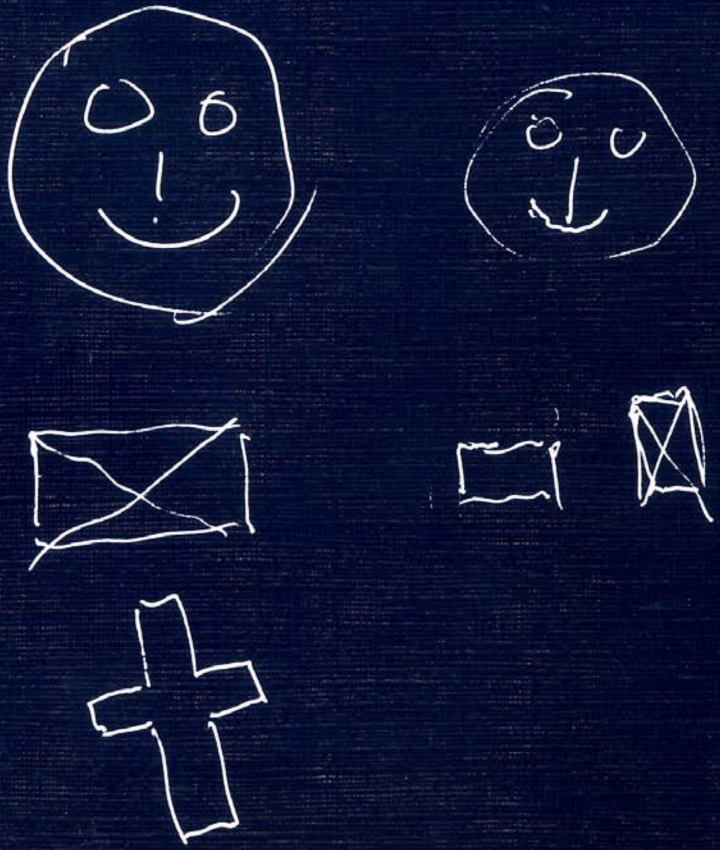
M. van den Heuvel-Panhuizen

ASSESSMENT AND REALISTIC MATHEMATICS EDUCATION



ASSESSMENT AND REALISTIC MATHEMATICS EDUCATION

The cover illustration is by Hans Freudenthal, drawn a few days before his death, in October, 1990. He made the sketch during a discussion on how to assess ratio with special education students. Freudenthal’s drawing emphasizes where the assessment of ratio should begin: at the visual roots.

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Heuvel-Panhuizen, Marja van den

Assessment and realistic mathematics education / Marja van den Heuvel-Panhuizen.
- Utrecht : CD-β Press, Center for Science and Mathematics Education. - (CD-β wetenschappelijke bibliotheek ; 19)
Also publ. as thesis Universiteit Utrecht, 1996. - With ref. - With summary in Dutch.
ISBN 90-73346-27-4
Subject headings: primary school education / mathematics education ; assessment.

Cover: OMI Utrecht University
Translation: Ruth Rainero, San Francisco
Hetty Olthuis Vonk, Naarden (first translation of Chapter 6)
Press: Technipress, Culemborg

ASSESSMENT
AND
REALISTIC MATHEMATICS EDUCATION

(with a summary in Dutch)

Marja van den Heuvel-Panhuizen

Also published as a doctoral dissertation at Utrecht University, The Netherlands

Supervisors: prof. dr. A. Treffers
prof. dr. J. de Lange
Co-supervisor: dr. K.P.E. Gravemeijer

For Gerard and Geert

Preface

This dissertation is the fruit of a number of years of research and development work involving assessment in Realistic Mathematics Education. Characteristic of how this book came about is that it did not *begin* as a dissertation but, rather, eventually *became* one. Freudenthal would have called it ‘scientific residues’, as this was how he viewed our work at the institute that came to bear his name.[1]

The roots of this dissertation are deeper than its actual inception. It all began with Fred Goffree, who opened my eyes to Realistic Mathematics Education. Without him, I never would have met Hans ter Heege, and seen how he inspired children to think. And then there was Koeno Gravemeijer, who asked me to work on the MORE project, which is how I ended up at what was then called OW&OC. It is the work at this institute (now called the Freudenthal Institute) that has brought me to where I am today. It is rare to find so much inspiration packed into just two floors of one building.

The first concrete stimuli towards this dissertation occurred in the autumn of 1989. It was at that time that Freudenthal acquired the moniker of my ‘new secretary’. He had been found sitting at my computer, making some corrections in an article I had written on the MORE tests, which he had translated into English. Later, while munching a pastry, he remarked in passing that I should turn the article into a dissertation.

A unique trio provided support during the writing: Adri Treffers, Jan de Lange and Koeno Gravemeijer. Together, they formed the rich context, in which I could develop certain insights, and gave me the right help at the right moment to be able to elaborate on these insights.

The inspiring discussions I had with Leen Streefland in this respect deserve mention here as well. And I would also like to thank Marianne Moonen for her continual assistance in tracking down various books and articles.

Finally, the day arrived that the dissertation was finished. But finished is not necessarily done. Ruth Rainero then translated it into English – and sometimes did more than that as well. She was namely also my first critical reader from outside the Freudenthal Institute.

While the dissertation was being translated, there was still plenty to take care of within the institute. The sense I had already had of involving the entire institute, because of the three supervisors, only became stronger as the book neared production. Many people contributed in one way or another: Han Hermsen and Wim van Velthoven in providing computer support, Ada Ritzer in scanning the illustrations, Ank van der Heijden, Ellen Hanepen and Betty Heijman in taking care of assorted tasks, and Els Feijs and Martin van Reeuwijk in helping solve all sorts of problematic translation issues. My very special gratitude goes to Sylvia Pieters, as she deserves most of the credit for the production of this book.

Finally, I would like to thank the home front: my parents who, in the past, offered me the possibility to study, and all my friends and relatives who, in the recent years, gave me the room to immerse myself somewhat obsessively in the work on this dissertation. Especially, I think here of my friend Thea and the mental support she gave me.

Most of all, I would like to thank my son, Geert, who grew up with the test problems, and my husband, Gerard. Both experienced this work at very close quarters. Too often, I was not around. Fortunately, they were always there.

Note

[1] See Freudenthal, 1987b (p. 13), where he quoted from a lecture given in 1973: "We of the IOWO regard ourselves as engineers [...]." According to him, this meant that "... we are making something – something that requires a scientific background, but that is itself not a science." Later, he added to this: "However – chopping wood produces chips, and the practical work of engineers may eventually also provide you with scientific residues."

Table of contents

Introduction

Part I

Part II

Introduction

Overview

The history of mathematics education reform in The Netherlands, begun in the late nineteen-sixties, would seem to bear out McLean's (1990, p. 12) statement that

> "Education is slow to change, but testing is slower."

As far as the first half of the statement is concerned, the events in The Netherlands clearly show how lengthy the process of realizing a new instructional approach can be. For approximately the last 25 years, much time and effort has been invested in the development and implementation of Realistic Mathematics Education (RME). As of today, this process is still unfinished. Certain aspects of the curriculum still require further examination, and a number of important questions yet remain with regard to the implementation of RME in educational practice. Consequently, the RME theory, which includes the overarching and connecting principles that guide decisions in classroom practice, is still in a process of development as well.

The educational situation in The Netherlands confirms the second half of this statement as well. For many years, anything having to do with assessment remained unchanged within the educational reform process. While the existing tests were rejected, it took a great deal of time – at least, in primary education – before systematic attention was paid to developing alternatives to these tests. The dissertation research of De Jong (1986), for instance, which made a detailed examination of the degree to which all sorts of RME characteristics could be found in the various primary school mathematics textbooks, did not include a different method of assessment among these characteristics.

Without intending any discredit to all the past research that has been conducted with respect to assessment in RME, it may be stated that the study at hand is, in a certain sense, the first to discuss the implications of RME for assessment, in any case in primary education.

Clearly, the situation in secondary education has not been the same. More or less obliged to do so by legislatively required exams, the development of assessment appropriate to RME was begun in the early nineteen-eighties, simultaneously with the reform of the secondary education mathematics curriculum, in order to secure the desired curricular changes. This secondary mathematics education reform was carried out in the framework of the HEWET project, which was established for this purpose. The dissertation of De Lange (1987a) gives a detailed account of the results of the related assessment research. Later, alternatives to the existing methods of assessment in primary education were also sought along the same lines.

Just as the HEWET project heralded developmental research on assessment for secondary education, so did the MORE research project play an important role in similar research for primary education. This project was in fact a study of the implementation and effects of mathematics textbooks (see Gravemeijer et al., 1993). The development of tests, which were needed in order to compare the learning achievements of students who had used different textbooks, gradually began to expand and to focus more on the implications and possibilities of assessment within RME. In this respect, the MORE research also became a field of exploration into RME assessment. The ideas and discoveries that emerged from this research eventually led to the present study. This study, which is based on assessment development, a study of the literature, and reflection on both, hopes to make a further contribution to the development of assessment within RME by highlighting various theoretical concepts and providing concrete examples.

Even though it is this study's intention to give a comprehensive survey of RME assessment, by no means has each and every aspect of assessment been covered. Certain choices were made right from the start. As mentioned above, the principal focus of this study is the *assessment of mathematics in primary education*.

Another self-imposed limitation is that the study chiefly covers what is described here as '*didactical assessment*'. This is assessment that is intended as a support to the teaching and learning process. It is closely linked to the instruction and, in principle, is part of the daily educational practice. This is what distinguishes didactical assessment from the kind of assessment that focuses on classifying students and evaluating education. It should be noted, to avoid misinterpretations, that the decision of this study to specifically examine didactical assessment does not imply that other purposes of assessment are rejected in RME. Both the HEWET project, which undertook, among other things, the development of a new method of administrating exams, and the MORE research, which concentrated on educational evaluation, attest to this.

A third limitation is the decision to focus on *paper-and-pencil tests*, and on *short-task problems* in particular. Although this might seem at first glance to contradict the tenets of RME, it will become clear as the study progresses that such tasks can indeed be compatible with the fundamental standpoints of RME, and can even become a source of inspiration for its further development.

By the late nineteen-eighties, the time was ripe in The Netherlands for a fresh examination of the implications of RME for assessment in primary education. Moreover, an international reform movement in the area of assessment was building up steam at the same time. However, as these developments outside The Netherlands occurred more or less concurrently, and the publications on this issue were few and far between, these new viewpoints on assessment from outside the RME circle did not in fact play a role on the developmental research on assessment that was conducted within RME. This does not mean, however, that the present study did not take these international developments into account. Since that time, numerous articles have been published on this topic, with 1992 being the most prolific year to date. Therefore, with a kind of hindsight, this study has, wherever possible, made links between international developments and the RME viewpoints on assessment, the objective being either to support, enrich or adjust the latter.

Outline

The book that has resulted from this study comprises seven chapters, the first four of which constitute part I of this book and are the core of the work. These four chapters discuss, in the following order: (1) the role of assessment in the early stages of RME, (2) the development of tests within the MORE research, (3) the current state of affairs with respect to assessment in RME, (4) the potential for written assessment in RME. The three chapters that follow in part II should in fact be regarded as appendixes. They involve, respectively: (5) an arithmetic test that was administered at the beginning of first grade, (6) a ratio test that was administered in special education, (7) (part of) a test on percentage that was administered to middle school students.

- Part I

Although, as stated above, the present study is the first in the area of primary education to devote specific attention to assessment within RME, it is pointed out immediately in *Chapter 1* that there was no need to start from scratch. This first chapter retraces the early years of the mathematics education reform movement in The Netherlands, and provides a survey, based on a study of the literature, of how assessment was regarded at that time. While the emphasis in this chapter is chiefly on primary education, secondary education is discussed as well.

The chapter begins by briefly elucidating the predominant characteristics of RME. This is followed by an examination of the place of assessment within RME that covers the period from the late nineteen-sixties, at the beginning of the development of RME, up through 1987. How assessment within RME developed further after that date is discussed later in the book.

The story of assessment within RME during the early years was primarily one of anti-assessment. That is to say, one may easily acquire this impression from the often fierce campaign waged at that time against the existing tests. Upon closer examination of contemporary publications, however, it becomes clear that, alongside opinions on what *not* to do, there were also some very clear ideas in those early years about what *should* be done. These ideas, together with the RME alternatives to the existing tests that were developed at that time, can be regarded as the basis for the further development of assessment within RME.

Chapter 2 focuses on the MORE research. The development of tests that took place during this research project provided an important stimulus for the further development of a theoretical framework for RME assessment. Although the MORE research only specifically required the development of evaluative tests, the experiences gained from these tests led to a renewed reflection on the significance of assessment for mathematics education. In Chapters 3 and 4, the results of this are discussed in detail. Chapter 2, on the other hand, concentrates more on how the foundation for the further development of assessment within RME was laid in the MORE research project. Among other things, an account is given of how the developmental research on assessment in this project gradually evolved into an independent project with its own research issues. In addition to describing the basic principles of the test development and the methods involved, Chapter 2 also devotes attention to certain crucial events that occurred during the test development, and which served more or less to determine its path. These include the surprising results of particular test problems, new ideas and discoveries, and evolving theoretical concepts. The concluding section of this chapter takes a closer look at the development context as a determinative factor for the test development in the MORE research. The chapter begins with a brief summary of the MORE research, in order to provide a backdrop for the test development.

Chapter 3 provides a general orientation of the present state of affairs within RME and constructs a theoretical foundation for RME assessment. The intention of the chapter is to present a further elaboration of the RME theory for assessment. The concepts that were developed during the early years of the RME educational theory with respect to assessment in primary education, as well as De Lange’s work on assessment in secondary education provide the cornerstones for this elaboration. Because RME is in a constant process of further development, this chapter should be regarded as a work in progress, based on current ideas and views within RME.

The beginning of the chapter deals briefly with the transition with regard to assessment that was made over the years within RME, and with the current need felt both within and without RME for a theoretical foundation for assessment. This is followed by a detailed description of the two determinating characteristics of assessment in RME: its ‘didactical’ nature and the crucial role played by the problems used in the assessment. Specific attention is devoted here to the role in assessment problems played by the context. The second half of Chapter 3 regards from a broader perspective the topics that were handled in this general orientation. The RME views on assessment are then, in a certain sense, held up to the mirror of international assessment reform.

In *Chapter 4*, as a supplement to the general orientation, the focus is shifted to one particular instrument, namely, written tests, that can be used for assessment within

RME. Paper-and-pencil short-task problems and the potential enrichment of such problems through application of the RME theory are discussed here. The core issue is how these problems can be made more informative. Numerous examples of such problems are provided, followed by a retrospective look at the implications of the RME views for written assessment. In this retrospection, it becomes apparent that the boundaries of traditional written assessment have been broken through in a number of places by RME. The retrospection is followed by a brief discussion on how views on assessment can influence views on mathematics education as well. In the final section of the chapter, the actual objective of this study – that is, the further incorporation of assessment into the RME theory – is re-examined, in order to evaluate what this incorporation has produced.

- Part II

The three chapters which follow describe three separate studies that served as the foundation for what was propounded in Chapters 3 and 4. These studies are examples of developmental research on assessment, in which the implications of the RME theory were made concrete through an iterative process in actual tests and problems. The test results and the experiences gained from these tests and problems were then linked to the theory and were used to obtain a better understanding of assessment. In other words, these studies served both as fields of application and as sources for the development of a theory for assessment.

The first study, as described in *Chapter 5*, involved the entry-test that was developed for the MORE research. The test was intended to provide an initial standard for the research, and was administered at the beginning of first grade. Due to the large number of students participating in the research, there was no other choice but to administer a written test. Chapter 5 first provides background information on the test that was developed for this purpose. This is then followed by a detailed examination of the unexpected results of this test. To the astonishment of everyone involved, the students turned out to already possess substantially more mathematical knowledge and skills than had been expected. These surprising results, discovered by means of a written test, were what led to further investigation of the potential of written assessment. The study also stimulated a renewed interest in the initial level of first-graders. Parts of the test were later administered in Germany and Switzerland as well, and these results are incorporated in the analysis of the test results contained in this chapter.

Chapter 6 gives an account of a study into the opportunities for RME in special education. The occasion for this research was the gap in The Netherlands between the mathematics instruction given in regular primary schools and that given in schools for special education. The reform of mathematics instruction along the lines of RME as it was implemented in regular education has not been accepted in special education. The chief argument has been that the RME approach is too difficult for children in special education, because of their limited abilities. An attempt was made in this study to push back the assumed boundaries. This was carried out by means of a written test on ratio, similar to the MORE tests. This chapter provides an outline of the study that describes how the test was constructed, to which students it was administered, and what results it produced. The test results showed that, even though the topic of ratio had not been taught at school, most of the participating students in the upper grades (grades 5 and 6) at two schools for mildly mentally retarded children were quite able to deal with the context problems on ratio. Furthermore, what they wrote on the scratch paper revealed how they had arrived at their solutions. This stands in stark contrast to the generally held assumptions regarding the potential abilities of these children. The conclusions that close the chapter stress that, at the very least, some reflection on the special education mathematics curriculum and its instructional approach is warranted. This study also demonstrated, once again, the potential of written assessment.

Chapter 7 covers a developmental research project on assessment that was conducted for the 'Mathematics in Context' project. This project involved the development of a new middle school mathematics curriculum (for grades 5 through 8) in the United States. The purpose of the assessment research was to improve the quality of short-task problems as used in written assessment. The research focused on investigating what kind of problems would be in tune with the RME philosophy that was behind this project. This chapter focuses on one particular characteristic of this philosophy, namely, that students should be given open problems so that they have the opportunity to demonstrate their abilities. The chapter gives an account of the two-stage developmental research that was conducted in this project, in which one particular assessment problem on percentage was designed, field tested, revised, and then field tested once again. The findings of the first stage of the study revealed a disadvantage of open problems, namely, that they can be answered in ways that do not provide sufficient certainty about whether or not the students have achieved a certain level of understanding. The second stage of the study, however, showed that this problem can be overcome by using what is called the 'safety-net question'. It is clear from the results to the problems that this safety-net question is a reliable way of increasing the certainty with regard to student understanding without, however, making the problems any less open. The chapter closes with a brief discussion on how interview techniques can also be applied in other ways in written assessment.

Part I

1 Assessment within Realistic Mathematics Education – from its inception through 1987

1.1 A new approach to mathematics education in The Netherlands

1.1.1 Developments at the confluence of four currents

During the late nineteen-sixties, the first steps were taken in The Netherlands in the direction of what would later be called ‘Realistic Mathematics Education’ (RME). Although still under development, and not yet entirely implemented in the classroom practice, the reform of mathematics education begun at that time has left its mark upon today’s primary school mathematics education. More than three-quarters of the Dutch primary schools now use a mathematics textbook that was inspired to a greater or lesser degree by this reform movement.

The first steps in this direction were taken approximately twenty-five years ago, at a time when new curricula for mathematics education were also being developed in other European countries and in the United States (see, for example, Kilpatrick, 1992). The reform of Dutch mathematics education was provoked to a great extent by the kind of material being exported to The Netherlands from the American reform movement. The Dutch reform movement, which held an aversion to the prevailing home-grown mechanistic approach to arithmetic education, particularly wished to offer an alternative to the American ‘New Math’ approach that was threatening to intrude on Dutch education by way of translated textbooks. The same was true to a lesser degree with respect to the structuralistic methods originating in France and Belgium and the empirically oriented educational materials from Britain.[1]

It was at this confluence of four separate currents – the prevailing mechanistic trend in Dutch education, the empirical trend, the structuralistic trend, and the New Math approach – that the Dutch reform movement developed (Treffers, 1978, 1987a). Each one of these currents, according to Treffers, left its mark on the development of the new Dutch approach to mathematics education. In this respect as well, the Dutch reform movement was no isolated development. Both in terms of time and of content, this movement must be regarded in relation to other currents.[2]

The actual impulse for the reform movement was the inception, in 1968, of the ‘Wiskobas’ project[3], initiated by Wijdeveld and Goffree. Wiskobas was a project of the CMLW (Mathematics Curriculum Modernization Committee), which was initially established by the government, in 1961, to modernize mathematics education in secondary schools. With the inception of the Wiskobas project, attention turned to primary education as well. In 1971, the establishment of the IOWO (Institute for

Development of Mathematics Education)[4], which provided the facilities needed by the Wiskobas project to develop in a professional manner, further confirmed this development. Here, too, developments in The Netherlands remained in step with international activities. Around the same time, similar institutes of research and development were being established in other countries as well (Kilpatrick, 1992). The British Shell Centre and the Institut für Didaktik der Mathematik in Bielefeld, Germany, for instance, as well as a great number of American institutes, also date from this period.[5]

Although the foundations of the Wiskobas work had already been laid by Wijdeveld and Goffree, it was Freudenthal, director of the IOWO who, by his resistance to the New Math movement, gave the initial impetus to the Dutch mathematics reform movement (Treffers, 1993a). The role played by assessment in this reform movement was also greatly influenced by Freudenthal’s ideas on assessment.

The following section provides a brief description of the most significant characteristics of RME. The rest of the chapter is then devoted to an extensive examination of the early stages of RME assessment.

1.1.2 The predominant characteristics of Realistic Mathematics Education

The development of the RME and its underlying educational theory continues even now. Refinements continue to be made and emphases altered on the basis of new developmental research. One can therefore only offer, at best, a picture of the state of affairs up to the present, rather than a finished portrait. In order to present this as comprehensively as possible, one should follow the development of RME through the years, as did Goffree (1993) and Treffers (1993a) with respect to the contributions of Freudenthal. As this type of description would be beyond the scope of this chapter, however, a concise overview will be presented here, which will serve both to support and introduce the following section on assessment.

As mentioned above, Freudenthal’s views were determinant for the direction taken by mathematics education reform in The Netherlands.[6] One of the most important characteristics of this reform was the assumption of a particular viewpoint regarding both people and mathematics (Freudenthal, 1977). According to Freudenthal, mathematics must be connected to reality, stay close to children and be relevant to society in order to be of human value. This viewpoint involves regarding mathematics not as subject matter but, rather, as a human activity, and this was also the message conveyed by Freudenthal in his 1968 lecture entitled ‘Why [...] teach mathematics so as to be useful’. As Goffree (1993) remarked, one of the essential passages of this lecture referred to mathematization as a major characteristic of RME:

> “What humans have to learn is not mathematics as a closed system, but rather as an activity, the process of mathematizing reality and if possible even that of mathematizing mathematics.” (Freudenthal, 1968, p. 7)

It was Treffers (1978, 1987a) who formulated in an educational context the idea of two types of mathematization, by distinguishing ‘horizontal’ and ‘vertical’ mathematization. In broad terms, these can be described as follows: in horizontal mathematization, the students come up with mathematical tools to help organize and solve a problem located in a real-life situation. Vertical mathematization, on the other hand, is the process of a variety of reorganizations and operations within the mathematical system itself. Or, as Freudenthal (1991) put it, horizontal mathematization involves going from the world of life into the world of symbols, while vertical mathematization means moving within the world of symbols. Finding shortcuts and discovering connections between concepts and strategies and then applying these discoveries is implicit in vertical mathematization. Freudenthal emphasized, however, that the differences between these two worlds are far from clear cut. In addition, in his eyes, the two forms of mathematization were of equal value and he stressed the fact that both activities could take place on all levels of mathematical activity. In other words, even on the level of counting activities, for example, both forms may occur.

The concept of horizontal and vertical mathematization is one of the salient features of the RME teaching methods[7]. It contains, in fact, all of the important aspects of the RME educational theory.

1.1.2a Students’ own activities and contributions

This idea of mathematization clearly refers to the concept of mathematics as an activity which, according to Freudenthal (1971, 1973), can best be learned by doing.[8] The students, instead of being the receivers of ready-made mathematics, are treated as active participants in the educational process, in which they themselves develop all sorts of mathematical tools and insights. Freudenthal (1973) called this the ‘reinvention principle’.[9] In his opinion, using scientifically structured curricula, in which students are confronted with ready-made mathematics, is an ‘anti-didactic inversion’. It is based on the false assumption that the results of mathematical thinking, placed in a subject-matter framework, can be transferred directly to the students. Besides the fact that such an approach – where students are simply required to cough up pre-digested material – is inhuman, it simply doesn’t work. Even in the most trivial situations, students who have learned mathematics in this way are unable to apply it. According to Freudenthal, this comes from placing the cart before the horse: failing to allow the students the opportunity to develop the mathematics themselves. Mathematics, in other words, must be taught in the order in which the students themselves might invent it (ibid., 1971).

The essential importance of this self-discovery is evident, for instance, in the topic of fractions where, traditionally, the children’s own activities are often omitted, with all the resulting consequences (ibid., 1979d). In the newly developed way of teaching fractions (Streefland, 1988, 1991), the students are therefore confronted with problem situations (in this case involving fair sharing) in which they can produce the fractions themselves. Another example in this context is the gradual discovery of multiplication and division algorithms through cleverly repeated addition and subtraction (see Treffers (ed.), 1979; Dekker, Ter Heege, and Treffers, 1982).

1.1.2b The link to reality and the focus on application

In addition to the students' own activities, great significance is also ascribed to the link to reality (see, for example, Streefland, 1985b). Just as mathematics arose from the mathematization of reality, so must learning mathematics also originate in mathematizing reality. According to Freudenthal (1973, p. 77):

> "...reality [is] the framework to which mathematics attaches itself."

When children learn mathematics in an isolated fashion, divorced from experienced reality, it will be quickly forgotten and they will not be able to be apply it (Freudenthal, 1971, 1973, 1986). Rather than beginning with certain abstractions or definitions that are to be applied later, one must start with rich contexts demanding mathematical organization or, in other words, contexts that can be mathematized (Freudenthal, 1979b, 1986). This also means, therefore, that one must begin with material that can contribute to putting this mathematization into practice (Freudenthal, 1984a). Just as one must avoid beginning with abstractions, so should one avoid pre-structured material (Freudenthal, 1978a, 1979b). Otherwise, one is again faced with a case of anti-didactic inversion, in which learning content is derived from the structure of the subject matter and is then rendered appropriate for education by means of embodiments.

In contrast to this top-down approach, Freudenthal (1978a, 1983a) proposed a 'didactical phenomenology' in which an analysis is made of the real-life sources of the mathematics. The point here is to determine which actual phenomena (in the past) contributed to particular mathematical concepts, how the students can come in contact with these phenomena, and how these concepts appear to the students.

This analysis can be used to help locate contexts that can serve the students as a source for developing mathematics. By this means, moreover, it can be discovered which mathematics is worthwhile learning. The contexts thus serve not only as a source, but as an area of application as well. The students must consider the problems worth solving. By providing strategies, the contexts can thereby be of significant assistance for arriving at a solution. Take, for example, a problem involving calculating the total rent for a piece of land (see Freudenthal, 1981a). The students who remained within the context while performing their calculations all arrived at the correct answer, while the students who left the context and created a bare multiplication problem did not. In addition to providing context-related strategies, the contexts can also elicit short cuts, as occurred in a problem involving soccer fans transported by bus who receive a discount for every ten buses (see Gravemeijer, 1982).

Teaching mathematics in a realistic context also means offering contexts in which students are confronted both by unsolvable problems and by problems that can be solved in a variety of ways. The familiar textbook contexts, in which everything is already supplied, must be dispensed with (Freudenthal, 1980, 1982a). It is important that students learn to think within the context and that they make use of experience gained in other, extracurricular contexts when solving problems such as: if a car gets 10 km to the liter, how far will it go on 50 liters of gas? (see Freudenthal, 1979a).

The contexts need not necessarily refer, however, to real life situations. The important point is that they can be organized mathematically and that the students can place themselves within them. The students must be aware of both the situation and the corresponding problem, and must image themselves in the situation.[10] It is this aspect – the 'imagining themselves' – that gave RME its name (Van den Brink, 1973a, 1989; Wijdeveld, 1980).[11]

1.1.2c Levels of understanding

As mentioned earlier, mathematization can occur on different levels. These levels of mathematization are connected to the various levels of understanding through which students can pass: from the ability to invent informal context-related solutions, to the creation of various levels of short cuts and schematizations, to the acquisition of insight into the underlying principles and the discernment of even broader relationships. The essence of this level theory, which Freudenthal (1973) borrowed from the observations and ideas of the Van Hieles, is that the mathematizing activity on a lower level can later become the object of analysis on a higher level. In other words, the students can conduct all sorts of operations involving fractions on an informal level and then, later, formalize them on the following level. The condition for arriving at the next level is the ability to reflect on the activities conducted. This reflection can be elicited by interaction and by the students' 'own productions' (see also Sections 1.2.3e and 1.2.5c).

In contrast to what is often believed, namely, that learning is a continuous process of small steps (which may indeed be true for certain drill activities), it is, in fact, one of leaps and discontinuity (Freudenthal, 1978b). Once again, it was the work of the Van Hieles that brought Freudenthal's (1973) attention to discontinuity in the learning process. This process, rather than developing smoothly and steadily, stands still at times, only to start up again on a higher level. It would appear that, during the lull, the student undergoes a kind of maturation process.[12]

This level theory dovetails with the educational approach initially developed by Wiskobas for learning column arithmetic, which is called 'progressive schematization' (see Treffers (ed.), 1979; Dekker, Ter Heege, and Treffers, 1982). Characteristic of this approach is that students use fairly complex problems when they first start learning to calculate, but they work them out on a low level of schematization. At a later stage, they begin to apply all sorts of short cuts based on their own constructions, so that each student follows his or her own path to eventually arrive at the standard algorithm. This standard algorithm, however, need not be attained by everyone (see also Freudenthal, 1986). Through this progressive schematization, differences between students can be taken into account (Freudenthal, 1981a).

Freudenthal's (1979c) skepticism regarding fixed patterns of development applies not only to learning column arithmetic but to development in general. Children are individuals, each following an individual learning path. Education must therefore be adapted to the children's distinctive learning processes. The children themselves, indicate, to a great extent, how this should be done, through their own constructions and informal strategies. Rather than repressing such activities, one should use them as footholds for learning the more formal strategies (Freudenthal, 1986). Models can serve as an important device for bridging this gap between informal, context-related mathematics and the more formal mathematics (Streefland, 1985b; Treffers, 1987a, 1991a; Gravemeijer, 1994; Van den Heuvel-Panhuizen, 1995b). The strength of these models is the fact that, while they are often rooted in concrete situations, they are also flexible enough to be introduced into higher levels of mathematical activities. They provide a foothold during the process of vertical mathematization, without obstructing the path back to the source. One of the most powerful examples of this is the number line, which begins in first grade as a beaded necklace, and which, by sixth grade, has become a double number line for supporting work with fractions and percentages.

To summarize, RME takes the perspective of mathematics as a human activity, while focusing on meaningful applications. An important role in RME is played by the students who, by using contexts and models, can pass through various levels of mathematization and thereby develop their own mathematics. As will be seen later, the same three cornerstones that support RME, namely, the views on mathematics, how children learn, and how mathematics should be taught, are also indicative of the viewpoints on RME assessment.

1.2 The focus on assessment

1.2.1 The influence of then prevailing viewpoints on assessment

At the time the first steps were being taken towards RME, a general optimism prevailed regarding achievement tests:

> "There is one field in which a considerable sophistication has developed since 1920: the field of achievement testing. It is possible now to study the degree and nature of a student's understanding of school subjects with a subtlety not previously available. Modern objective achievement tests, when properly developed and interpreted, offer one of the most powerful tools available for educational research. Findings have been made through their use that rise far above common sense" (Bloom and Foskay, 1967, p. 65).

This optimism regarding the potential of achievement tests determined, to a great extent, the standpoint taken by RME with respect to assessment. The above quote from an IEA[13] report was often used by Freudenthal (1975b, 1967a, 1991) to illustrate how much he disagreed with this optimism.[14] The then prevailing notion that testing would make nearly everything possible and would resolve any and all problems in education led to a focus among proponents of RME on what should *not* be done rather that what *should* occur with respect to assessment. Freudenthal and the other IOWO staff members presented considerable opposition to what they saw as unsound testing both at home and abroad. The impression, however, that besides criticism, no attention was paid to assessment within the developing RME teaching methods, is, upon closer consideration, clearly not correct.[15] In actual fact, the foundation for assessment within RME was laid early on, concurrently with the development of the RME teaching methods. This foundation was perhaps illuminated insufficiently, however, due to the fact that assessment considerations were integrated into educational development as a whole.

The rest of this chapter is therefore an attempt to shed more light on the viewpoints and ideas regarding assessment by proponents of RME. Although it is not the intention to rake up old controversies over then prevailing opinions and methods of assessment, they cannot be entirely ignored either. After all, it is here that the RME standpoint is so clearly defined.

1.2.2 Assessment within RME – an initial orientation

In spite of the battle waged by the proponents of RME against assessment, the quantifiability of learning results was never discussed (Freudenthal, 1978a), nor was assessment within RME ever dismissed (Treffers, 1983). The idea that assessment constitutes an important part of education was expressed early on in the development of RME, as can be seen in the following quote from ‘Mathematics as an Educational Task’:

> “Examining is a meaningful activity. The teacher should be able to check the influence of the teaching process, at least in order to know how to improve it. The student has the right to know whether he has really learned something (...). Finally there are others who are interested in knowing what somebody has learned” (Freudenthal, 1973, p. 83).

It should be noted that assessment is not viewed here in the narrow sense of determining what the student has learned, but that it is also regarded from the viewpoint of educational evaluation and educational development.[16] Another striking aspect is the important role played by the teacher. This viewpoint recurs repeatedly (see, for example, Treffers (ed.), 1979). The following quote also reflects this idea quite clearly:

> “Such reflective moments in education, in which the teacher contemplates what has passed and what is still to come, are important” (Streefland, 1981b, p. 35; translation of original Dutch text).

This indicates, moreover, that assessment is not only intended for looking back, but also for looking forward. Another aspect that soon arose was the preference for observation as a method of assessment:

> "... we know that it is more informative to observe a student during his mathematical activity than to grade his papers" (Freudenthal, 1973, p. 84).

This partiality for observation does not mean, however, that other forms of assessment, such as the administration of tests, was not considered suitable for RME. In spite of the objections to the then current tests, assessment was still seen as something indispensable:

> "He who wishes to impart something to someone else will also want to find out what the other already knows, in order to build further upon this. And, if he has taught something, he will want to find out whether this has taken root [...] and whether something in the instruction should be altered. [...] One would be blind to the reality of the world and society should one contend that assessment is unnecessary. By this I mean assessment in the very broadest sense: questioning, individual oral testing, class written tests and exams, as well as the strict, so-called objective multiple-choice tests. The point is to test sensibly, [...] to test better and more efficiently with each experience, and this means that the function, rather than the form of the assessment is of primary importance" (Freudenthal, 1976a, pp. 68-69; translation of original Dutch text).

What strikes one in this quote is the broad interpretation of assessment and the emphasis on sensible assessment. Elsewhere, too, Freudenthal (1976b) indicates that, while the choice of form is not a principle issue, one must, in all circumstances, choose the soundest means of assessment.

To summarize, an initial orientation suggests that assessment is considered important in RME, that it is regarded in a broad sense and viewed in relation to education, that the teacher plays a crucial role, and that, while a preference is held for observation, there is also room for tests.

1.2.3 Certain preferences in assessment – a more specific orientation

1.2.3a A high priority assigned to observation

Notwithstanding the broad viewpoint expressed in the above section, RME indeed gives a high priority to observation. Or, in the words of Freudenthal (1981a, p. 137):

> "I stressed observing learning processes against testing learning products."

This preference for observation is closely connected to the points of departure of RME. To start with, it emanates from the RME viewpoints on mathematics. Because mathematics is viewed as a student's individual activity, in which he or she uses certain mathematical insights and devices in order to get a grip on a given problem situation, it is clear that the goal of assessment in RME is the solution procedures themselves, rather than the results. Assessment must provide, as it were, insight into the students' mathematization activities.[17]

In addition, the high priority attached to observation is closely linked to the discontinuity that occurs during learning processes. Awareness of this discontinuity is considered to be of crucial importance for comprehending such processes (Freudenthal, 1978a). It is precisely through this discontinuity – which may manifest itself, for instance, in the form of a spontaneous short cut or the taking of a different standpoint (Freudenthal, 1991) – that one can see that a student has achieved a certain level of comprehension. According to Freudenthal (1979c), the observation of learning processes should focus primarily on the leaps the students take.

In order for this discontinuity to be seen, students must mainly be followed individually. Cross-sections and group averages are not particularly useful, as these tend to erase the discontinuity (Freudenthal, 1973, 1978a, 1978b).

In addition, the type of education in question is important as well. Education given in a traditional manner[18], for instance, will not provide much information on the students' learning processes (Freudenthal, 1978a).

A plea to exercise great restraint is, lastly, another characteristic of this manner of observation (Van den Brink, 1973b; Treffers, 1979). The observer must, as it were, stand in the child's shoes and listen to what he or she has to say.

1.2.3b The continuous and integrated nature of assessment

The consequence of focusing on discontinuity is that the observation must then be continuous (Freudenthal, 1978a). Freudenthal (1985) even suggested that the educational process be seen as a permanent process of assessment, in which the teacher must constantly sense what the next step should be. One effect of emphasizing this approach to assessment, therefore, is the integration of education and assessment. In RME, in fact, the instructional activities and the instances of assessment go hand in hand (Ter Heege and Goffree, 1981). This integration is expressed most clearly in the 'test-lessons'[19] (Ter Heege and Treffers, 1979).

1.2.3c The important role of the teacher

Implicit in the above is that the teacher must play an important role in RME assessment. It is, after all, the teacher who conducts the daily observations (Treffers (ed.), 1979). Observing, administering tests, diagnosing and providing remedial work are all simply part of skilled teaching (Ter Heege and Treffers, 1979). Even assessment development is regarded primarily as the teacher's domain, because it constitutes a significant moment of reflection on the instruction given (Treffers, 1980a). One therefore wonders:

> "Can such a test, that mirrors this reflection, be designed outside the actual educational environment by someone who did not follow it closely?" (Streefland, 1981b, p. 35; translation of original Dutch text).

1.2.3d A holistic approach

Assessment in RME not only evaluates a student's acquisition of certain skills, but also attempts to acquire as complete a picture of the student as possible. This is yet another reason for this predilection for observation:

> "Observations, even though they are mere impressions caught by the expert teacher during a lesson, can provide a rather complete picture of the learning process" (Ter Heege, 1978, p. 82).

In addition to noting the acquired skills, observation involves paying attention to approach behavior, mathematical attitude, solution level, type of errors made, manner of collaboration, need of support, reaction to hints, emotional aspects, motivation and concentration (see Ter Heege, 1978; Treffers, 1978;1987a; Ter Heege and Treffers, 1979; Ter Heege and Goffree, 1981; Treffers and Goffree, 1982). Written tests are clearly inadequate in this respect. A written test on mental arithmetic, for example, cannot assess daring and flexibility, although these attitudinal aspects are indeed important for mental arithmetic (Ter Heege and Goffree, 1981).

Besides an expansion in breadth, this attempt to acquire as complete a picture of the student as possible also signifies an increase in depth. The assessment must not be merely a superficial test.

> "Particularly in the case of mathematics, with some people is it sometimes necessary to delve deeply in order to verify whether your educational resolutions have indeed been translated into learning processes" (Freudenthal, 1985, p. 304; translation of original Dutch text).

At times, this may result in the need for continued questioning, in spite of the fact that things seem to be going smoothly. Freudenthal offers the example of a student who solved one equation after another correctly:

> "The better things went, the more urgently I wondered whether she really understood anything." (ibid., p. 305; translation of original Dutch text)

Thanks to Freudenthal's continued questioning, the student was able, with the support of the empty number line, to make the leap to inequalities. Only then was it clear that she was able to do more than just perform an acquired trick when solving equations.

1.2.3e Choosing an open-ended test format

It is obvious that a closed type of test question, in which the student must simply mark the correct answer, would never have led to this discovery. If assessment, as stated above, is to offer insight into the students' mathematization activities, then these mathematization activities must be as visible as possible. This can only occur with open-ended questions, in which the students work out a problem and formulate an answer on their own.

Another reason for choosing an open-ended test format is to avoid seeing the students' development as following a fixed pattern (see Section 1.1.2c). The students being tested may, for instance, be on different levels of short cuts and schematizations. Closed types of tests lack the flexibility necessary to make this apparent.

This lack of flexibility becomes quite clear when one wishes to use the students' own insights and strategies as footholds for further instruction, as Freudenthal urged (see Section 1.1.2c), or, as formulated by Streefland (1985a, p. 285), one desires:

> "...to foresee where and how one can anticipate that which is just coming into view in the distance" (translation of original German text).

