1997) but also exploit the function of evaluation in stimulating student effort. In other cases, students are challenged in a passive way. The standards are formulated in such a way that too many failing grades are circumvented. In this respect, teachers appear to acknowledge what Crooks (1988) and Gagne (1977) found, namely, that evaluation affects student motivation and their perception of competence and that self-efficacy in a particular domain is built by repeated success on tasks. Perceptions of self-efficacy appeared to have a strong influence on effort and persistence with difficult tasks or after experiences of failure (Bandura, 1982; Schunk, 1984, 1985). Thomas, Iventoch, and Rohwer (1987) found self-efficacy to be a better predictor of school achievement than the selected measure of academic ability. The 23 teachers who use adjustment strategies acknowledge these pedagogical aspects of evaluation. They do not temper justice with mercy, but they temper justice with pedagogics.

Now that we have concluded that adjustment strategies can improve the pedagogical quality of classroom evaluation, we can rank the pedagogical quality of classroom evaluation across a spectrum (see Figure 1). At one extreme, labeled None, evaluation has no pedagogical goal; it only plays a role in selection decisions, which is the case for only 2 out of the 25 teachers. At the other extreme, labeled Active, evaluation is used to actively challenge low-achieving students to improve their levels of achievement by using adjustment strategies. The teacher whose way

![Figure 1: Pedagogical quality of classroom evaluation in the Netherlands on a spectrum.](image-url)
of combining various adjustment strategies we mentioned is an example of such a teacher. The point labeled Passive represents a situation in which students' actual results are accepted, while students are not discouraged by failing grades.

**REFLECTION ON THE ADJUSTMENT STRATEGIES FROM A MEASUREMENT PERSPECTIVE**

Are the adjustment strategies also appropriate from a measurement perspective? We first take a close look at them, then present a view of measurement and answer the question whether the adjustment strategies are appropriate from this measurement perspective.

The first adjustment strategy concerns instruction, which, as a part of the learning process itself, is usually seen as being separate from evaluation. That teachers adjust instruction to improve student results affirms the function of measurement in fostering better results.

Viewed from a measurement perspective, the other two adjustment strategies—adapting information gathering and adapting judging—are more problematic. Within a psychometric approach, the instrument of measurement should be constructed beforehand, and it should be clear how the results are judged. Adjustments are not appropriate, particularly not when such adjustments take place to flatter the results. The frequent use of tests on small amounts of content, for instance, could in itself fit within a measurement approach, but that it occurs on the spot to improve the test results is contradictory.

In classroom evaluation, however, no formally defined reference points exist, neither for the way in which information should be gathered nor for the way in which the results should be judged. The construction of classroom evaluation tests requires an estimation of its level, amount of content to be tested, the appropriate moment, and an appropriate way of judging. Teachers who use adaptation strategies fill this gap with pedagogics.

Evaluation generally does not occur on an individual basis, except if the tasks are different for each individual anyway (such as turns or book reports). After the test has been constructed (and the pedagogical goals have been implemented), the same test is given to every student. If procedures (such as counting the amount of spelling mistakes) can be used for judging, the same procedures are applied for everybody, although they may deliberately have been developed in such a way that failing grades are circumvented for just a few students. Thus, teachers who use adjustment strategies remain fair to the whole group by giving the same tests and by applying the same judgment procedures to everybody. Therefore, it is unlikely that students notice the use of adjustment strategies.
If a formal reference point does not exist, however, can we speak about measurement, or is the method of classroom evaluation totally arbitrary? Mrs. Akkermans gives an answer.

Mrs. Akkermans:

Because of your experience and what you’ve agreed upon with your colleagues about tests you’ve designed together, you know the requirements a student has to meet.

Also Mr. Visser relates the test and the related scores to his knowledge about the requirements that students must generally meet.

Mr. Visser:

You want to regularly give children the impression that they can do better than a failing grade. So you give them a decent grade. Of course, this means that if you want to remain fair to the group as a whole, the other students are going to wind up with even higher grades. I have to keep this in mind that you correct this later in the school year. You have to keep in mind that you don’t promote children into a grade or level they cannot yet cope with scholastically.

And Mr. Morssink also compares student achievement on the test to his knowledge about the requirements that must be met.