In addition to demanding open and constructive education (Streefland, ibid.), this requires an open-ended test format, in which the students are offered the opportunity to show what they can do – which may have a different appearance with each child.

This same open attitude can be seen in the idea of offering help on a test (Ter Heege and Treffers, 1979). Rather than maintaining a static approach focused on what the child is able to *produce* at a given moment, in RME it is deemed more important to find out what the child is able to *learn* (Freudenthal, 1979c). In this respect, assessment in RME strongly resembles the standpoint found in Eastern Europe (Streefland, 1979; Ter Heege and Treffers, 1979; Treffers and Goffree, 1982), where, following Vygotsky's (1978) concept of the 'zone of proximal development', help is even offered on tests in order to see what the next step will be in the child's development.

This open attitude can be seen most clearly, however, in the students' own productions. Elaboration of this concept (which has played an increasingly important role in RME) can be found in Treffers and Goffree (1985), Treffers (1987a), Streefland (1987, 1990a) and Van den Brink (1987). Although the students' own productions were primarily viewed in these publications in the light of their function for mathematization (viewed from the student's standpoint) and for instruction (viewed from the teacher's standpoint), the potential for using students' own productions in assessment was recognized as well:

> [On the one hand,] "...producing simple, moderate, [and] complex problems means that the pupil reflects on the path he himself has taken in his learning process..." [and, on the other hand,] "...the pupil's production, as the result of instruction, functions as the mirror image of the teacher's didactic activity" (Treffers, 1987a, p. 260-261).

Not only was there an awareness of this potential, but it had in fact already been applied some time earlier. An example of this can be found in an 'test-lesson' that was given at the end of a series of lessons on base eight (De Jong (ed.), 1977; Treffers, 1978a; see also Section 1.2.5b).

1.2.3f A preference for true application problems

RME requires problems involving rich, non-mathematical contexts that are open to mathematization. In order for the children to be motivated, the problems, according to Treffers (1978, 1987a), must be formulated in a challenging fashion and not dictate to the children what they should do. In other words, it must be obvious to the students why an answer to a given question is required (Gravemeijer, 1982).

These application problems should not, however, be confused with so-called ‘word problems’, which were often presented as application problems in traditional, mechanistic arithmetic education. Word problems are rather unappealing, dressed up problems in which the context is merely window dressing for the mathematics put there. One context can be exchanged for another without substantially altering the problem (Treffers and Goffree, 1982). Problems involving marbles, for instance, might just as well be problems involving pounds of ham (Freudenthal, 1980) (see Figure 1.1).

Jan has 16 marbles and wins 10 more. *How many does he have now?*	*The butcher has 16 pounds of ham in his shop and orders 10 pounds more.* *How much does he have now?*

Figure 1.1: Examples of word problems (from Freudenthal, 1980, p. 14)

In fact, the reality referred to by these contexts has been replaced by a mathematics textbook context, in which each problem has but one answer. True reality, with its unsolvable or multi-solvable problems, has actually been excluded (Freudenthal, 1980, 1982a; see also Goffree, 1984). Indeed, the children are apparently not even supposed to place themselves in the situation (Gravemeijer, 1982; Treffers, 1990, 1991b). The aim of RME, by contrast, is to place oneself in the context and learn to think within it (Freudenthal, 1979a). Streefland, Hartings, and Veldhuis (1979) offer an example that illustrates this well (see Figure 1.2).

> *Mr. Jansen lives in Utrecht.*
> *He must be in Zwolle at 9:00am Tuesday morning.*
> *Which train should he take?*
> *(Check the train schedule.)*

Figure 1.2: Example of a context problem
(from Streefland, Hartings, and Veldhuis, 1979, p. 6)

This problem is nearly unsolvable if one does not place oneself in the context. It is also a problem where the students need not marginalize their own experiences. At the same time, this example shows that true application problems can have more than one solution and that, in addition to written information, one can also use drawings, tables, graphs, newspaper clippings and suchlike. Characteristic of this kind of problem is the fact that one cannot learn to do them by distinguishing certain types of problems and then applying fixed solution procedures. The object here is for the student to place him or herself in the context and then make certain assumptions (such as how far Mr. Jansen lives from the station and how important it is that he arrives at his destination on time). According to Freudenthal (1979a), problems such as this one require the student to learn a solution *attitude* rather than a solution *method*.

With the arrival of the first RME textbooks, traditional word problems were increasingly replaced by true application problems. Examples of the latter can be found in Treffers and Goffree (1982) and in Goffree (1984). In addition to the characteristics mentioned above, these problems can also be distinguished from existing word problems by the fact that they are constructed out of a number of sub-problems that are grouped around a theme, thereby forming a coherent unit.

1.2.4 Objections to earlier tests and to their underlying standpoints

As mentioned above, the initial phase of the development of RME was characterized by controversy regarding assessment. This dispute, in which a number of ugly things were said[20], may have long obstructed any kind of collaboration with the people responsible for the tests. On the other hand, it probably also helped save Dutch education from the potential negative effects of such tests.

1.2.4a Unsoundness of the taxonomies as a means of test construction

Freudenthal's struggle against the existing tests was primarily a struggle against the taxonomies. Bloom's taxonomy, in particular, was the receptor of his criticism. The categories of educational goals that Bloom distinguished for the cognitive domain (which were intended as a way of simplifying test construction and making it more objective), had, according to Freudenthal (1975b, 1978a), a detrimental effect on test development. In addition to the artificial nature of these categories and the absence of certain categories, this detrimental effect is mainly due to the fact that the level classification attached to the categories is linked to the problems, rather than to how the students solve the problems. According to Freudenthal, *how* the students solve the problems is the whole point. In a nutshell, Bloom sees the capacity to solve a given problem as being indicative of a certain level, while, in Freudenthal's eyes, it is the way in which the student works on a problem that determines the level. The latter illustrates this viewpoint using the following example:

> "A child that figures out 8 + 7 by counting 7 further from 8 on the abacus, acts as it were on a senso-motoric level. The discovery that 8 + 7 is simplified by 8 + (2 + 5) = (8 + 2) + 5 witnesses a high comprehension level. Once this is grasped, it becomes mere knowledge of the method; as soon as the child has memorized 8 + 7 = 15, it is knowledge of facts. At the same moment figuring out 38 + 47 may still require comprehension; later on, knowledge of method can suffice; for the skilled calculator it is mere knowledge of facts" (Freudenthal, 1978a, p. 91).

Moreover, Freudenthal points out that, if the 'anti-didactic inversion' is rejected, and one follows instead the 'didactical phenomenology', then one may arrive at an entirely different order of hierarchical learning levels. He clarifies this on the basis of the taxonomy of De Block (1975). The hierarchical learning levels distinguished here, which ascend in the order of knowledge, recognition, application, integration, could also run in reverse order:

> "Let us follow the pattern using the knowledge of 3 < 4; this is integrated by, for instance, comparing one's own family of three with the family of four next door; it is applied to other pairs of quantities; it is recognized by means of a one-to-one-relationship ('because they have one more child'); and it becomes 'knowledge' through the counting sequence 1, 2, 3, 4 ..." (Freudenthal, 1981b, p. 5; translation of original Dutch text).

Freudenthal objects to these taxonomies because, instead of having been derived from a 'phenomenological' or 'didactical' position they are *a priori* constructions postulated on logical grounds (Freudenthal, 1981b).

This same objection is raised with regard to the matrices used to represent learning hierarchies. Take, for example, the matrix accompanying a CITO[21] test on column arithmetic, in which the entries are the size of the number and the number of times one must carry over. There is, however, little reason to assume that adding three-digit numbers lies on a higher level than adding two-digit numbers (Treffers, 1980b; Freudenthal, 1981a). The only reason, according to Freudenthal, that the larger numbers might be more difficult, is that there is more chance of making a mistake.[22]

1.2.4b One-sidedness of the psychometric foundation

Another objection to the taxonomies is that they are used to validate assessment, as was the case with Bloom's taxonomy in the IEA research that was so fiercely criticized by Freudenthal (1975b). Consequently, the formal characteristics are validated while the subject matter and educational content are ignored (Freudenthal, 1978a).

This disregard for the content can also be found in the excessive attention devoted to the reliability of the test instruments (regarded by Freudenthal (1978a) as a kind of ritual), instead of to the actual validity. Freudenthal (1976a, p. 64) speaks here of the:

> "...researcher passing his responsibility from validity to reliability" and of the "...flight from validity to reliability."

According to Freudenthal (1976a), there are countless appalling examples resulting from this. He mentions one, taken from a research project on arithmetic. This research is considered as highly reliable, however, decisions concerning the subject matter were taken that threatened the validity. In analyzing the student work, for instance, a solution procedure for 'find 50% of 200' that involved calculating 1% was considered insightful, while an approach involving '50% means half' was not. In Freudenthal's (1975b) words, the meddling of psychometrics with test development has resulted in insufficient contemplation of subject matter content.

1.2.4c Inadequacy of the goals and goal descriptions

The same can be said for the instructional theory of two decades ago that propagated the formulation of concrete behavior goals as a guideline for instruction (see Gronlund, 1970; Popham and Baker, 1970). This led, around 1980, to the introduction in The Netherlands, of the 'Criterion-Referenced Tests'[23] developed by CITO. Criticism of these tests by proponents of RME focused mainly on the atomization of the goals and the elimination of the process goals. The tests therefore inadequately reflected the educational intentions, which caused repercussions for the education itself (Freudenthal, 1984b). This compartmentalization in sub-goals threatened to shift the emphasis in education more towards partial achievements and 'small products', instead of emphasizing the 'larger process', which was the point of it all (Treffers, 1980a).[24]

The recurring question posed here was whether the tests – criterion-referenced and otherwise – did indeed test the essence of the subject matter that they aspired to test. And the recurrent answer was that the tests did not correspond to the RME view of mathematics education, that they did not cover the domain in question, and that they displayed deficiencies in terms of the teaching methods. The list of the criterion-referenced test's instructional objectives for the topic of measurement, for instance, failed to include a number of essential elements, such as the relation between surface and circumference (Treffers, 1980a). The criterion-referenced test for column arithmetic was sorely inadequate due to the fact that the topic of schematization had been entirely ignored and significant sources of error had been missed (Treffers, 1980b, 1983; Freudenthal, 1981b). The criterion-referenced test for decimal numbers was the focus of criticism as well, because the entire emphasis had been laid on the characteristics of form, while the substance of decimal numbers – namely, how they arise – had been ignored completely (Streefland, 1981a, 1981b, 1982).

For the test on mental arithmetic, which was part of the 'General Test'[25], the objection was raised that a number of test problems had nothing to do with mental arithmetic, but, instead, were about measurement and geometry. Another objection concerned the fact that, because of the format of the test, nothing could be seen of flexible arithmetic. In other words, an extremely superficial aspect of mental arithmetic, namely, that of memorized arithmetic, had been selected (Ter Heege and Goffree, 1981).

Criticism was also raised regarding a BOVO test[26], where arithmetic comprehension was tested by a series of traditional word problems, to the exclusion of true application problems (Treffers and Goffree, 1982).

The following, and final, example is from a test developed by CITO, which was intended to evaluate a first-grade Wiskobas unit involving 'bus problems' (Van den Brink, 1989). Here, again, it is clear that the point of view of the unit's designer did not mesh with the viewpoint of whomever had the final responsibility for the test:

> "Item 5, for instance, consisted of a column of bare 'missing addend problems' that needed to be solved first before being linked to the pictures. The RME viewpoint, however, is that the pictures can serve as contexts to help solve the bare problems.

> This will not, however, be possible if the bare problems have already been done" (ibid., p. 79; translation of original Dutch text).

In other words, the objections to the traditional tests were based on numerous differing viewpoints with regard to the subject matter and the educational goals. At the heart of the matter was the question of whether one can disengage assessment from education. While the test designers of that time answered this question affirmatively (see Bokhove and Moelands, 1982), proponents of RME disagreed:

> "If educational goals are isolated from the educational process as a whole, and if they do not contain the aspects essential for comprehension, insight and application, how can they then be used [for example] to check progress?" (Treffers, 1983, p. 56; translation of original Dutch text).

1.2.4d Objections to the formalized nature of the tests

Aside from objections to the content, there was also dissatisfaction with regard to the formalized design of the tests, although it should be noted that these objections were not as severe as those regarding the content. By 'formalized tests' was meant the so-called objective or multiple-choice tests (Freudenthal, 1976b). The objections raised were that these tests neither provided any means of interpreting students' errors nor any footholds for further instruction (Ter Heege, 1978). The harshest criticism was reserved for the process of diagnosis. According to Freudenthal (1978b, p. 6), the point of assessment was to:

> "...broaden and sharpen the teacher's awareness of the presence (or absence) of learning processes. The formalized tests are absolutely inadequate for this purpose. Information set in a rigid framework is useless for making a diagnosis. Moreover, the object is to open the eyes of the evaluator, which cannot be done by handing him mechanized tests" (translation of original Dutch text).

This, then, becomes a kind of 'blindfolded diagnosis' – a way of making a diagnosis where all that counts is *who* got it wrong (Freudenthal, 1978a). Freudenthal later stated that, rather than saying *what* went wrong:

> "true diagnosis tells you *why* something went wrong. The only way to know this is observing the child's failure and trying to understand it" (Freudenthal, 1981a, p. 135).

In addition to the objection that multiple-choice tests focus on results and provide no information on the strategy applied, the point was also made that some educational goals are fundamentally untestable through multiple choice. Freudenthal (1984b) offers an example here of a multiple-choice question whose purpose was to ascertain whether students were able to simplify products such as $3^2 . 3^5$. But all that was verified, in Freudenthal's opinion, was whether students were able to pick out the correct simplification from a list of four possibilities. This choice is all one can see, in fact; for the rest, one can merely hypothesize about a thought process that had passed unobserved.

Proponents of RME also refused to endorse the assertion that multiple-choice questions had the advantage of objective scoring[27]:

> "In mathematics, open-ended questions can be evaluated just as objectively by people as closed questions can be evaluated by the computer. Perhaps even more objectively. With open-ended questions one at least knows what one is testing. The rigidity of the closed test format is both objectively (incorrect formulation) and subjectively (irrelevant incomprehensibilities) a graver source of error than are the margins for interpretation when evaluating answers to open-ended questions" (Freudenthal, 1984b, p. 22; translation of original Dutch text).

In order to see whether open-ended and closed questions would produce the same results, Freudenthal (ibid.) once administered the final test for primary school – the so-called 'CITO test' – to a sixth-grade student a second time, a number of weeks after the official test. But this time the questions were open-ended. The results of this second test were considerably better than the first and, moreover, the student needed less time to complete it. One would think this sufficient reason to further investigate the difference between the open-ended and closed form of test questions. And, indeed, Freudenthal had intended his 'n = 1 research' as a stimulus for further investigation. Unfortunately, the designers of the test in question simply regarded his research as a good joke! This is ample indication of the distance that lay between the standpoints of that time.

1.2.4e An aversion to traps in test questions

The alternatives invented by the test designers as possible answers to multiple-choice questions were problematic as well. Often, these alternatives were misleading and managed to entrap great numbers of students (Freudenthal, 1978a). This situation needed to be avoided:

> "Immunity against traps is an enormously useful capacity, but should it be tested along with arithmetic?" (ibid., p. 86).

Another, related problem was the ambiguity of the tasks themselves. According to Freudenthal (1978a), the consequence of the artificial way in which the test questions were sometimes formulated and the type of language used was that the skill of understanding what the test designer had in mind was tested, rather than the arithmetic itself. He lamented that the additional educational goal most frequently built into tests but left inexplicit was that of understanding test logic (Freudenthal, 1984b). An example of this is the so-called 'misconception' children are said to have of conservation. According to Freudenthal (1973, 1978b, 1979c, 1983a), this is due purely to the way in which the children are questioned[28]. Often, for instance, questions were posed that required a formal answer, while the children (as was evident from their answers) had no idea what a formal answer was:

> "If I enter a meeting room and ask 'are all chairs occupied' and somebody answers 'no, you can sit on a bench', then this is an answer, not to the formal question, but to the intentional question 'where can I sit?' " (Freudenthal, 1973, p. 674).

The importance ascribed to posing a good question can be seen in the following quote (Freudenthal, 1985, p. 304).[29]

> "If you want to entrap someone, then pose the wrong question. Whoever invented the question is responsible for incorrect answers" (translation of original Dutch text).

While out walking with his grandson Bastiaan, Freudenthal (1979c) asked him a question about the size of the church clock (at least, that was his intention). But, when he was asked how high the church clock was, Bastiaan – after some thought – answered with great self assurance that the clock was 30 meters high. This answer – in spite of being 'incorrect' – was immediately accepted. Streefland (1980), too, has had similar experiences with posing incorrect questions. He, too, believes that the children's answers will tell you when something is wrong with the question. This will only happen, however, if the questioner is aware and keeps his eyes open which, considering the following point of objection, is not always the case.

1.2.4f Criticism of how answers are evaluated

The communication problem described above also exists in reverse. Not only is there a problem when questions are posed in such a way that the students have difficulty understanding them. Even worse is when a student's correct answer is misunderstood by the poser of the question. Freudenthal (1973) offers an appalling example of this, taken from the experiments on conservation mentioned above. These were experiments conducted on the conservation of distance. They involved two dolls placed at a certain distance from one another and at different heights. The child was asked whether it was farther from A to B or from B to A. The answer given by the student was as follows:

> "...it is not the same distance, because the little man below looks to the feet of the man above, and the man above looks to the eyes of the man below..." (Freudenthal, 1973, p. 676).

Even though the student did not consider the two distances to be equal, there was certainly no question here of non-conservation. The questioner, however, did not understand this!

Mention should be made here as well of the often rigid scoring rules for tests. An example of this can be found in the above-mentioned test developed by CITO for a first-grade Wiskobas unit on bus problems. The test consisted, among other things, of a series of 'arrow problems', in which the students had to indicate what had taken place (see Figure 1.3).

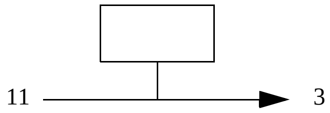


Figure 1.3: Example of an 'arrow problem'

According to the scoring rules, the only correct answer here is – 8, so –10 + 2 would be marked wrong (Van den Brink, 1989).

1.2.5 Realistic alternatives

1.2.5a Suggestions for improving assessment

The suggestions for improving assessment offered by proponents of RME are, on the one hand, closely connected to the RME preferences regarding assessment and, on the other hand, are the mirror image of the RME objections to the existing tests. Here, too – as was the case with respect to the preferences and objections – the suggestions reflect the standpoint of subject matter content and of the learning child.

- help teachers observe

  If one wishes to improve assessment, one must, according to Freudenthal (1976b), begin in the micro-environment by first helping teachers learn to observe learning processes. The teachers must become aware of when learning processes are taking place and when they are not. For this reason, learning to observe learning processes is regarded as the principal part of all courses in mathematics education (Freudenthal, 1976a).

- use observation as a point of departure for test development

  In order to develop tests with a diagnostic purpose, the designer should observe both the teacher and him or herself in the classroom. In this way, the designer can get a sense of what criteria an observer of learning processes uses to ascertain progress (Freudenthal, 1976b).

- conduct discussions with the students

  In order to truly fathom what the students know and understand, one should discuss their answers with them (Freudenthal, 1979c).

- place more emphasis on formative assessment

  With respect to the nature of the assessment, Freudenthal (1976a, 1976b) argues strongly in favor of formative assessment, which is understood as informal assessment. Here, again, he emphasizes that formative assessment should serve to broaden and sharpen the teacher's awareness of learning processes. According to Freudenthal, the virulence of summative assessment could be neutralized by shifting the test load more towards formative assessment. A satisfactory system of process-assessment could render product-assessment unnecessary.

- conduct domain analyses and improve the goal description

  In order to improve the content of tests, an argument was presented in favor of conducting 'mathematical-didactical analyses' (Treffers (ed.), 1979; Treffers, 1980a, 1980b; Streefland, 1982) or, as Freudenthal (1978a, 1983a) put it, carrying out 'didactical-phenomenological analyses'.[30] The point of such analyses is that the entire structure of a given domain would be laid bare, thereby exposing all significant points of learning (Treffers, 1980a). Basing the tests on these analyses could help prevent tests being too imbalanced and superficial.

Another foothold for improving assessment, related to the previous one, concerns goal formulation. According to Freudenthal (1978a), educational goals must not be formulated behind a desk but, rather, through a dialogue with students, teachers, supervisors, parents and other interested parties. Freudenthal (1984b) also argued in favor of using paradigmatic examples when formulating the educational goals. The same plea was made by Treffers (1980a). Sample tests are, in his opinion, an excellent way to indicate the purpose of certain higher goals. But, in order to clarify what exactly is intended, the goal formulation should also be expanded to include an educational description (Treffers, 1980a; see also Streefland, 1981b). By providing a description of the educational process in mind, both the underlying purpose and the choice of instructional activities can become clear, resulting in a better harmony between education and assessment. Treffers (1978, 1987a) introduced the ‘three-dimensional goal description’ for this purpose. In addition to the components of behavior and content, these goals also contain an instructional component. This last component means that all aspects of the educational process that can assist the clarification of the intended process and product goals (such as posing fundamental questions, providing instructional hints, indicating solution levels, points of observation and suchlike) will be included in the goal description.

1.2.5b The test-lesson

Aside from suggestions for improvement, the proponents of RME also developed a number of alternatives for the existing tests. The most detailed alternative was that of the ‘test-lesson’. A salient feature of this test-lesson is its bilateral nature: it is both a lesson in which a test is administered and, simultaneously, a test situation in which a lesson is given (Ter Heege and Treffers, 1979). The three examples of test-lessons that follow reveal that such test-lessons may deal with different types of problems and can be used for a variety of purposes.[31]

- the Column Arithmetic test-lesson

  The first test-lesson involves a kind of ‘status quo’ assessment in which the administered test consisted of bare multiplication problems (Ter Heege and Treffers, 1979). This test-lesson focused on the technical aspect of doing multiplication. The objective was to ascertain how advanced the children had become in doing column multiplication after one year of learning progressive schematization (see Section 1.1.2c). The test was constructed in such a way that it incorporated all degrees of difficulty. It consisted of a total of 14 problems increasing in difficulty from 6 × 18 to 314 × 207. During the administration of the test, every effort was made to avoid an atmosphere of nervous apprehension. The children were well-informed about the purpose of the test. Although the objective was, indeed, to view individual work, they were allowed to sit in groups while working on the test. Enough blank space was provided for working out the problems. Due to the specific manner of notation that results from progressive schematization, each child’s level of schematization could be seen from the worksheet. In other words, the paper revealed the solution behavior without any other form of communication being necessary. Ter Heege and Treffers (ibid., p. 128) speak of “solidified student behavior [that] has become visible on paper”.[32] In addition to the important information provided by the written calculations, information was also gathered by observing how the students worked on the test and by offering them help when needed. Affective and emotional aspects, motivation, concentration, need for support, reaction to hints and collaboration all received attention during this observation. After the test, the students' work was analyzed, primarily with respect to the level of schematization and the nature of the errors. The combination of this information and that acquired during the observation determined which instructional aids the students required. The idea was to administer this type of test-lesson regularly – at least once a month – in order to keep track of each student's progress (Treffers, 1979).

- the Winter Supplies test-lesson

  The second example is an introductory test that was administered when the children were just learning to do multiplication (see Ter Heege, 1978; Dekker, Ter Heege, and Treffers, 1982). Although this test did, in a certain sense, examine whether the students were ready for this new subject matter, it should not be viewed as a preconditional test. Instead of investigating whether the students already possessed the requisite prior knowledge, access points were sought to introduce new instructional activities for learning the multiplication algorithm. At the point this test was administered to a third-grade class, the students had already had considerable experience with the addition algorithm, but had not yet begun multiplication. This test consisted of one sole problem that was not presented in the form of a multiplication problem. The problem involved a squirrel family that was busy gathering a supply of acorns for the winter. When they had finished, Grandfather Squirrel would then count how many acorns had been collected. Each of the 8 squirrels gathered 23 acorns. The idea was for the students to work out the problem on paper. They were left entirely free to work as they wished, and they could use manipulatives as well. In spite of the fact that these students had not yet mastered the multiplication algorithm, they were nonetheless able to solve such a large multiplication problem as 8 × 23. They did this by using solution strategies such as repeated addition and repeated doubling. These informal solution strategies could then serve as starting points for the new learning process. The results revealed that this educational approach of progressive schematization dovetailed perfectly with the children's own natural approaches.

- the Kingdom of Eight test-lesson

  The third example was part of the last lesson in a series of four lessons on doing arithmetic in base eight (see De Jong (ed.), 1977; Treffers, 1978, 1987a). The purpose of this last lesson was to evaluate the previous ones. The exceptional aspect of this test-lesson is that the students truly had to apply something new, namely, calculation in other number systems. The first part of the lesson introduced the Kingdom of Six. The context here was a Walt Disney film in which the characters had only three fingers on each hand. The students were asked a number of questions about the Kingdom of Six, such as how a familiar board game would look in that country. In the second part of the lesson, the students were presented with an open-ended problem in which they could earn a passport to self-invented countries, providing they were able to present a number of documents that they had created. These documents consisted of a segment of a counting sequence, a not too simple addition problem, a not too simple subtraction problem, and, optionally, one or more multiplication tables. The students worked in small groups and were provided assistance when needed. By observing the students at their work and discussing the solutions afterwards, it could be seen which children were independently able to transfer the essential properties of the positional system to other number systems. It not only became clear how the children reasoned and the specific difficulties they experienced but, also, what they were able to achieve with some assistance. An example of this was a student named Yvonne, who, during a class discussion on the Kingdom of Twelve, discovered that two extra single-digit numbers were needed. She suggested that two Chinese symbols be used for this.

By concluding the series of lessons with an test-lesson, the evaluation remained entirely within the educational process, rather than occurring after the fact. In Treffer's words, an internal process evaluation thereby takes place, rather than an external product evaluation (Treffers, 1978, 1987a).

1.2.5c Other new ideas for written tests

In addition to their innovative form, these last two test-lessons were also innovative in terms of content. This has to do with the problems that were used for them.

- a context problem as an introductory test of new subject matter

In the Winter Supplies test-lesson, a context problem was used as a way of forging links to new instructional activities for learning the multiplication algorithm.

- students' own productions

In the Kingdom of Eight test-lesson, students were tested on their insight and skills with respect to other number systems through their own productions in the form of 'foreign documents'. In a number system of their own choosing they had to notate a segment of the counting sequence and invent a few problems.

A second example of students' own productions is Van den Brink's (1987, 1989) invention of 'students as arithmetic book authors'. Here, the first-graders' assignment was to make an arithmetic book for the students who would be entering first grade the next year. This idea was used in order to ascertain to what extent differences in instructional approach between two classes would be conveyed in the students' own productions. In addition to the fact that their arithmetic books revealed distinct differences between the two classes, the students' new role as authors also provided a great deal of information on the skill level of individual students.

- doing the forbidden on a mental arithmetic test

Another example of an altered perspective can be seen in a mental arithmetic test that was part of the General Test (see Section 1.2.4c). Ter Heege and Goffree (1981) allowed a student named Marieke to do something that is usually forbidden on mental arithmetic tests: use scratch paper. The reason behind this was not to turn mental arithmetic into column arithmetic, but rather to display the flexible thought processes that are essential to mental arithmetic.

- a new form of assessment for Mathematics A

And then there were the developments in secondary education where, within the framework of the HEWET[33] project, a new curriculum of applied mathematics was designed for the highest grades of VWO[34]. This new curriculum, which acquired the name 'Mathematics A'[35], created the necessity of designing new forms of assessment as well (De Lange, 1985b). The goal of Mathematics A, which was planned according to the principles of RME, was the mathematization of real-life problems. This meant that an alternative would have to be found for the traditional restricted-time written tests. While knowledge and skills can be assessed using a restricted-time test, such tests are inadequate for displaying the mathematization process, critically evaluating and refining the models used, and so forth. Although it was obvious to all concerned that other means of assessment would be necessary, designing tests that would fit the curriculum was not easy.[36] The first steps in this direction were taken in 1981-1982, concurrently with the first on-site experiments at the schools. The following assumptions were essential to the design of these alternative tests: (i) the tests must contribute to the learning process, (ii) the students must have the opportunity to show what they know, (iii) the tests must cover the goals of Mathematics A, (iv) the quality of a test is not determined in the first place by the ability to score it objectively, (v) the tests must fit fairly well within the classroom practice (De Lange, 1987a; see Chapters 3 and 4 for more on these RME principles for assessment problems). In the course of time, four alternative assessment tasks were developed and put into practice at various schools (see also Section 1.4.1).

1.2.5d Observation and interview techniques

Observation and interview techniques occupy an important place in the concepts relating to assessment that were developed within RME. The RME viewpoint regarding learning processes can be recognized clearly here. Rather than focusing on the registration of externally perceptible behavior, these observation and interviewing techniques are primarily intended to display the students' underlying thought processes and insights.

- Freudenthal's walks with Bastiaan

The pioneering research for these observation and interview techniques was conducted by Freudenthal. His 'Walks with Bastiaan' (1975a, 1976c) provided, as it were, an alternative for the clinical interview method developed by Piaget. Although it was Piaget's intention that this method would examine children's thought processes through a flexible, unstructured and open-ended manner of questioning (Ginsburg, 1981; Ginsburg et al., 1983), Freudenthal (1973, 1978b, 1979c, 1983a) raised vehement objections to the way in which this took place. As indicated earlier (see Sections 1.2.4e and 1.2.4f), he objected, among other things, to the artificial nature of the questions, the language used, and the rigid manner in which the answers were evaluated. If the student failed to use a certain word, for instance, the answer might be marked incorrect, so that it became the verbal skills, rather than the mathematical content of the thought processes, that were actually being assessed. Freudenthal, on the contrary, consistently attempted to see beyond a student's incorrect formulations. Another significant point of difference was that Freudenthal's observations, rather than taking place in an artificial laboratory environment, occurred in a more or less impromptu fashion in a natural setting, such as during a meal or while taking a walk. Freudenthal (see, among other things, 1979c) was thereby able to delve deeply into children's thought processes and trace the first signs of the formation of mathematical concepts – what he called the 'constitution of mental objects'. Inspired by Freudenthal, Streefland (1978) used observation in the same way in his research into the mental constitution of the concept of fractions.

- interview-lessons for the benefit of educational development

  Interview-lessons are related to the above, but contain more initiative on the part of the interviewer, who has a particular goal in mind when presenting specific problems to the students. Van den Brink (1980) used these interview-lessons in order to ascertain which contexts were appropriate for instructing a given subject matter. Brief, individual interviews, having the character of a lesson, were conducted in order to determine whether the invented contexts formed a close link with the children's mental world and to find footholds for instruction. For the benefit of the analysis, the behavior of both the students and the researcher was recorded.

- innovative elements in observation

  During the test-lessons a number of innovative elements can be identified in the technique of observation as well. Firstly, there was the increased number of points to be given attention, as mentioned earlier. In addition to the applied solution strategies, attention was paid to affective and emotional behavioral aspects and to the manner in which students collaborated (Ter Heege and Treffers, 1979). Secondly, the observation was characterized by the exercise of considerable restraint – which should not be interpreted as detached observation. The principle here was to listen closely to the children and not to intervene too quickly. Or, in other words, when intervention was necessary, it was done as subtly as possible and the footholds for intervention were sought, whenever possible, in the child's own thought processes. Among other things, the interviewer could use the following techniques to assist the children: raising their awareness, guiding them back to a lower level of activity, having them verbalize the situation, making them aware of the point of the arithmetical activity, having them invent a concrete application for the bare problem, and urging them to find short cuts (Treffers, 1979).

- discussions with underachievers

  Ter Heege's discussions with underachievers (Ter Heege 1980, 1981-1982; Ter Heege and Treffers, 1979) provide an example of how arithmetic problems can be diagnosed in another way than by simply counting right and wrong answers. His method is, in fact, diagnosis and remediation in one: in an individual interview, an attempt is made to fathom a student's problems and to find links to remedial activities. In addition, however, prods in the right direction are already given during the discussion. Alongside this bilateral quality of diagnosis and remediation[37], the discussions are also characterized by a high degree of reflection on the part of the interviewer. The interviewer is astonished by certain observations and answers, wonders what these might signify, invents alternative approaches and observes the student's reactions.

- the technique of reciprocal observation

Of these new interview and observation techniques, one which proved to be extremely productive – and was also developed in the most detail – was Van den Brink's (1981a, 1981b, 1989) technique of 'reciprocal observation'.[38] This technique arose from assessment discussions between Van den Brink and first-grade children. When he found that the children were having to wait rather long while he took his research notes, he began to read these notes back to them. To his surprise, the children reacted by correcting him and even by helping him take his notes. They began of their own accord to think out loud about their own activities and reasoning, and became aware of what they had or had not done or thought. Together with the researcher, they became a kind of co-observer. As test subjects, the children were closely involved in the research work and could provide the precise information needed by the researcher. The richness of this technique is due to the fact that it produces much more information than can be acquired by simply asking children 'why' they did an activity in a certain way; many children simply answer such questions with 'just because'. By being involved in the research, on the other hand, not only do the children comprehend much more clearly what it is the researcher wants to know, but they also have the opportunity to provide a more complete answer.

1.3 RME-related developments in assessment

This picture of assessment in RME would not be complete without a description of a number of other developments in the area of assessment. Although these developments were closely related to the RME viewpoint, they did not arise within the framework of the IOWO or its successor, the OW&OC. Specifically, these were the Kwantiwijzer instruments, the Pluspunt gauges and the test designed for 'Arithmetic in a Second Language'. In contrast to the RME alternatives described in the previous section[39], these instruments all appeared in a more or less commercial edition and were therefore more widely disseminated.

1.3.1 The Kwantiwijzer instruments

The Kwantiwijzer is a set of diagnostic instruments whose purpose is to track down and tackle arithmetic problems in the lower grades of primary school (Van den Berg and Van Eerde, 1985). The history of this set of instruments goes back to 1975. Recently, an abbreviated version of the Kwantiwijzer was published that is more manageable for teachers (Van Eerde, Lit, and Van den Berg, 1992).[40]

One of the most important principles in the development of the Kwantiwijzer was that of activity theory, which regards behavior as a focused activity and not, for instance, as a reaction to stimuli (Van Parreren, 1981). This foundation in activity psychology is expressed, among other things, by a great interest in the activity structure, that is, in the activities performed by children while solving a given problem. In addition to activities employing manipulatives, these may also include mental and verbal activities. The work of Wiskobas was also influential in the development of the Kwantiwijzer, and a number of problems in the Kwantiwijzer are based on Wiskobas work (Van den Berg and Van Eerde, 1983b).

A great number of research techniques for discovering how children work are described in the Kwantiwijzer. Among these are

- observation,
- introspection (asking the student to think out loud),
- retrospection (asking the student after the fact to describe what was done or thought),
- continued questioning (repeating the question in another way or attaching a new question to an incomplete answer),
- mirroring (encouraging reflection by demonstrating the student's own activity or that of another student),
- problem variation (offering a different problem of the same degree of difficulty, a more difficult problem or a less difficult problem),
- offering assistance (providing the student with material, solving the problem together and then having the student solve a similar problem, pre-structuring the solution strategy, drawing attention to potential errors, etc.).

Many of these research techniques, albeit arranged much less systematically, can also be found in the interview and observation techniques propagated by the RME approach.

A characteristic of the Kwantiwijzer approach is that it does not follow a standardized interview procedure but, rather, allows the choice and order of questions to be determined by what the child says (Van Eerde and Van den Berg, 1984). The instruments do, therefore, place demands upon the user with regard to diagnosis and subject matter (Van den Berg and Van Eerde, 1983a). For this very reason, in fact, a special version was developed for teachers. Another significant characteristic of the Kwantiwijzer instruments is that diagnosis and remediation lie in one line: during the diagnosis, as much information as possible is gathered that could be useful for remediation (Van den Berg and Van Eerde, 1985). Moreover, the instruments are viewed in the broader framework of diagnostic instruction and of reflection by the teacher on his or her own instructional activities.

1.3.2 The Pluspunt gauges

Another example of such RME-related developments in the area of assessment are the gauges designed by Ter Heege for the NOT (Dutch Educational Television) program 'Pluspunt' (Scholten and Ter Heege, 1983-1984; Ter Heege, Van den Heuvel-Panhuizen, and Scholten, 1983-1984). This television program on mathematics for grades 2 through 5 was developed in order to bring Dutch teachers in contact with examples of RME. The program focused on improving arithmetical skills, flexible arithmetical behavior, insight into numbers, and on applying all of this to daily life. For each grade, in addition to the three television broadcasts, the program contained a set of worksheets, a gauge and instruction booklet for the teacher, and a book of background information on the teaching methods (De Moor, 1984).

Each gauge consisted of a variegated collection of approximately eight problems, which sometimes involved a series of similar problems. For second grade the gauge contained, among other things, the following: a problem involving counting a flock of birds; a series of 'machine problems'[41] on multiplication; a problem involving a boy who had saved 42 nickels; a series of bare problems that all resulted in the number 40; and a word problem in a foreign language with the accompanying bare problems, where the student's task was to make up a story in Dutch for these bare problems. In order to acquire as much information as possible on the applied strategies, the teachers were expected to observe their students as they worked on the gauge.

Moreover, the instruction booklet suggested that the students be encouraged to write down their thoughts on the test paper and that they also use the test pages themselves, rather than separate scratch paper, for notating interim solutions.

The exceptional characteristic of this gauge is that it was administered prior to the television program under the motto 'Know Your Class'. This decision to evaluate beforehand was due to the fact that this program's RME approach differed strongly in many aspects from the textbook then customarily used in the classrooms. At the time this program was broadcasted, RME textbooks had not yet gained as large a share of the market as they have today. The gauge was therefore also intended to open teachers' eyes to how children solve problems that deviate from more traditional ones. In order to intensify the potentially revelatory effect of the gauge, the teachers were advised to make a prognosis of their expectations before administering the gauge.

1.3.3 The test for 'Arithmetic in a Second Language'

The third example of RME-related developments in the area of assessment is the test entitled 'Arithmetic Puzzles' developed by Van Galen et al. (1985) for the 'Arithmetic in a Second Language' project. The occasion for developing this test was the issue of whether RME would also be suitable for students whose mother tongue was not Dutch. In other words, would these students have difficulty understanding problems presented in a context, and could they apply their own strategies.

The test developed for this purpose consisted of ten context problems and a series of bare problems. Following the lead of Dekker, Ter Heege, and Treffers (1982), the chosen context problems lay above the level of what the students had previously done, but could still be solved within the chosen context using already mastered skills.

The innovative aspect of the test was that, prompted by the unusual character of the target group, an attempt was being made to design a test in which language played as small a role as possible. Situations were sought that could be described in a few words. Illustrations, rather than text, played a major role. They formed a significant part of the test problems, either because they were essential to the situation description (see, for instance, the Bead problem in Figure 1.4), or because the illustration could be used to find the answer (see, for instance, the Candy problem in Figure 1.5; examples of solutions to this problem can be found in Van Galen and Meeuwisse, 1986). In addition to the space left in the illustration for making drawings, a strip of scratch paper was also reserved along the margin of each test page, on which the students could write and draw what they wished.

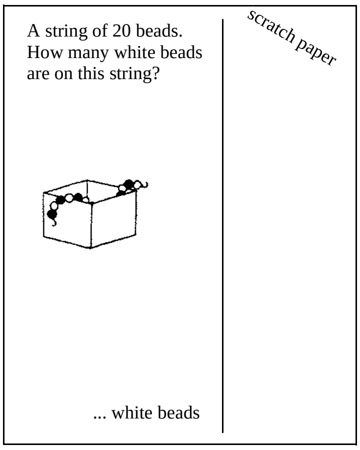


Figure 1.4: Bead problem

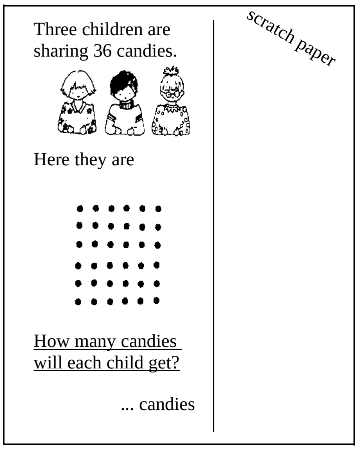


Figure 1.5: Candy problem

This test was administered in January/March, 1985, to 157 third-grade students, some of whom were native Dutch speakers and some not. The results revealed that the relation between bare problems and context problems was no different for students whose mother tongue was not Dutch than it was for those who spoke Dutch as a first language. Moreover, the additional information provided by the children's scratch paper scribbles revealed that no difference could be perceived between the approaches taken by the two groups.

1.4 1987: A noteworthy year

Although the development of RME is marked by a few crucial dates – the establishment and discontinuance of the IOWO, for instance – the innovation of mathematics education has, in fact, always taken place in a rather informal manner (see also Gravemeijer, 1994). Although an institute was indeed involved in this innovation[42], there was never a question of a strict, formal structuring in sub-projects, in which previously determined aspects and sub-sectors of the RME theory were investigated. The development of the RME teaching methods, therefore, also eludes formal classification. There were, however, certain instances during its development in which the emphasis was shifted or in which an approach to a particular domain was initiated or concluded.[43]

In terms of the issue of assessment, such an instance was 1987. Not that a specific development was concluded in that year or that a new breakthrough occurred. On the contrary, business proceeded as usual. Nevertheless, seen in retrospect, there are a number of reasons to justify regarding 1987 as a noteworthy year for assessment: the publication of the dissertation of De Lange, the administration of the PPON tests and the OW&OC conference on assessment.

1.4.1 The dissertation of De Lange

The first research project into new forms of assessment appropriate to RME reached its conclusion with the publication of De Lange's dissertation (1987a). The HEWET project referred to here produced four alternatives to the traditional restricted-time written tests (see also Section 1.2.5c).

The most obvious alternative, considering the RME predilection for observation, was the *oral task*.[44] Various types of oral tasks were developed, such as an oral task on a mathematical topic announced to the students one week earlier, and an oral task on a written task already completed by the students.

In addition, three alternative written forms of assessment were developed by the HEWET project: the 'essay task', the 'take-home task' and the 'two-stage task'.[45] For the *essay task* the students had to write, for example, an analytically supported reaction to a newspaper article. For the *take-home task,* the students were given a choice of a number of essay-like assignments to be completed at home, either individually or in groups; if they wished, they could use the textbook and ask others for help.

The most startling new alternative was the *two-stage task*. As described by De Lange (1985b), it was Van der Blij who, concerned by the issue of objective scoring with respect to the take-home task, invented the idea of the two-stage task. This task was first administered at school, corrected by the teacher who supplied written comments, and then returned to the student, who took it home to improve upon it. The results were impressive. The students who had fallen short in the first round seized the opportunity to make corrections.

On the whole, these alternative forms of assessment fulfilled their purpose, that is, they satisfied the principles formulated earlier (see Section 1.2.5c). Because of the participating schools' concern regarding the objectivity of the scores, a separate research project was conducted into scoring objectivity. The various evaluators of the students' work were in sufficient agreement with one another. Evaluation of this open form of assessment, however, did prove to be exceedingly labor intensive, as, indeed, was the designing of the test problems.

1.4.2 The PPON tests

In 1987, the first PPON (National Assessment of Educational Achievement) study[46] was carried out on the topic of mathematics (see Bokhove, 1987; Wijnstra (ed.), 1988). The tests used for this study signified an important step in the direction of a better harmony between the tests and the altered viewpoints with respect to mathematics education. These PPON tests, which were designed in close consultation with experts on mathematics education, signaled the advent of a new generation of CITO tests. In contrast to previously administered primary school tests, these tests mainly contained open-ended questions, in which the students could formulate the answer themselves.

In addition, more than half of each test consisted of application problems, which were designed to contain meaningful contexts recognizable to the students. Moreover, as a trial, another research project into the students' solution strategies was conducted that examined the students' notes and their responses to individual questioning. The educational experts who were asked for their opinions regarding the PPON tests (see Huitema, 1988; Teunissen, 1988; Treffers, 1988) were, on the whole, fairly satisfied with the quality of the problems, although the presentation – particularly the linguistic aspect – did receive some criticism. With the exception of a few sections that were considered too difficult, the consensus was that the attainment targets for primary school mathematics had been adequately covered by the subject matter content.

1.4.3 The OW&OC conference on assessment

Another event that took place in 1987 was the conference organized by the research group OW&OC entitled 'Assessment, Attainment Targets and Viewpoints on Mathematics Education'. This was the first time the proponents of RME had taken the initiative at a conference to focus explicitly on the subject of assessment. And there were, indeed, those who saw this as a sign that RME was bowing to the pressure from the educational scientists to place assessment in a central position in education (see Van 't Riet, 1987). Their fears were soon allayed, however, as it became clear that the viewpoints on assessment had not been altered. The occasion for the conference had, in fact, been the dissatisfaction with secondary school exams and the hope that these could be improved through the formulation of attainment targets.

This dissatisfaction had primarily to do with certain already administered Mathematics A exams, that had turned out to be entirely incompatible with the purpose of Mathematics A (De Lange, 1987b; Van der Kooij, 1987). There was dissatisfaction as well with the MAVO and LBO exams[47] because they, in contrast to the exams for upper secondary education, consisted almost entirely of multiple-choice questions. Besides the familiar objections to multiple-choice, which were expressed by others as well (see, among others, Van der Blij, 1987; Broekman and Weterings, 1987; Van Hoorn, 1987; Zwaneveld, 1987), the specific criticism here was that this type of assessment laid an enormous extra burden on weaker students in particular. And this burden could have been avoided by formulating the problems in a less complicated fashion, as had, indeed, occurred during the lessons (Querelle, 1987; Tessel, 1987).

The objection to the Mathematics A exams concerned their failure to test the higher-order goals, a failure which was seen as a great threat to the continued existence of the subject as originally designed (De Lange, 1987b). According to De Lange, these inadequacies were caused by the fact that the HEWET experiments had been conducted too hurriedly and because of the lack of opportunity, when designing the exams, for applying experience gained from the research into alternative forms of assessment.

Prior to introducing Mathematics A, it was felt that a three-dimensional goal description should have been created that included examples of test questions. At the conference, it was Treffers (1987b) who argued strongly in favor of this, as he had done before with respect to primary education (see Section 1.2.5a). He pointed out, moreover (as had De Lange), that such a three-dimensional goal description can only be possible once the development of the education has reached a certain stage:

> "If attainment targets and tests are introduced at an early stage of a fundamental innovation, this will inevitably lead to an unfocused innovation" (Treffers, 1987b, p. 149; translation of original Dutch text).

With reference to the above, the conclusion may therefore be reached that criterion-referenced tests, which were introduced into primary education in the early nineteen-eighties, arrived too early in the development of RME. They were thus regarded as a threat to its further development and implementation.[48] The fact that the situation, with respect to primary education, had clearly changed by 1987 could be seen from the decrease in opposition to these tests – in spite of the fact that they were still on the market.[49]

At the conference, too, it was clear that primary and secondary education had experienced differing histories of assessment with regard to RME. Considering the fact that the elaboration of RME had solidified earlier within primary education than within secondary education (take, for instance, the 'National Plan for Mathematics Education'[50] and the increasing number of RME textbooks), this is not particularly surprising. Secondary education, moreover, carries the burden of final exams, which explains its increased focus on the issue of exams. Primary education, in contrast, thanks to the absence of a true final exam[51], is in the enviable position of being able to concentrate on 'didactical assessment', that is, assessment for the express purpose of educational decision making.

Although, in 1987, the final CITO test[52] for primary school did meet with some criticism, particularly with respect to the breadth of the areas tested (Treffers, 1985, 1987b), a fair degree of satisfaction prevailed regarding the PPON tests. The same is true of the assessment in the new RME textbooks. These were considered to contain a broad spectrum of forms of assessment in order to afford the teachers both insight into the skill level and into the students' approach and solution behaviors which, in accordance with the RME viewpoint, could occur on various levels (see Buys and Gravemeijer, 1987; Huitema and Van der Klis, 1987; Nelissen and Post, 1987). In the secondary school textbooks, on the other hand, tests played a distinctly subordinate role. They only appeared in segments intended for lower grades and were occasionally included in the teacher's guides in the form of suggestions for test questions (Van 't Riet, 1987).

1.4.4 Conclusion

Considering the number of events that took place in the mid-nineteen-eighties, one would assume that the topic of assessment had by now acquired its own spot within the Dutch reform movement in mathematics education. But not so much had changed after all. It is worth noting that De Jong's (1986) voluminous dissertation (on the degree to which 'Wiskobas' characteristics[53] can be found in primary school mathematics textbooks) makes no mention of any different form of assessment that would be more appropriate for RME. Perhaps this is a kind of legacy from the early years of RME.

Nor with the exception of the Kwantiwijzer project, is anything on the development of new assessment methods to be found in Goffree's (1985) overview of 1984 research projects on mathematics education.[54]

In Freudenthal's opinion, too, nothing had changed in 1987. His comments upon browsing through a manuscript on mathematics education that he had written in 1942 but never published were as follows:

> "At that time I had been willing to allow test development the benefit of the doubt; now nothing remains but the doubt" (Freudenthal, 1987a, p. 338; translation of original Dutch text).

In addition to De Lange's dissertation, the PPON tests and the OW&OC conference on assessment, there is one more reason to view 1987 as a noteworthy year, namely, the MORE research project, which was begun during that year. The MORE project involved research into the implementation and effects of mathematics textbooks in primary education (see Gravemeijer et al., 1993). Tests had to be developed to measure these effects – which meant that this research was originally intended to deal only obliquely with the issue of assessment. However, as the research progressed, this marginal interest for assessment changed substantially, as will be described in the following chapter.

Notes

1 For a description of these different trends, see Treffers (1978, 1987a, 1991a).

2 This is still true, in fact. The continued development of the Dutch reform movement has much in common with, for example, the recent constructivistic approach to mathematics education.

3 Wiskobas is a Dutch acronym for ‘Mathematics in Primary School’.

4 In 1981, this institute was officially disbanded by the government. Its numerous responsibilities (research, development, pre-service and in-service education) were divided among various other institutes. The research work was set forth by a newly founded institute entitled OW&OC (Mathematics Education Research and Educational Computer Center). In 1991, a year after the death of Freudenthal, the name of the institute was changed to ‘Freudenthal Institute’.

5 Kilpatrick failed to include the establishment of the IOWO in his survey.

6 Without underestimating his contribution, it should nevertheless be noted that the ideas on, for instance, themes and projects, already existed within the Wiskobas group before Freudenthal became a participant (Treffers, 1995).

7 And this concept is what distinguishes RME from the other trends at this confluence of four currents (see Treffers, 1978, 1987a).

8 This is a variation on a comment by Comenius: “The best way to teach an activity is to show it” (see Freudenthal, 1971, p. 414).
As will later be apparent, this activity principle of RME should not be confused with such things as ‘hands-on’ mathematics, in which manipulatives play a major role in conveying certain insights and procedures to the students (see also Gravemeijer, 1994).

9 Later, Freudenthal (1991) would call this ‘guided re-invention’.

10 In Dutch ‘zich realiseren’ also means ‘to realize’ in the sense of ‘to picture or imagine something concretely’.

11 The term ‘realistic mathematics education’ dates, moreover, from 1979 (see Treffers (ed.), 1979; Treffers, 1992).

12 Freudenthal questions, moreover, why development is always seen as progression and not regression. In his opinion, this question has never been posed (Freudenthal, 1978a), even though numerous examples indicate that students may be able to do something at one time that they are later unable to do (see, for instance, Freudenthal, 1973).

13 IEA stands for the International Association for the Evaluation of Educational Achievement, a society of prominent educational researchers that, since 1964, has been engaged in comparing student achievement internationally. At the present time, the third of such research projects, entitled ‘TIMSS’, is being conducted by the IEA.

14 The last time Freudenthal used this quote, he added that others involved in this research project did not hold such high expectations of achievement testing.

15 This idea certainly played a role in the planning of this chapter.

16 Take, as an example, the interviews with kindergartners and first-graders during the early years of the IOWO. The purpose was not to register the students’ errors, but to penetrate the children’s mental world and thereby understand how certain insights could be conveyed to them (see Van den Brink, 1980).

17 According to Oonk (1984) it is not at all coincidental that the first RME textbooks attached a great value to observation.

18 In other words, mechanistic, rule-driven education that is based on the principle of ‘demonstrate and imitate’, in which “...the pupil can be taught to parrot the ready-made mathematics he learned” (Freudenthal, 1973, p. 117).

19 These are lessons in which both the assessment and the teaching takes place by working through a test. See also Section 1.2.5b.

20 The extent of this distaste can be seen in Freudenthal’s (1978a, p. 88) reaction to the handbook by Bloom, Hastings, and Madaus (1971) on evaluating learning results that appeared

around that time: “The chapter on mathematics contains stuff (...) that comes straight from the horror and lumber cabinets of old mathematics instruction.” Freudenthal’s (1984b, p. 21) comments regarding the home-grown CITO tests (see Note 21) were no softer: “The ‘Criterion-Referenced Tests’ for primary education [see Note 23] are so bad that not even a blind horse would let himself be hitched behind this cart” (translation of original Dutch text). Humor was not lacking, by the way. The chapter in which Freudenthal (1978a) so fiercely criticized Bloom was given the lyrical title ‘In Full Bloom’.

21 CITO stands for National Institute for Educational Measurement.

22 Here Freudenthal (1981b) makes the comparison with spelling, in which more errors may be made in a long sentence than in a short one.

23 In Dutch these are called ‘Leerdoelgerichte Toetsen’.

24 Although both Freudenthal and Treffers point to the absence of process goals, their emphasis differs on the interpretation of such goals. When Freudenthal (1984b, p. 22) speaks of process goals, he means: “... educational goals in which the stepwise acquisition of certain concepts by means of networks of short cuts, generalization, abstraction, compilation and schematization should be understood and formulated as a goal in itself” (translation of original Dutch text). Treffers (1978, p. 155-156), on the other hand, links process goals more closely to the necessary instructional activities: “Such a process goal can only be described by naming the activity itself (...): working mathematically together. (...) A test in which only a number of product goals were tested, would never fully cover the intentions of the education” (translation of original Dutch text).

25 This test was developed by the Utrecht School Consultant Service (SAC) and was administered at the conclusion of primary school in the district of Utrecht. In Dutch the test is called ‘Algemeen Proefwerk’.

26 BOVO tests are regional tests developed by school consultant services to be administered at the conclusion of primary school.

27 Or, as Freudenthal (1984c; see also 1991, p. 155) would say: ‘cheap’.

28 Freudenthal, (1979c, p. 1) wondered how it was possible that: “... say, Piaget and his collaborators managed to interview thousands and thousands of subjects, without even being asked the question ‘what do you mean?’, even not by subjects who obviously did not understand anything. (...) An interviewer who would allow such questions would be lost. Answering questions would mean putting one’s cards [on the table] and losing the game. The game against the child.” With respect to the children’s alleged ‘misconception’ of conservation, (that has arisen in many conservation experiments), research conducted by Donaldson (1978) has revealed that this has to do with which questions are put to the children.

Seen in the light of Freudenthal’s comment, the following quote from Ginsburg, Jacobs, and Lopez (1993, p. 158) is certainly a bit peculiar: “Piaget’s investigations led to the conclusion that children’s thinking is often different from the adults and that even a wrong answer may result from interesting – and indeed ‘logical’ and sensible – thought processes.”

29 Although this quote refers to questions put to Freudenthal, himself, rather than to children, it expresses well where the responsibilities of assessment lie. Take note of the exchange of perspective applied here, which is so characteristic of RME.

30 Although these two analyses are often lumped together, they do exhibit notable differences. While Freudenthal’s didactical-phenomenological analysis lies close to mathematics and the mathematical phenomena in reality, the mathematical-didactical analyses as interpreted by Treffers lie closer to educational experiences and educational theory.

31 Treffers and Goffree’s (1982) suggestion that test-lessons be developed on a national level was never taken up, although the Pluspunt gauges (see Section 1.3.2) do come close.

32 With respect to this issue, see also Freudenthal, 1983b.

33 HEWET is a Dutch acronym for ‘Reshuffling Mathematics I and II’.

34 VWO means pre-university secondary education.

35 The topics here were applied analysis, matrices and applications, probability and statistics, and data processing (De Lange, 1985b).

36 This was first revealed while tests were being developed (see De Lange and Verhage, 1982), became clearer during the designing of school examinations, but was the most obvious during the construction of nationwide exams (De Lange, 1985).

37 As Treffers (1993b) pointed out, these discussions could also lead to the development of a new course.

38 In the past, also described as 'mutual observation'.

39 This is primarily true of the alternatives for primary education. The tests developed in the framework of the HEWET project were available to the schools via the series of publications entitled 'Hewet & Toets' (Hewet team, 1984; De Lange (ed.), 1985a, 1985b; Kindt (ed.), 1986a, 1986b).

40 Due to the breadth and depth of the instruments, the earlier versions were primarily used by child psychologists and other researchers of children.

41 These are problems in which a 'machine' operates on the numbers. A 'times 8' machine for instance, will change 3 into 24.

42 After 1981, when the IOWO was disbanded, there were in fact numerous institutes and research groups involved in the reform of mathematics education: OW&OC (now called Fi), SLO, the PANAMA project of the SOL (now called HMN), the NVORWO and its work groups.

43 An example of this is the development of RME itself. According to Treffers (1992), the emphasis during the first decade lay more on the development of a viewpoint and its horizontal elaboration in the thematic component of mathematization, while, in the second decade, the focus turned to the vertical structuring of learning strands.

44 The Netherlands has a long tradition of administering oral tests at the conclusion of secondary education. Only in 1974 were these discontinued.

45 De Lange calls these 'tasks' but, where an assignment consists of more than one task, these might also be called 'tests' (see Section 4.1.2a).

46 This is a large-scale assessment project into the level of primary education in The Netherlands that examines third-grade and sixth-grade students every five years. In addition to mathematics, the assessment covers a number of other subjects as well.

47 MAVO and LBO are Dutch acronyms for, respectively junior general secondary education and junior secondary vocational education.

48 Treffers (1983) mentions an interview with the well-known Dutch soccer player, Van Hanegem, to make his point. Just as, according to Van Hanegem, permissible behavior during a soccer match depends on the score at that moment, so, in Treffers' opinion, does permissible assessment in mathematics education depend on the state at that moment of educational development.

49 And they are still on the market (see NICL, 1992; NICL, 1993).

50 See Treffers and De Moor (1984). From this first initiative for a national curriculum for primary school mathematics education later arose the series of publications entitled 'Design of a National Curriculum ...' (see, among others, Treffers, De Moor, and Feijs, 1989). In Dutch this series is called: 'Proeve van een Nationaal Programma voor het rekenwiskundeonderwijs op de basisschool'.

51 The final test administered by most schools at the end of primary school (in most cases this is the CITO Test) cannot be compared with the exams taken by students at the end of their secondary education. Only if one wishes to be admitted to HAVO or VWO – the higher levels of secondary education – a test at the end of primary school is required. In other words, the stakes are not nearly as high for the final test for primary school.

52 Although this test still consisted of multiple-choice questions with four options, it nevertheless was not subjected to such heavy criticism (see Treffers, 1980a, 1980b, 1983).

53 These are the characteristics of RME that were developed within the Wiskobas project (see Section 1.1.1).

54 Only Treffers (1985) is indicating in his contribution to this publication that one might expect that the already existing friction between the final CITO Test for primary school and mathematics education may increase.

2 The MORE research project as a field of exploration into assessment

2.1 A summary of the MORE research project

The new developments in mathematics education in The Netherlands (see Chapter 1) and the corresponding shift in school textbooks from the mechanistic manner to the RME approach (see De Jong, 1986) raised the issue of the implementation and effect of the textbooks in question. One of the research projects that was established to investigate this issue was entitled the MORE project.[1] This project was subsidized by the Institute for Educational Research (SVO) (project number SVO-6010). It was conducted by the OW&OC research group (renamed 'Freudenthal Institute' in 1991) in collaboration with the Educational Science Department (VOU) and the Interdisciplinary Research Institute for Social Sciences (ISOR), all are part of Utrecht University. This research project began in 1987.

The following sections offer a brief summary of the design of the research and the results that emerged. This discussion is based on the final report of the MORE project (Gravemeijer, Van den Heuvel-Panhuizen, Van Donselaar, Ruesink, Streefland, Vermeulen, Te Woerd, and Van der Ploeg, 1993). The summary will serve as background information for the description of assessment development within the MORE research project which follows.

2.1.1 The goals and design of the research

The project involved a comparative study of two different domain-specific educational theories: realistic and mechanistic. The RME theory, as described in detail in the previous chapter, implies an approach to education characterized by, among other things: use of contexts and models, students' own constructions and productions, interactive education, and links between learning strands. It is, in many respects, the antithesis of the mechanistic approach which entails: step-by-step construction, bare problems before applications, instrumental instruction, fixed instructional approach, and extrinsic motivation. This theoretical framework of the mechanistic approach to mathematics education strongly resembles the theory of learning and instruction based on task analysis propounded by Gagné (1965).

The goal of the research project was to determine whether and to what extent the approach in question – respectively realistic or mechanistic – had actually been implemented in the classroom, and which factors were influential in this implementation. The research also intended to portray the potential consequences of textbook choice or type of mathematics education for the learning results. The issue here was not whether one textbook produced *better* results than the other, but to what extent the results *differed* with respect to the commonly accepted (communal) mathematics goals.

A longitudinal design was implemented for this research, involving approximately 430 students and their respective teachers, who were followed for three years, from first through third grades. Two groups of primary schools participated in the research. The first group consisted of eight schools that were using the mechanistic textbook ‘NZR’.[2] The second group comprised ten schools whose realistic textbook was entitled ‘WIG’.[3] In most of the schools, a single class took part in the research, but some schools had more than one participating class. The project took place in an actual educational setting, whereby the complexity typical of educational reality was also reflected in the research. Not all the data could be fully discovered according to the research precepts, and the gathering of information was at times disturbed by unintended changes in the student body or teaching staff. Rather than remove these imperfections by artificial means, an attempt was made to do justice to the complexity of education by approaching the education from numerous perspectives. At certain times a quantitative approach was chosen, at other times a qualitative one.

During the three years that the research project was being conducted, data was collected on the education at hand (the implemented curriculum), the teachers’ viewpoints and the learning results. In order to answer the question of effect as accurately as possible, supplementary information was also gathered on a number of background characteristics pertaining to the students and the educational conditions.

2.1.2 The instruments developed and the data they collected

In nearly all cases, the instruments used to collect the data were developed specifically for this research project.

For ‘type of education’, for instance, an analytical instrument consisting of rating scales was designed to measure the realistic and mechanistic characteristics mentioned above. A group of experts on mathematics education used this instrument to rate the mechanistic and realistic calibre of the lessons. For this purpose, an audio tape was made of at least three of each teacher’s mathematics lessons per year, which were then transcribed. In order to discover the teachers’ viewpoints, a list of written questions was designed, containing both general questions on the subject and the teaching methods, and specific questions concerning the micro-didactical approach. With respect to the learning results, both the students’ skill level and the strategies they applied were examined. Written tests for the entire class and individually administered oral tests were developed to measure these results. The written tests[4] were administered by the teachers, and the oral tests by the researchers themselves. The written tests were administered four times per year. Only four students per class participated in the oral tests, which were administrated less frequently than the written ones (see Figure 2.2).

One of the written tests served as an entry test for determining the students' numerical knowledge and skills at the beginning of first grade. It was administered three weeks after the start of classes. The students' intelligence was also measured, using a standard test (Raven's Progressive Matrices). This test was administered twice, once in first grade and then again in third grade. The students' socio-economic status was also determined as an indication of their background.

With respect to the educational conditions, a list of questions was used to determine how much time was allotted to mathematics instruction and the way in which the class was organized. The amount of time spent on real learning activities during each lesson was also investigated using an observational instrument that produced a score every ten seconds.

A textbook analysis was also conducted alongside this data collection. This was done in order to gain insight into the subject matter contained in each of the two textbook series.

2.1.3 The research results

The quantitative and qualitative analyses revealed that, while a different textbook did lead to different education, the implementation of RME was far from ideal. Moreover, as was indicated by the analysis of the lessons, it was not at all easy to instruct RME using a teacher's guide that is rather general. Instruction of RME demands specific instructional skills and knowledge. Aside from the scant support for putting RME into practice, another cause of the disappointing implementation may have been the fact that the WIG version used in the research project was clearly a first generation RME textbook. More recent RME ideas such as the role and place of basic skills had not yet been included. The mechanistic textbook, NZR, on the other hand, was indeed, for the most part, implemented according to the intent. Moreover, the character of the first three sections of NZR – intended for first and second grade – clearly differed from those which followed. These initial sections did offer a rich variety of learning activities on number concept.

The data on the teachers' viewpoints indicated that these did generally correspond with the intended curriculum. There was a striking discrepancy, however, between the more general viewpoints and the viewpoints on specific aspects of education. Many teachers appeared to have difficulty transferring their general realistic viewpoints to a more specific level of instruction. These findings clearly indicate that in-service education must focus primarily on micro-didactics.

The content analysis of the two textbooks revealed considerable differences in subject matter content. Moreover – as mentioned before – the composition of the textbooks did not prove to be homogenous: not all NZR sections were entirely mechanistic, nor were all WIG sections thoroughly realistic.

The differences in subject matter content was, indeed, revealed in the students' achievements. Viewed across the entire three years (see Figure 2.1),

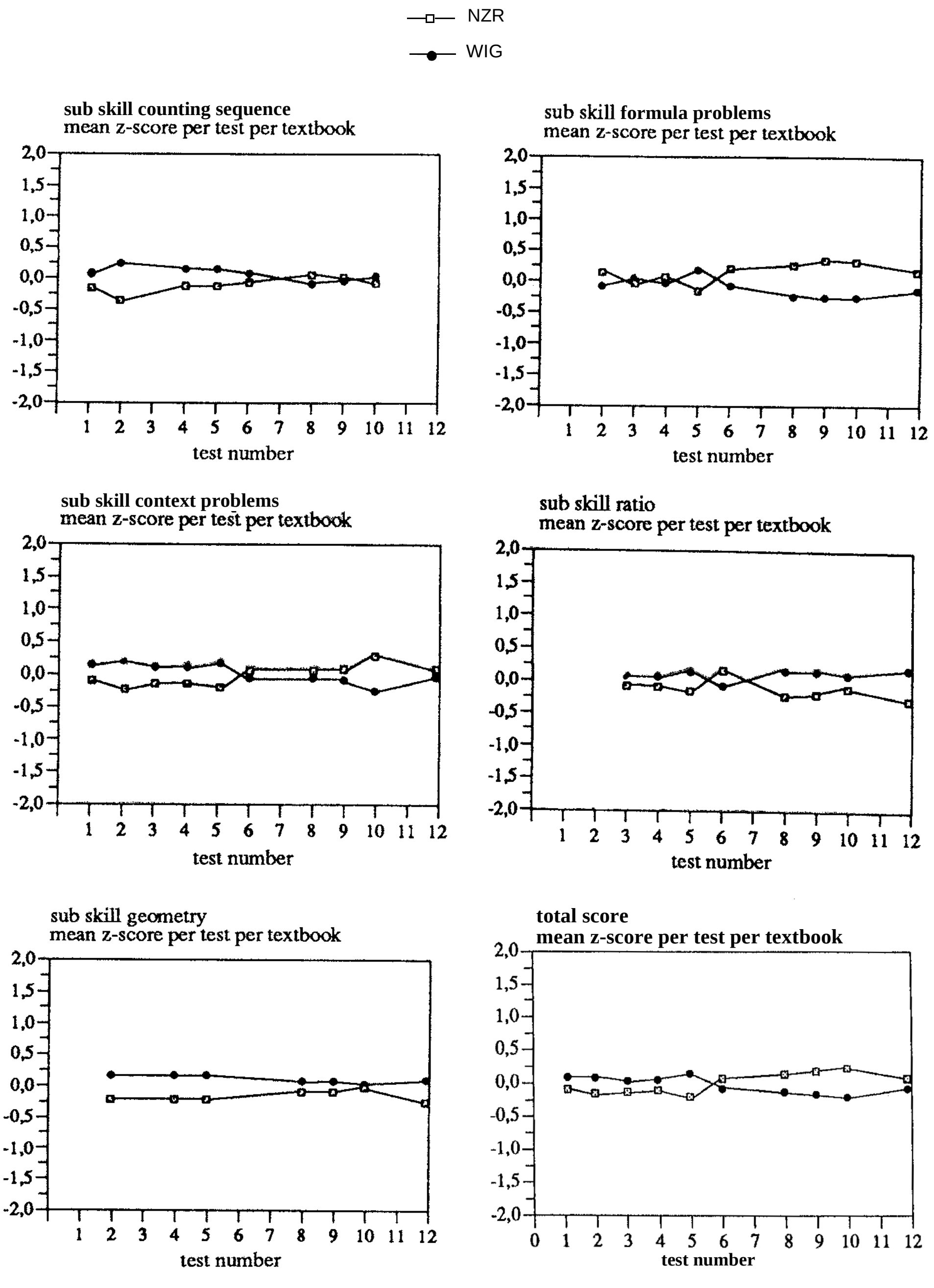


Figure 2.1: Overview of the test scores for each textbook during the three years[5]

the NZR students did better with formula problems while the WIG students were more competent with the sub-skills of geometry and ratio, and were more familiar, at first, with the counting sequence. The achievements with respect to context problems varied. The WIG students did better at first and the NZR students later on. By the end of the research period, the scores were almost even.[6]

That the content of a textbook can indeed influence results was also clearly revealed in a separate analysis of certain specific arithmetic topics. The WIG first-graders, for example, did better on problems using numbers greater than 20, while their NZR colleagues achieved higher scores on problems involving ‘bridging’ the ten. The NZR students also demonstrated a better knowledge of the basic arithmetic facts. Which textbook was used can also be seen, albeit to a lesser degree, in the kind of strategies applied. On the whole, the results clearly indicated that choice of textbook is significant. The conclusion here is that the textbook can be a powerful instrument of innovation with regards to the educational content.

The research failed, however, to produce unequivocal results with respect to the influence of the type of education. The meager realization of RME in classes using the WIG textbook made it difficult to determine its effects.

2.1.4 The instruments’ crucial role

Aside from the problems inherent in the first version of the WIG textbook, the instrument with which the lessons were analyzed may also have been responsible for the absence of realistic characteristics. Because the characteristics included in this instrument (see Section 2.1.2) were derived from the characteristics of RME courses (see Treffers and Goffree, 1985; Treffers, 1987a), they are of a fairly general nature, in the sense that they are not linked to specific educational topics or the specific nature of the lesson.[7] In hindsight, one may well wonder whether these characteristics were in fact the best choice. A RME characteristic such as the use of context, for instance, takes an entirely different form in an exploratory lesson on making and describing buildings made of blocks than it does in a practice lesson on the multiplication tables. Although, in an early stage of the research, it had become clear that evaluating the implemented education using the categories in question would not always be easy, only during the analysis of the second-grade lessons did it dawn on the researchers that the general character of the categories might actually be responsible for these difficulties.[8] That this had not become clear earlier is not really so surprising, considering the fact that from that point on the diversity of topics steadily increased and the evaluation often involved extremely contrasting lessons. The research was by that time so far advanced, however, that it was no longer possible to adapt the instrument to encompass characteristics formulated on the lesson level.[9]

With respect to the measurement of the learning results, on the other hand, it was clear from the outset that the question of the effects of the two textbooks could only be adequately answered if the instruments were sufficiently specific to allow the illumination of potential differences. This viewpoint was reinforced further by disagreement with the research of Harskamp and Suhre (1986), whose final conclusion was that no significant differences in learning results existed between traditional and modern mathematics textbooks. The problem with the latter conclusion – which was in fact later contradicted by the results of the PPON research (see Wijnstra, 1988)[10] – was that it was based on analyses of total scores on tests administered at the very end of primary school. Moreover, the participating students had not all been taught using the same realistic or mechanistic textbook. The danger of making such crude comparisons is that the potential differences between students and textbooks will be averaged away. Consequently – at least with respect to measuring learning results – a more finely meshed design was chosen for the MORE research. This finer mesh was constructed by regarding specific topics as well as the long-term progress of scores, both of which can be seen in the graphs in Figure 2.1.

The following section will look closely at the instruments which collected this data on learning results and at how these instruments were developed.

2.2 Developing tests within the MORE research

This section examines in detail the test development within the MORE research project. An explanation is given of why the project needed to develop its own tests, followed by an overview of the tests that were developed. Lastly, the starting points and procedures used during test development are outlined briefly. This outline concentrates on the development of the group written tests.

2.2.1 The necessity of developing its own tests

Formulating a response to the question of the effects of the two textbooks on the learning results placed a number of demands on the assessment instruments. As the issue of effect was not intended to show higher achievement by one group of students over the other but, rather, to examine to what extent the results between the two groups differed, the administration of norm-referenced tests would not suffice. Such tests, which place the scores achieved by the research group against the scores of a control group, do not clearly indicate in what respect the students' skills and strategies differ. As a consequence, norm-referenced tests are unsuitable for curriculum comparison (see also Popham, 1975). An additional problem was posed by the fact that then existing norm-referenced tests, such as the 'Arithmetic Tests' developed by CITO and the 'Schiedam Arithmetic Test' were mechanistically inclined with respect to content.[11]

This same drawback applies as well to most criterion-referenced tests. In themselves, these tests are more suitable for curriculum comparison, as they do indicate whether or not the students have mastered certain topics. But, as stated above, most criterion-referenced tests, such as those of CITO mentioned earlier (see Section 1.2.4c), are based solely on the repertoire of the traditional arithmetic textbooks. The PPON tests, which were administered for the first time shortly before the start of the MORE research, are an exception to this. As was discussed in the previous chapter (see Section 1.4.2), these tests contain a much broader repertoire and cover the goals of RME more comprehensively. However, as these tests were not publicly available, they could not be used for the MORE research. Moreover, the PPON tests were intended for grades 3 and 6 and therefore would not have provided a solution for first-grade and second-grade data collection. Nor did either the ‘Pluspunt gauges’ (see Section 1.3.2) or the test ‘Arithmetic Puzzles’ from the ‘Arithmetic in a Second Language’ project (see Section 1.3.3) contain problems for first grade.

So nothing remained but the tests and assessment suggestions included in the two textbooks in question. These, however, were also unsuitable. As the purpose was not to ascertain for each textbook individually whether the specific textbook goals had been met but, rather, to determine the differences with respect to the commonly accepted (communal) mathematics goals, the tests for both groups of students needed to be equally new and accessible. Constructing a test on the basis of material gathered from the textbooks would not, in this case, have been a satisfactory solution.

Another, final, problem was that most of the tests included in the textbooks focused on the answers and not on the applied strategies.

In other words, the inevitable conclusion was that the existing tests were inadequate for the research at hand and that the MORE project would therefore have to develop its own tests.

2.2.2 An overview of the instruments developed

Three types of tests were developed for the MORE research in order to measure the mathematical skills of students in first through third grades[12]:

- class written tests on general mathematics;
- class written tests on number facts;
- individual oral tests on general mathematics.

The written tests were developed to measure the students’ achievements. The general mathematics tests concentrated on a broad repertoire of mathematical skills, while the number-facts tests ascertained to what extent basic number facts had become automatized.

The main purpose of the oral tests – also called ‘student interviews’ – was to ascertain which strategies the students applied. The problems included on these tests were basically the same as the problems on the written general mathematics tests.

The entire battery of tests (see Figure 2.2) consisted of ten written general mathematics tests (TG) (two tests, TG2.1 and TG3.1, were repeat tests), two number-facts tests (TN) and eight oral tests (I). The tests were distributed throughout the three school years in such a way that testing took place four times per year. The test numbers shown in parentheses in Figure 2.2 correspond with the test numbers under each graph in Figure 2.1.

	grade 1		grade 2		grade 3	
	Written tests	Interviews	Written tests	Interviews	Written tests	Interviews
Sep	(1) **TG**1.1	**I** 1.1	(5) **TG**2.1 = **TG**1.4		(9) **TG**3.1 = **TG**2.4	
Nov	(2) **TG**1.2	**I** 1.2	(6) **TG**2.2	**I** 2.2	(10) **TG**3.2	**I** 3.2
Feb	(3) **TG**1.3	**I** 1.3	(7) **TN**2.3		(11) **TN**3.3	
Apr/May	(4) **TG**1.4	**I** 1.4	(8) **TG**2.4	**I** 2.4	(12) **TG**3.4	**I** 3.4

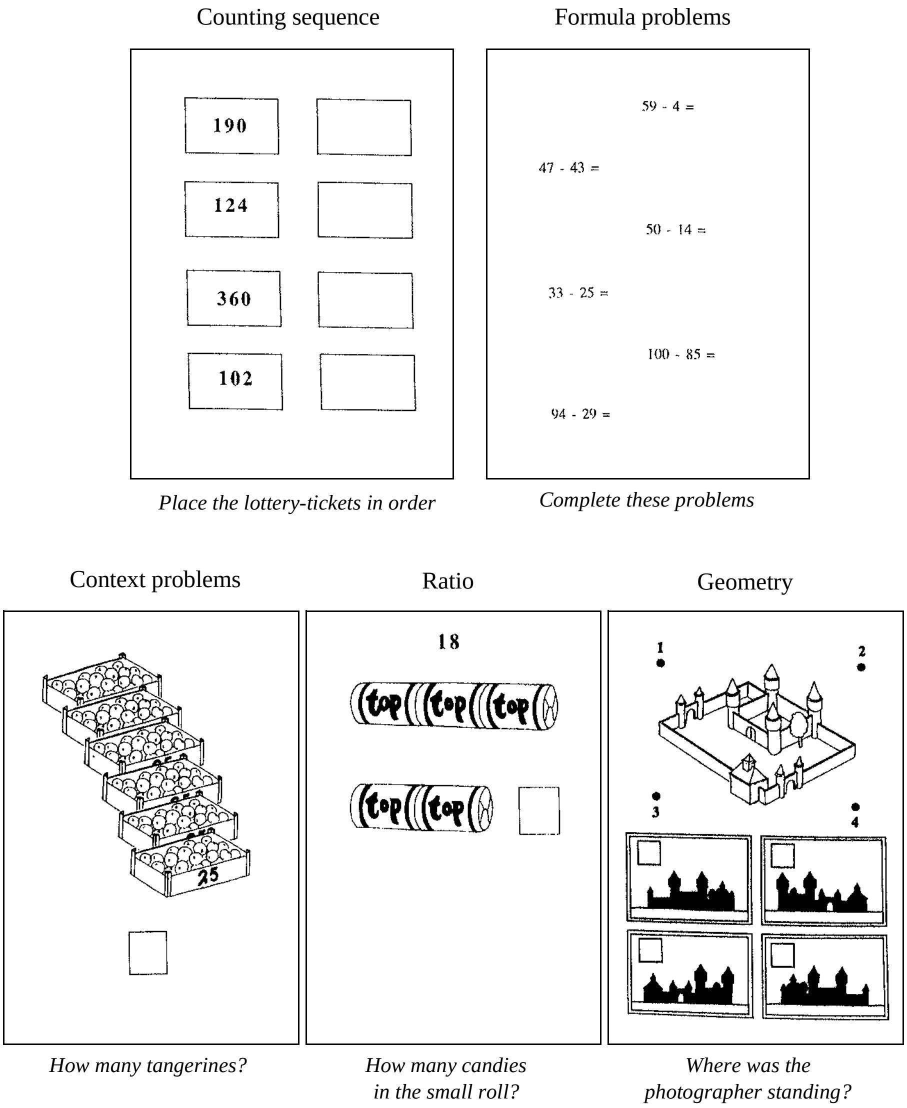
Figure 2.2: Overview of the MORE tests

2.2.2a The class written tests on general mathematics

The class written tests on general mathematics were constructed across a broad spectrum and contained both context problems and bare formula problems. Aside from involving traditional arithmetic operations, the problems also dealt with elementary relationship concepts, number symbols and knowledge of the counting sequence. In addition, these tests also included ratio problems, and assignments where the students had to make up their own problems (‘own-production problems’). The broad composition of these tests was expressed primarily by the inclusion of problems involving geometry as well as arithmetic.

The problems followed the general line of the curriculum for grades 1 through 3, so each test did not contain every type of problem. Problems assessing knowledge of relationship concepts and knowledge of symbols, for instance, only appeared on the initial first-grade test, while column arithmetic problems were only included on the last third-grade test.

The problems were grouped in a variety of ways for analyzing the results. For the detailed analysis per grade, the problems were arranged, for instance, according to the size of the numbers (addition and subtraction problems with and without bridging tens). For the comparison of the developmental progress over the entire three-year period, the problems were categorized in five sub-skills, i.e.: counting sequence, formula problems, context problems, ratio and geometry. The results of this latter comparison have already been discussed in Section 2.1.3. An example of how each sub-skill was measured is shown in Figure 2.3.[13]



Counting sequence

Place the lottery-tickets in order

Formula problems

Complete these problems

Context problems

How many tangerines?