Mr. Morssink:

Students are also given grades for drawings they do of microscopic images. This has a moderating effect. You are not just testing their intellectual level, but also their observation skills and their ability to commit these observations to paper. Often, an average student will be quite good at this. When I have to advise on their promotion to the next grade, however, I often look more closely at how they did on the written tests. I don’t just look at their final grade but primarily—and this is something that becomes quite clear on a test—at their ability to deal with large quantities of information, how they have incorporated this into their own body of knowledge. I believe this to be a good criterion for making such a decision. Not just the average, which is often inflated by grades on drawings and the easier quizzes which are given so that weaker students can get a good grade.

These teachers appear to have some kind of framework that functions as a reference. They pretend to know which level students should reach to be able to be promoted to the next year, and they use this knowledge while constructing tests. It is very likely that this knowledge has developed along with the external standards of the final year. Starting from these standards,
the level of the first years is gauged. The final standards are also the basis of
the discussions with fellow teachers about tests; these tests, after all, should
check whether students are well on their way to reach the standards. The
final standards are also a reference point for deciding which cognitive skills
should be emphasized in tests at a certain point during the learning process.
During their career, teachers observe numerous students progressing
toward the final standards. They experience, for instance, that the scores
improve when only small amounts of context are tested; however, during
the final year they also experience that students must have developed the
ability to deal with much larger quantities of information. In other words,
the teachers’ point of reference for measuring student results does not lay
in an instrument but in a culture or a tradition that has been developed
along with the external standards: They have a historical reference that
consists not only of standards but also of experience about how students
have progressed toward these standards.

Interestingly, conviction that (new) traditions should be formed can be
found in modern literature on evaluation and testing. Frederiksen and
Collins (1989) believe that subjective direct tests, for which judging is
necessary, are unavoidable for measuring higher order thinking skills.
Basing educational assessment on subjective scoring requires that scorers
become familiar with a library of examples of student work representing
different levels of the desired traits. Along with these examples, assessors
should practice scoring until they have internalized the criteria. In this way,
the assessment system provides a basis for developing a meta-cognitive
awareness of the important characteristics of good problem solving,
good writing, good experimentation, good historical analysis, and so on.
Frederiksen and Collins recommend no longer confining the practice of
assessment to the testing organizations; it should become more decentra-
lized, as teachers and students are taught to internalize the standards of
performance for which they are to strive. Frederiksen and Collins advocate
passing down a framework for evaluating students, consisting of a set of
internalized criteria and a meta-cognitive awareness, which functions
as a reference. They do not advocate a view of objectivity based on
impersonal, unequivocal judgment, but one based on connoisseurship
(Eisner, 1998), which enables teachers to conduct interpretive measurement
(see Moss, 1996).

The teachers quoted previously refer to a similar framework. It is likely
that their direct or indirect experience with the external standards has been
important in constituting this framework. The students who take the final
examinations supply indeed a library of examples for which marking
schemes are available. This has helped the teachers internalize the criteria.
Moreover, during the preceding years, teachers observe numerous students
progressing (or not) towards these standards. Referring to their experience,
they believe they understand the requirements that students should meet during the first years. Therefore, they feel free to temporarily deviate from these criteria to encourage students to achieve better.

Does the deviation from the internalized standards—the use of adaptation strategies—interfere with proper measurement? It depends on one’s view of measurement. As we said before, it does not fit within a psychometric approach. A developmental perspective of measurement, however, acknowledges that the results of measurement influence the learning process. Within this view, the aim of the selective function is not to confine students to their present level of development but to contribute to proper selection by the end of the period. A developmental perspective acknowledges that humans are open systems: Their development is an ongoing nonlinear process, sometimes unpredictable. The active use of adjustment strategies can be viewed as on the spot experiments (Schön, 1983) to find out whether students can be encouraged to achieve better. As a result of a temporary lowering of the requirements, for instance, some children may begin to flourish or at least retain motivation. It the teacher knows what he or she is doing, such experiments add to the informativeness of the grades and increase the likelihood of proper placement. In this respect, the use of adaptation strategies reinforces the measurement goal of evaluation for placement decisions.

Had the external standards not existed and had teachers not had a framework of criteria, the use of adjustment strategies would have lowered the standards and would have resulted in improper placement. However, it all takes place in a context in which the final standards are not adjusted. This limits the room for adaptation. Therefore, the use of adjustment strategies does not always result in the actual prevention of failing grades: Students who do not meet the standards on a long-term basis either have to repeat the year or are transferred to a lower stream. This illustrates that, although many teachers have reduced the tension between the pedagogical goal and the measurement goal of evaluation considerably, the tension does not disappear completely. The room for variation that exists, however, is used to implement adjustment strategies.