Ratio

How many candies in the small roll?

Geometry

Where was the photographer standing?

Figure 2.3: Some examples of problems from the written tests on general mathematics

These examples were taken from the first two third-grade tests. The actual size of the test pages was 12 × 17 cm, so they have been considerably reduced for inclusion here. The complete test-instructions were included in the teacher's guide that accompanies the tests. While the students were looking at the page in question in their workbook, these instructions were read aloud to them. No time limit was set for each problem. Only when each student had finished did the teachers go on to the following problem. This does not mean that they had to wait endlessly for students who were unable to solve the problem. Instructions in the event of this occurrence were included in the teacher's guide as well.

2.2.2b The class written tests on number facts

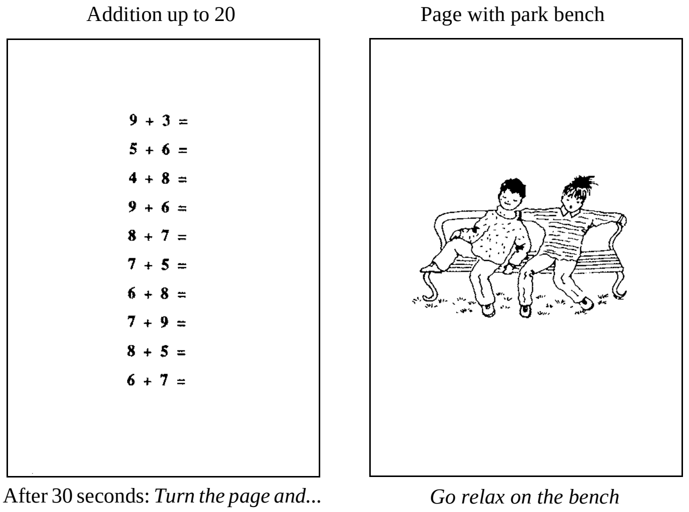
The number-facts tests, consisting entirely of bare arithmetic problems, were used to collect data on the students' skill level with respect to the basic arithmetic facts. A time limit was in effect during this test, as the only way that skill level can be determined on a group test is by limiting the solution time, so that the students are virtually unable to count out or work out the problem at length. This manner of test administration did imply a broad definition of 'automatized' arithmetic facts. An arithmetic fact was considered to be automatized either when the result was known immediately, or when it could be calculated by means of a very few intermediate steps. Only in the first case is there actually a question of memorized knowledge. In practice, however, this is nearly indistinguishable from quick calculation, which is why a broader definition was chosen.

After taking the test *with* a time limit, the students were permitted to take it again *without* any time constraint. They could then use a different colored pen to correct any errors or to complete problems they had left unfinished. In this way, the students' ability to solve the problems correctly in the absence of any time constraint could be seen as well. In the MORE research, this second round only took place on the third-grade test.

The number-facts tests for second and third grade also differed slightly from one another with respect to content. The second-grade test (TN 2.3) consisted of addition and subtraction problems up to 10 and 20, while the third-grade test (TN 3.3) also included multiplication and division tables. Both tests also included a number of introductory problems alongside the problems that formed the heart of the test. These were intended to familiarize the students with the test. The problems were all grouped in rows of ten. The students were given 30 seconds to do each row, but less for the introductory problems. The purpose of this was to prepare the students for the quick tempo.

These test booklets were so designed that all the left-hand pages were left blank. The right-hand pages alternated between a page of problems and a page containing an illustration of a park bench (see Figure 2.4).

This bench[14] played an important role in the test design. Since the students were only given a very limited amount of time to complete each row, it was extremely important that all children would stop and start at exactly the same moment. When the 30 seconds were up, the students were told to turn the page and take a seat on the bench. While they were resting, the test administrator could introduce the next row of problems.



Addition up to 20

9 + 3 =
5 + 6 =
4 + 8 =
9 + 6 =
8 + 7 =
7 + 5 =
6 + 8 =
7 + 9 =
8 + 5 =
6 + 7 =

After 30 seconds: *Turn the page and...*

Page with park bench

Go relax on the bench

Figure 2.4: Example of two consecutive test pages from a number-facts test

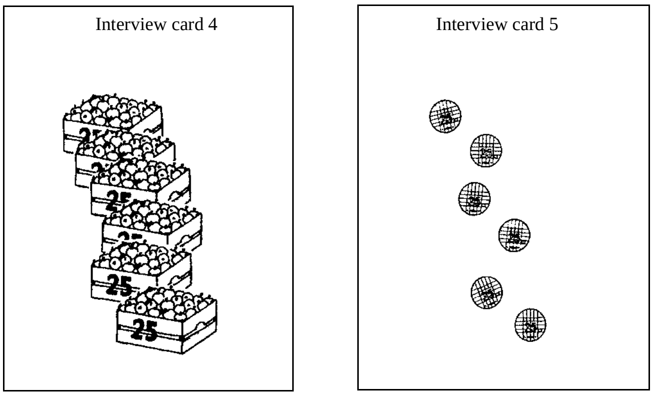
2.2.2c The individual oral tests on general mathematics

Data on the students' solution strategies could be collected by means of the individually administered oral tests. In this sense, the oral tests were clearly intended to supplement the written tests, as their content was obviously similar. The questions and problems included pertained to the following subject matter topics: counting and knowledge of the counting sequence, formula problems (addition, subtraction, multiplication and division), context problems (addition, subtraction, multiplication and division), knowledge of real-life measurements, and own-production problems.

Considering the potential of the interview method for gathering information on matters that are less suitable for written assessment, an expansion of the content did take place in a number of respects. This was particularly true of more process-oriented arithmetic skills, such as the various ways of counting. The interview method was also applied where individual opinions on problems were at issue, as with the students' own productions. In addition, these interviews were also used to collect information on the numerical knowledge and insights acquired by the students with respect to all sorts of everyday matters.

Most of the problems on the oral tests, however, were a kind of shadow of the written ones. An example of this is shown in Figure 2.5. This example was part of the first third-grade interview (I3.2, see Figure 2.2) and basically resembled the context problem on tangerines included on the corresponding written test (TG3.2, see Figures 2.2 and 2.3).

Using the interview techniques as described in the interview guide, the interviewer attempted to determine how the student had calculated the problem at hand.



How many apples are there in all?

How many cents in all?

Interview instructions:

Show the students Interview card 4 and ask the CORE QUESTION:
"It's the school principal's birthday and she's going to treat all the students to an apple. 6 boxes of apples have been brought to school. There are 25 apples in each box. How many apples are there in all?"

*If there is a long silence and one cannot tell what the student is doing, then say: **"What are you doing? Do it out loud. Then I can hear how you're doing it."**
*If it is not clear how the student arrived at the answer, ask: **"How do you know? How did you figure it out?"** or **"Do it once more, but out loud."**
*If the student immediately gives the answer, ask: **"How did you know that so fast?"**

*If the student is unable to solve the problem, show Interview card 5
and ask the accompanying EXTRA QUESTION: **"Can you tell me how many cents this is in all?"**

Solution strategies:

1-1c	The objects/units are counted one by one.
multiplication table	The student recognizes a multiplication problem and calculates it by reciting the multiplication table in question.
via 100	The student calculates via 4 × 25.
via 50	The student calculates via 2 × 25.
via splitting t/o	The student calculates 25 × 6 or 6 × 25 by splitting the tens and ones.
m	The student has the result memorized.
os	Other strategies.
unk	Strategy unknown.

Figure 2.5: Example of an oral test

The interview was always conducted in approximately the same manner. Noteworthy here is that a certain prioritization was attached to the applied interview techniques (see Figure 2.6).

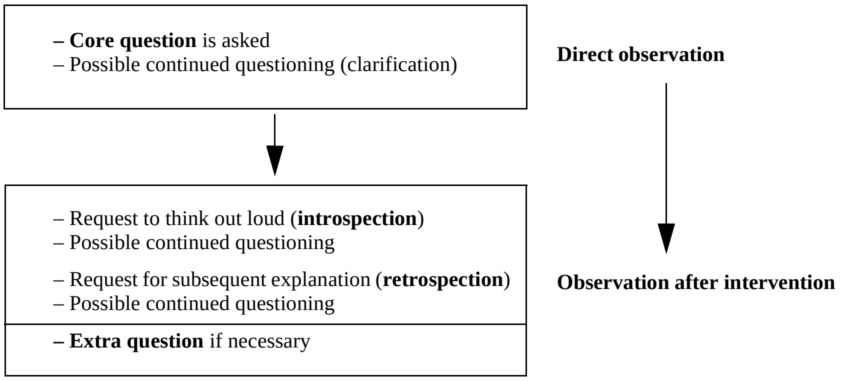

– **Core question** is asked
– Possible continued questioning (clarification)
Direct observation
– Request to think out loud (**introspection**)
– Possible continued questioning
– Request for subsequent explanation (**retrospection**)
– Possible continued questioning
Observation after intervention
– **Extra question** if necessary

Figure 2.6: Interview techniques

The core question was posed first. If the student did not precisely understand the question, another question was then posed which, for example, formulated the first question in a different way. After posing the core question, the interviewer was to exercise restraint and await the student's reaction. Only when the desired information could not be observed might the interviewer intervene. Two successive interview techniques were involved here: introspection and retrospection. The advantage of this prioritization was that it enabled as accurate an image as possible to be formed of the way in which students approached solutions to different problems. Accompanying protocol forms, on which a number of possible answers and strategies had already been noted, were used to report the findings.

2.2.3 Starting points and method for developing the tests

This section examines in more detail the method followed for developing the MORE tests and the starting points which thereby served as a guideline. Because of the key role played by the written general mathematics tests – both in the MORE research and in further concept formation and elaboration of RME assessment – this section concentrates on the development of the written general mathematics tests. This is, in the first place, primarily a technical description of the procedures followed. Subsequent sections will then examine other aspects of the development process.

2.2.3a A pragmatic approach

Although the written general mathematics tests, as a product of their time, were grafted to then prevailing RME viewpoints on assessment and to examples of assessment instruments inspired by the RME approach to mathematics education (see

Chapter 1), to a great extent this occurred implicitly. This was because the ideas on RME assessment had not yet been articulated explicitly – not, at any rate, with respect to assessment in primary education. The sole certainty was that assessment problems (keeping in mind the goals of RME) needed to be relevant and that priority should be given to observation and to conducting discussions with individual students. With respect to these discussions, however, the MORE research had to set limitations. The large number of students participating in the research made written testing simply inevitable. And, because of the lack of realistic examples of first-grade written tests, development of these tests had to be started more or less from scratch.[15] Moreover, the approach to assessment development had to be broader (or narrower) than the RME perspective alone. Here, too, the research made certain demands. In actuality, the research questions and the research design were what determined the method used. The following starting queries were used as a guideline for developing the tests:

1 What should (must) the students be able to do at a given moment, considering:
   - the general prevailing opinions?
   - the specific content of the textbooks NZR and WIG?
2 How can these skills be assessed through written tests so that:
   - accurate and thorough insight is obtained into what the students are able to do?
   - differences between students or between groups of students can become apparent?
   - the method of assessment is fair to both groups?
   - the tests can be easily administered by the teachers themselves?

These queries also reflect the principles for developing alternative assessment problems for RME as formulated by De Lange (1987a). In this respect, assessment development within the MORE research project elaborated upon the work of De Lange, as discussed earlier (see Sections 1.2.5c and 1.4.1).

2.2.3b A general outline of the procedures followed

The development of the MORE tests may be characterized as an iterative process of generation, selection and adjustment of assessment problems, in which the two queries stated above constantly served as a guideline. Roughly speaking, a number of steps could be distinguished in this process, which, rather than proceeding according to a fixed plan, at times overlapped one another.

- step one: demarcation of the subject matter to be assessed

The first step involved a rough demarcation of the subject matter that would be included on the tests. It was deemed important that this subject matter be related to commonly accepted (communal) mathematics goals. Reference points here were the ‘Design of a National Curriculum ...’ (Treffers, De Moor, and Feijs, 1989; Treffers and De Moor, 1990) and its precursors, which had been published in various journals. Data produced by the textbook analysis conducted for the MORE research was also taken into account in choosing the subject matter. This initial demarcation of the subject matter primarily involved organizing the sub-topics. The decision was made, for instance, to have the initial first-grade test contain problems on number symbols, the counting sequence, resultative counting, and performing addition and subtraction operations in a context format using small amounts and numbers.

- step two: inventing problems

The next step was to invent a vast number of problems suitable for inclusion on the tests. Here, the demarcation of sub-topics was often temporarily set aside. No emphasis was laid, for instance, on a strict operationalization of the previously distinguished sub-topics. Instead, the focus was more on finding interesting, attractive problems that would both appeal to the students and reveal their capabilities.

Although the subject matter contained in the two textbooks was certainly taken into account when inventing the problems, the problems included on the tests were not textbook-specific. The idea was to invent assessment problems that would be fair to both groups of students. There was to be no difference in procedural prior knowledge between the two groups of students with respect to the problems. For this reason, neither the WIG ‘bus problems’ nor the NZR ‘split-problems with broken lines’ were included on the first-grade tests. Instead, problems were invented that would be *accessible* to both groups. For the bare problems, too, all textbook-specific forms of notation were avoided as far as possible. As a rule, only types of problems that appeared in both textbooks were included. Not that the tests were devoid of any problems more familiar to one group of students than the other. However, the students were not to be confronted by an unknown type of problem, i.e., a problem where certain procedural knowledge is required in order to solve it. Consequently, no long-division was included in the third-grade tests, for example, because WIG did not introduce this procedure until fourth grade. The first-grade tests, on the other hand, did contain problems involving bridging ten (NZR) and problems above 20 (WIG) because both of these can also be solved without any specific procedural knowledge.

- step three: thinning out

After as many problems as possible had been invented, they were subjected to a strict selection procedure. At this point, the focus of attention was mainly on whether the test in question and the entire test series had been constructed evenly: had the various subject-matter sub-topics all been covered sufficiently, was there enough variety in the types of problems, and did the tests provide adequate insight into the developmental progress of the skills?

- step four: refining problem presentation and formulation

After the composition of a test had been roughly determined, the problems themselves were once again examined: what exactly would be printed on the test paper, and what instructions would be included? Here, too, accessibility was of primary importance. An assessment of reading comprehension rather than of mathematical skills had to be avoided at all costs. This is a clear and present danger in context problems, which describe an entire situation. The decision was made to use problems having few words, but, instead, containing illustrations needing little or no explanation.[16] The test administrator would then read aloud the instructions accompanying the illustration. In order to ensure that the students' memory would not be overburdened, numerical information pertaining to the problem was always printed on the test page.

- step five: trial version

  The next step was to put the trial version into practice. This took place at two schools that were not participating in the research. During this trial run, the focus was mainly on whether the instructions were clear, the students understood the intentions right away, the length of the test was satisfactory, and whether the problems were sufficiently inviting and the appropriate degree of difficulty.

- step six: final version[17]

  According to the degree of success of the trial version, alterations were made where necessary (see Chapter 5 for examples). Experience gained from the trial version was also used in designing the final version of the test instructions and the teacher's guide. An unavoidable problem inherent in the pilot testing was that the trial version had to be administered six to eight weeks before the actual administration of the final version.[18] This time lag, together with the limited scope of the trial run, made it difficult to estimate the degree of difficulty of the tests.

- step seven: scoring scale

  After administering the tests, the students' answers (based on a sample of two classes) were used to create a scoring guideline for each problem.

- step eight: subsequent empirical verification

  In order to acquire more clarity with regard to the psychometric quality of the tests used in the research project, the data from the final version was used to calculate a number of psychometric values for each test.[19]

  Table 2.1 indicates the internal homogeneity (Cronbach's alpha) of the tests in their entirety and of the sub-skills distinguished within each test. The assignment of problems to these sub-skills was conducted after the fact and based on intrinsic characteristics. Each problem was assigned to one sub-skill only, with the exception of ratio problems. These were grouped with the context problems but, because of their specific mathematical nature, were also classified under a separate sub-skill. The internal homogeneity of the tests was extremely high. The alpha-values lay between 0.85 and 0.91. There was a fairly high homogeneity for most of the sub-skills as well, with the exception of the sub-scores for ratio and geometry. Some of the tests, however, contained only a few problems in these areas.

Table 2.1: Psychometric data from the written tests on general mathematics (I)

MORE written tests on general mathematics										
	number of problems per test and per subskill									
sub-skills	TG1.1	TG1.2	TG1.3	TG1.4	TG2.1	TG2.2	TG2.4	TG3.1	TG3.2	TG3.4
counting sequence	8	6		7	7	5	3	3	2	
formula problems		6	8	20	20	18	20	20	23	19
context problems	12	15	16	10	10	14	16	16	14	14
ratio			2	5	5	2	6	6	2	3
geometry		1		6	6		2	2	2	4
test total	28	28	27	43	43	38	42	42	44	37
	alphas per test and per subskill									
sub-skills	TG1.1	TG1.2	TG1.3	TG1.4	TG2.1	TG2.2	TG2.4	TG3.1	TG3.2	TG3.4
counting sequence	.67	.66		.74	.73	.64	.58	.46	.21	
formula problems		.79	.86	.87	.87	.85	.88	.88	.87	.83
context problems	.89	.83	.79	.63	.63	.76	.80	.79	.73	.75
ratio			.05	.45	.47	.31	.62	.58	.06	.39
geometry		--		.44	.53		.20	.10	-.20	.35
test total	.88	.89	.85	.89	.90	.90	.91	.90	.89	.88

Aside from the internal homogeneity, each test was examined with regard to the nature of the score distribution, the average score, the variability of scores and the minimum and maximum scores (see Table 2.2). Most of the tests revealed a fairly normal distribution of the total scores. With the exception of the initial first-grade test, none of the tests displayed a serious ceiling or basement effect.

Table 2.2: Psychometric data from the written tests on general mathematics (II)

MORE written tests on general mathematics					
tests	**average % correct**	**stand dev in %**	**min % correct**	**max % correct**	**n total**
TG1.1	75	19	21	100 (n=31)	441
TG1.2	53	23	0 (n=3)	96	443
TG1.3	72	18	0 (n=1)	100 (n=3)	440
TG1.4	58	19	2	89	439
TG2.1	65	19	5	100 (n=1)	432
TG2.2	57	20	0 (n=1)	95	427
TG2.4	60	20	5	98	432
TG3.1	67	19	2	100 (n=1)	425
TG3.2	55	18	5	91	416
TG3.4	52	19	3	95	419

It is striking to what extent the tests vary with respect to the degree of difficulty. The fluctuating pattern of high and low average total scores clearly reflects the parallel processes of test development and data collection. The scores from one test obviously determined to a considerable extent the appearance of the following one. The adjustments made as a result of the trial run evidently produced little change. This was particularly true of the tests developed for first grade. Gradually, the fluctuating pattern did even out, at least if one takes into account the fact that TG 2.1 and TG 3.1 were repeat tests. It would seem that the difficulty level began to be predicted more accurately as experience was gained in developing the tests.

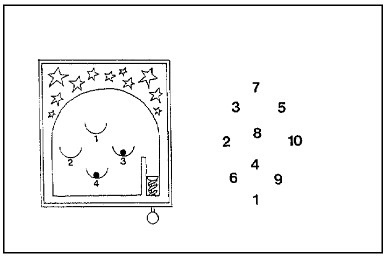
Little by little, other aspects of assessment were also discovered and clarified during these cycles of design, trial run, adjustment and definitive administration. This will be examined in more detail in the following section.

2.3 Crucial events during the development process

This section describes certain crucial events that occurred during the development of the written general mathematics tests. These occurrences involved a number of discoveries and experiences that more or less guided the development process. They also especially contributed to the development of tests becoming more than only a concrete development task, i.e., the development of instruments on behalf of the MORE research.

2.3.1 Unexpected initial findings

The first round of testing took place three weeks after the start of first-grade classes. The test used here and its results are discussed in detail in Chapter 5. Of particular significance in the context of this section are the surprising results. At the start of mathematics education, the children turned out to be in possession of more mathematical knowledge and skills than the content of the textbooks or the judgment of experts had led anyone to expect.[20]



Instructions to be read aloud:

“We’re going to play a game of pinball. Shoot two pinballs and see how many points you’ve gotten.
Cross off the correct number next to the pinball machine.”

Actual % correct (n = 441): 43%
Estimated % correct: ≈10%

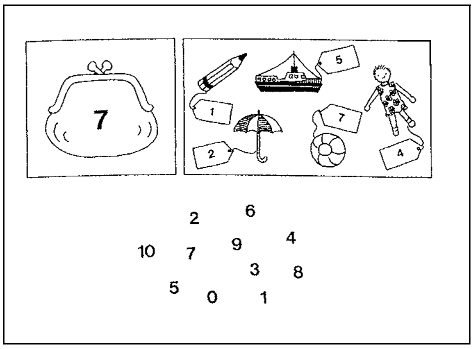
Figure 2.7: One of the Pinball problems from the initial first-grade test

Take, for instance, the Pinball problem involving non-countable amounts (the values are given by means of symbols) illustrated in Figure 2.7.[21] It had been estimated that approximately 10% of the students would be able to solve this problem correctly. As it turned out, however, more than 40% of the students were able to do so.

The salient feature of the high scores is that they occurred on a written test that was administered to the class as a whole.[22] This experience stood in stark contrast to the initial reluctance of the test developers to design written tests. The assumption had namely been that this kind of test would not be suitable for eliciting existing knowledge and skills – particularly in the case of younger children. Nevertheless, these written tests had indeed revealed the abilities of beginning first-graders. It was simply a case of inventing assessment problems that involved all kinds of real-life situations, which the students either had experienced themselves or could at least imagine. Aside from this, certain measures had been taken to deal with the fact that most beginning first-graders cannot yet read. Other than that, however, this test had no especially distinguishing characteristics.[23] Aside from the link to application contexts, the mathematical content of this test corresponded to what is generally presented on first-grade mathematics tests. Nor did the answer format deviate particularly from what is customary on written mathematics tests, as the research conditions had necessitated a fairly closed answer format. Although this first test was not so unusual in itself, the revelatory nature of the results provided a stimulus for further investigation into the potential of written tests, thereby initiating a more unusual manner of written assessment.

2.3.2 Open-ended problems are also possible

Aside from the unexpectedly high percentage of correct answers, this first test also produced another significant experience, involving the potential of open-ended problems. A few of these problems had been included in the test, more or less as a counterbalance for the apparently inevitable closed problems. It was hoped that these open-ended problems would help clarify what the students were able to do on their own. The prevailing opinion, however, was that these problems would be considerably more difficult – both for the students and with respect to processing the results. But the wish to include open-ended problems surmounted these objections. The Pinball problem shown in Figure 2.7, for instance, was followed by another Pinball problem in which the students had to shoot two pinballs themselves and then determine the total points. In the same way, a closed Shopping problem was followed by another in which the students first had to choose what they would buy (see Figure 2.8).



Instructions to be read aloud:

"You have seven guilders in your purse. Choose one thing that you'd like to buy. Cross off that problem."
(Wait until all the children are ready.)
"How many guilders do you have left in your purse?
Cross off the correct number underneath the picture."

Figure 2.8: Example of an open-ended test problem from the initial first-grade test

The assumption that open-ended problems would prove more difficult for the students proved false. Table 2.3 shows that the differences were, in fact, only slight.

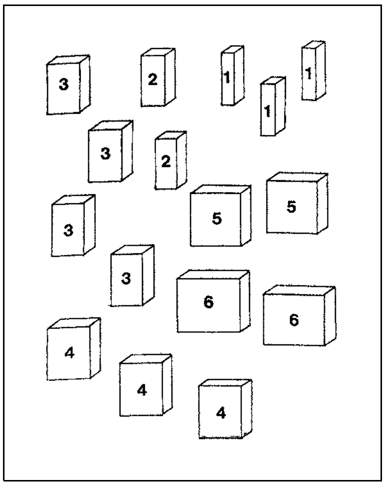
Table 2.3: Correct answers in percentages to open-ended and closed problems

Pinball problems	% correct (n = 441)	Shopping problems	% correct (n = 441)
closed		closed	
1 + 3	53%	5 – 2	60%
3 + 4	43%	10 – 8	44%
open-ended		open-ended	
. + .	49%	7 – .	42%
. + .	52%	9 – .	39%

Moreover, open-ended problems were evidently able to provide more information than just a simple answer to whether a student could or could not solve a problem. Thanks to the choice of numbers, each student could solve the problem on his or her own level. And, at the same time, the student's level could become apparent to an observer. In the Shopping problem (see Figure 2.8), some students bought the pencil (7 – 1), while others decided to buy the doll (7 – 4). Important consequences of this open-ended design are that the problem then becomes accessible to more students and that the students can show what they *can* do – rather than what they *cannot* do.

Lastly, the processing of the answers proved to be less of a problem than had been predicted. Depending upon the goal of the test, a rough or a fine criterion can be applied. With a rough criterion, one merely ascertains whether the students are able to solve an open-ended problem within a given range of numbers, but with a finer criterion, the choice of numbers is considered as well. When comparing two groups of students, as was the case in the MORE research, one can suffice with the former. The finer criterion is desirable when footholds for further instruction are being sought on an individual or a class level.

After these first experiences with open-ended problems, the way was paved for other such problems that could provide more information than simply whether the answer was right or wrong. This was partly put into practice by conferring on the student a certain responsibility in designing the problems and partly by allowing the student more freedom in the type of response. An example of the former is an own-production problem, where the students themselves had to invent an easy and a difficult problem. The latter is illustrated by the Candle problem, shown in Figure 2.9. Both problems are part of the third first-grade test.



Instructions to be read aloud:

"The candles in this store have been packaged in all different ways. Some boxes contain only one candle. Other boxes contain two or three candles. And there are also boxes containing four, five or six candles. You want to buy twelve candles. You can choose which boxes to take.
Cross off the boxes that you want to take, but be sure that you'll have exactly twelve candles."

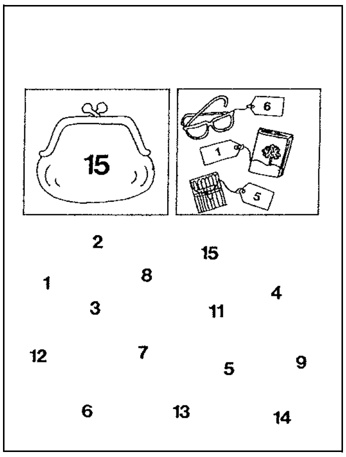
Figure 2.9: Example of a problem where various answers are possible

2.3.3 An unintentional error in an assessment problem

Another crucial event originated in an error that snuck onto a test page. One of the possible answers to a Shopping problem on the second first-grade test was left out (see Figure 2.10a), making it impossible for the student to buy the package of markers. Figure 2.10b shows how two students handled this error. Student (a) canceled the first choice in order to buy something that corresponded to an answer on the test page. Student (b) simply added the missing answer to the ones already present.

Aside from the differences in approach behavior reflected by the different solutions – information that can be crucial to the teacher – the contrasting reactions to this error also awoke the developers to the fact that written tests need not be static; they, too, can have a dynamic character.

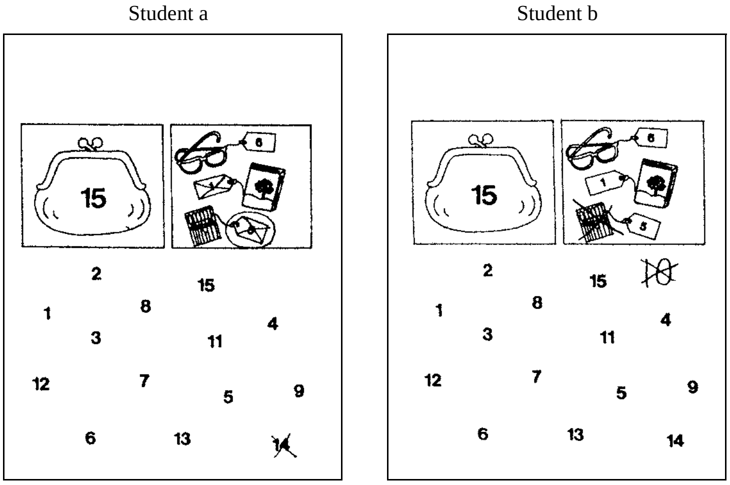
Nevertheless, it was clear that such test-page errors hovered on the brink of disaster. Intentional inclusion of errors can easily result in traps for the students, which should never be done deliberately (see also Section 1.2.4e).



Instructions to be read aloud:

"You have 15 guilders in your purse.
Choose one thing that you'd like to buy.
Cross off that problem."
(Wait until all the children are ready.)
"How many guilders do you have left in your purse?
Cross off the correct number underneath the picture."

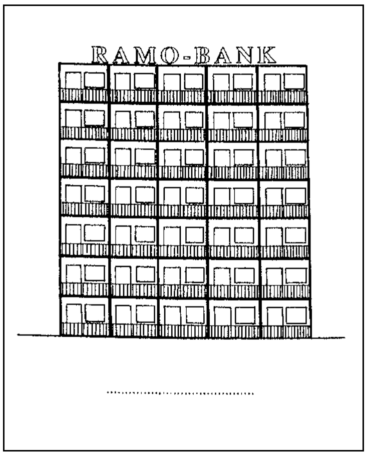
Figure 2.10a: Open-ended Shopping problem


Student a

Student b

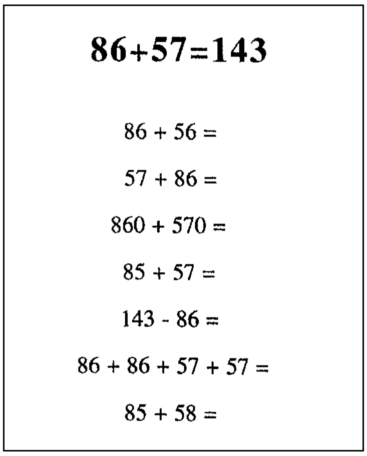
figure 1Figure 2.10b: Two solutions to the test-page error

Although, indeed, no errors were intentionally included during the further development work, the experience did awaken interest in the potential of providing more dynamic forms of written assessment. In addition, this event contributed to experimentation with related matters, such as supplying extra information and having the students fill in missing information. An example of each can be seen in Figure 2.11.



Instructions to be read aloud:

"How large do you think the letters on top of this building are?
Write your answer on the dotted line."



86+57=143

86 + 56 =

57 + 86 =

860 + 570 =

85 + 57 =

143 - 86 =

86 + 86 + 57 + 57 =

85 + 58 =

Instructions to be read aloud:

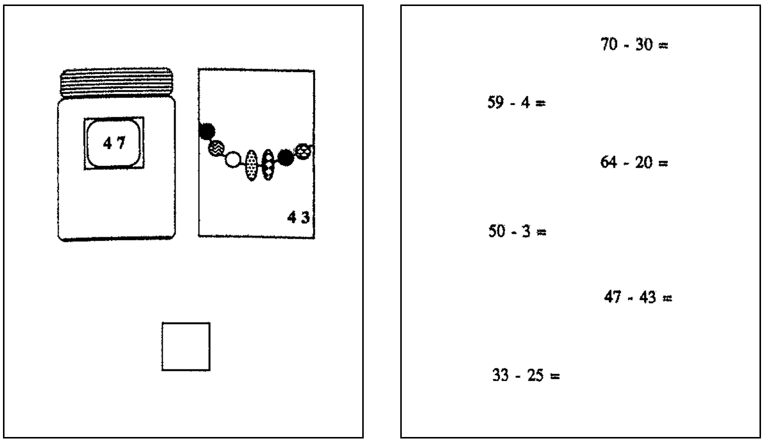
"At the top of this page is problem that has already been completed. Underneath, there are more problems that are kind of similar. Try to find the answers to these problems by looking carefully at the one that is already finished. You can do them in any order. Go ahead."

Figure 2.11: Examples of test problems: one with missing and one with extra information

2.3.4 Context problems versus bare operations

The variation in how the problems were presented (with or without context) was intended, firstly, as a way of providing equal treatment to the two textbooks[24], in order to trace possible differences in learning results. However, the second second-grade test unexpectedly exposed an interesting phenomenon that came to light thanks to this variation, namely, that considerable differences may exist between context problems and their corresponding bare operation. An example of this is the Bead problem and the corresponding formula problem 47 – 43 (see Figure 2.12). The accompanying scores can be found in Table 2.4.[25]

The students evidently found it easier to determine how many beads would be left over from their necklace than to perform the bare calculation 47 – 43. The facilitating role of the context apparently led them to calculate differently here than in the case of the formula problem. At the time this test was administered, neither textbook had as yet dealt extensively with such problems involving compound numbers (consisting of ones and tens). In other words, something was being assessed that had not yet been taught, i.e., 'advance testing': the testing occurred in advance of the new subject matter rather than after the fact.[26] This Bead problem opened the door – or re-opened it, if one considers the Winter Supplies test-lesson from 1978 (see Sections 1.2.5b and 1.2.5c) – to using context problems to test informal knowledge as a foothold for further education.



Instructions to be read aloud:

"Here's a jar with 47 beads. Do you see it? You're planning to make a pretty necklace. To do this you need 43 beads. How many beads will you have left? Write your answer in the empty box."

Instructions to be read aloud:

"There are a number of subtraction problems on this page. Some are easy and some are a little harder. See if you can find the answers. You can choose which one to do first."

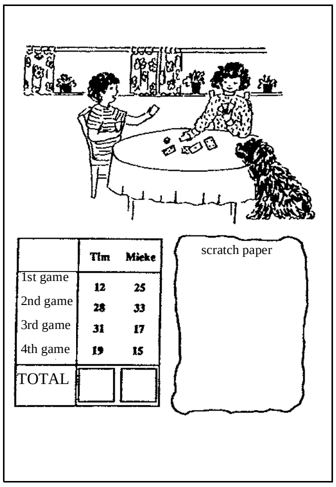
Figure 2.12: Example of variation in presentation: context problem versus formula problem

Table 2.4: Correct answers in percentages to the two presentations

type of problem	NZR	WIG	total (n = 427)
context problem 47 – 43	61%	59%	60%
formula problem 47 – 43	38%	39%	38%

2.3.5 Scratch paper, and other ways of tracing strategies

In order to use informal knowledge as a foothold for further education, it is essential that one learn more about the strategies children apply. To assist this endeavor, scratch paper was included in the fourth second-grade test, which was a test on general math[27] (see Figure 2.13). Here, too, a long tradition was being set forth (see Chapter 1). However, although Ter Heege and Treffers had already spoken of 'solidified student behavior' that 'becomes visible on paper' in 1979 (see Section 1.2.5b), scratch paper and similar materials for collecting data on student strategies during testing had never really been applied systematically.[28]



	Tim	Mieke
1st game	12	25
2nd game	28	33
3rd game	31	17
4th game	19	15
TOTAL	□	□

Instructions to be read aloud:

“Tim and Mieke just finished a game. Do you know what the final score was? How many points did Tim get in all, and how many points did Mieke get?
Write your answer in the empty boxes.

If you like, you can use the scratch paper to figure out how many points they scored.”

Figure 2.13: Example of an assessment problem containing scratch paper

The ‘Arithmetic Puzzles’ test from the ‘Arithmetic in a Second Language’ project (Van Galen et al., 1985) was an exception to this. On this test, a strip of scratch paper was included along the right-hand margin of each test page (see Figures 1.4 and 1.5).

In spite of the fact that strategy data was being collected for the MORE research through student interviews, this scratch paper still appeared on the written tests. This occurred more or less spontaneously, thanks to a context in which scratch paper is customary, such as a word game in which one must keep track of the score. The numbers were so chosen that they would add up easily.

Scratch paper also fit in with the initiative – stimulated by earlier experiences – to investigate the potential of written tests. Not that the collection of strategy data through interviews was therefore discontinued. Such an intervention in the design of the research was not directly feasible after one single trial, nor did it make much sense later on in the research.[29]

Although, from the students’ point of view, the scratch paper was introduced in a natural fashion and in an appropriate play context, a conscious decision was made to keep the use of scratch paper voluntary. In order not to frustrate the students with something unfamiliar and to avoid making the problems more difficult for them, it was clearly stated that they could use the scratch paper if they wished, but were not obliged to do so. Because experience had shown that some students find it difficult to mark up a clean book, the scratch paper was purposefully drawn with ragged edges, as if it did not belong on the test page.

The first run with the scratch paper produced such a good crop of strategy information (see Figure 2.14) that the decision was made to proceed with it on a voluntary basis. With a few exceptions, no explicit explanation was requested with any of the scratch paper. It was there purely for the students’ own use.

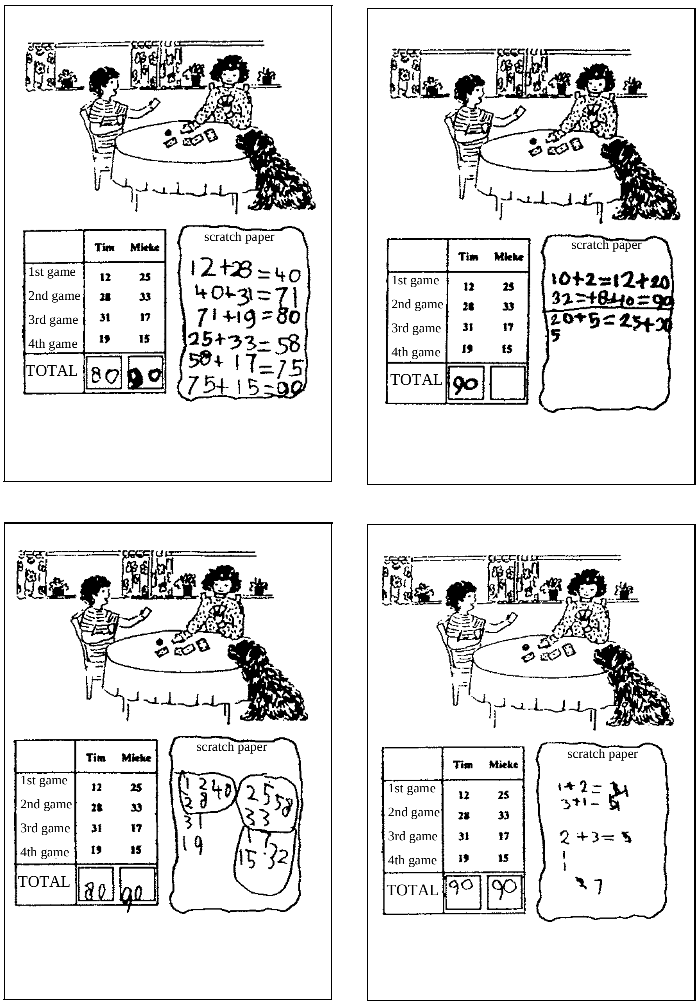
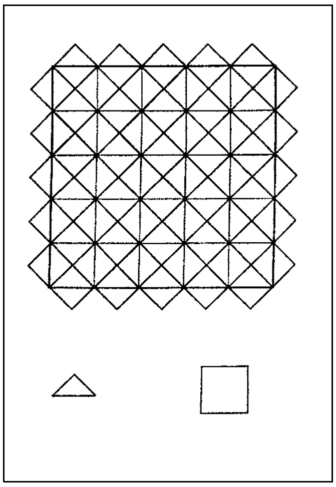


Figure 2.14: Examples of strategy traces found on scratch paper

Additional strategy information was also gathered by other means besides the scratch paper. In a certain sense, the open-ended problems mentioned earlier (such as the Candle problem) also sometimes revealed aspects of the applied strategy. This occurred even more directly, by including ‘context drawings’ on the test page that the students could then use as ‘working drawings’ for solving the problem. Here, again, the source of inspiration for this type of problem was the ‘Arithmetic Puzzles’ test with its Candy problem (see Figure 1.5). An example of such a working drawing can be seen in the Tiled Floor problem on the final third-grade test (see Figure 2.15).[30]



Instructions to be read aloud:

“How many of the small triangular tiles were used to make this floor?
Write your answer in the empty box at the bottom of the page.”

Figure 2.15: Example of a test problem containing a ‘working drawing’

2.3.6 Exchange of perspective: assessment problems with ‘elasticity’

Aside from the fact that strategies can be revealed by these working drawings (as shown in the two examples in Figure 2.16), the students’ reactions made it clear that more was going on in this type of problem.

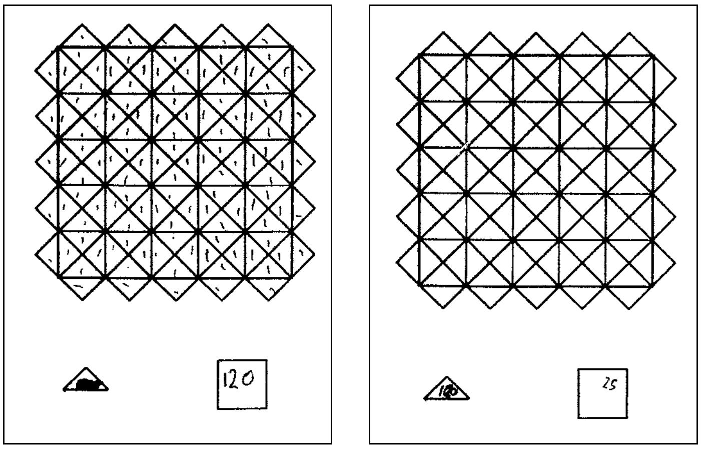


Figure 2.16: Examples of strategy traces in a working drawing

In order to discover what was going on, an exchange of perspective was necessary. A break had to be made with the prevailing psychometric idea that assessment problems should be unequivocal with respect to the solution strategy. Not only are these types of problems extremely informative for teachers and others who wish to know more about the applied strategies, but this manner of presentation also has significant consequences for the students themselves. Thanks to the context – in this case the working drawing – these problems can be solved on various levels. In other words, the problems have a certain elasticity. Another way to express this is that there is a kind of stratification in the problems. The problems are therefore accessible to more students. Both 'little solvers' and 'big solvers' can take part. This is also true, of course, of the open-ended problems, the own-production problems, problems such as the Candle problem (in which different answers are possible), and problems in which students must supply certain missing information.

Thanks to this idea of 'assessment problems with elasticity', a certain coherence arose in the search for new opportunities for written tests. These (re)discovered new opportunities proved to be interconnected, and some problems turned out to contain more than one of the newly discovered opportunities. This was most striking in the Polar Bear problem, which will be discussed further as a prototype of a RME assessment problem in Chapter 3.

These new opportunities for written tests were taken, first and foremost, from the RME approach to mathematics education and inspired by examples of tests that had been developed in connection to this. In addition, however, they arose from the concrete development work itself and from reflection on the work that arose from experiences gained in this work and from the analysis of the collected data. In the following section, particular attention will be paid to the development context as a determinative factor in the development process.

2.4 The influence of the development context on the developed tests – a postulated awakening

To conclude this chapter, the tests and their incorporated concepts of RME assessment will be regarded once more – this time in the light of the development context. While a more technical account of the test development process was set forth in Section 2.2, followed by a description of a number of crucial events in this process in Section 2.3, the present section consists of a reflection on the development process in relation to the development context. Taking the tests produced by the development process as a point of departure, an attempt will be made to answer the question of why these tests turned out as they did, and to what extent this may have been influenced by the development context.

2.4.1 The development context as an influential factor

- the size of the research group

The size of the research group may have been the most determinative factor for the direction taken by the development of tests for the MORE research. Had there not been such a large group of students to assess, developmental research into new forms of assessment appropriate to altered insights into educational theory with respect to primary school mathematics education probably would not have begun with written tests.[31] Considering the preferences inherent in RME, such research, had it been conducted independently of the MORE research, would certainly have concentrated more on observation and individual interviews.

- the limitations of beginning first-graders

  The fact that the development work started at the beginning of first grade was another determining factor in the choices made. These circumstances contributed to the decision to include so much illustrative material and to have the text be read aloud rather than printing it in its entirety on the test page. Most beginning first-graders are namely not yet able to read and write. For this reason, only pre-determined possible answers were used at the beginning, from which the students had to cross off the correct one. The initial first-grade test was, in fact, simply a multiple-choice test. The test development thus began using an extremely closed format.

  Along with the lack of reading and writing skills, beginning first-graders have not yet received any formal arithmetic instruction. As a consequence, a test at the beginning of first grade cannot consist of the customary problems. If the purpose is to find out more about the students' mathematical knowledge and skills, then it will be necessary to turn to situations from everyday life from which the students may have assimilated some information and in which they can apply what they have learned.

- two different types of textbooks

  The fact that the two groups of students were using strongly contrasting textbooks also clearly left its mark on the development of the tests. On the one hand, the assessment had to take place as close to the textbooks as possible in order to expose the differences in the learning achievements. On the other hand, since communal goals were the objective, the assessment problems needed to be equally accessible to both groups of students. In order to achieve the latter, everyday situations were found, in which the students could imagine themselves, rather than problems that referred to 'the book'. As a result, problems could be presented that the students 'hadn't had yet'. Aside from making the assessment problems equally accessible to both groups of students, this type of problem also exposed the informal foundation of the skills to be learned, thereby leading to the 'advance testing'.[32] And, had two similar textbooks been involved in the research, not as much attention would have been devoted to the difference between context problems and bare problems.

- the longitudinal character of the research

  This study of new opportunities for written tests would have been inconceivable, however, without the longitudinal character of the MORE research. Thanks to the repeated pattern of developing a test, administering it, analyzing the answers and then developing the following test, an opportunity was always available to try out ideas, adjust them based on the experiences, and then try them out again. The adjustment process took place quite naturally and the tests were able to grow along with the students' abilities.

Another consequence of this longitudinal character was that 'long-lasting' assessment problems had to be used for the sake of comparison. These problems can often be solved on different levels, and they also contributed to the discovery of the importance of having 'elastic' assessment problems.

- the simultaneous development of oral tests

The fact that each written test administered to the class was developed at the same time as an individually administered oral version probably influenced the test development in two ways: directly, by leading to reciprocal influence, e.g. the awareness of the potential for a more dynamic approach to written assessment and related content; and more indirectly, although not less importantly, by decreasing the test load. Because the strategies had already been measured by means of the individual oral tests, this was not necessary on the written tests, thereby creating some room for experimentation.

- the low-stake level of the assessment

Another factor enabling more room for experimentation was the absence of high-stake assessment. Although the research context did have some limitations (large number of subjects, limited time for instrument development), a certain liberty to try out different ideas prevailed. As the results were used for no other purpose but to answer the research questions, not much was at stake either for the teachers or the students. Moreover, many things were assessed that had not yet been handled in the classroom. This, too, contributed to the fact that the students – and especially the teachers – did not feel themselves judged.

2.4.2 From developing tests for the MORE research to developmental research on assessment with the focus on didactical assessment

As indicated earlier (see Section 2.2.3a), the attitude with regard to the development of the tests was pragmatic: as there were no suitable tests available and the group of research subjects was quite large, written tests would have to be developed. At the same time, the prevailing opinion was that, if one wished to know exactly what the students could do, then observation and individual oral tests would serve better. This explains the certain reluctant attitude which characterized the outset of the development of the written tests.

Surprising results and experiences, new ideas and discoveries, and newly arising concepts with respect to assessment, alongside a constant desire (fostered by the RME educational theory) to approach assessment differently – all these contributed to the development work becoming more than simply the production of a new series of tests. Along with research activities that focused on answering the research questions, a parallel path of developmental research was gradually emerging within the MORE research project. This path concentrated on a search for alternatives to the existing written tests, which, in the opinion of many, were entirely inadequate (see also Section 4.1.1). While working on the tests for the MORE research, new opportunities for class written tests on mathematical skills were examined, based on the principles of the RME instructional theory (see Van den Heuvel-Panhuizen, 1990a and Van den Heuvel-Panhuizen and Gravemeijer, 1991a). The primary concern was to find new opportunities for didactical assessment: assessment on behalf of educational decisions taken by the teacher.

This expansion towards didactical assessment based on RME, together with the strict time schedule, resulted in the impossibility of making full use of every opportunity that appeared during the developmental research to develop instruments for the MORE research. One example of this is the use of scratch paper described earlier (see Section 2.3.5). Not knowing how students and teachers would react to this, and unsure of whether it would produce enough information, the test developers decided to let scratch paper use be optional at first. Later on, although it had by then become clear how revelatory this scratch paper could be, this optional use (with a few exceptions) nevertheless remained. In a nutshell, while the MORE tests were certainly not the best tests that were developed – or perhaps could have been developed – they did set something in motion. The following chapter will discuss in detail what was produced by this departure from the MORE research.

Notes

[1] The acronym ‘MORE’ stands for ‘MethodenOnderzoek REkenen-wiskunde’, which can be translated as ‘Mathematics Textbooks Research’.

[2] NZR stands for ‘Naar zelfstandig rekenen’ (‘Towards independent arithmetic’) (Kuipers et al., 1978).

[3] WIG stands for ‘De wereld in getallen’ (‘The world in numbers’) (Van de Molengraaf et al., 1981).

[4] The class written tests referred to in this chapter are tests in which the test-instruction is given orally. More on these tests can be found in Section 2.2.2a.

[5] Figure 2.2 indicates when the tests were administered.

[6] Considering the small number of problems which determined certain sub-scores (see Section 2.2.3b), some caution is advised regarding these conclusions.

[7] This general character is relative, however. In other researches into the implementation of specific theoretical innovations, characteristics are even used that are in no way connected to the content of the innovation. Harskamp and Suhre (1986), for instance, used the number of tasks completed as a measurement of implementation.

[8] The mechanistic characteristics, too, were of course of a general character. This, however, did not pose such a problem in the evaluation of the lessons. The fact that this evaluation did prove difficult for the realistic characteristics indicates, once again, the important role played by the content in RME.

[9] That this was no longer possible had to do with the framework in which the research was being conducted. On the one hand, the budget limitations did not allow a thorough run of

all the developed instruments beforehand. On the other hand, such interim adjustment of an instrument on the basis of insight acquired during the collection and analysis of data is inadmissible within the empirical analytical research tradition in which this project was assumed to take place. Interim discoveries such as these are attributed in this research tradition to an inadequate plan of research. Moreover, interim adaptations of instruments are regarded as an inadmissible disturbance of the research, making the drawing of scientific conclusions impossible.

10 See also Section 1.4.2.

11 Both tests are still available (see NICL, 1992; NICL, 1993).

12 These tests were developed by Van den Heuvel-Panhuizen and Gravemeijer (1990a).

13 The key words of the instructions are printed below the pictures of the test pages. In their complete form, these instructions were read aloud by the test administrator.

14 This idea was taken from the ‘Entry Tests’ developed for the Kwantiwijzer instruments (see Lit, 1988).

15 This is not true of the written tests developed later on. For the second-grade tests, a number of problems could be taken from, for instance, the test ‘Arithmetic Puzzles’ from the ‘Arithmetic in a Second Language’ project (Van Galen et al., 1985). The situation with respect to the number-facts tests and the oral tests was clearly different as well. The design of the number-facts tests strongly resembled the ‘Entry Tests’ developed for the Kwantiwijzer instruments (Lit, 1988). Moreover, the Kwantiwijzer instruments (Van den Berg and Van Eerde, 1985) themselves were an important source of inspiration for the choice of interview techniques. The same is true of the publications of O’Brien and Richard (1971), Lankford (1974), Schoen (1979), Scheer (1980), Rudnitsky, Drickamer, and Handy (1981) and Labinowicz (1985). In addition, the research of Groenewegen and Gravemeijer (1988) was used for choosing solution strategies.

16 The test entitled ‘Arithmetic Puzzles’ clearly served as an example here.

17 This is the version of the tests that was used in the MORE research. That these tests are called the ‘final version’ does not imply, however, that there is no room for improvement.

18 Due to the lack of time for an independent development phase, the test development progressed concurrently with the data collection.

19 Similar verification also occurred after each trial run.

20 Nor had this been expected by the developers of the tests themselves.

21 The instruction included in Figure 2.7 is the literal text that was read aloud during administration of the test.

22 In itself, the fact that beginning first-graders already possess more knowledge of arithmetic than is often assumed, is not entirely new. See Chapter 5 for more on this topic.

23 This qualification is relative, however. A domain-specific test for measuring children’s learning potential (Tissink, Hamers, and Van Luit, 1993), which included problems taken from this particular MORE test, was described as ‘a new branch on the assessment tree’ (Hamers, 1991).

24 NZR consists almost entirely of bare operations, while WIG is much more application oriented.

25 It should be noted that the similarity between the two groups was not as strong in the other pair of problems (consisting of 33 – 25 as a bare problem and as a context problem) included on this test. There, too, however, the results of the group as a whole differed strikingly between the context problem and the bare operation.

26 As NZR was ahead of WIG with respect to these types of problems, this was even more the case for WIG than for NZR.

27 The third second-grade test was a number-facts test.

28 Nor was this really necessary here as the tracks are almost always visible in column arithmetic if the problem is figured out on the test paper itself.

29 Due to a lack of time, no thorough comparative study was ever made of the strategy data collected during the interviews and the data that surfaced on the written tests. Because the

instrument development and the data collection ran according to a strict timetable (as soon as the tests results had been analyzed, the interviews had to start), it was not possible to develop corresponding categories of solution strategies for both types of data collection.

30 The test problem illustrated here is the trial version of this problem. In the final version, the answer box was replaced by an answer line. The reason for this was that the answer box interfered with the tile pattern. An example of this interference can be found in the right-hand example of Figure 2.16.

31 Here, “altered insights into educational theory ...” is purposely spoken of in general terms as it does not apply specifically to RME, but applies as well to other reform movements that stem from a constructivistic approach to mathematics education (see Sections 3.3.2 and 4.1.1).

32 Instead of providing feedback, such assessment problems give a kind of ‘feed-forward’!

3 Assessment within RME – an up-to-date general description

3.1 Further steps towards a theoretical basis for RME assessment

3.1.1 A shift in opinion within RME regarding assessment

In comparison with the early period of RME – when Dutch primary education was threatening to become awash in a flood of tests and test-oriented education (see Chapter 1) – a recent shift has occurred in the attitude of RME developers with respect to assessment. Then as well as now, there has always been considerable interest in assessment. In contrast with two decades ago, however, when most of the assessment was imported and was held in considerable disregard by the pioneers of RME, more interest has been shown in recent years from within. A significant catalyst was provided by the endeavour to secure the new secondary school mathematics curriculum through appropriate exams. In addition, the further development and crystallization of RME has led to a viewpoint that regards assessment as an integral element of education, rather than as a necessary reaction to an undesirable development. As a result, the essential question:

> “Does assessment reflect the theory of instruction (and learning)?” (De Lange, 1992b, p. 54)

is now also inverted, that is to say:

> what does the RME educational theory signify for assessment?

Moreover, thoughts on the consequences for assessment have also influenced the RME educational theory itself (see Section 4.2.2a).

This shift in attitude first began during the mid-nineteen-eighties, which is about as precise as one can be about this process. Moreover, no collective shift in attitude took place, covering all those involved in the development of RME. Freudenthal, for instance, never participated in this shift and continued to maintain his suspicious attitude. In his last book he remarked, derisively:

> “Tests should be trustworthy, and what is more trustworthy than numbers, obtained by measurement (...)?” (Freudenthal, 1991, p. 154).

Freudenthal continued to hold the opinion that mathematics education was an unsuitable area for applying physical science models. Nevertheless, an enormous expansion of psychometrics was taking place in education and, in his opinion, measuring instruments were being developed that relegated people to the position of objects. According to Freudenthal, quantitative measurement was being regarded as the sole objective, and no consideration was given to the matter of whether this type of measurement was indeed meaningful for the understanding and explanation of certain phenomena. On the other hand, Freudenthal did admit that this criticism of assessment – seen in the light of recent developments – might appear outdated. He stated namely that:

> “There is an unmistakable trend away from [formal testing techniques] among educationalists with a strong mathematical background, who have good reasons to do so and are courageous enough to no longer be afraid of being reprimanded by methodologists for neglect of ‘objective’ methods. Open testing, interview, and observation of learning processes are gaining ground, and well-designed intersubjectivity is given a chance to supersede ill-conceived objectivity. Little attention, however, has been paid to the methodology behind the new approach” (Freudenthal, 1991, p. 155).

Freudenthal (ibid.) did make one exception regarding this methodology, namely, the work of De Lange (see Chapter 1). In Freudenthal’s opinion, this work demonstrated the presence, in any case, of promising new developments, although he did wonder whether these developments would be successful in the long run.

In spite of his reticence, Freudenthal still felt that assessment needed to be afforded it’s own place in developmental research whose focus was the further elaboration and implementation of the RME educational theory.

The shift from resistance to a more constructive, open attitude to assessment may have to do with the increasing acceptance of RME in recent years. It is, therefore, no longer necessary to brace oneself against inadequate tests that could discredit RME (see also Treffers, 1987b and Section 1.4.3). Positive experiences with new methods of assessment – for instance, the new secondary school exams and the PPON tests for primary school – have certainly also stimulated this further exploration from within.

Developments elsewhere may have contributed to this shift in attitude as well. In contrast to twenty years ago, recent ideas on assessment being formulated abroad show considerable similarity to those ideas within RME. Many of the objections raised earlier by proponents of RME are now being heard again, only now from other quarters, primarily in the realm of (socio-) constructivism and cognitive psychology. Moreover, mathematics educators in various countries – particularly Great Britain, Australia and the United States – are now also working on a mathematics education reform that requires a different method of assessment.

To sum up, there is now a ubiquitous litany – although the reciters come from a variety of backgrounds – that assessment in mathematics education must be changed on all levels. Partly thanks to this litany, the time is now ripe for a new evaluation of RME from the perspective of assessment.

3.1.2 The necessity for a theory of mathematics assessment

One of the points now emphasized abroad is the necessity for developing a theory of assessment in mathematics education (see Glaser, 1986, cited by Grouws and Meier, 1992; Collis, 1992; Grouws and Meier, 1992; Webb, 1992, 1993; MSEB, 1993a/b). Grouws and Meier (1992), for instance, stress that the development of alternative forms of assessment must proceed simultaneously with the development of an appropriate theory. The new developments could then be combined to lay a foundation for further research and development.

A major step towards such a theory has since been taken by Webb (1992, 1993). Through an analysis of relevant literature and research results, he has attempted to identify a number of theoretical principles for assessing mathematical knowledge[1], which, in turn, can lead to a theory of mathematics assessment. The result is a complex unit consisting of the definitions of assessment and related concepts, an overview of the various assessment purposes, the different concepts of mathematics that can form the foundation for a given approach to assessment, and a detailed description of the various components of assessment. In addition, Webb also indicates areas in which a theory of assessment in mathematics education must devote attention and provide answers.

Although this theory of assessment as propounded by Webb is intended to connect to assessment in other subject areas and, moreover, is intended to be concerned with the relation between specific subject matter knowledge and general cognitive skills, in Webb's opinion the subject of mathematics is specific enough to justify a separate theory of assessment. Aside from the fact that certain mathematical skills (such as proving something) require specific assessment techniques, the inadequacy of a general theory of assessment is primarily due to that which is assessed (the subject content). Moreover, assessment in an educational setting clearly differs in and of itself from the psychometric approach to assessment. Webb does not stand alone here. In the report entitled 'Measuring What Counts' (MSEB, 1993a), considerable emphasis is laid on the importance of the role played by subject content. The theory advocated by Webb is therefore a theory of mathematics assessment that can:

> "...unify the principles of assessment with the unique characteristics of the structure of mathematics. Such a unification could serve as a framework for assessing mathematical knowledge and performance, both within the classroom and across large groups" (Webb, 1992, p. 680).

In other words, Webb is seeking a kind of universal theory of mathematics assessment. [2] But is there any point in gathering all these different standpoints on learning and teaching mathematics under one umbrella of assessment theory? Can such a general theory provide, for example, the specific support appropriate to a given view of mathematics education? The reaction of Collis (1992) to a particular test problem indicates how difficult this may be.

Collis, who may be regarded as a representative of cognitive psychology, is a strong proponent of constructing a better theoretical foundation for assessment problems. In order to add weight to his argument, Collis offers an example of how things may go wrong when, due to the lack of such a foundation, one loses sight of the representational level (sensorimotor, iconic, concrete-symbolic, formal, post-formal) one wishes to assess.

The assessment problem in question involves two simple geometrical figures that one student must describe to another over the telephone.[3] The 'flaw' in this problem is that, because of the chosen context, some students do not solve this problem using mathematical language (concrete-symbolic) but, rather, in an informal manner using everyday language (iconic).

The degree to which this problem is flawed (or the degree to which assessment should occur on one level) is dependent upon the underlying viewpoints on the subject of mathematics and on learning and teaching this subject. Within RME, such a problem would be evaluated much more positively, and potential improvement would be sought in something other than a more unambiguous interpretation of the representational level (see Sections 4.1.3 and 4.2).

Webb (1992), too, agrees that considerable differences may exist with regard to the underlying viewpoints. In his opinion, however, it is the task of the theory to make this explicit and to point out the consequences. It is clear that Webb is seeking a kind of meta-theory. And his aspirations extend further than the elaboration of the local RME educational theory for assessment, which is the focus of this chapter. Nevertheless, Webb's extensive inventory of everything available in the area of assessment of mathematics education can certainly be of service here. Moreover, his meta-outline can also be used to assess the elaboration of the RME theory for assessment after the fact, by determining whether it has satisfied the criteria set by Webb (see Section 4.2.2b).

3.1.3 Elaboration of the RME educational theory for assessment

The goal of the present chapter is to examine the RME educational theory anew from the perspective of assessment, in order to arrive at an elaboration of this theory on behalf of assessment. This entails providing a description of the nature of assessment in RME, which then leads to a number of ground rules for this assessment. These ground rules clarify which choices are made by RME with respect to assessment and how these choices link up with and arise from the starting points of the RME theory of learning and teaching mathematics. The purpose of this elaboration is to lead to an improvement of assessment practice through giving certain instructions. In this way – whether practically or theoretically – a contribution can be made to optimizing RME.

With regard to the development of a theory, the theory for assessment would appear to be proceeding in the same way as did the RME educational theory itself. A retrospective description of the underlying educational theory was not articulated until a later phase in RME development (Treffers, 1987a).[4] In assessment, too, a similar situation has arisen involving the retrospective registration of viewpoints acquired through research experiences and subsequent reflection on them. This manner of integrated development of education (in the broadest sense of the word) and theory (where the theory is established in retrospect) is characteristic of theory development within RME (see Freudenthal, 1987b, 1991; Gravemeijer, 1992, 1994; Van den Heuvel-Panhuizen, 1992a; Streefland, 1993). Gravemeijer (1994, p. 12) speaks of:

> “a bottom-up approach, where deliberation on practice is taken as a starting point for the constitution of a description on a more theoretical level.”

This approach clearly differs from that of Webb (1992, 1993), who would seem to be furnishing a previously established framework, using the various purposes and components of assessment to form the general organizational principles.[5] The RME educational theory, in contrast, was developed (or formulated) for the most part after the bulk of the curriculum development had already taken place (Treffers, 1987a). This does not at all imply, however, that this development took place without any theoretical foundation.

> “All the time the theory was still implicit in the action, that is in the creation of instruction and only as the work went on could a theoretic[al] framework be constructed” (Treffers, 1987a, p. 242).

Gravemeijer (1994, p. 115), too, is quite clear in his reference to RME as a retrospective theory which, in his words, means that:

> “...it is the reconstruction of a theory in action.”

It should be mentioned here that, before this reconstruction occurred in the form of Treffers’ book ‘Three Dimensions’, Freudenthal (1973, 1978a) had already published on this topic. The theory was thus only implicit up to a certain level, which was also why Gravemeijer (1992) emphasized so strongly the theoretically-guided aspect of educational development. In the words of Streefland (1993, p. 20):

> “The cyclical process of development and research, besides being theory-*oriented*, is therefore theory-*guided* as well.”

Similarly, the present reconstruction of the theoretical framework with respect to assessment in RME need not begin from the beginning, but can build further on earlier initiatives in this area.

These earlier initiatives can be found both in assessment ideas from the early years of RME, which were mainly developed for primary education, and in De Lange’s pioneering work, which focused on secondary education assessment (see Chapter 1). The principles formulated by De Lange (1987a) for RME assessment and the levels in assessment problems he later distinguished (De Lange, 1995) form a major part of the foundation for further elaboration of the RME educational theory for assessment as discussed in this and the following chapters.

This elaboration is fostered, in the first place, by the RME educational theory itself and by the insights acquired through developmental research on assessment that took place in several research projects: the MORE research project (see Chapters 2 and 5), a small research project in special education (see Chapter 6) and the 'Mathematics in Context' project (see Chapter 7 and Van den Heuvel-Panhuizen, 1995a).

After an initial general examination of current viewpoints on assessment within RME, this chapter will look at the theoretical ideas and research findings from sources outside the RME circle. It will examine to what extent these outside ideas and findings diverge from or correspond to the RME viewpoints on assessment, and whether they might contribute to the foundation, enrichment or refinement of the RME viewpoints.[6]

3.2 RME assessment

3.2.1 RME requires a different method of assessment

As was described earlier, in Chapters 1 and 2, an adapted method of assessment was gradually developed in The Netherlands in the wake of the development of RME. This means that assessment, like education: must regard mathematics as a human activity, while focusing on meaningful applications. Moreover, as in education, an important role in assessment is played by the students who, by using contexts and models, can pass through various levels of mathematization and thereby develop their own mathematics (see Section 1.1.2). In other words, if assessment is to be appropriate to RME, then it must be tailored to the three pillars of RME, to wit: the viewpoints on the subject matter, on how instruction should be given, and on the manner in which learning processes progress. Together, they determine what, why, and how assessment occurs.

The choices made by RME have already been described, for the most part, in Chapter 1. Although not entirely crystallized, and still bearing the traces of the era in which it was developed, even during the early years of RME it was clear how the characteristics of the intended education were reflected in the assessment in mind. The development of tests in the framework of the MORE research project (see Chapter 2) also clearly demonstrated how these characteristics served as guidelines for the tests that were developed. The present chapter takes stock of the situation with respect to assessment within RME, taking into account the earlier initiatives (see Chapter 1), and is supplemented by a reflection on the RME theory and developmental research on assessment (see Chapters 5, 6 and 7).

The next section presents a general survey, in which two determining characteristics of RME assessment are discussed, namely, its 'didactical' nature and the crucial role played by the chosen problems. The role of contexts is also discussed in connection to this crucial role. This is followed by an expansion in the field of vision in order to examine developments outside the RME circle with respect to these characteristics.

3.2.2 Assessment within RME means: didactical assessment

The assessment most appropriate to RME can best be described as ‘didactical assessment’. This assessment is closely linked to the education, and all aspects of it reveal this educational orientation. This means that the purpose of the assessment as well as the content, the methods applied and the instruments used are all of a didactical nature.

3.2.2a The purpose is didactical

Assessment within RME is primarily assessment on behalf of education. Its purpose is to collect certain data on the students and their learning processes, in order to make particular educational decisions. These decisions may involve all levels of the education and may vary from local decisions on suitable instructional activities for tomorrow’s mathematics lessons, to broader decisions on whether to pass or fail, on which students need what extra assistance, on whether or not to introduce something new, on a given approach to a given program component, or on whether to take certain large-scale measures regarding the design of the mathematics education. The didactical nature of the purpose of assessment is expressed most clearly in the ever present focus on educational improvement. Even when the purpose of the assessment is to reach a decision in a pass or fail situation, the education, as well as the student, is evaluated. As was stated earlier (see Chapter 1), this assessment purpose – which focuses on educational evaluation and educational development – was consistently present from the very start of RME development (see Freudenthal, 1973). The closer assessment lies to actual education, the more self-evident this will be. Then, in the classroom, direct assessments are made in order to determine how the education can best dovetail with the students’ already present knowledge and abilities. When assessment is not directly linked to educational practice there is the danger of ‘anti-didactical’ assessment. For this reason, large-scale assessment in RME is always examined critically with respect to its potential contribution to educational improvement.

Characteristic of RME, moreover, is the bilateral nature of this focus on improvement. Not only must assessment lead to good education, but it must simultaneously improve learning by giving the students feedback on their learning processes. De Lange articulated this latter point explicitly. The first of the principles formulated by him for developing RME assessment problems (see Section 1.2.5c) reads as follows:

> “The first and main purpose of testing is to improve learning” (De Lange, 1987a, p. 179).[7]

3.2.2b The content is didactical

Choosing didactical assessment means that the content of the tests is closely linked to the prevailing viewpoints within RME on the subject of mathematics and to the goals aspired to by the education. This implies that the assessment may not be restricted to particular easily assessed isolated skills, but, instead, that the entire range of goals must be covered, both in breadth (all curriculum components and the links between them) and in depth (all levels of comprehension).

Notwithstanding the emphasis (from the perspective of innovation) often laid by RME on the higher-order goals, this in no way implies that the students no longer need acquire the basic knowledge and skills, nor that these no longer need be assessed in RME. In other words, 'assessment of the entire range of goals' means just that: the entire range.

The didactical nature of that which is assessed emerges even more clearly in the priority given to learning processes. Because, as mentioned earlier (see Sections 1.1.2 and 1.2.3a), mathematics is viewed in RME as a student's own activity, in which he or she uses certain mathematical insights and devices to grasp a given problem situation, the main focus of assessment in RME is obviously not on the results (with the exception of the assessment of the number-facts), but on the solution procedures themselves. Assessment must provide insight into the students' mathematization activities.

The breadth, depth and coherence of the product-aspect, as well as the focus on the process-aspect can also be found in the principles formulated by De Lange:

> "The third principle is that the task should *operationalize* the goals as much as possible. [...] This also means that we are not interested in the first place in the *product* (the solution) but in the process that leads to this product" (De Lange, 1987a, p. 180).

3.2.2c The procedures are didactical

This didactical nature is clearly recognizable once again in the procedures applied in RME assessment. The most distinctive procedure in this respect is the integration of instruction and assessment. A striking example of this integration are the test-lessons discussed earlier and their accompanying informal forms of assessment, such as observation, oral questioning and written assignments (see Section 1.2.5b). This integration of instruction and assessment also means that assessment will play a role during each phase of the teaching and learning process. Moreover, it implies that assessment will look forward as well as backward. Looking *backward* involves determining what the students have learned, in the sense of educational results. Although this, too, can produce certain indications for instruction, it is looking *forward* that concentrates on finding footholds for further instruction. This looking forward or 'advance testing' is inextricably linked to the emphasis laid by RME on the students' own contributions and on building further on what the students already know. The major role played by the teacher is yet another aspect of the integration of instruction and assessment. Just as the learning process in RME depends to a great extent on how the teacher builds on what the students present, so does the teacher also determine what, how and when will be tested.

To summarize, one can state that the method of assessment must be appropriate to the educational practice and must be able to be conducted within it. This, too, is one of the principles formulated by De Lange:

> "Our fifth and last principle is that, when developing alternative ways of evaluating students, we should restrict ourselves to tests that can readily be carried out in school practice" (De Lange, 1987a, p. 183).

3.2.2d The tools are didactical

Because RME requires as complete a picture of the students as possible, assessment in RME involves using an extensive variety of tools for collecting the necessary information. The closer these tools lie to the education and its goals the better, as they will then produce information that can be applied directly to education. In RME, aside from a teacher's specific purpose for assessment, the assessment tools are often indistinguishable from the tools used to initiate certain learning processes. A question may, for instance, be intended both to effect a given reflection and its resulting rise in level, and to assess a given insight. What is important, in any case, is that the assessment tools expose the learning process and that they provide insight into the students' repertoire of knowledge, skill and insight at a given moment. This requires an open method of assessment, in which the students are able to demonstrate their abilities (see Section 1.2.3e). This focus on what the students are able to do is also expressed in the principles formulated by De Lange:

> "The second principle is [...]: Methods of assessment [...] should be such that they enable candidates to demonstrate what *they know* rather than what they do not know" (De Lange, 1987a, p. 180).[8]

Because assessment focused for too long on determining whether the students knew something specific, one mostly received, in return, information on what students did not know. The emphasis now is on 'showing what they *do* know'. This does not mean, however, that 'knowing what they *do not* know' has become unimportant. The point – as will be discussed later (see Section 4.2.1e) – is not so much to distinguish between knowing and not-knowing, or ability and inability, but to distinguish on which level they are able to do or know something.

Finally, the consequence of the didactical nature of the applied tools is that a distance is clearly maintained from a psychometric approach in which the overpowering pursuit of objective and reliable assessment often occurs at the expense of the content (see Section 1.2.4b). This, again, is expressed in the principles formulated by De Lange:

> "The fourth principle is that the quality of a test is *not* defined by its accessibility to objective scoring" (De Lange, 1987a, p. 180).

3.2.3 Assessment within RME means: the problems play a crucial role

Another characteristic of RME assessment is the crucial role played by the assessment problems.[9] In RME, *what* is being asked is more important than the format, or the way in which something is asked. After all, a poor question in written form will not immediately improve simply by being presented orally, even though an oral format offers the opportunity to observe the students as they answer the question. For this reason, improvement of assessment in RME is principally sought through the improvement of the assessment problems (see, for instance, De Lange, 1995; Van den Heuvel-Panhuizen, 1993b, 1994c).

This focus on the problems themselves is a salient feature of RME in its entirety. Not only the assessment, but also the manner of instruction is problem-centered to a great extent (see also Goffree, 1979; Gravemeijer, 1994). Starting with a particular problem, students develop mathematical tools and insights with which they can then solve new problems. These problems are viewed as situations requiring a solution, which can be reached through organization, schematization and processing of the data or, in other words, by mathematization (see Section 1.1.2). An essential aspect here is the reflection on the mathematical activities that can enable a rise in level; for example, the discovery of a certain short cut that can then be applied to new problems. The problem situations may also differ considerably with respect to complexity and level and may involve real-life situations as well as mathematical structures.[10]

The didactical phenomenology or mathematical-didactical analyses (see Section 1.1.2b and Section 1.2.5a, Note 30) play an important role in the development of education. They play an equally important role in determining content in the development of assessment problems: what is worth knowing about the students' insights and skills, and which situations (or problems) are suitable for providing this knowledge? Aside from the content of the assessment – which can set its own content-specific requirements for the problems – there are two general criteria which the problems must fulfill if they are to be suitable for RME assessment: they must be meaningful and informative.

The following section will examine these criteria and discuss how they are linked to the characteristics of RME. This is followed by a separate discussion on the significance of the use of contexts.

It should be mentioned beforehand that, although the didactical nature and the problem-centeredness were mentioned as separate characteristics of RME assessment, they cannot be regarded independently of one another. They are reciprocally connected and run to some extent parallel, as has already been demonstrated by the overlap between the principles formulated by De Lange (1987a) and the various aspects of didactical assessment. Reference to these principles in the previous sections was, in fact, a prelude to the requirements placed by RME on assessment problems, which will now be discussed. Here, again, De Lange's principles are apparent.

3.2.4 RME requirements for assessment problems

3.2.4a The problems must be meaningful

In RME, which is based on the idea of mathematics as a human activity (Freudenthal, 1973), the primary educational goal is that the students learn to do mathematics as an activity. This implies that one should “teach mathematics so as to be useful” (see Freudenthal, 1968). The students must learn to analyze and organize problem situations and to apply mathematics flexibly in problem situations that are meaningful to them. From the point of view of the *student*, the problems must therefore be accessible, inviting, and worthwhile solving. The problems must also be challenging (Treffers, 1978, 1987a) and it must be obvious to the students why an answer to a given question is required (Gravemeijer, 1982) (see also Section 1.2.3f). This meaningful aspect of the problems may also entail allowing the students to pose or think up questions themselves (see, for instance, Van den Heuvel-Panhuizen, Middleton, and Streefland, 1995). Another significant element is that the students can mold a given problem situation so that they, themselves, are in a certain sense master of the situation (‘owning the problem’). This is the case, for instance, in the percentage problems, where the students may decide whether someone has passed or failed a given exam (see Streefland and Van den Heuvel-Panhuizen, 1992; Van den Heuvel-Panhuizen, 1995a), or in the problems in which the students may decide what to buy (see Section 4.1.3d and Figure 2.8).

In order for the problems to be meaningful with respect to *subject matter*, they need to reflect important goals. If something is not worthwhile learning, then neither is it worthwhile assessing. The problems should also be correct mathematically and, furthermore, should not be restricted to goals that can be easily assessed, but, rather, should cover the entire breadth and depth of the mathematical area. This means that assessment should cover all topics of the subject matter and should include problems on each level: from basic skills to higher-order reasoning (De Lange’s *third principle)*. The emphasis on higer-order reasoning implies that the problem situations should be fairly unfamiliar to the students, as this will then offer them an opportunity for mathematization. In other words, problem solving in RME does not mean simply conducting a fixed procedure in set situations. Consequently, the problems can be solved in different ways. This aspect is present in the next requirement as well.

3.2.4b The problems must be informative

In RME, the students are expected to play an active role in constructing their own mathematical knowledge (see Sections 1.1.2a and 1.1.2c). The education is designed to dovetail as closely as possible with the students’ informal knowledge, and therefore help them to achieve a higher level of understanding through guided re-invention. In order to support this process of guided re-invention, the assessment problems must provide the teacher with a maximum of information on the students’ knowledge, insight and skills, including their strategies. For this to succeed, the problems must again be accessible to the students. In addition to what has been said of accessibility in the previous section, accessibility with respect to the informative nature of the problems implies that the accompanying test-instructions must be as clear as possible to the students. Another point is that the students must have the opportunity to give their own answers in their own words (see Section 1.2.3e). Moreover, there must be room for the students' own constructions, which means that the problems must be of a kind that can be solved in different ways and on different levels. In this way, the problems must be able to make the learning process transparent – to both the teachers and the students. The students (as stated in De Lange's *first principle*) are active participants, and as such should also receive feedback on their learning progress. Furthermore, the problems should reflect 'positive testing'. They should allow the students to demonstrate what they know, rather than simply revealing what the students do not yet know (De Lange's *second principle*). Again, this means that accessible problems must be used that can be solved in different ways and on different levels. Contexts play an important role in fulfilling these requirements. This point is dealt with further in Sections 3.2.5, 3.2.6, and 3.2.7.

Another requirement is the quality of the information received. If the information is to be usable, then one must be able to count on it providing an accurate picture of the student. Aside from the fact that one must beware of certain inaccuracies which may arise as the result of inequitable differences in administration and analysis procedures, this means, more importantly, that justice must be done with respect to the students (De Lange's *second* and *fourth principles*).[11] In other words, one must test fairly.

Lastly, there is the practical side of the problems. If the problems are to produce usable information, then their administration and the analysis of the students' reactions must also be feasible (De Lange's *fifth principle*).

It is interesting to compare the requirements formulated here for problems used in RME assessment with other RME requirements for problems.

According to Goffree (1979), for example, a problem that is designed as an 'entry problem' for a lesson, or series of lessons, must inspire the students to engage in active investigation. What is essential for such an investigative attitude is that, on the one hand, the problem situation encompass something mathematical, or a specific mathematical activity, and, on the other hand, the students will be able to recognize this situation and imagine themselves within it. Another of Goffree's criteria for entry problems is that they be accessible on different levels.

According to Dekker (1991), working with heterogeneous groups requires problems that are (i) realistic, in the sense that the students are able to imagine something in them, (ii) complex, in that they demand different skills, (iii) contain the potential for construction and (iv) contain the potential for level increase. With the exception of the complexity – whose purpose is to stimulate interaction between students – these characteristics can all be found in the above-mentioned requirements placed by RME on assessment problems.

The high degree of consistency between the various aspects of RME is once again apparent here. Problems suitable for instructing heterogeneous groups are, on the whole, suitable for assessment as well. Or, put more broadly, good assessment problems in RME have much in common with good instructional problems. This is particularly true of the requirement that the problems be meaningful. The requirement that the problems be informative, on the other hand, is more applicable to assessment.

3.2.5 Use of contexts in RME assessment problems

As mentioned earlier (see Section 1.1.2 and, especially, Section 1.1.2b), use of contexts is one of the salient features of RME. According to Treffers and Goffree (1985), contexts play a major role in all aspects of education: in concept forming, model forming (see also Gravemeijer, 1994; Van den Heuvel-Panhuizen, 1995b), application, and in practicing certain skills. Aside from the fact that contexts – as a source and as an area of application (see Streefland, 1985b) – form the backbone of the curriculum, they also fulfill an important function in assessment.

In RME, contexts are viewed in both a broad and a narrow sense. In a broad sense, they may refer to situations in everyday life and to fantasy situations, but also to bare mathematical problems. What is important is that these be situations or problems which are suitable for mathematization, in which the students are able to imagine something and can also make use of their own experiences and knowledge. Bare problems, too, may suffice here (see Section 1.1.2 and Treffers and Goffree, 1985). In a narrow sense, context means the specific situations to which they refer.

Various kinds of contexts can be distinguished, depending upon the opportunities they offer. De Lange (1979), referring to the opportunities for mathematization, distinguishes three types of contexts. ‘First-order’ contexts only involve the translation of textually packaged mathematical problems, while ‘second-order’ and ‘third-order’ contexts actually offer the opportunity for mathematization. Third-order contexts are understood to be contexts that allow students to discover new mathematical concepts.