Our material shows that teachers use some kind of historic reference, both for gauging the appropriate level and the room for adjusting evaluation. Our material is not suitable for judging the quality of the framework that the interviewed teachers use. Whether the framework of all the teachers is appropriate, however, is not the essence of our line of thought. What is important is that external standards do not frustrate the development such a meta-cognitive awareness of criteria of good work but support it. These standards function as a reference, both in judging the quality of work for which higher order skills have to be used and in judging which adaptation strategies are appropriate. Moreover, the external
standards offer a basis for communication with fellow teachers about teacher-made tests and the quality of students’ work.

Concerning the measurement goal of evaluation, we conclude that the use of adjustment strategies is not appropriate from a psychometric perspective. If teachers have a proper framework of criteria, however, the use of adjustment strategies fits within a developmental and interpretive view of measurement. From a developmental perspective, the quality of measurement lies not in refraining from the use of adjustment strategies but in the richness and adequateness of teachers’ historical framework, their awareness of the cognitive skills that students have to develop to finally achieve the external standards, and of their awareness of the way in which evaluation can be used to encourage students to achieve the external standards. From a developmental perspective, the quality of measurement is not separated from the way in which measurement can be used to achieve the external standards. Classroom evaluation offers teachers the flexibility to implement a pedagogical goal in a measurement situation. Teachers who use adjustment strategies play with pedagogical opportunities that are intrinsic aspects of classroom evaluation. Those teachers who do not use these strategies miss a pedagogical opportunity and are not better measurers.

CONCLUSION

Between the pedagogical goal of supporting the learning process and the measurement goal for selection decisions, a tension exists that has given rise to an intense conflict concerning which goal should be given priority. This tends to be an either/or discussion. The teachers’ perspectives showed that an either/or discussion unnecessarily partitions the two camps and fails to reconcile the advantages of both aspects of evaluation.

Our work shows that many Dutch teachers, who work with external standards during the final year and use classroom evaluation during the preceding years, implement adjustment strategies for evaluation to encourage better student performance in the long term. They thus implement a pedagogical goal in a measurement situation. Adjustment strategies enrich the teacher’s toolbox of pedagogical instruments—if used actively, they are an instrument to drag students across the finish line, so to speak—and offer a way to deal with differences.

The use of adjustment strategies, however, can only be done properly if teachers have an adequate meta-cognitive awareness of characteristics of good work and, thus, know what they are doing. The external standards are likely to reinforce the development of such a framework. Teachers use this framework, both for gauging the criteria for classroom evaluation and for
gauging the room for adjustment strategies. The standards during the last years, thus, function as a reference in gauging the room for implementing a pedagogical goal to a measurement situation during the first years.

The external standards, moreover, benefit another pedagogical goal as well: They encourage the function of evaluation to stimulate student effort to reach certain goals. At the same time, external standards encourage a measurement goal because they counterbalance grade inflation. The active challenging of students by using adaptation strategies increases the informativeness of the test results, which benefits a proper selection process. The passive use of adjustment strategies to circumvent failing grades is part of a policy: Scales and judgment policy are adapted to avoid unnecessary discouragement. In such cases, it is clear what the grades stand for. Thus, to some extent, the pedagogical goal of classroom evaluation does not conflict with the measurement goal and sometimes even reinforces it.

We have mentioned the following factors that can reduce the tension between the pedagogical and the measurement goal of evaluation: the use of classroom evaluation during the first years, which allows authentic meaningful tasks and offers room for adjustment strategies; the use of a developmental and interpretive perspective of measurement, which acknowledges that student achievement is influenced by the measurement itself and that connoisseurship is important for judging higher order skills; and the development of a proper meta-cognitive awareness for classroom evaluation that teachers use, both for judging higher order skills and for gauging the room for adaptation strategies and for gauging how evaluation can be used to support the learning process of students. External standards are likely to be an important reference in the forming of this framework.

These factors, however, cannot prevent that between the pedagogical goal and the measurement goal some tensions remains. All qualitative or quantitative criteria require decisions to be made about whether the standards are reached or not. The adjustment strategies can only be used on a temporary basis. They may encourage student achievement in the long term; however, they do not guarantee that all students will actually reach the minimum goals. In the context of national standards and within classroom evaluation, the two goals of evaluation are compatible, albeit to a limited extent.