Contexts also differ with respect to their degree of reality. Here, too, De Lange (1995) distinguishes three categories: (i) no context, (ii) ‘camouflage contexts’ (also called ‘dressed up’ contexts) and (iii) ‘relevant and essential contexts’. In the first case, no real context is present, but, instead, a bare mathematical problem or, stated more positively, a mathematical context. The camouflage contexts correspond to the above-mentioned first-order contexts. Here, the contexts are not actually relevant but are merely dressed up bare problems. A relevant and essential context, on the other hand (the word says it), makes a relevant contribution to the problem. Although one's first thought here may be of a rich topic, presented in the form of an extensive task, according to De Lange (1995) even very simple problems may have a relevant and essential context. This can even be true of multiple-choice problems. As an example, De Lange offers a multiple-choice problem in which one must estimate the width of the classroom. He also shows how artificial contexts, too – such as a story about a new disease in the 21st century – can be relevant. The disease in question was, in fact, AIDS, but was changed, for emotional reasons, to a science-fiction disease. This example clearly demonstrates that it is more important that the context stimulate and offer support for reflection than that the data and the situation be real. Moreover, the degree of reality of a context is relative. De Lange (1995) wonders how close to the students' world contexts must necessarily be. How suitable is a context involving being an airplane pilot, for instance, if most of the students have never had such an experience? In De Lange's experience, such contexts do indeed work, and with girls, too. This context was used in a booklet on trigonometry and vectors, which was pilot tested at an almost exclusively girls' school. This example, too, indicates the complexity of the contextual aspect. One single rule cannot always be found for choosing contexts, but one should at least try and create a balance between a good context and a good mathematical problem (De Lange, 1995).

In contrast to De Lange's experiences described above, which primarily involved extended assessment tasks, experiences with contexts in briefer tasks took place in the MORE research project (see Chapter 2). In short-task problems, too, the contexts may be more or less real, may stand in various degrees of proximity to the students, and may offer more or less opportunity for mathematization. The single point of difference, however, is that the contexts in short-task problems are indicated by minimal means.

3.2.6 Illustrations as context bearers

In the tests for the MORE research project, illustrations, supplemented by a few, short sentences in which the assignment is formulated, are used as context bearers. The function of these illustrations is more important, however, than the typical obligatory picture accompanying a problem, whose usual purpose is to make the problem more attractive. Besides a (i) *motivating function*, illustrations also have a number of other functions, as follows:

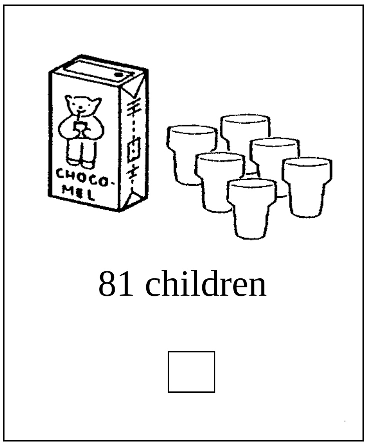
(ii) *situation describers*; one illustration can tell more than an entire story (see, for example, the Pinball problem – Figure 2.7);

(iii) *information providers*; the necessary (supplemental) information can be derived from the illustration (see, for example, the Shopping problem – Figure 2.8);

(iv) *action indicators*; a given action is elicited that has the potential to lead to a solution (see, for example, the Chocolate Milk problem – Figure 3.1);

(v) *model suppliers*; the illustration contains certain structuring possibilities that can be used to solve the problem (see, for example, the Tiled Floor problem – Figure 2.15);

(vi) *solution and solution-strategy communicators*; the solution and aspects of the applied strategies can be indicated in the drawing (see, for example, the Candle problem – Figure 2.9).



Instructions to be read aloud:

"One carton of chocolate milk will fill 6 glasses. There are 81 children on the playground. Each one gets a glass of chocolate milk.
How many cartons will you need?
Write your answer in the empty box."

Figure 3.1: Chocolate Milk problem

It is obvious that each illustration does not always fulfill each function and that the function of an illustration is not always easily labeled. The above categorization is therefore not intended as a classification scheme for differentiating types of illustrations in context problems, but rather as a way of acquiring a sense of the potential for using illustrations as context bearers in assessment problems.

3.2.7 Principal functions of contexts in assessment problems

The same analysis can also be made of the role that the contexts themselves (in the narrow sense described above) play in assessment. The following functions of contexts, specifically related to assessment, arose from the experiences gained from the MORE project tests and the developmental research on assessment that followed.

3.2.7a Contexts enhance accessibility

In the first place, contexts can contribute to the accessibility of assessment problems. By starting from easily imaginable situations which are presented visually, the students will quite quickly grasp the purpose of a given problem. The advantage of this direct, visual presentation is that the students need not wrestle through an enormous amount of text before they can deal with the problem. In addition to making the situations recognizable and easily imaginable, a pleasant, inviting context can also increase the accessibility through its motivational element.

3.2.7b Contexts contribute to the latitude and transparency of the problems

Compared with most bare problems, context problems offer the students more opportunity for demonstrating their abilities. In bare problems – such as long division – the answer is either right or wrong. In a context problem – for instance, where one must figure out how many buses are needed to transport a large contingent of soccer fans – the students can also find the answer by applying an informal manner of division, namely, repeated subtraction. Because the problem can be solved on different levels, its elasticity is increased. Not only can the quick students solve the problem, but the slower students as well – on a lower level. This reduces the 'all-or-nothing' character of assessment.

By giving the students more latitude in the way they approach the problems, the contexts further increase the transparency of the assessment. In bare problems, the operation to be performed is generally fixed in the problem in question, so one can only verify whether students are able to perform certain procedures that they had learned earlier. For this reason, bare problems are not suitable for advance testing. One cannot present a long division problem to a student who has never done one and expect to find footholds for further instruction.

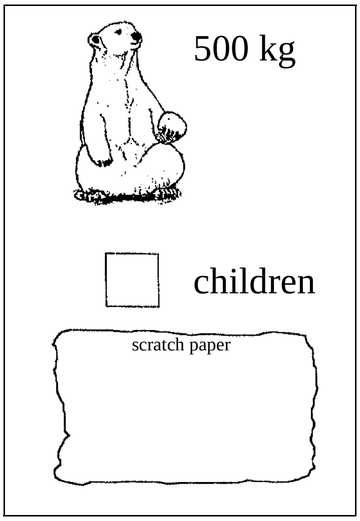
3.2.7c Contexts provide strategies

The most important aspect of contexts in assessment (assuming they are chosen well, which is also true of other functions) is that they can provide strategies (see, for instance, the Bead problem, Figure 2.12, discussed in Section 2.3.4). By imagining themselves in the situation to which the problem refers, the students can solve the problem in a way that was inspired, as it were, by the situation. Sometimes this will mean that the students use an accompanying drawing in a very direct way as a kind of model, while, at other times, it is the action enclosed within a given context that elicits the strategy. How close the students stick to the context with their solution depends upon the degree of insight and the repertoire of knowledge and skills they possess. This role of strategy provider is not only important with respect to expanding the breadth of assessment problems and the potential this creates for advance testing, but it touches the core goal of RME as well: the ability to solve a problem using mathematical means and insights. An essential element of this is formed by the ability to make use of what the contexts have to offer.

3.2.8 The Polar Bear problem as a paradigm of a context problem

The following example of an assessment problem demonstrates the viewpoint of RME quite well. It illustrates concretely the points that have been made here about assessment problems. This specific example does not mean, however, that such problems (consisting of a short, contextual task) are the only kind used in RME. It should be clear by now that RME uses a wide variety of assessment problems. For this reason, an example of a bare assessment problem is also included in the following section.

For the Polar Bear problem, one must put oneself in the shoes of third-grade children in November of the academic year. They have already become rather proficient at multiplication and division tables up to ten, and have some experience as well in doing multiplication and division with two-digit numbers. Written algorithms, however, have yet to be introduced. This is an ideal moment for assessment in RME; it includes, as always, looking backward to see what the students have already mastered and looking forward in search of footholds for further instruction. The Polar Bear problem, which was developed for this purpose, is shown in Figure 3.2.



Instructions to be read aloud:

“A polar bear weighs 500 kilograms.
How many children together weigh as much as one polar bear?
Write your answer in the empty box.

If you like, you may use the scratch paper.”

Figure 3.2: Polar Bear problem[12]

The Polar Bear problem is a meaningful task as it is both worthwhile solving and familiar. Children often encounter these kinds of problems in television programs, magazines and shops. They arouse children’s natural curiosity. Furthermore, the problem represents an important goal of mathematics education. In addition to applying calculation procedures, students are also required to solve real problems in RME, that is, problems where the solution procedure is not known beforehand and where not all the data is given. Often, such problems require students’ own contributions, such as making assumptions on missing data. This is why knowledge of all sorts of measurements from daily life is so important in RME. Due to the absence of information on the children’s weight, the Polar Bear problem becomes a real problem. The problem gives no indication of what kind of procedure must be performed, and it implicitly requires the students to think up the weight of an average child first. In both areas, the students are given room to make their own constructions, which may be considered the most important aspect of such problems. On the one hand, the students are provided the opportunity to show what they know, and, on the other hand, the teachers thus acquire information on how their students tackle problems, on their knowledge of measurements, on which ‘division’ strategies they apply, and on which models and notation schemes they use to support their thought processes in a hitherto unfamiliar division problem.

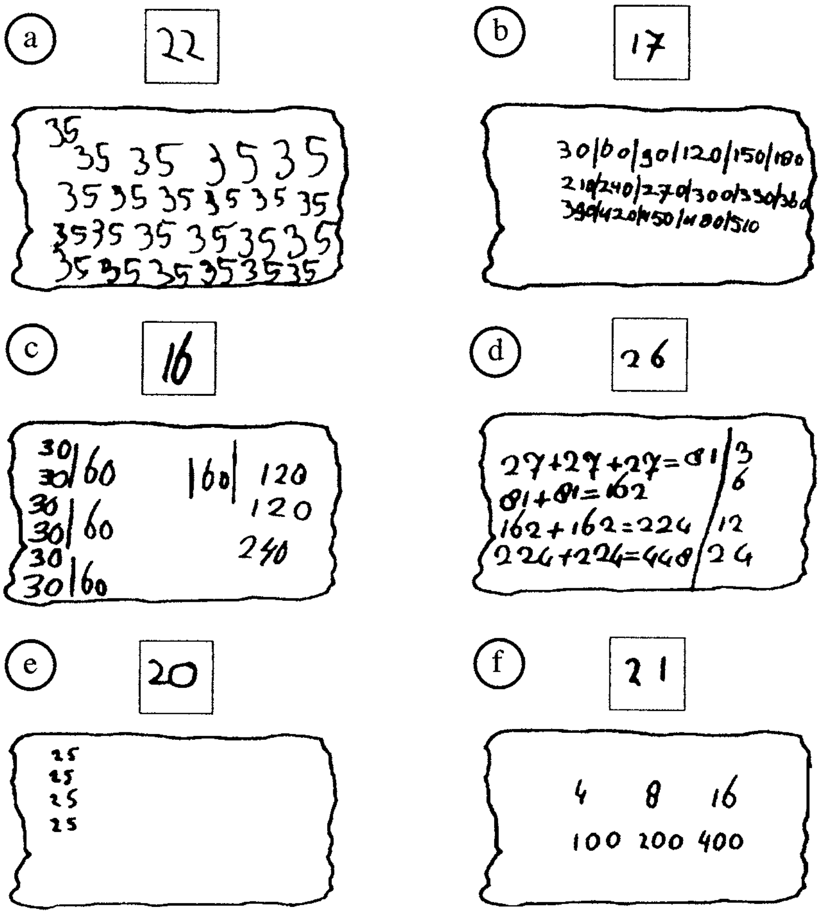


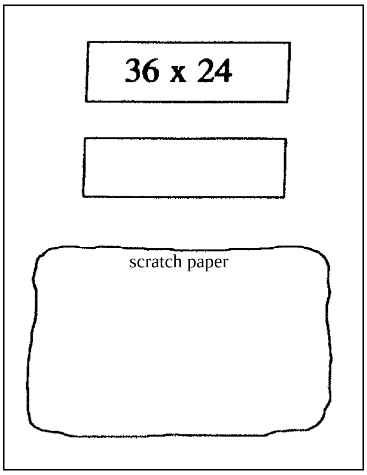
Figure 3.3: Student work on the Polar Bear problem

Figure 3.3 illustrates some examples of student work collected in the MORE research project. The scratch paper reveals, in the first place, that the students used different average weights, varying from 25 to 35 kg. Some used round numbers and some did not. There was also a considerable variety in the solution strategies applied.

As was to be expected, none of the students performed a division operation. Instead, they all chose a form of repeated multiplication: keep adding a given weight until you get near 500 kg. This was done in a great variety of ways: Student (a) kept adding up 35 mentally, and kept track by writing down 35 each time. The ‘35’ in the upper left-hand corner is probably intended to indicate the choice of 35 kg. Student (b) also kept adding up, but wrote down the new result each time. Student (c) applied a more structured form of adding up by combining this with doubling; the final doubling to reach the solution was done mentally: 8 children is 240 kg, so you’ll need 16 children. Student (d) also applied a doubling strategy; moreover, a clever way was found to keep track of how many times the given weight had been counted. The last part of the solution was found by mental arithmetic but, because of a mistake made when doubling 162, the final result was a bit high. Student (e) changed to mental arithmetic halfway through: 4 children are 100 kg, so you’ll need 20 children. And student (f) used a notation in the form of a ratio table to solve the problem.

3.2.9 The Same Answer problem as a paradigm of a bare problem

The following example demonstrates how bare problems can also be suitable assessment problems in RME. Alongside context problems, which can be mathematical contexts involving mathematization (see also Section 3.2.5), bare test problems can also be used to acquire information on insights and strategies. The example included here (see Figure 3.4) was designed for grade 5. Its purpose was to ascertain the extent of students’ insight into the operation of multiplication. More precisely, it concerned insight into the associative property and the ability to apply it without actually using this term in the test instructions.



Instructions to be read aloud:

“It is not necessary to fill in the result of this problem. The only thing you have to do is to think up a different multiplication problem that will give the same result as this one. You may not use the same numbers.
Write your multiplication problem in the empty box.

If you like, you may use the scratch paper.”

Figure 3.4: Same Answer problem[13]

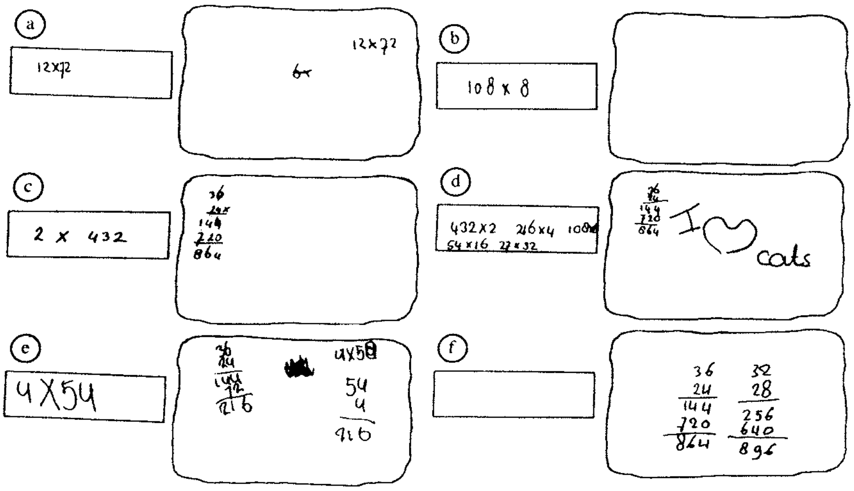


Figure 3.5: Student work on the Same Answer problem

Figure 3.5 again illustrates some examples of student work.[14] The scratch paper played a major role in this problem. Where the student came up with an alternative multiplication, but left the scratch paper empty, one may assume that the associative property was used.[15] This was evidently the case with students (a) and (b). The other students arrived at a different problem via the result, or attempted to do so. Here, too, the scratch paper reveals considerable differences in approach. Students (c) and (d) arrived at a multiplication giving the same result by halving the result of the given problem. Student (d) even generated a number of multiplications giving the same result. Student (e) went wrong due to an error in calculating the given multiplication. Student (f) did calculate correctly, but went no further with the result. Instead, this student thought up a new multiplication problem to see whether it would produce the same result. The procedure this student used to come up with this new multiplication (subtracting four from one number and adding four to the other) reveals that the student was thinking additively rather than multiplicatively.

3.3 Assessment from a broader perspective

The description of assessment in RME presented in the preceding sections is the result of a reflection on previously established theoretical ideas and on experiences gained from developmental research conducted within the RME circle. This perspective will be broadened in the second half of this chapter through the inclusion of developments and research data from outside this circle.

3.3.1 Different education requires a different method of assessment

Not only in The Netherlands are changes in mathematics education taking place. Worldwide, the conclusion has been drawn that traditional curricula no longer fulfill today's requirements and that, therefore, the curricula and their attendant assessment must change (Romberg, Zarinnia, and Collis, 1990). A striking feature of the current reform movement in mathematics education is the extent to which there is general agreement about the direction in which this new education should develop (AFG, 1991). This is attested to by numerous reports that have appeared in different countries.[16]

On the whole, the suggested changes – which correspond to the RME characteristics described in Chapter 1 – boil down to the following: mathematics education should no longer focus solely on the ready knowledge of various facts and the ability to perform isolated skills in routine problems[17], but should be principally concerned with developing concepts and skills that can be used to solve various non-routine problems, as are often found in reality. The new goals are: problem solving[18], higher-order thinking, reasoning, communication, and a critical disposition with respect to using mathematical tools. Because of these altered viewpoints on the goals of mathematics education, a different method of assessment is needed. Aside from the fact that the content of the assessment may no longer be restricted to easily measurable, lower-order skills, the assessment must also produce a different *kind* of information. Results alone are no longer sufficient. The assessment must also provide insight into the students' thought processes behind these results.

Although there would seem to be less of a consensus for assessment than for the direction the new education should take (see AFG, 1991; Lesh and Lamon, 1992a), nevertheless, wherever education is undergoing reform throughout the world, new assessment is being developed as well (Lesh and Lamon, 1992b; Marshall and Thompson, 1994). An important underlying motive is that the assessment method is greatly significant for the success of the reform. How the assessment is conducted is determinative for the kind of instruction that is given, and this is true of *what* is taught as well as the *way* in which it is taught (Romberg, Zarinnia, and Williams, 1989; Madaus et al., 1992).[19] Because traditional tests do not conform with the intent of the new education, and, therefore, send an incorrect message about what is deemed important, mathematics education reform does not stand a chance without assessment reform (Romberg, 1992).[20] Sometimes (as is the case within RME – see Van den Heuvel-Panhuizen and Gravemeijer, 1991a, and De Lange, 1992a), the inverse reasoning is followed: a different method of assessment is viewed as a springboard to different education. An interesting consequence of conformity between the tests used and the education desired is that what was previously regarded as a disadvantage and threatening to innovation, namely, 'teaching to the test', now becomes an advantage and an aid to innovation (Wiggins, 1989a; see also Gronlund, 1991; Resnick and Resnick, 1991). Examples of this are provided by Kulm

(1993) and Stephens, Clarke, and Pavlou (1994).[21] Education reform in Great Britain has even been described as 'assessment-led': many schools have delayed implementing changes until required to do so by the new exams (Brown, 1993; Joffe, 1990; Wiliam, 1993).[22] Schwartz (1992) points out that high-quality assessment items will have the most positive effect on teaching and learning if they are made public after administration rather than, as is so often the case, being kept a secret.

Regardless of the more pragmatic reason of securing or effecting desired changes in education through different assessment, a different method of assessment will inevitably arise from the altered viewpoints on education (see also Clarke, 1986). The method of assessment is, namely, determined to a great extent by the ideas and viewpoints on mathematics and learning mathematics (Galbraith, 1993). This does not alter the fact, however, according to Galbraith, that great discrepancies may still exist between the method of assessment and the kind of education being advocated. Curricula developed from the constructivistic paradigm, for instance, may still be accompanied by assessment programs containing positivistic traces of traditional, mechanistic mathematics education. When it comes to testing, suddenly a context-related interpretation of the problem is no longer permitted, and the students must give a standard answer to a standard test problem. This lack of consistency between education and assessment arises, in Galbraith's opinion, from insufficient attention being paid to the paradigms underlying both education and assessment.

As has been clearly demonstrated in Chapter 1, there has always been a high level of awareness of these underlying paradigms in assessment development within RME. Indeed, it was because the ideas behind the existing tests did not agree with the education being designed that these tests were rejected and alternatives were sought that would better conform to the viewpoints of RME.

3.3.2 Didactical assessment as a general trend

Much of what has been described in the first part of this chapter as characteristic of RME assessment is also characteristic of assessment reform outside The Netherlands. This is particularly true of the didactical focus of the assessment. In this respect, there is no lack of consensus regarding assessment reform.

In the American 'Curriculum and Evaluation Standards', for instance, the didactical function of assessment is strongly emphasized:

> "Assessment has no *raison d'etre* unless it is clear how the assessment can and will be used to improve instruction" (NCTM, 1987, p. 139, draft version; cited by Doig and Masters, 1992).[23]

In the accompanying 'Assessment Standards' (NCTM/ASWG, 1994), many elements of assessment can, indeed, be found that are also advocated by RME.[24] Among other things, the following are mentioned: alignment with the new curriculum, the importance of thinking skills and problem solving in meaningful situations which arises from this alignment, the contribution assessment must make to learning, the opportunities students must be given to show their abilities, fairness of assessment, integration of assessment and instruction[25], informal assessment, the major role played by the teacher in assessment[26], the active role of the student, variety in methods of assessment in order to obtain as complete a picture of the student as possible[27], the instruments that make the student's work visible and that are open to various solution methods.

In the assessment reform movement in Australia, too, a clear shift can be seen in the direction of didactical assessment:

> "This strong sense of assessment informing instructional practice is also evident in the materials arising from the Australian Mathematics Curriculum and Teaching Program" (Clarke, 1992, p. 154).

Here, too, assessment as a formal, objectivity oriented, product-measurement is avoided. Instead, assessment is viewed as an integral component of the teaching and learning process, in which the teacher attempts to acquire as complete a picture of the student as possible through all sorts of informal assessment strategies, such as class observation, questioning, practical assignments, constructing work-folios and having the students keep journals. These activities guide the educational process and provide both the students and the teacher with information on the learning process at hand (Clarke, 1988; Clarke, Clarke, and Lovitt, 1990; Clarke, 1992).

It should be noted that, although the above examples describe how mathematics educators and educational communities currently regard assessment in mathematics education, it does not mean that this has necessarily been put into practice (see, for instance, Joffe, 1990).

The situation in Great Britain is unusual in this respect. There, the recent assessment reform was, in a sense, coerced by a governmental decree requiring a national assessment to be conducted by the teachers. This assessment, which was linked to the National Curriculum, first took place in 1991.[28] The earliest versions of the Standard Assessment Tasks (SATs), which were developed for this purpose, encompassed a broad spectrum of assessment methods, including extensive projects, oral and practical work, and – for primary education – cross-curricular theme-based tasks. It was expected that the assessment activities would be conducted like other classroom activities. The tasks were generally well-received by the teachers, some of whom felt the open-ended tasks had enhanced their teaching, and were greeted enthusiastically by the students. One year later, however, the government turned back the clock. Almost all the SATs for mathematics are now in the form of worksheets consisting of routine pencil-and-paper items, many of which are simply bare arithmetic problems lacking any familiar context. Moreover, there is no longer any assessment of the learning process (Brown, 1993).

The widely recognized trend towards didactical assessment is closely related to the fact that the basis for mathematics education reform consists, to a great extent, of constructivistic ideas on learning. As is the case in RME, this involves regarding the students as active participants in the educational process, who construct their mathematical knowledge themselves. And this, in turn, has significant consequences for assessment. An essential aspect of the socio-constructivistic approach (Cobb, Wood, and Yackel, 1991), in which learning is seen as the social construction of knowledge, is that the teachers are aware of how their students think. Informal methods of assessment, such as observing, questioning and interviewing, are particularly important here. Only through these kinds of procedures can teachers quickly acquire the information needed for decision making. For this reason:

> "Assessment from this [socio-constructivistic] perspective is intrinsic to the very act of teaching" (Yackel, Cobb, and Wood, 1992, p. 70).

A similar attitude towards assessment can be found in the 'Cognitively Guided Instruction' (CGI) project, which is based on constructivistic principles. In this project, the teachers design a curriculum supported by knowledge from the cognitive disciplines and based on their own analyses of what the students know, understand, and are able to do (Carpenter and Fennema, 1988; Loef, Carey, Carpenter, and Fennema, 1988; Chambers, 1993).

The integration of instruction and assessment, aside from being a result of underlying viewpoints on teaching and learning processes, can, according to Galbraith (1993), also contribute to a decrease in potential inconsistencies between education and assessment. In Galbraith's opinion, the important thing is that this integration also be present on the level of educational development. In RME, where educational development has always been regarded in the broadest sense (Freudenthal, 1991), this is certainly the goal. An example of this can be found in the assessment development connected to the 'Mathematics in Context' project (see, for instance, Van den Heuvel-Panhuizen, 1995a, part of which is described in Chapter 7).

3.3.3 The 'didactical contract' versus the 'assessment contract'

In spite of the close relation between instruction and assessment described above, one must not ignore the differences between them. Instructional activities do have a different purpose than assessment activities. Aside from the fact that this is expressed by the kind of problems used – particularly by the amount of information they provide (see Section 3.2.4b) – there are also consequences for the nature of the corresponding interaction between teacher and students. In other words, alongside the interest held by RME for social interaction in educational processes (see Treffers, 1987a)[29] – which is recently again receiving attention (see Gravemeijer, 1994)[30] under the influence of the socio-constructivists Cobb, Yackel, and Wood (1992)[31] – there is now also the awareness that attention must be paid to the interaction between teacher and student in an assessment situation. Elbers (1991b, 1991c, 1992), in particular, has pointed out that, aside from the 'didactical contract' (Brousseau, 1984)[32], which consists of often implicit agreements that form the foundation of class communication, there is also a question of an 'assessment contract'. An important difference between an instruction situation and an assessment situation, according to Elbers, is that the teacher is prepared to offer assistance in the former but not in the latter. In an assessment situation, students must find the answers on their own. The problem here, however, is that:

> "Children are not familiar with the 'contract' of a testing situation, and adults do not realize that children make assumptions about the interaction different from their own" (Elbers, 1991b, p. 146).

As Elbers and Kelderman (1991) have demonstrated by means of a conservation experiment, this ignorance about both the purpose of the assessment and the rules of the corresponding interaction may have considerable repercussions for the assessment results. In the research group in which an assessment contract was drawn up, the percentage of children who gave the correct answer lay much higher than in the control group. It is, therefore, of the utmost importance, in Elbers' opinion, that the children be well informed regarding the purpose and the rules of the assessment situation.[33] De Lange (1987a) had previously argued the necessity of this matter, the importance of which – particularly in the case of a new method of assessment – is now being emphasized by others as well (see MSEB, 1993a; NCTM/ASWG, 1994).[34] Woodward (1993) goes even further. She advocates constructing a 'negotiated evaluation' and is of the opinion that the teachers need to negotiate with students about what they know and understand.[35]

Alongside the more social aspects of how a teacher and student interact and what they may expect from one another in a given situation, Elbers (1992) also mentions the language aspect. Many words acquire a specific connotation in the mathematics lesson, and the students must become familiar with these words through interaction in the classroom. Hughes (1986) illustrates this by a splendid example:

> "I can recall the day when Sally, my 7-year-old stepdaughter, came home from school and showed me her two attempts to answer the subtraction problem 'What is the difference between 11 and 6?' Her first answer had been '11 has two numbers', but this had been marked wrong. Her second attempt was '6 is curly', but this had also been treated as incorrect" (Hughes, 1986, p. 43).

Another example, given by Maher, Davis, and Alston (1992), is about a teacher who introduced fractions by posing the following question:

> "Suppose I call this rod 'one' [the red one, which is 2 cm long]. Which rod should I call 'two'?" (ibid., pp. 259-260).

The teacher was then astonished to find that a third of his class has come up with the 'wrong' answer, namely, the 3 cm green rod instead of the 4 cm purple one. These children had, evidently, not simply assumed the issue to be the length of the rods, but had thought that the question referred to their order.

Socio-constructivists, too, such as Cobb et al., believe that the common knowledge which forms the basis of a mutual understanding between teacher and student is extremely important. In their opinion (see, for instance, Cobb, Yackel, and Wood, 1992), the 'taken-as-shared' interpretations are what constitute a basis for communication. These interpretations comprise, in their eyes, more than just the specific meaning of the words used in the mathematics lesson and involve, instead, an entire entity of expectations and obligations. The students may believe, for instance, that the objective in the mathematics lesson is to quickly give the right answer, and that, if the question is repeated, it means their answer was incorrect (see Yackel, 1992). In order to avoid such misinterpretations, Cobb et al. are endeavoring to develop 'social norms' that conform to the new education through negotiating with the students. The students must become aware of the change in what is now expected of them during mathematics lessons. And this is also true, of course, of assessment. Here, too, the students' expectations may differ from what is intended. Moreover, these norms may even differ from the norms that apply during the lesson situation. Aside from the matter of whether or not to provide assistance, as mentioned by Elbers, it is also extremely important to know what counts as a solution in assessment. For instance, must circumstances from the context in question be included in the solution? As long as no explicit negotiations are held with the students regarding the assessment contract, they will have to keep guessing when it comes to such matters.

3.3.4 The focus on problems

In recent assessment reform in mathematics education, particularly in the United States, most of the attention has been devoted to the formats and modes of assessment and the organizational aspects of assessment, rather than to the actual problems to be used. A variety of assessment tools and methods is advocated, such as portfolio assessment[36], open assessment, oral presentations, practical assessment, project assessment, group assessment, classroom observations, individual interviews[37], and student journals and other forms of student self-assessment.[38] Later, however, with the emergence of performance assessment [39] , interest in the problems to be used in assessment has increased. Another factor that has contributed to this focus on problems is the shift in attention from external accountability to informed instruction (see the quotation from the 'Curriculum and Evaluation Standards' mentioned in Section 3.3.2). Both in the 'Assessment Standards for School Mathematics' (NCTM/ASWG, 1994) and in the report 'Measuring What Counts' (MSEB, 1993a), explicit attention is devoted now to posing good questions and using good assessment problems (see also Romberg and Wilson, 1995). It is urged that assessment problems be invented which encompass all that good instructional problems also contain: they should be motivating, provide opportunities to construct or expand knowledge, and be flexible so that students can both show what they can do and have the opportunity to receive feedback and thereby revise their work. The publication of a booklet containing prototypes for fourth-grade assessment tasks provides a good illustration of this (MSEB, 1993c).

In general, however, there is a difference between the developments in the United States and the approach to assessment in RME. In RME, the problems have always been considered to be the core of the assessment development, and inventing new methods of assessment has always been linked to thinking up suitable assessment problems. Nor has RME remained alone in this respect. Particularly in the case of materials from the British Shell Centre (see, for an overview, Bell, Burkhardt, and Swan, 1992a, 1992b) and from the Australian 'Mathematics Curriculum and Teaching Program' (see Clarke, 1988), the problems to be used in assessment have always been the focus of attention. Developments in the United States, however, do also conform with these, as follows: the activities in assessment development conducted on the basis of the SOLO taxonomy (see Collis, Romberg, and Jurdak, 1986), the California Assessment Program (see Pandey, 1990), and the work of Lesh et al., who have been involved for the past two decades in inventing problems that will lead to meaningful mathematical activities (see Lesh and Lamon, 1992b). The next section contains an overview of these authors' ideas of what good assessment problems should look like.

3.3.5 Characteristics of good assessment problems

3.3.5a Problems should be balanced

On the whole, it can be stated that the tasks must reflect important mathematical content and processes (Magone et al., 1994), and that they must provide a good reflection of the knowledge and skills intended to be covered by the curriculum (see, for instance, Stenmark (ed.), 1991; Bell, Burkhardt, and Swan, 1992a; Collison, 1992; Lesh, Lamon, Behr, and Lester, 1992; MSEB, 1993a; Swan, 1993). Bell et al., and Swan particularly emphasize the need for a 'balanced assessment'. This means that the problems used for assessment must involve both lower-order and higher-order skills, may contain applied as well as pure mathematics, and may range from short, written tasks to extended practical problems.

It is interesting to note that Bell et al. (1992b, 1992a) not only pay attention to the extended problems one associates with assessment reform, but, also, to the improvement of short tasks such as 'fill-in-the-blank' problems. The two catchwords here are 'flexibility' and 'practical relevance'. The former stands for creating the opportunity for more than one correct answer; this point will be discussed in more detail shortly. The latter refers to placing the question in a real, meaningful context (see Section 3.3.7).

3.3.5b Problems should be meaningful and worthwhile

Although, according to Bell, Burkhardt, and Swan (1992a), there are a number of dimensions which can be used in distinguishing assessment tasks[40] (and which can serve in designing and evaluating them), it does not make much sense, in their opinion, to set up classification schemes for types of tasks, as these simply cannot capture the nature and the variety of mathematical performance.[41] Moreover, such classification schemes may prove harmful, as – in the opinion of Bell et al. – the past has revealed. One should, therefore, not be surprised that, on the whole, they advocate:

> "...tasks that lift the spirit, that people will talk about afterwards because they are interesting" (ibid., 1992a, p. 123).

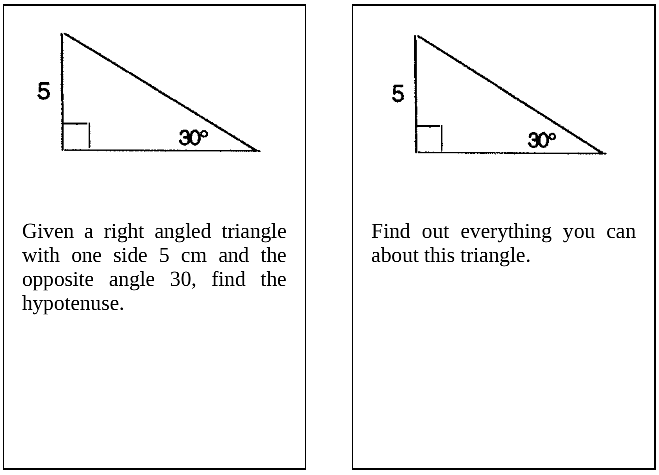
Others, too (see, for instance, Stenmark (ed.), 1991; Collison, 1992; Schwarz, 1992; Wiggins, 1992), emphasize that problems must be mathematically interesting and engaging. Moreover, according to Winograd (1992; see also Silverman, Winograd, and Strohauer, 1992), this is what the students believe as well.[42] Fifth-graders, for instance, found that a 'good' story problem had to be challenging, had to include interesting content from everyday life, and had to contain non-routine characteristics, such as extra information.[43] Nor do the students stand alone here. For Lubinski and Nesbitt Vacc (1994), for instance, a task is worthwhile if it requires problem solving. Then, according to them, comes the question: what makes a problem a problem? For some, a problem is a situation in which no ready-made solution method is available[44], while others find that the student must also have a reason to solve the problem. According to Lubinski and Nesbitt Vacc, real problems are not contrived but arise from actual classroom events. Wiggins (1992), too, greatly values the way in which children view the problems. In his opinion, meaningful tasks need not necessarily be directly relevant or practical, but they must be appealing and challenging to the children. Silverman, Winograd, and Strohauer (1992) mention the degree to which the students can take ownership of the problems. This corresponds to the importance attached in RME circles to students being in control or in charge of a given problem situation (see Section 3.2.4a). Christiansen and Walther (1986) also point out the importance of students' involvement with the problem.[45]

3.3.5c Problems should involve more than one answer and higher-order thinking

For Schwarz (1992), mathematically interesting problems are primarily those in which more than one correct answer is possible. This, too, is widely regarded as a characteristic of a good problem (see, among others, Quellmalz, 1985; Stenmark (ed.), 1991; Collison, 1992; Lamon and Lesh, 1992; Lesh and Lamon, 1992b). According to Lesh and Lamon (1992b), problems having only one level and/or type of correct response are lacking everything that problem solving is about – such as interpreting the problem, and testing and revising hypotheses. Put another way, the problems must have a certain degree of complexity (Magone, et al., 1994) or, in still other words, the problems must be rich (Stenmark (ed.), 1991; Collison, 1992).[46]

Other requirements set by Lesh and Lamon (1992b) for posing problems, particularly in order to assess higher-order thinking, are as follows: encourage making sense of the problems based on students' personal knowledge and experience, and allow students to construct their own response. In addition, they offer a number of suggestions for stimulating students to construct significant mathematical models: have the students make predictions, describe patterns, and provide explanations.

Clarke (1988) and Sullivan and Clarke (1987, 1991) have also placed great emphasis, as did Lesh et al., on posing good questions. Besides providing many examples, they have also formulated a number of requirements that these examples must fulfill. On the whole, these correspond to those requirements mentioned earlier. In their opinion, questions should possess three features in order to reveal the level of student understanding: (i) they should require more than recall of a fact or reproduction of a skill, (ii) they should have an educative component, which means that both students and teacher will learn from attempting to answer them, and (iii) they should be open to some extent, which means that several answers may be possible. They suggest, with respect to this last characteristic, for instance, that a 'Find-out-everything-you-can' question be created in place of the conventional type of problem, where one missing bit of information must be derived (see Figure 3.6).



Given a right angled triangle with one side 5 cm and the opposite angle 30, find the hypotenuse.

Find out everything you can about this triangle.

Conventional task

Open-ended task

Figure 3.6: Conventional and open-ended tasks (from Clarke, 1988, p. 31)

In addition to these three chief characteristics, Sullivan and Clarke have also set a number of supplementary requirements, some of which have been described above: a question must be clear; a question must be posed in such a way that everyone can at least make a start in formulating a response; it must be clear if insufficient knowledge is involved; a question must, in any case, be answerable by students possessing sufficient knowledge; and a question must provide opportunities for delving deeper. The following section describes the tensions that can arise between these requirements and the preference for open-ended problems.[47] More attention is devoted to this matter in Chapter 7.

3.3.5d Concerns about open-ended problems

In spite of his advocacy for open-ended questions, Clarke (1993b) later began to wonder whether open-ended problems were, in fact, able to accurately portray students' learning and indicate directions for future actions. His concerns are based on his research experience, which revealed that students are rather leery of giving more than one answer, even when this is encouraged by education. Only by explicitly asking for more answers did a significant rise occur in the number of answers given. Aside from the fact that this demonstrates the absolute need for clarity in what exactly the students are asked, it is also apparent here that the problem itself must be included in the discussion. The examples Clarke provides of problems used in his research are such, in fact, that one would not immediately expect the students to give more than one answer. A question such as

> "A number is rounded off to 5.8, what might the number be?" (Clarke, 1993b, p. 17).

is not inherently a problem which does require more than one answer. If that had been the goal, then the question could better have been formulated as follows:

> Many children in our class got 5.8 in answer to a problem. But that was after they had rounded off their first number. What might their 'first numbers' have been?

Other concerns Clarke expresses with regard to open-ended tasks involve the following: the time-consuming aspect (whereby only a small part of what has been learned can be tested), the complexity of the information, the legitimacy of grading, the reliability of the scoring, the pressure open-ended tasks may put on students[48], the level of literacy demanded (see also, in this respect, Clarke, 1993a)[49], and the complicating factor of context.[50] Although RME also regards most of these issues as significant points for attention in the assessment of open-ended problems, there is a clear difference of opinion with regard to the final point. While Clarke regards context as a potential complicating factor, in RME context is seen as a facilitating component. Here, again, it is extremely important which contexts are chosen. Moreover – to return to the cause of Clarke's objections to open-ended problems – one may wonder whether a student may be 'required' to provide more than one answer to a question. Whether this is 'obligatory' depends entirely on what one wishes to assess.

3.3.5e Problems should elicit the knowledge to be assessed

Clarity in terms of assessment objectives is also what Webb and Briars (1990) stress so strongly in their requirements for good assessment tasks. In their opinion, the tasks must, in any case, involve what one wishes to assess, provide information on the extent to which students possess this knowledge, and express as much as possible about the extent to which students have integrated this knowledge and can apply it to new situations.

These kinds of requirements can also be found in Magone et al. (1994). According to these authors, the problems must assess the specific skill or content area and higher-level cognitive process for which they were designed.[51]

3.3.5f Problems should reveal something of the process

It is also important, according to Magone et al. (1994), that tasks elicit certain strategies and that they expose the solution techniques. For Lamon and Lesh (1992), too, it is important that good problems reveal something of the underlying process; in their opinion, there is a general consensus on this matter. They do caution, however, that one cannot simply tack on an extra sentence such as

> "Explain how you got your answer" (Lamon and Lesh, 1992, p. 321).

This same exhortation can be found by Stenmark (ed.), 1991. If no thinking is necessary for the original question, because nothing needs to be found out, linked up, or interpreted, then one should not – according to Lamon and Lesh – expect such a request to provide one with insight into a student's thinking. That Foxman and Mitchell (1983) did have a positive experience with a simple request for clarification probably has to do with the original question. Moreover, the students were clearly told beforehand, in Foxman and Mitchell's research, that the request for explanation did not mean that the answer was wrong. In other words, the 'assessment contract' (see Section 3.3.3) was explicit.

One should also keep in mind when posing questions that – as put by Yackel, Cobb, and Wood (1992) – 'genuine requests for clarification' are submitted (see also Bishop and Goffree, 1986). The requests must be 'information-seeking questions' rather than 'known-information questions'. In other words, pseudo-questions, where one asks for the sake of asking should be avoided in assessment situations. These, according to Elbers (1991c), are questions that are often posed in instructional situations in order to get the educational communication off the ground. One may also wonder, therefore, about the suitability of the 'counter-suggestions technique', which is often used after a child has given the correct answer in order to check the stability of the understanding (see Ginsburg et al., 1992; Ginsburg, Jacobs, and Lopez, 1993). These are, in fact, potential traps[52], which were objected to by Freudenthal (see Section 1.2.4e), and which may threaten the atmosphere of trust that is necessary if one is to encourage students to express their thinking (Yackel, Cobb, and Wood, 1992).

3.3.5g More concerns about open-ended problems

Back to Lamon and Lesh (1992). They also have concerns, as does Clarke, about open-ended problems. Although assessment reform – in terms of the assessment problems – is principally sought in open-ended problems, Lamon and Lesh believe these should not be regarded as a panacea. It is interesting to note that, in their view, the search for reform in this area has to do with the fact that the focus of the current reform is on the format, rather than on the substance, of assessment. As an illustration of the limitations this presents, they give an example of a typical textbook problem that was 'improved' by making it more open:

> Typical textbook problem: "The ratio of boys to girls in a class is 3 to 8. How many girls are in the class if there are 9 boys?"
>
> Improved problem: "For every 3 boys in a class, there are 8 girls. How many students are in class?" (ibid., p. 321).

The teachers that were involved in the project agreed that the second question would be better because more than one correct answer was possible. The results were very disappointing, however. The teachers failed to acquire the information on the students they had been expecting, and, therefore, obtained insufficient footholds for further instruction. The students who had given '11' as the answer certainly knew that more answers were possible, didn't they? Or was this the result of 'mindlessly' adding up the two given numbers? In order to obtain more certainty, it was decided to narrow the question. One of the teachers suggested:

> "...Could there be 25 people in the class? Why or why not?" (ibid., p. 322).[53]

Lamon and Lesh also offer another example as an illustration of why making a question open-ended does not always improve matters (see Figure 3.7). The multiple-choice problem here involved numerical relations, but, by turning it into an open-ended question (fill-in-the-blank), it was trivialized, and became merely a task requiring the application of a procedure in which only one correct answer is possible.

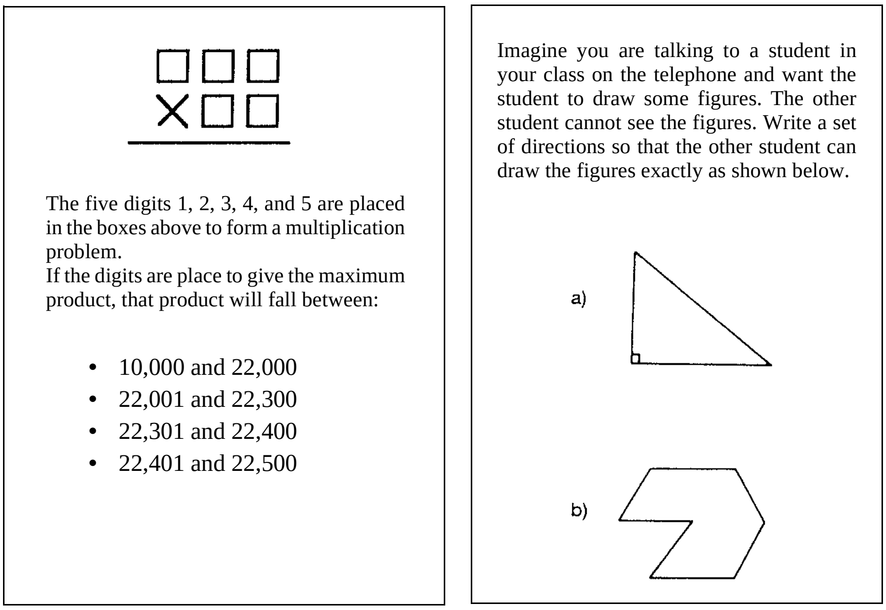
5/15 = 3/9. Suppose I want these fractions to remain equal. If I change the number 15 to 24, does anything else have to change? (a) The 3 or the 9. (b) The 3 and the 5. (c) The 9 or the 5. (d) The 5 and the 9. (e) None of the other numbers.	5/15 = 3/9 and ?/24 = 3/9
The multiple-choice problem	The 'improved' version

Figure 3.7: Closed problem, made open-ended (from Lamon and Lesh, 1992, p. 323)

3.3.5h Good problems can have different appearances

The conclusion drawn by Lamon and Lesh (1992) is that multiple-choice questions need not necessarily be bad and that, in order to judge the 'goodness' of a problem, one must match the question format with one's assessment purpose (see also Section 4.1.3b; and Van den Heuvel-Panhuizen, 1994a/b, in which a similar conclusion is drawn). Also of extreme importance is the distinction made by Lamon and Lesh between question formats that *elicit* higher-order thinking and question formats that *expose* higher-order thinking. In their opinion, multiple-choice questions can sometimes elicit[54] but rarely expose.[55]

The viewpoint – that the format does not in itself determine the quality of assessment problems and that multiple-choice problems may also require higher-order thinking – can be found in the 'power items' developed for the California Assessment Program (see Pandey, 1990). In this project, an endeavor was made to develop items in which understanding and insight were assessed in connection with mathematical ideas and topics. Moreover, the items were intended to serve as examples of good teaching practice. Characteristic of these 'power items', however, is that their format may be either multiple-choice or open-ended. An example of each is illustrated in Figure 3.8. Both items can be solved in different ways, depending upon the student's mathematical sophistication. Only in the open-ended item on geometrical figures, however, is this level revealed, thanks to the reactions on the worksheet.



The five digits 1, 2, 3, 4, and 5 are placed in the boxes above to form a multiplication problem.
If the digits are place to give the maximum product, that product will fall between:

- 10,000 and 22,000
- 22,001 and 22,300
- 22,301 and 22,400
- 22,401 and 22,500

Multiple-choice power item

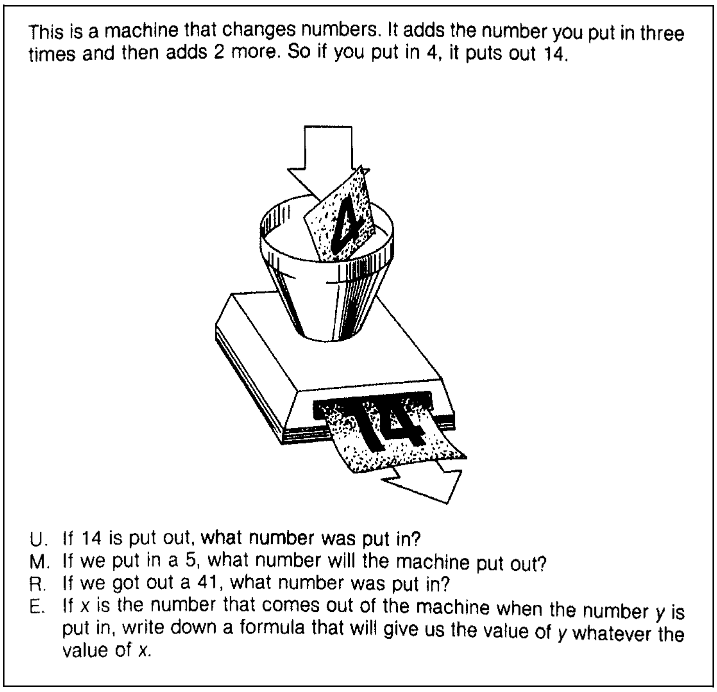
Imagine you are talking to a student in your class on the telephone and want the student to draw some figures. The other student cannot see the figures. Write a set of directions so that the other student can draw the figures exactly as shown below.

Open-ended power item

Figure 3.8: Examples of power items (from Pandey, 1990, pp. 47-48)

One student, for instance, will use mathematical terms and concepts to describe the figures, while another will use more informal language.[56]

In addition to this open approach, in which the student's level is exposed through his or her own work, a more structured approach was also developed to acquiring this information. This involved the ‘superitems’, which were based on the SOLO[57] taxonomy (see Collis, Romberg, and Jurdak, 1986). A superitem consists of introductory information (the ‘stem’) and a series of questions, each of which indicates a certain fixed level of reasoning (see Figure 3.9). The levels were identified by interviewing students. They refer to how a student deals with the information provided in the stem, and each successive correct response requires a more sophisticated use of the given information than its predecessor.[58] This is a rather general level classification, which would not seem of particular assistance to teachers looking for footholds for further instruction. Although the correct answers do provide an indication of a certain level of reasoning, these are still all-or-nothing questions. The answers can only be right or wrong, and different solution levels – if there are any – cannot be seen, resulting in an absence of footholds for further instruction.[59]

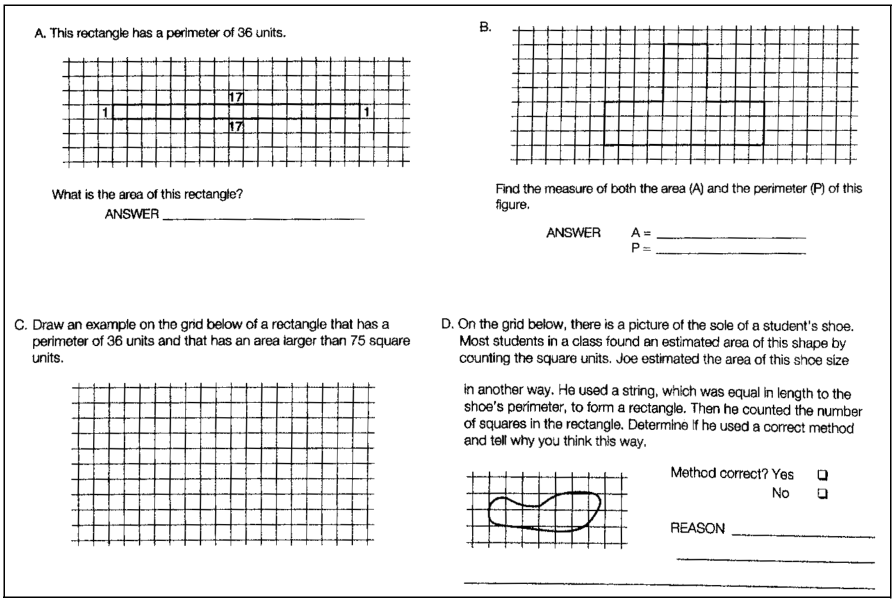


This is a machine that changes numbers. It adds the number you put in three times and then adds 2 more. So if you put in 4, it puts out 14.

U. If 14 is put out, what number was put in?
M. If we put in a 5, what number will the machine put out?
R. If we got out a 41, what number was put in?
E. If x is the number that comes out of the machine when the number y is put in, write down a formula that will give us the value of y whatever the value of x.

Figure 3.9: Example of a superitem: Machine problem (from Collis et al., 1986, p. 212)

In this respect, a more recent example of a superitem (described by Wilson and Chavarria, 1993) offers more possibilities (see Figure 3.10). This superitem involves measurement and is much more informative than the previous example. This is principally due to the final question. In contrast to the final question in the Machine problem, this question is not attached to a certain fixed level, but can be answered on different levels, as is the case in the open-ended power items.



A. This rectangle has a perimeter of 36 units.

What is the area of this rectangle?

ANSWER ____________________

B.

Find the measure of both the area (A) and the perimeter (P) of this figure.

ANSWER A = ______________
P = ______________

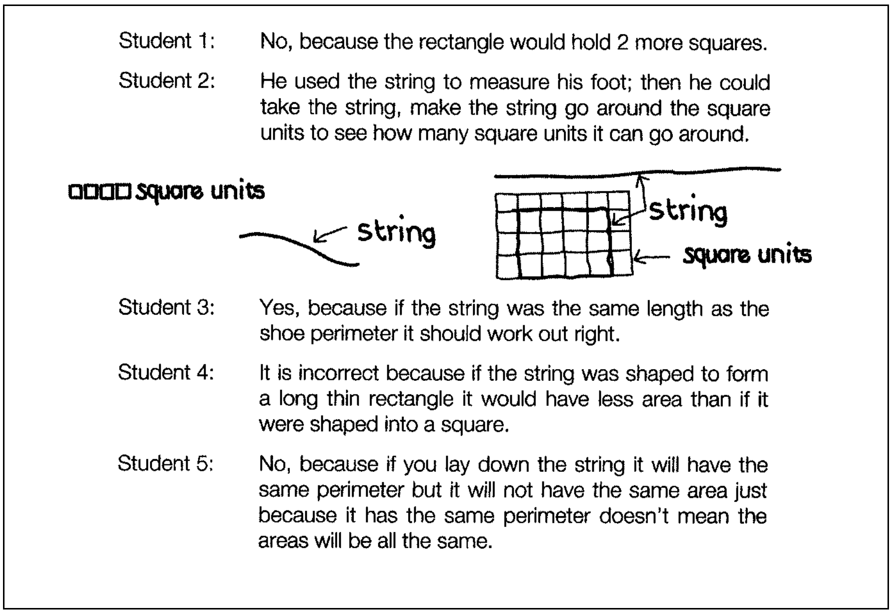
C. Draw an example on the grid below of a rectangle that has a perimeter of 36 units and that has an area larger than 75 square units.

D. On the grid below, there is a picture of the sole of a student's shoe. Most students in a class found an estimated area of this shape by counting the square units. Joe estimated the area of this shoe size in another way. He used a string, which was equal in length to the shoe's perimeter, to form a rectangle. Then he counted the number of squares in the rectangle. Determine if he used a correct method and tell why you think this way.

Method correct? Yes □
No □

REASON ______________
____
____

Figure 3.10: Example of a superitem: Area/Perimeter problem
(from Wilson and Chavarria, 1993, pp. 139-140)



Student 1: No, because the rectangle would hold 2 more squares.

Student 2: He used the string to measure his foot; then he could take the string, make the string go around the square units to see how many square units it can go around.

□□□□ square units
string
string
square units

Student 3: Yes, because if the string was the same length as the shoe perimeter it should work out right.

Student 4: It is incorrect because if the string was shaped to form a long thin rectangle it would have less area than if it were shaped into a square.

Student 5: No, because if you lay down the string it will have the same perimeter but it will not have the same area just because it has the same perimeter doesn't mean the areas will be all the same.

Figure 3.11: Answers to question D from the Area/Perimeter problem
(from Wilson and Chavarria, 1993, p. 141)

The assorted answers to the Area/Perimeter problem (see Figure 3.11) clearly show the extent of the students' insight into the relation between area and perimeter.[60] As opposed to the all-or-nothing answers to the final question in the Machine problem (which indicate whether or not the student is able to transpose the operation conducted by the machine into a formula), these answers provide many more footholds for further instruction.

3.3.5i 'The task' does not exist

In all these attempts to track down good assessment problems, one must keep well in mind – as is emphasized by socio-constructivists such as Cobb, Yackel, and Wood (1992) – that there is no such thing as 'the task' (see Gravemeijer, 1994). Or, as Schoenfeld puts it:

> "the key issue is how students will interpret what you show them rather than how much of it they will absorb. ... your job begins rather than ends with a clear presentation of the material" (Schoenfeld, 1987, pp. 20-30; cited by Galbraith, 1993, p. 79).[61]

The interpretation of the task – whether or not arrived at through explicit classroom interaction – determines, in fact, how the task is construed and what is perceived as a solution.[62] Nevertheless, each student can still interpret a given task in an individual manner, depending upon his or her background, and this interpretation may diverge from the intentions held by the administrator or designer of the tasks (see Elbers, 1991b; and see Grossen, 1988, cited by Elbers, 1991b; see also the examples given in Sections 1.2.4e and 3.3.3). As a result, it cannot always be determined beforehand which skills the students will demonstrate (Joffe, 1992). This is why Elbers (1991a) advocates always ascertaining whether the students have interpreted the task in the correct – or, better stated, intended – manner. Moreover, according to Cooper (1992) one must be prepared, in terms of the interpretation of the tasks, for differences between students with respect to their various social backgrounds. A research project conducted by Holland (1981, cited by Cooper, 1992; see also Section 3.3.7h) revealed that an assignment to classify illustrations of food in a certain way was interpreted entirely differently by eight and ten-year-old middle-class children than by their working-class counterparts. The latter group appeared to stick more to the 'everyday' frame of reference, while the middle-class children seemed readier to spot and/or to accept the rules of the testing exercise, and to switch to the required approach. Linn, Baker, and Dunbar (1991) and Lesh, Lamon, Behr, and Lester (1992) also caution that the introduction of new types of problems is not sufficient in itself for enabling a wider variety of students to show what they can do. If no attention is concurrently paid to the fairness of the problems for different groups of students, then, in their opinion, the more privileged students will benefit more from these improved problems than will the less privileged.

3.3.6 A different interpretation of the traditional psychometric quality requirements

This attention to the fairness of the problems reveals that, alongside the different viewpoints on assessment, there is also a difference of opinion about the traditional psychometric quality controls of reliability and validity.[63] This is clearly the case in RME. As has been described in Chapter 1 (see Section 1.2.4b), objections were raised from the very outset of RME to the one-sided manner in which these quality controls were defined (see, for instance, Freudenthal, 1978a). Another sign of this critical attitude can be found in the *fourth principle* formulated by De Lange (1987a), namely, that the quality of a test is not defined by its accessibility to objective scoring and that the focus should first be on the content of the test.

In recent years, a fairly general consensus has also arisen among mathematics educators outside the RME circle on the concept that the new understanding about the nature of mathematics, mathematics instruction, and mathematics learning also involves new ways of thinking about traditional assessment issues, such as reliability and validity (Lesh, Lamon, Behr, and Lester, 1992; Webb, 1992; see also Wiggins in Brandt, 1992). Tests that may once have been valid and reliable may become invalid and counterproductive when goals and conditions shift (Lesh, Lamon, Behr, and Lester, 1992).

3.3.6a Reliability

In recent years, educators have become increasingly troubled by the traditional interpretations of reliability when it comes to open-ended, meaningful and integrative assessment tasks. This is a concern that was expressed by Freudenthal (1976a, 1976b) nearly two decades ago. The corresponding narrow focus on technical criteria[64] is now even regarded as a counterproductive factor for good assessment (MSEB, 1993a). An important reason behind this discomfort with the existing notion of reliability is that the responses acquired with the new methods of assessment are typically quite complex (Lesh, Lamon, Behr, and Lester, 1992). To use the words of Bell, Burkhardt, and Swan (1992c), these responses are too complex to fit the simplistic definition of reliability, which, according to them, is an artifact that is statistically sophisticated but educationally naive. Moreover, the principle that assessment must contribute to learning is not compatible with the requirement that students must find the same results in a repeated testing (Lesh, Lamon, Behr, and Lester, 1992).

Another characteristic of this shift in attitude with respect to reliability is that the once prevailing notion of *objectivity* is being increasingly dispensed with – complex answers cannot, after all, always be scored unequivocally – and is being replaced by the much broader criterion of fairness (see, for instance, MSEB, 1993a; Linn, Baker, and Dunbar, 1991; Bell, Burkhardt, and Swan, 1992c). This criterion involves taking into account and doing justice to differences between the students, and providing them with opportunities to show what they can do.[65] In contrast to the past, when teachers were expected to disregard any knowledge they held of the students while correcting their work, they are now increasingly urged to use this knowledge (Moss, 1994). As a result, the notion of *standardization* – which is now regarded as a kind of mistrust of the teachers' ability to make fair, adequate judgments of their students' performance (MSEB, 1993a) – is up for review as well. In addition to allowing the students more room in which to interpret and solve the problems in their own way (Moss, 1994), it may be deemed necessary, for fairness' sake, to provide certain students with more assistance. This is something that, until recently, was inconceivable (MSEB, 1993a).

In a nutshell, this increasingly popular different interpretation of reliability (which, however, does not imply that matters such as generalizability and comparability are no longer considered important[66]), has much in common with the viewpoints on this subject in RME. Take, for instance, the shift in emphasis to fairness and the strive for 'positive testing' (see Section 3.2.4b), and the fact that offering assistance is regarded as an enrichment of assessment (see Section 1.2.3e) – all of which are present in RME.

3.3.6b Validity

Another striking point of similarity involves the relation between reliability and validity. Others, too, outside the RME circle, now increasingly regard validity as more important than reliability (Linn, Baker, and Dunbar, 1991; Bell, Burkhardt, and Swan, 1992c; MSEB, 1993a; Moss, 1994). The former situation, when – in the words of Freudenthal (1976a) – responsibility was shifted from validity to reliability, is clearly past. [67]

Moreover, nearly everyone agrees that the notion of validity needs to be expanded (Linn, Baker, and Dunbar, 1991; Lesh and Lamon, 1992a; Baker, O'Neil, and Linn, 1993; MSEB, 1993a; Wiliam, 1993). According to Linn, Baker, and Dunbar (1991), validity is usually viewed too narrowly and too little attention is paid in technical presentations to evidence supporting the assessment procedure. Even though the content frameworks and correlations with other tests and teacher assessments, that are customarily presented, do contain valuable information for judging validity, they do not, in the opinion of these authors, do justice to the notion of validity.

In the minds of others, the expansion of the concept of validity primarily involves broadening the domain of important mathematics. This means, for example, that skills like communication and reasoning, which were previously classified as non-mathematical skills, must now be integrated into mathematics assessment (MSEB, 1993a). For this reason, Bell, Burkhardt, and Swan (1992c) give great weight to the face validity of the set of tasks as a way of providing a balanced reflection of what the students should be able to do.[68] They also believe that pilot testing, including a detailed study of student responses, can offer important information to this end.

Linn, Baker, and Dunbar (1991) are in complete agreement with this last point. In their opinion, evidence must demonstrate the technical adequacy of the new assessment.[69] They do wonder what sort of evidence is needed and by what criteria the assessment should be judged. The difficulty in answering this question, however, is that few proponents of assessment reform have addressed the question of criteria for evaluating the measures. It is often simply assumed that assessment is more valid as a matter of course with direct measurement, as occurs in performance assessment (see Section 3.3.4), than on a multiple-choice test. Even though Linn, Baker, and Dunbar find that direct assessment of performance does appear to have the potential to enhance the validity, they also believe that much more research must be conducted into the criteria used to evaluate these new forms of assessment. Considering the comment in his final book (see Section 3.1.1), this is certainly a standpoint that Freudenthal would have endorsed.

With an eye on gaining a broader view of validity, Linn, Baker, and Dunbar (1991) formulated eight criteria, some of which clearly contain elements of reliability as well. If assessment is to be valid, then, in their opinion, there must be evidence of: (i) the intended and unintended consequences for teaching practice and student learning, (ii) fairness, (iii) transfer and generalizability[70], (iv) cognitive complexity, (v) content quality, (vi) content coverage, (vii) meaningfulness, (viii) cost and efficiency. The list is not intended to be exhaustive and, moreover, certain criteria may be more or less important, depending upon the goal of the assessment.

This broader validity also lies in the procedures used to acquire the desired evidence. According to Linn, Baker, and Dunbar, these must not only be statistical procedures, but qualitative ones as well. With regard to cognitive complexity[71], for instance, they suggest that analyses of open-ended responses that go beyond judgments of overall quality can be quite informative. And, with respect to content quality, they advise that systematic judgements are needed from subject matter experts.

These suggestions were taken into account in a research project recently conducted by Magone et al. (1994), which investigated the validity of assessment problems that had been developed for the QUASAR project.[72] For the validation of the cognitive complexity and content quality of the assessment tasks, a number of procedures were employed, including the following: logical analyses of the task content and expected performance, internal and external expert review, and qualitative analysis of actual student responses that had been collected in pilot testing. There was a continual interplay among these procedures as the tasks were developed. The validity evidence was used for making decisions on whether tasks should be discarded, revised, or included as is in the assessment.

A similar approach, consisting of a mathematical-didactical analysis, discussions with colleagues, pilot testing, and an analysis of student response was taken for the test on percentage that was developed in the framework of the ‘Mathematics in

Context' project (see Van den Heuvel-Panhuizen, 1995a). Furthermore, the method proposed by Collis (1992) is also closely aligned with this approach. He suggests a careful analysis of both the activity to be assessed and the item devised for this assessment, supplemented by detailed interviews to determine how students approach the task.

3.3.7 The role of contexts

3.3.7 a Authentic assessment

The call for 'authentic assessment'[73], whereby the students can demonstrate in an authentic setting that they are capable of something and have understood it, is one of the most characteristic aspects of current assessment reform in the United States. The term 'authentic' involves more, however, than simply using real-life problems. Wiggins (1989b), for instance, has composed a list of authenticity characteristics. In addition to characteristics dealing with the nature of the problems in question (such as the fact that authentic tests are contextualized and that they may sometimes contain ambiguous and ill-structured problems[74]), a large number of these characteristics have to do with the conditions in which the assessment takes place. These latter characteristics include the following: authentic tests require some collaboration with others, no unrealistic and arbitrary time limitations may be set, the assessment may not be restricted to recall and one-shot responses, there must be a multi-faceted scoring system, and all students must have the opportunity to show what they can do.[75] For Stenmark (ed.) (1991) and Collison (1992), on the other hand, 'authentic' means that the problems require processes appropriate to the discipline and that the students value the outcome of a task.

Baker, O'Neil, and Linn (1993), too, indicate that the term 'context' has different, and somewhat conflicting, meanings. In their view, some use the term to indicate that the assessment involves domain-specific knowledge, while others mean that the problems must be authentic for the student and that the problems must involve a situation taken from everyday life or must involve the assessment of skills in application situations.

3.3.7 b The context of the problem

Although, in RME, too, the notion of 'context' can be interpreted in different ways, and contexts may assume different guises (see Section 3.2.5), the term refers primarily to the situation – in a broad sense – in which the problems are placed. The description of a context given by Borasi (1986) comes close to this interpretation:

> "We define 'context' as a situation in which the problem is embedded. The main role of the context seems to be that of providing the problem solver with the information that may enable the solution of the problem. [...] the context may be fully given in the text of the problem itself. However, it is important to realize that this is not the case most of the time. [There may be] considerable differences in terms of the amount and

> quality of the information given. [...] The absence of the detailed formulation present in the text could have even opened the problem solver to a wider range of possible solutions" (Borasi, 1986, pp. 129-130).

For Borasi, too, contexts need not necessarily refer to reality. Alongside real-life problems such as

> "The Nelsons wish to carpet a small room of an irregular shape. In doing this they want to estimate the amount of carpet they will need to purchase" (ibid., p. 137).

she also distinguishes, among other things, exercises, word problems, puzzle problems and situations.[76] The emphasis is laid more on the type of problem than on the context. With respect to the context, a distinction is only made as to whether this is entirely, not at all, or partially contained in the text.

Swan's (1993) classification according to (i) pure mathematical tasks, (ii) applications and (iii) real-life tasks, on the other hand, focuses more on the nature of the context. His classification greatly resembles that of De Lange (see Section 3.2.5). In the pure mathematical tasks, the focus is on exploring the structure of mathematics itself. In the applications, the emphasis is still on a mathematical idea, but this is embodied in a realistic or pseudo-realistic application. In the real-life tasks, the point is clearly to gain new insight into the real world. These tasks require the integration of mathematical and non-mathematical skills, may involve contexts containing too much or too little information, and produce solutions that are often of practical value.

According to Swan (1993; see also Bell, Burkhardt, and Swan, 1992a), however, the problem is that the context problems (which have recently begun to increasingly appear on exams) are often merely 'dressed up' bare problems. The reality they encompass is largely cosmetic. An obvious example of this is a problem he presents involving an ironing board, which, when examined closely, is simply a formal exercise in trigonometry. Why in the world would anyone want to know the angle between the legs of an ironing board? Swan does point out that, with a slight alteration – namely, by asking where one should place the stops on the legs in order to obtain a good height for ironing – the problem can be made realistic.

3.3.7c The importance of contexts and their influence

One of the most important cornerstones for improving the quality of assessment is constructed, according to Bell, Burkhardt, and Swan (1992b), by focusing on authentic mathematical knowledge, realistic problem-solving situations, and diverse mathematical abilities that are productive in realistic situations. They caution against using decontextualized 'vanilla' problems that fail to provide the students with the opportunity of making sense of the situations through using their own, everyday knowledge and experience. Numerous research projects have confirmed how important this 'making sense' of the situation is, and how contexts can contribute to providing insight into what children can do.

Pioneering work in this area was conducted by Donaldson, who, in collaboration with McGarrigle, demonstrated how context-determined the results were of Piaget's conservation experiments (see McGarrigle and Donaldson, 1974; Donaldson, 1978). Through a slight alteration of the context – whereby a mischievous teddybear shifted the blocks instead of the researcher[77] – the number of conservers suddenly increased in comparison to the original version of the experiment.[78]

A similar confirmation of the influence of context is provided by Carraher et al. (Carraher, Carraher, and Schliemann, 1985; Carraher, 1988). These authors showed that many children who were unable to perform certain calculations at school had no trouble doing so when selling candy at the market. Research by Foxman et al. (1985, cited by Joffe, 1990), too, showed that some students, who did poorly in certain addition problems, were even able to multiply and divide when asked to organize a party. Joffe (1990) points out as well that, when the same calculations are presented in different situations, these are evidently not interpreted as being similar tasks. Carpenter and Moser (1984) further discovered that young children could solve word problems through informal strategies before they had learned to do bare arithmetic problems (see also Section 5.2.1). The same results were found in research conducted by Hughes (1986), in which a group of 60 children, varying in age from three to five, was presented with a number of addition problems – some in a game situation and others in a formal presentation. As can be seen in Table 3.1, the differences in achievement were enormous, depending on the conditions.

Table 3.1: Results of the research by Hughes (1986, p. 31)

% correct	Box closed	Hypothetical box	Hypothetical shop	Formal code
Small numbers (< 5)	83	56	62	15
Large numbers (Š≥ 5)	28	20	23	6

When an actual box was used, where blocks were put in and taken out, or when this box or a shopping situation was referred to, the achievements were much higher than in the formal situation, where no reference was made to reality.

3.3.7d Context problems versus bare problems on written tests

The discovery made by Clements (1980) with regard to different results in assessment problems was also quite revealing. The unusual aspect here is that the two test items in question were part of the same written test. One item consisted of a bare arithmetic problem and the other of a context problem, but both held the same mathematical content (see Figure 3.12).

Question 5: Write in the answer $1 - \frac{1}{4}$ = ________ (Answer).

Question 18: A cake is cut into four equal parts and Bill takes one of the parts. What *fraction* of the cake is *left*?

Figure 3.12: Two problems from the test by Clements (1980, p. 19)

The difference in results between these two items presents the same picture as the difference in results between certain problems from the MORE tests (see Section 2.3.4). Of the 126 sixth-grade students, 45% answered the bare arithmetic problem correctly, while 78% found the correct answer to the context problem.[79] The interviews conducted afterwards with the students revealed that they had had support from the imagery in the cake problem.[80] Evidently, it did not matter that the context problem involved more reading, comprehension and transformation.[81]

3.3.7e Influence of the language used

Nevertheless, the amount of text is often regarded as an impeding factor. This is, indeed, supported by research results such as, for example, that of Mousley (1990), which involved the participation of 186 fifth-grade and sixth-grade students. Mousley used a test here consisting of 20 open-ended problems, each of which also existed in a modified version. Aside from the fact that the achievements correlated with the amount of text, they were also related to the sentence structure and to whether certain difficult terms were used.[82] With respect to difficult terms, the research of Foxman (1987, cited by Joffe, 1990) also demonstrated the relevance on a written test of replacing a definition, such as ‘line of symmetry’ with a more familiar word like ‘reflection’. In the latter case, the success rate rose by 20%.

3.3.7f Limitations of the context

Using contexts familiar to the students does not always provide support, however. Van der Veer and Valsiner (1991, cited by Elbers, 1992) have shown that certain limitations may also ensue from calculating within a given context. When an anthropologist attempted to ascertain how far a certain people were able to count and suggested that they count pigs, they counted no further than 60, as so many pigs was inconceivable to them. Another example, taken from the research of Wood (1988, cited by Elbers, 1992), involves a problem in which candy had to be distributed in such a way that one student would get four more candies than the other. The children to whom this problem was presented refused to do it, however, on the grounds that it wasn’t fair. Gravemeijer (1994), describes a similar experience involving a problem in which 18 bottles of coca-cola must be shared fairly by 24 students at a school party. These students, too, refused to interpret the problem as it was intended because, they said, some students didn't like coca-cola and, moreover, not everyone drank the same amount.

3.3.7g Ignoring the context

The opposite can occur as well, namely, that the students ignore the context entirely. Balloons to be shared fairly are then neatly cut in half (Davis, 1989; Elsholz and Elsholz, 1989). Greer (1993), for instance, discovered great differences between the achievements on straightforward items, where no limitations are presented in the situation referred to by the problem, and more complex items where this does occur. The fewest errors were made on items of the first type. The complex problems, by contrast, were often answered using stereotypical procedures that assumed a 'clean' modeling of the situation in question. The students' expectations as to how problems should be solved resulted in a lack of openness to potential limitations. They, therefore, simply took the problem out of its context and solved it as a bare arithmetic problem.[83] In order to discourage this, according to Greer, students must be confronted with various types of problems and must learn to regard each problem according to its own merits.

Verschaffel, De Corte, and Lasure (1994), who repeated Greer's research along broad lines, also had to acknowledge a strong tendency by the students to exclude real-world knowledge and realistic considerations when solving word-problems in the classroom.[84]

3.3.7h Sometimes the context 'must' be ignored

The paradox in this situation of 'dressed up' problems (as referred to by Swan (1993), is that they often do have to be 'undressed'. Cooper (1992), using a number of test problems that were part of the 'Summer 1992 National Pilot' for testing at 'Key Stage 3' of the National Curriculum for England and Wales, shows that the students who were able to marginalize their knowledge of reality stood the best chance of answering the problems correctly.[85] One of the examples he gives involves an elevator that has to take 269 people to the upper floors during the morning rush hour. A sign states that no more than 14 people may take the elevator at one time. The question put to the students is how many times the elevator must go up and down. The accompanying 'marking scheme' indicates that the only answer that may be considered correct is 269 divided by 14 is 20. Neither 19 nor 19.2 is acceptable. In order to come up with this answer, the students must forget reality and assume that the elevator is always full, that nobody decides to take the stairs, and that, for instance, no one in a wheelchair needs to use the elevator. The difficulty is that there is nothing in the question to indicate that the students may not use their knowledge of reality. The only way the students can be aware of this is from their own knowledge of testing. According to Cooper, this knowledge, and the ability to interpret these ‘real-life’ problems as they are apparently intended, is not distributed evenly among the students of varying social classes. Sociological studies have revealed that children from working-class families have a greater tendency to stick to the reality indicated by the problem (see Section 3.3.5i). As a consequence, in Cooper’s opinion, teachers should caution their students to beware of the practical, real and everyday world. Moreover, teachers must make it clear to the students that, while the real world must sometimes be used in solving problems, at other times it must consciously be ignored. What is remarkable about Cooper’s suggestion, is that he does not simultaneously address the issue of the grading chart. After all, if other answers to the problem were also considered correct, then the students would have the opportunity to solve the problem from their own interpretive perspective (see also Section 4.1.5).

3.3.7i Many issues still remain to be resolved

This does not mean, however, that the issue of context problems has been resolved. There are other problems, too, concerning the use of contexts. For instance, according to the report entitled ‘Measuring What Counts’ (MSEB, 1993b), there is the danger that the knowledge of situations appealed to in a problem is not distributed evenly among students from different backgrounds. In addition to the class differences, there is also the difference between boys and girls. Research conducted by Graf and Riddell (1972, cited by M.R. Meyer, 1992), for instance, revealed that female college students had more difficulty answering problems involving contexts familiar to males than problems involving contexts familiar to females. Furthermore, they needed more time to solve the problems involving male contexts.

Lesh and Lamon (1992b), too, believe that a great many issues still remain to be resolved about the nature of ‘real’ mathematics, realistic problems, realistic solutions and realistic solution processes. Although they do still regard ‘reality’ as a praiseworthy goal, it is no longer their primary criterion for distinguishing ‘good’ problems from ‘bad’ ones. Instead, they now believe that:

> “The real purpose of emphasizing realistic problems [is] to encourage students to construct and investigate powerful and useful mathematical ideas (that is, models and principles associated with models) based on extensions, adaptations, or refinements of their own personal knowledge and experience” (ibid., p. 39).

They also emphasize that the implementation of ‘realism’ is not so simple and straightforward. After all, that which is real to the person who invented the problem is not necessarily real to the person for whom the problem is intended. There is also the issue of trendiness and the above-mentioned class issues. Lesh and Lamon also point out that the students, when viewing videotapes in class, adopt the same, divorced from reality, attitude as when watching films and television. Another important point is that, in their opinion, much manipulative material is just as abstract to students as written symbols. By contrast, certain computer-based explorations are very real to them, even though these have nothing to do with their daily life.

3.3.8 RME assessment and the recent reform movement in the area of assessment

If the present situation in RME assessment is compared with recent international developments in the same area, it is clear that the international tide is now moving with RME. In contrast to the early days of RME, when it was necessary to go against the current of then prevailing viewpoints on assessment, new methods of assessment are now being internationally advocated that agree in many respects with the assessment propagated by RME. In connection with these worldwide changes in goals and approaches to mathematics education, assessment can no longer be restricted to lower-order goals and must reveal more than results alone. Just as has always been the case in RME, the emphasis internationally has now also shifted in a certain sense to 'didactical assessment'. This is assessment that lies adjacent to instruction and that is intended to directly support the educational process. The paradox inherent in this current similarity, however, is that it is also a source of potential differences. Wherever there is a question of different teaching methods, didactical assessment will also be interpreted differently. An example of this has already been given in Collis' reaction to a particular geometry problem (see Section 3.1.2). In his view, the problem will only be suitable if the students are addressed on one particular level and if they, as a result, also answer on one level. As another example, whether or not the teaching methods in question link up with students' informal knowledge, or use certain models, will also certainly affect assessment. And then there are the specific interpretations of the subject content, which, too, can greatly influence the appearance of this didactical assessment. If, for example, elementary geometry is primarily seen as the ability to identify geometrical figures and the knowledge of their properties, then this will obviously lead to different assessment problems than if elementary geometry at a certain level is mainly regarded as spatial orientation. Most of the differences that arise from the varying points of departure for subject content will only come to the surface, however, if more content oriented studies are conducted in the vein of: 'how do you assess percentage?'[86] and 'how do you know what students know about early algebra?'[87]

Whereas, in RME, improvement of assessment has principally focused on improving the assessment problems themselves, reform elsewhere, particularly in the United States, has concentrated on developing new formats and means of organization. The attention paid to portfolio assessment, for instance, is illustrative of this focus. There is also, however, a significant stream of international publications in which the emphasis, as with RME, is on the problems themselves. The requirements set for assessment problems in these publications show a great similarity to those set by RME. The problems must be meaningful to the students, there must be opportunities for student ownership, the problems must cover the entire range of goals, and they must afford insight into how the students think and how they solve the problems. But – and this caveat has mainly surfaced in constructivist circles – these requirements are not enough. It is important that the teachers be aware of how the students view the problems, in order to be able to interpret their answers. Another point of concern is the development of new quality controls for assessment. Within RME, too, this is in need of further elaboration, particularly the aspects of task specificity, selection of problems within a domain[88], and fairness of the assessment. This last point, in particular, places a heavy burden on the role of the contexts, which are considered so important by everyone.