PRACTICAL IMPLICATIONS

Our culture has a strong preference for objective, unequivocical methods of representation. The educational community should teach society that the evaluation of meaningful and higher order skills requires judgment rather than calculation.
At teacher training colleges, student teachers should understand the purposes of the two goals of evaluation and the ways in which they intertwine, reinforce, or work against each other. Authentic evaluation offers a way to direct the learning process toward the development of meaningful and higher order skills. Classroom evaluation requires teachers to make decisions on the factors influencing the outcomes: the amount and type of instruction needed before evaluation takes place, the way in which information is gathered, and the way in which information is judged. This offers an opportunity to take the concrete situation in the classroom into account, to temper justice with pedagogics. Student teachers should become aware that there is some room for playing with the factors of evaluation also when external standards exist. This playing should occur consciously to foster the learning process, while acknowledging that the room for playing is limited.

To be able to gauge the limits for playing and to evaluate higher order skills, student teachers need to develop an adequate framework. Qualitative and quantitative external standards are likely to be an important reference in the development of such a framework. Formal criteria only, however, do not suffice. By studying a library of examples in the context of external standards, student teachers can learn to understand the criteria for judgment and the required meta-cognitive skills. Classroom evaluation requires a culture in which teachers understand what can be expected of a student, given his or her stage of development. During teacher training, students should become familiar with this culture. Participation in the marking of external exams is important. Discussions with more experienced teachers about criteria will sharpen the sense of what the external standards stand for. Connoisseurship and a sense for the room for adjustment, thus, develop in real settings, rather than in courses in which teachers are taken from their natural setting. This shows the primacy of experience over the command over formal knowledge and abstract procedures (Eisner, 1988).

National standards encourage the body of criteria to be shared within schools, between schools, and to form a tradition. Nevertheless, it is not clear whether the framework of the Dutch teachers interviewed in this study is appropriate enough. As our material shows, teachers have discussions with fellows about the requirements of some tests. However, the use of adaptation strategies is not always discussed openly, as Mr. Visser says.

Mr. Visser:

So you are constantly correcting your grades. At least I do; I don’t know how my colleagues do this. I can’t recall that we’ve discussed this, but I think that, basically, that’s how it works.
Dutch educational experts and school boards, like many abroad, encourage the formalizing of subjective judgment into objective judgment. Instead, they should encourage teachers to develop a proper framework. Open discussions about adaptation strategies are necessary.

Finally, our study has implications for the relationship between researchers and practitioners. As we showed, authentic evaluation traditionally exists in the Netherlands. Reflection on this kind of assessment, however, is lacking. Many researchers assume that teachers do not implement pedagogical goals in their evaluation practice (Alberts, 1987, 1991; Kremers, 1984; Verloop & Van der Schoot, 1995). This assumption arose because teachers appeared not to implement a measurement-driven approach. None of these studies, however, took the perspectives of teachers into account. They failed to discover that teachers had their own ways of implementing the pedagogical goal in their evaluation practice. Meanwhile, the measurement-driven approach has become obsolete (Shepard, 1991). Although the practitioners are often characterized as traditional, our research shows that many of them implement a view of the relationship between the pedagogical goal and the measurement goal of evaluation that concurs with modern research literature, even before that literature existed. This underscores the importance of integrating teachers’ experience into the formulation of theory.

References


JACQUELIEN BULTERMAN-BOS is completing a dissertation about teaching diverse learners. Her research interests include the relation between theory and practice in education. She is a former teacher and a teacher trainer. Her recent publications include “Observation in Teaching: Toward a Practice of Objectivity,” Teachers College Record, 104(6), with J. Terwel, N. Verloop, and W. Wardekker.

NICO VERLOOP is professor of education and dean of ICLON Graduate School of Education at Leiden University, The Netherlands. He is immediate past president of the Dutch Educational Research Association. His research interests include research on teacher education and teacher cognition. His recent publications include “Student Teachers’ Beliefs about Mentoring and Learning to Teach during Teaching Practice,” British Journal of Educational Psychology, 71, with A. Zanting and J. D. Vermunt.

JAN TERWEL is professor of educational psychology at the Faculty of Psychology and Education, Vrije University Amsterdam. His research interests include adaptive learning and instruction, social interaction and group composition. His recent publications include “The Effects of Integrated Social and Cognitive Strategy Instruction on the Mathematics Achievement in Secondary Education, Learning and Instruction, 9(5), with D. Hoek and P. Van den Eeden.

WIM WARDEKKER is assistant professor at the at the Faculty of Psychology and Education at Vrije Universiteit. His research interest is in developing theories and practices of education based on a sociocultural viewpoint. His recent publications include “Schools and Moral Education: Conformism or Autonomy? in the Journal of Philosophy of Education, 35(1).