Although much of what has been said in the international literature about the role of contexts in assessment does dovetail with the RME ideas on this topic, and is therefore extremely relevant to RME, it is clear, nonetheless, that differences do exist as well. One of these, for instance, lies in what is understood to be a context. This is why the role of contexts, as interpreted by RME, is not put on entirely the same par as the recent strive for authentic assessment. Taking everything into consideration, at least three different perspectives can be distinguished within the authentic assessment movement:

(i) In the circles of cognitive psychology, emphasis is laid on the alignment between knowledge and context (see, for instance, Brown, Collins, and Duguid, 1989). The activity and the context in which learning take place are regarded as fundamentally inseparable from what is learned.[89]

(ii) From society comes the need for a mathematically literate workforce (see, for instance, NCTM, 1989). The emphasis is, therefore, on the utilitarian aspect, which, in turn, goes hand in hand with a strong emphasis on the realistic component.

(iii) Among mathematics educators, it is stressed that mathematics is not about being able to perform separate and distinct skills, but about being able to adequately use mathematical tools and apply insights in complex problem situations (see, for instance, Clarke, 1993b).

In most instances, these perspectives are not easily distinguishable from one another. Something of all three perspectives can be found in every endeavor to achieve authentic assessment, including RME. In RME, where the final perspective is principally apparent, contexts are not exclusively used as a goal, but also as a source for developing certain tools and insights.[90] It is this use that is specific to RME. Therefore, alongside the realistic component, RME contains other important criteria, such as the extent to which students can imagine themselves in a context and the mathematical potential of a context as an action indicator and model supplier. Moreover, because mathematization is viewed vertically as well as horizontally, the nature of the contexts can also be mathematical.

Notes

[1] For brevity's sake, the word 'knowledge' is used here rather than mentioning all the modalities of learning results and learning levels. Unless otherwise stated, 'knowledge' is continually used in this broad sense.

[2] This idea is also presented in the report entitled 'Measuring What Counts', where the viewpoint is that of today's vision of mathematics instruction.

[3] This problem was taken from the California Assessment Program (see Pandey, 1990). It is discussed further in Section 3.3.5h.

[4] An earlier version of this had been previously published (see Treffers and Goffree, 1985).

[5] It should be noted here that Webb (1992) does find that practical assessment problems needing answers now cannot wait for a theory to emerge and must be solved at once.

[6] Here, too, one finds a situation analogous to Treffers' retrospective description of the RME theory, where a confrontation with the 'outside' was also described (see Treffers, 1987a).

[7] De Lange borrowed this principle from Gronlund (1968). According to Gronlund, the learning aspect of assessment lies not only in the fact that students can obtain information from the tests regarding the educational goals and can thereby receive feedback on their own development, but also in the guiding function that tests can perform with respect to the higher-order goals. It is striking, however, that Gronlund perceives the learning effect of this primarily in connection with the greater degree of retention and the greater degree of transfer that are inherent in higher goals. He, therefore, in fact remains within the then prevailing viewpoints on assessment. De Lange, on the other hand, attaches a new interpretation, namely, that an assessment situation is not only a situation in which students' skills are measured, but that it is also a learning situation in which new knowledge is constructed. This interpretation of the learning aspect of assessment can also be found, for instance, in Elbers (1991a, 1991b), Grossen (1988), and Perret-Clermont and Schubauer-Leoni (1981), both cited by Elbers (1991b). It is also a central topic in the report entitled 'Measuring What Counts' (MSEB, 1993a/b) (see Section 3.3.2, Note 24). Bell, Burkhardt, and Swan (1992a), in contrast, approach the learning aspect of assessment from a more pragmatic perspective. Because the new methods of assessment take up so much classroom time, in their opinion it is extremely important that said assessment provide valuable learning experiences for the students.

[8] De Lange borrowed this second principle, which he calls 'positive testing', from the report entitled 'Mathematics Counts' (Cockcroft, 1982).

[9] The word 'problems' includes all types of questions, tasks and assignments that can be presented to students in order to gather information on their learning processes.

[10] The problem-centered nature of RME diverges considerably from the distinction according to type of problem, which is characteristic of the mechanistic approach, where the students must handle each type successively. Lenné (1969, cited by Christiansen and Walther, 1986) labeled this approach 'Aufgabendidaktik'. The CGI approach (see Sections 3.3.2 and 5.2.1), too, distinguishes problems according to type, although the order in which they are presented is not determined in this case by a textbook. An objection to such arrangements according to type of problem is that they are based on bare problems and on only one (formal) feature of these problems. This feature may be the size of the numbers ('two-digit addition' or 'three-digit addition') or the structure of the problem ('join – change unknown' or 'join – result unknown'). The result is that there is usually nothing left to mathematize, as each type of problem has its own individual solution strategy. Other aspects of the problem (if there are any), such as the underlying context, are ignored, even though this may confuse the hierarchy with respect to the degree of difficulty (see Sections 2.3.4 and 3.3.7c).

[11] Smaling (1990) speaks of the positive and the negative aspect of 'doing justice'. The former entails giving the research object its due and the latter involves not allowing it to

distort. De Lange's *second* and *fourth principles* mainly emphasize the positive aspect of 'doing justice'. This does not mean that avoiding distortion is therefore neglected. When the open test format was introduced, research was simultaneously conducted into the extent to which this might lead to differences in evaluation. The results of this research were positive, however (see De Lange, 1987a).

12 This problem was included on test TG3.2, which was developed in the framework of the MORE research project (see Chapter 2).

13 This problem was part of a test (see Van den Heuvel-Panhuizen and Gravemeijer, 1991b) that was developed for an intended sequel to the MORE research. This sequel never took place due to the absence of financial support. The test in question was administered, however, and to the same students who had participated in the MORE research. The analysis of the test results was conducted for the most part by Aad Monquil.

14 The letters do not refer to the same students whose work was illustrated in Figure 3.3.

15 The students were not permitted to use a calculator during the test.

16 See, among others, the following reports: 'A Nation at Risk' (NCEE, 1983) and the 'Curriculum and Evaluation Standards for School Mathematics' (NCTM, 1989), published in the United States; 'Mathematics Counts' (Cockcroft, 1982), later followed by 'Mathematics in the National Curriculum' (DES/WO, 1989), published in Great Britain; 'Mathematics Study Design (VCAB, 1990) and 'A National Statement on Mathematics for Australian Schools' (AEC, 1991), published in Australia.

17 The limited scope of other subjects has been acknowledged as well; history, for instance, is more than the knowledge of dates, and literature is more than the names of famous authors (Schwarz, 1992).

18 The concept of 'problem solving', now present on every list of goals for mathematics education, first gained attention at the 1980 ICME conference. It was listed under the heading 'unusual aspects of the curriculum' (Schoenfeld, cited by Bell, Burkhardt, and Swan, 1992c).

19 An interesting paradox occurs here: on the one hand, the objection is raised that traditional ability and achievement tests do not provide information for instruction, while, on the other hand, there is the concern that these tests do, in fact, influence instruction, but in a negative way (Campione, 1989).

20 The success of the current reform movement in mathematics education is thus attributed to the widespread consensus on the viewpoint that the method of assessment must change concurrently with the education. During previous reform movements, scarcely any attention was paid to the consequences for assessment (Lesh and Lamon, 1992a; MSEB, 1993a).

21 The influence on policy and instruction at school and classroom level by the introduction of new assessment practices into existing high stakes assessment is described by Stephens, Clarke, and Pavlou (1994) as the 'ripple effect'.

22 Bell, Burkhardt, and Swan (1992c) disagree that this has taken place. In their opinion, mathematics education reform in Great Britain has occurred in the same way as in the United States. In the USA, however, the influence of assessment was not seen until a later stage.

23 This viewpoint is also clearly evident in the report entitled 'For Good Measure' (MSEB, 1991).

24 The same is true of the report entitled 'Measuring What Counts'(MSEB, 1993a), which provided the foundation, in a certain sense, for the Assessment Standards. This report emphasizes that assessment must involve mathematical content that is relevant to the students (the 'content principle'), that it must improve learning and must contribute to good education (the 'learning principle'), and that it must offer all students the opportunity to show what they can do (the 'equity principle').

25 See also Webb (1992). Even the renowned American assessment institute, the 'Educational Testing Service' (ETS), which, for decades, has been the foremost producer of stan-

dardized tests, is now publishing educational programs that, rather than indicating scores, promote a new method of instruction. In these programs, the educational activities and the tests are inextricably intertwined (Grouws and Meier, 1992). A similar development is also taking place in The Netherlands, where the National Institute for Educational Measurement (CITO) has recently become involved in the development of educational activities (see Kraemer et al., 1995).

26 Rather than making the assessment 'teacher-proof', the teachers become connoisseurs of student learning (AFG, 1991). See also NCTM, 1991; Graue and Smith, 1992; Cain, Kenny, and Schloemer, 1994.

27 Due to the dynamic nature of understanding, student understanding cannot be captured by one single method of assessment (Ginsburg et al, 1992; Marshall and Thompson, 1994).

28 Moreover, new methods of assessment were already being designed in Great Britain from 1977 through 1988, namely, by the 'Assessment of Performance Unit'. In addition to written components, the APU tests consisted of practical oral tests, some of which were administered individually and some to small groups. With regard to the content, these tests not only assessed insights and skills, but paid attention to strategies and attitudes as well. Moreover, some of the problems also dealt with everyday contexts (Foxman, 1993; see also Ernest, 1989).

29 Treffers (1987a) mentions Bauersfeld's interactional theory as an example of a theory which is significant for developing and realizing mathematics education on the level of textbooks and teaching methods.

30 Moreover, it has also become an object of research in The Netherlands (see Wijffels, 1993; Elbers, Derks, and Streefland, 1995).

31 Cobb, Yackel, and Wood, in their turn, were influenced by Bauersfeld.

32 Elbers does not, in fact, mention Brousseau, who was the first to speak of a didactical contract in the framework of mathematics education. Instead, Elbers refers to Rommetveit, who speaks of contracts as premises for communication in general. Brousseau (1984, p. 112), in contrast, views the didactical contract as "...that part of the contract which is specific for the content, i.e., the mathematical knowledge. Therefore [,] we do not deal here with all aspects of the reciprocal obligations."

33 One must take into account here the fact that Elbers is speaking from a rather strict and research-linked perspective with regards to assessment, in which the test administrator may offer no assistance. The question is whether the assessment contract he has in mind would be valid for RME assessment, where offering assistance can sometimes form an explicit component of assessment; see the Kwantiwijzer instruments (see Section 1.3.1), the two-stage test (see Sections 1.4.1 and 4.1.2a), and the standby sheet (see Section 4.1.4c).

34 In addition to the purpose and the rules of the assessment situation, such an assessment contract should also indicate the 'zone of free construction', so that the students are not only informed of what they *must* do, but also of what they *may* do.

35 Woodward would also like to involve the parents as, in her opinion, parents possess extremely valuable information pertaining to the development of their child. It should be mentioned here, however, that Woodward was principally speaking of instruction in reading and writing.

36 In portfolio assessment, a collection is made of each student's work, in order to acquire an impression of his or her development (see, for instance, Wolf, 1989; Mumme, 1990; Herman and Winters, 1994).

37 See, for instance, Long and Ben-Hur (1991).

38 These various tools and methods overlap one another a great deal.

39 Performance assessment concentrates on organizing the assessment in such a way that the students can demonstrate, as far as possible, the actual behavior to be assessed. This contrasts with situations involving, for instance, multiple-choice questions, in which *indicators* of this behavior are used (see, for instance, Wiggins, 1989; Linn, Baker, and Dunbar,

1991; Collison, 1992; Baker, O'Neil, and Linn, 1993). Performance assessment has its roots in athletics and the arts (Resnick and Resnick, 1992), and is sometimes called 'authentic assessment' (see Section 3.3.7a). A confusing aspect of the label 'performance assessment' is that it is also used as a collective term for the in the chapter mentioned formats and modes of assessment. Portfolio assessment, for example, is then described as a special form of performance assessment (see, for instance, Resnick and Resnick, 1992; O'Neil, 1992).

40 Bell, Burkhardt, and Swan (1992a) mention the following: length of the task, autonomy (degree to which the student may present his or her own solution), unfamiliarity, focus on application, context, mathematical content. There also exist more formal classifications according to type of questions, an example of which is given by Nesbitt Vacc (1993). She distinguishes between: factual questions, reasoning questions (closed reasoning, recalled sequences; closed reasoning, not recalled sequences; open reasoning, more than one acceptable answer), and open questions (questions that do not require reasoning, but that present an opportunity for students to describe observed phenomena for which they have not yet learned a name, and that can be used in introducing new concepts). Other classifications according to type of problems are those of Borasi (1986) and Polya (1981, cited by Borasi, 1986). These classifications, in contrast to those of Bell et al., do not explicitly refer to assessment tasks.

41 Christiansen and Walther (1986) also point out the subjective and relative nature of problems: what may present a problem during one stage of development may become a routine task at a later stage.

42 These authors, too, are not explicitly discussing assessment problems.

43 In a developmental research on assessment, conducted during the 'Mathematics in Context' project, the same student preferences with respect to assessment problems were found (see Van den Heuvel-Panhuizen, 1995a). In the same research, the students were also involved in the development of assessment problems through their own productions (see also Van den Heuvel-Panhuizen, Streefland, and Middleton, 1994; Van den Heuvel-Panhuizen, Middleton, and Streefland, 1995).

44 Problems where such a solution method is available are called 'exercises' (see Christiansen and Walther, 1986; Kantowski, 1981, cited by Borasi, 1986).

45 It should be noted here again that both Lubinski and Nesbitt Vacc, and Christiansen and Walther are not specifically referring to the characteristics of assessment problems, but, rather to the suitability of problems for education in general.

46 Carter, Beranek, and Newman (1990) call these 'problems with possibility of development'. Although they, too, do not particularly focus on assessment, nevertheless, their list of characteristics of good mathematics problems displays considerable similarity to the characteristics that are mentioned with respect to assessment problems. Other characteristics mentioned by them are: stimulating mathematical curiosity, demanding inventions, multiplicity of access and representation, possibility to empower the problem solver, and possibility for independent exploration.

47 Instead of 'open-ended problems', one might better call them 'constructed-response problems'. This term shifts the emphasis from a task characteristic to a response feature (Romberg, 1995).

48 The findings of Telese (1993) contradict this. His research reveals, in fact, that alternative assessment techniques create a non-threatening atmosphere which may encourage all students to participate and to use higher-order thinking skills in mathematical discourse.

49 As demonstrated by Mousley (1990), the language aspect does influence the results (see also Section 3.3.7e), but this is unrelated to the open-ended nature of a problem.

50 More on this matter in Section 3.3.7.

51 These kinds of requirements do, in fact, involve the validity of the assessment problems. Attention is devoted to this issue in Section 3.3.6b.

52 The same can be said of problems that put the students on the wrong track in a different way, such as the Jam Jar problem (see Section 3.3b and 4.1c in Van den Heuvel-Panhui-

zen, 1995a). In this example, the visual presentation can form a distraction, so a kind of 'alarm question' was used to lessen the trap-like nature.

53 It is not clear whether this question was a substitute for the 'how-many' part of the improved problem, or whether it was only used as a follow-up question during a class discussion after the improved problem was administered.

54 See also Mehrens (1992).

55 Szetela and Nicol (1992), who believe that problems have to stimulate both thinking and written communication, have developed a number of interesting formats for this, such as that of confronting a student with another student's strategy.

56 This, then, is the reason why this open-ended item is rejected by Collis (1992) but not in RME (see Section 3.1.2).

57 SOLO stands for Structure of Observed Learning Outcome (see also Section 4.2.1e).

58 In order to answer the first question, the student need only use part of the information provided by the stem ('U' stands for 'Unistructural'); for the second question, two or more components of the information must be used ('Multistructural'); for the third question, the student must integrate information that is indirectly related to the stem ('Relational'); and, for the fourth question, the student must be able to define an abstract general principle that is derived from the stem ('Extended abstract').

59 One should keep in mind, however, that these superitems were not developed to provide footholds for instruction but, rather, as a means of having an instrument with concurrent validity with interview data.

60 There is still room for discussion, however, on which of the five students have indeed obtained this insight. According to Wilson and Chavarria (1993), this is only true of students (4) and (5). There is something to be said for including student (2) as well, however. If the form of the rectangle is chosen in such a way that it looks something like the shoe, then one can arrive at a fairly good estimation of the area using this procedure.

61 One of Stenmark's (1989, p. 32) objections to the existing written tests (multiple-choice, single-word answers, etc.) is that "students are forced to match the test-makers' formulation of an answer with the test-makers' formulation of the problem."

62 A good example of this is given by Booth (cited by Bishop and Goffree, 1986, p. 313), when referring to an observation described by Brown and Küchemann. The problem involves a gardener who has 391 daffodils. These are to be planted in 23 flowerbeds, and each flowerbed is to have the same number of daffodils. The question is how many daffodils will be planted in each flowerbed. According to Booth, "[t]o the experimenter (and the mathematics teacher) the question is really: 'here is a problem which can be modelled by an algorithm; which algorithm is it?' To the child, however, it is a specific problem concerning the planting of daffodils, and the 'trick' in the question is probably to decide what to do with the ones that remain[.]"

63 Although the new requirements for reliability and validity may not be left undiscussed in this chapter, the subject is so complex that an extensive discussion would be beyond its scope. On the other hand, the topic does not lend itself well to a brief synopsis. Nevertheless, in spite of the inherent defects of such a synopsis, that is the choice that has been made here.

64 It is interesting to note that the absence of expressly articulated educational principles is regarded as the cause of the fact that these technical criteria have become *de facto* ruling principles (MSEB, 1993a).

65 This corresponds to the 'two-sidedness of objectivity' distinguished by Smaling (1990) (see Section 3.2.4b, Note 11).

66 Matters such as generalizability and comparability are still deemed important (see, for instance, Linn, Baker, and Dunbar, 1991; Baker, O'Neil, and Linn, 1993). But the new forms of assessment sometimes require new approaches, such as, for example, the methods described by the latter authors for assuring comparability in the scoring of tasks (see also Section 4.1.5c). Experiences gathered in the HEWET project made it clear, more-

over, that inter-subjective scoring and proper scoring instructions provide enough guarantees for fair measurement: fair both to the student and to the curriculum (De Lange, 1987a).

67 Linn, Baker, and Dunbar (1991, p. 16) use approximately the same words as Freudenthal to indicate how things used to be. They state that: “Reliability has often been over-emphasized at the expense of validity.”

68 Wiliam (1993, p. 7) goes so far as to eventually regard validity as something subjective and personal: “A test is valid to the extent that you are happy for a teacher to teach towards the test.”

69 This, in fact, is reason enough for these authors to state that face validity is not enough.

70 The greatest problem here is task specificity. The variance component for the sampling of tasks tends to be greater than for the raters (Linn, Baker, and Dunbar, 1991; Baker, O’Neil, and Linn, 1993).

71 Cognitive complexity means that the tasks require the use of higher-order skills.

72 This project was designed to improve mathematics education for students attending middle schools in economically disadvantaged communities. An article by Magone et al. (1994) describes the development and validation of a number of assessment problems that were used in this project.

73 Although the terms ‘performance assessment’ and ‘authentic assessment’ are often used interchangeably and in relation to one another (see Section 3.3.4, Note 39), there is a difference, according to C.A. Meyer (1992). Performance assessment refers to the kind of student response to be examined, while authentic assessment refers to the context in which that response is performed.

74 Elsewhere, Wiggins (1992, p. 28) adds that “a context is realistic to the extent that we so accept the premises, constraints, and ‘feel’ of the challenge, that our desire to master it makes us lose sight of the extrinsic factors and motives at stake – namely, that someone is evaluating us.”

75 There are also different reasons for using authentic tasks in assessment. Wiliam (1993) mentions the following ones: to better represent the mathematical performance, to better predict further success, to focus on an important part of mathematics, to encourage teachers to incorporate such activities into their teaching.

76 These ‘situations’ are a kind of ‘no-question’ problem, consisting entirely of information which the students can use to construct their own problems.

77 This experiment was designed and conducted by McGarrigle.

78 In the teddy bear version, 63% of the children found the correct answer, while, in the original version, only 16% had done so. Less spectacular differences were found in a later repetition of the experiment by Dockrell, but these differences were still too large to be attributed to coincidence (Donaldson, 1978).

79 In a later administration of the same test, in which the order or the two items was reversed, the difference was even greater: in this case the percentages were, respectively, 52% and 91%.

80 The main difference between the two problems is that the first item involves a bare fraction and the second item a so-called ‘named fraction’; that is, the question is posed: ‘what fraction of the cake?’ The numbers used in the different conditions in Hughes’ research (see Section 3.3.7c) differ in a similar fashion. ‘Named numbers’ are used in the game and shopping situations, whereas the formal problems contain ‘unnamed numbers’.

81 The prevailing viewpoint at that time was that a verbal arithmetical problem is necessarily more difficult than the corresponding arithmetical problem involving the direct application of the relevant process skills. Clements used this discovery to point out that this is not always true.

82 While these research results do show the importance of the wording of assessment problems, they should not be used as a confirmation of the assumption that working in contexts is more difficult than doing bare arithmetic problems (see Section 3.3.5d). A distinction

must be made between the context itself and the way it is transmitted.

83 This is also illustrated by the research of Säljö (1991), which showed that students solved a problem differently during and outside the mathematics lesson. The problem in question involved determining the postage for a letter whose weight differed slightly from those shown on the postage rate chart. During the mathematics lesson, calculations were made to find out the 'real' price of this letter, while, outside the lesson, the rate was simply found by looking at the next higher weight on the rate chart.

84 It is worth noting, however, that in both these research projects, two problems stood out in a positive sense: the Balloon problem and the Bus problem. As Verschaffel, De Corte, and Lasure remarked, these were the only problems in which the result of the calculation had to be connected to the context of the answer, so that the students would immediately notice an unrealistic answer (for instance, 13.5 buses or 3.5 balloons). Nevertheless, the authors sought the explanation for the low level of realistic considerations in the stereotypical and straightforward nature of the word problems used in the instruction and the way these problems were instructed, rather than in the actual problems used in the research.

85 A well-known historical example of this is Terman's revision of Binet's test problems, in which he narrowed the problems by only permitting certain answers (Gould, 1981).

86 An attempt to answer this question can be found in Van den Heuvel-Panhuizen, 1995a.

87 The work presently being conducted by Van Reeuwijk endeavors to answer this question.

88 Noteworthy, on this point, is the discussion that has taken place in The Netherlands about the evaluation of primary school education (see Wijnstra, 1995, and Van der Linden and Zwarts, 1995).

89 The importance attached in cognitive psychology to the use of contexts is evident from the fact that the context sensitivity of assessment tasks is considered to be the heritage of cognitive psychology (Baker, O'Neil, and Linn, 1993).

90 Here, RME diverges from Clarke's (1993b, p. 24) view that "...it may not be realistic to expect students of any age to access recently-acquired skills in open-ended or in problem-solving situations." According to Clarke, the use of such tasks for assessment should be undertaken with caution. Experience within RME, however, has shown that students also reveal things in assessment that have not yet been dealt with in class (see Sections 2.3.4, 4.1.3e and 4.1.3f).

4 Written assessment within RME – spotlighting short-task problems

4.1 New opportunities for written assessment

In the previous chapter, an overall picture was sketched of assessment in RME, while spotlighting its didactical nature and the crucial role played by the problems. In the present chapter, attention is shifted to a particular assessment method, namely, that of written assessment. By this is meant assessment in which the tasks are presented on a test page and must be answered by the students in writing, in what ever way.

At first glance, this assessment would not seem such an obvious choice, considering the resistance to written assessment that prevailed during the early days of RME. As described in Chapter 1, more informal assessment methods such as observing and interviewing, were initially preferred. However, in Chapter 1 it was also pointed out how the subsequent need for exams suited to the new secondary education curriculum did lead to a search for new methods of written assessment. This search was later continued in primary education by the MORE project, as was described in Chapter 2. In contrast to the alternative written tests developed for secondary education, which mainly contained extensive tasks, the tests produced for the MORE research involved short tasks. The first section of the present chapter concentrates on the developments surrounding these tests, in which an attempt was made to make short paper-and-pencil tasks more informative. This is then followed by a more reflective section on the implications of the principles of RME for written assessment and vice versa.

4.1.1 Objections to the traditional written tests

The international appeal for new kinds of assessment, as described in the previous chapter, was mainly a reaction to the existing written tests.

> "Today, there are strong pressures to move away from the traditional multiple-choice or short-answer tests, toward alternative forms of assessment that focus on real-life situations, authentic mathematics and performance activities" (Lesh and Lamon, 1992a, p. 3).

4.1.1a A mismatch between what *should* be tested and what *is* tested

Many of the objections raised to existing written tests during the early stages of RME – described in detail in Chapter 1 – are now again being heard.

The widespread complaint heard nowadays concerns the mismatch between what the existing tests measure and the altered goals of and approach to mathematics education (Romberg, Zarinnia, and Collis, 1990; Resnick and Resnick, 1992; Romberg and Wilson, 1992). According to Joffe (1992), the weakest point often lies in the test content. Most written tests concentrate solely on simple skills while ignoring higher-order thinking (Grouws and Meier, 1992), and such tests cannot provide complete information about students' structures of knowledge (Webb, 1992). According to Resnick and Resnick (1992) the existing tests are characterized by an assessment of isolated and context-independent skills. In their opinion, this is due to two assumptions: that of 'decomposability' and that of 'decontextualization'; both of these assumptions, however, are more suited to a 'routinized curriculum' than to the 'thinking curriculum' currently advocated. Resnick and Resnick thus emphatically object to tests that consist of a large number of short problems demanding quick, unreflective answers. In other words, the general criticism of the existing tests is that they do not measure the depth of students' thinking (Stenmark, 1989) or – as stated in the 'Assessment Standards' (NCTM/ASWG, 1994) – that the traditional written work does not provide the students any opportunity to show their underlying thought processes and the way in which they make links between mathematical concepts and skills. This is especially the case with the younger children. In these same 'Assessment Standards', it is stated that observing young students while they work can reveal qualities of thinking not tapped by written or oral activities. According to Cross and Hynes (1994), disabled students make up another group that is particularly penalized by traditional paper-and-pencil tests[1].

More specific criticism of multiple-choice tests – which is the most restricted form of assessment (Joffe, 1990) – is that the students are not asked to construct a solution and that such tests only reflect recognition (Feinberg, 1990, cited by Doig and Masters, 1992; Mehrens, 1992; Schwarz, 1992). Consequently, these tests do not measure the same cognitive skills as in a free-response form (Frederiksen, 1984, cited by Romberg, Zarinnia, and Collis, 1990). Moreover, according to Schwarz (1992), multiple-choice tests transmit the message that all problems can be reduced to a selection among four alternatives, that mathematics problems necessarily have answers, that the answers can be stated briefly, and that correct answers are unique.

4.1.1b No information on strategies

In both multiple-choice and short-answer tests, the process aspect remains out of sight (MSEB, 1993a). Consequently, the tests cannot provide much information on the various strategies that students employ when solving problems (Ginsburg et al., 1992). The tests do not reveal, for instance, how an incorrect answer came about. This lacuna, of course, causes certain repercussions. It is no wonder that the 'Assessment Standards' (NCTM/ASWG, 1994) caution that evidence acquired exclusively from short-answer and multiple-choice problems may lead to inappropriate inferences. For this reason, Joffe (1990, p. 158) wonders:

> "...what kind of teaching would be guided by the results of tests which assess only the things accessible by timed multiple-choice tests."

The strongest criticism of written assessment is voiced by the socio-constructivists. Yackel, Cobb, and Wood (1992) stress that children's progress cannot be sensibly judged by looking at their answers to a page of mathematics problems, but requires, instead, repeated face to face interactions. Mousley, Clements, and Ellerton (1992; see also Clements, 1980) also find that the diagnosis of individual understanding necessitates informal and frequent dialogue between teachers and children, whereby the teacher can offer feedback and assistance. In this respect, these authors agree with Joffe (1990), who views this kind of 'all-or-nothing' situation as one of the shortcomings of traditional assessment: the students either can or cannot answer a question. By contrast, in the interactive mode – such as in an interview situation – students can be questioned about their methods and solutions.

4.1.1c Nevertheless, written assessment does have a future

In spite of the widespread criticism of written assessment, this method has not been ruled out – either in RME or elsewhere. Joffe (1990, p. 147), for example, begins her outline of how assessment methods might be improved with a section-heading that reads:

> "Getting more from existing written tests."

In her opinion, a more creative exploitation of the potential of such tests is necessary. The example she offers is taken from the APU Mathematics Survey by Foxman et al. (1985) and involves choosing a series of problems and answers on decimal numbers in such a way that the students' reactions will expose any misconceptions. Ginsburg et al. (1992, p. 286) also see opportunities for written assessment:

> "The point of the story is that even a dumb test can be exploited as a useful assessment technique, provided children are encouraged to reflect on and reveal the solution processes employed."[2]

Lesh, Lamon, Behr, and Lester (1992) point out that the shortcomings of the existing tests are more than merely superficial, and that they cannot be improved through simplistic techniques, such as converting multiple-choice questions to their 'fill-in-the-blank' counterparts.[3] According to these authors, there are three reasons to search for alternative assessment methods:

> "(i) to emphasize broader and more realistic conceptions about mathematics, mathematical problem solving, and mathematical abilities, [...]
> (ii) to identify talented students and to give special attention to targeted groups of minority students and women whose abilities have not been recognized [...] by traditional [...] tests, and
> (iii) to help to organize all students' opportunities for success by facilitating informed decision making, [...] to simultaneously develop and document their increasing knowledge and capacity" (ibid., p. 398).

One could also interpret these reasons as being footholds in a search for alternative assessment methods.

4.1.2 RME alternatives to traditional paper-and-pencil tests

This section offers a survey of the alternatives to traditional paper-and-pencil tests that were invented within RME.[4] Although the emphasis is chiefly on primary education, a synopsis of the alternatives developed for secondary education will be presented first, as the picture of RME assessment would otherwise not be complete.

4.1.2a Developmental research on assessment for secondary education

The above-mentioned points enumerated by Lesh, Lamon, Behr, and Lester (1992) for improving existing assessment are nearly the same as De Lange's (1987a) five principles for developing RME assessment problems, which he formulated during the HEWET project (see Sections 1.2.5c and 3.2.2). The learning principle, the principle of 'positive testing'[5], and the principle that assessment must fit the new goals of mathematics education can all be found in Lesh et al. Although expressed more implicitly, this is also true of the principle of fairness and the principle of practicality. One point mentioned by Lesh et al. does not appear in De Lange's principles, namely, the facilitation of informed (instructional) decision making. This is understandable, however, as De Lange's principles were primarily formulated with an eye to developing alternatives to the existing final exams.

In addition to these principles – which may be regarded both as result and guideline – this RME developmental research on assessment for secondary education[6] also yielded something different, along with the concrete assessment products. This took the form of new kinds of written test formats, problem formats, and problems organized according to level.

- other written test formats

The following assessment methods were developed within RME as alternatives to the traditional restricted-time written tests for secondary education:[7]

(i) *essay test*[8]

In this assessment method, the students were asked, for instance, to write a reaction to a newspaper article (see, for example, the Indonesia problem in De Lange, 1987a) or to give their advice on a problem taken from everyday life (see, for example, the Traffic Light problem in De Lange, 1995).

(ii) *take-home test*

Here, the students could do the test – usually an essay test – at home; they could work on the test individually or in groups, were permitted to use the textbook, and could even request assistance (see, for example, the Sparrow-hawk problem in De Lange, 1987a).

(iii) *two-stage test*

This assessment method combined various test formats. In one case, a written test was first completed at school, then corrected and commented on by the teacher, and subsequently returned to the student for additional work at home. Such a test might consists of long-answer as well as short-answer questions (see, for example, the Foresters problem in De Lange, 1987a).

(iv) *production test*
In this assessment method, the students' assignment was to design a test themselves (see one of the tests attached to the teaching unit Data-visualization in De Lange and Van Reeuwijk, 1993).

(v) *fragmented information reasoning test*
Here the students – either individually or as a group – were presented with certain information in a fragmented form. They were then asked to derive the relevant bits of information, combine them and, if necessary, supplement them with other information, in order to test a given hypothesis (see the problem on the Mesolithic period in De Lange, 1995).

- other problem formats

  Aside from the existing multiple-choice and short-answer problems[9], there were two additional kinds of problem formats (see De Lange, 1995):

  (i) *open-open questions*
  This format differs from the traditional short-answer problems with respect to the activities involved in obtaining an answer. Open-open questions, although also requiring a short answer, are not just triggering questions, but questions that needed some thought and understanding. Moreover, these open-open questions offer the students some opportunity to solve the problem in their own way (see, for example, the Equator problem in De Lange, 1995).

  (ii) *extended-response open questions*
  Compared with the previous format, the students have more freedom here to 'produce' an answer (see, for example, the Harvesting problem in De Lange, 1995). Characteristic of this format is that it is often impossible to predict the students' answers. Evaluating these answers is also rather difficult, particularly in terms of distinguishing between good, better and best.

- assessment tasks on different levels

  Three problem levels are distinguished in connection with the various goals aspired to by mathematics education (see De Lange, 1995)[10]:

  (i) *lower-level tasks*
  This level encompasses the basic abilities, such as the knowledge of number facts and definitions, technical skills and standard algorithms (for instance, solving an equation).

  (ii) *middle-level tasks*
  This level comprises problems in which the students themselves must make certain links, integrate different information, and invent a solution strategy (for instance, in figuring out whether one particular boxplot could be a combination

of two other boxplots).

(iii) *higher-level tasks*

The problems on the highest level are even more demanding in this respect. On this level, mathematization is truly present, for the following reasons: analysis and interpretation are necessary; creativity and one's own constructions are required; reflection, model forming and generalization must also occur; moreover, achieving this level also implies that the student has both the ability to communicate about the approach taken and is able to maintain a critical disposition (using a graphic illustration to show how the population of The Netherlands is aging would be a problem on this level).

The above examples, taken from De Lange (1995), are all intended for secondary education. In order to demonstrate that problems on different levels can be found at all educational levels, De Lange also provides examples that are appropriate for primary education. He regards the Chocolate Milk problem (see Figure 3.1), for instance, as a middle-level task, and the Polar Bear problem (see Figure 3.2) as a higher-level task. Similarly, the Pinball problem (see Figure 2.7) may be viewed as a lower-level task.[11]

The classifications made here of types of written tests, types of problem formats, and problems on different levels cannot be viewed independently, but are, in fact, interconnected to a great extent. The lower-level tasks usually occur in the form of short-answer questions and are largely found on traditional restricted-time written tests. The higher-level tasks, in contrast, require the production, integration and expression of ideas, which necessitates a freedom of response that can be better realized in the form of extended-response open questions. These, in turn, are often found on essay tests.

Furthermore, the classifications themselves should not be interpreted too strictly. It is often difficult to make a distinction between the various categories, which even overlap at times, as in the case of the essay test and the take-home test. Rather than being a fixed plan for categorizing types of problems and tests, these classifications are intended more as a supporting frame of reference in the search for written assessment suitable to RME.

De Lange's recently developed 'pyramid model' (see, for instance, Boertien and De Lange, 1994), which is intended as a guideline for composing a test, has a similar function. This model can be used to show how the problems on a given test have been distributed over the various topics, levels and degrees of difficulty (see Figure 4.1).

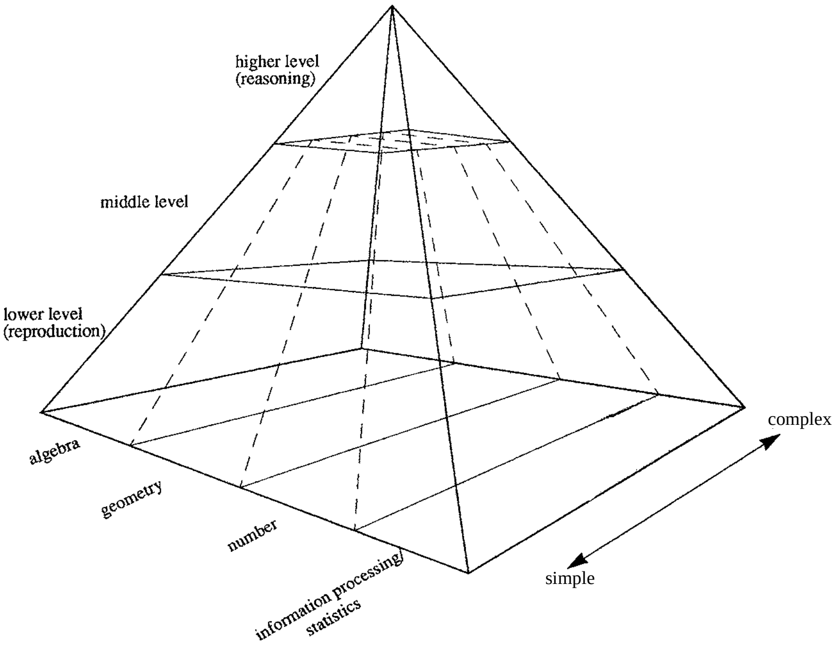


Figure 4.1: Pyramid model for composing tests

4.1.2b Developmental research on assessment for primary education

Alongside developmental research on assessment on behalf of secondary education, similar research was also conducted within RME for primary education. The latter took place more implicitly, however. The MORE research project in which this occurred was not, after all, intended in the first place as assessment research, but as research on textbook comparison, which necessitated the designing of a series of written tests. As was described in Chapter 2, the experiences gained here gradually led to the creation of a parallel current of developmental research in search of alternatives to the traditional written tests for primary education.

In contrast to secondary education, where the alternatives to existing tests were mainly sought in extended tasks, the alternatives that arose during the MORE research took the form of short tasks. Aside from the research context, which, as stated, did not initially focus on the development of alternatives, this also had to do with the more limited reading and writing skills of primary school children. The short-task format was therefore chosen to develop assessment problems that would, nevertheless, be without the drawbacks of traditional short-answer or multiple-choice problems. An attempt was made to invent problems that would both involve the entire breadth and depth of the mathematical area and provide information on the applied strategies (see Section 4.1.1). In other words, the assessment problems had to be both meaningful and informative (see Section 3.2.4).

The assessment problems developed during the MORE research were also characterized by a strategy of designing problems that would be as accessible as possible. This was done by using imaginable contexts that were communicated through illustrative material (see Sections 3.2.6 and 3.2.7), and by formulating the text as simply as possible. Simultaneously, the students were given a great deal of liberty with respect to their answers. The description of 'short' for these assessment problems really applies more to the question than to the answer; indeed, they might better be called 'short-question' problems than 'short-answer' problems.

Another characteristic of these unrestricted-time group tests was that they were actually a kind of hybrid of a written and an oral test. Although the assessment problems were presented to the students in the form of a booklet of test pages – which were also to be used for writing their answers – the instructions accompanying the test problems were read aloud by the teacher. Therefore, the students did not have to read large quantities of text. But, because all the essential information was present on the test page, no demands were made of the students' listening or memory abilities either.

The above is a brief sketch of the MORE tests. More can be found on this topic in Chapter 2.

Just as the developmental research on assessment for secondary education produced more than simply a number of concrete tests and test problems, so did similar research in the MORE research project yield more than the MORE tests alone. Aside from the tests that were inspired by this research and later developed in other projects[12], the MORE research stimulated renewed reflection on the place of assessment within RME, the crucial role of the problems, and the function of the use of contexts. The results of this reflection can be found in Section 3.2. Furthermore, the theme of the developmental research was primarily one of making short-task problems more informative. This will now be discussed in more detail.

4.1.3 Steps for making short paper-and-pencil tasks more informative

The first condition for creating more informative assessment tasks is to dispel three misconceptions about problems, that have long determined assessment. These may also be called 'the three taboos'.[13] By accepting and embracing these taboos, namely (i) that problems can be solved in a variety of ways, (ii) that problems may have more than one correct answer, and (iii) that the correct answer cannot always be determined in some problems[14], an open attitude is created that is necessary if what students can do and how they go about doing it is to become apparent. However, if students are to have an optimal opportunity for showing what they can do, then there is also a second condition, namely, that ‘all-or-nothing’ testing be avoided as far as possible. One manner of accomplishing this is by ensuring that test problems be ‘elastic’ in various ways. Yet this is perhaps considered the greatest assessment taboo of all, particularly within the world of psychometric assessment experts. As will be discussed later (see Section 4.2.1d), although a portion of the assumed certainty is indeed lost with the introduction of this elasticity, such elasticity does, on the other hand, provide a wealth of information – particularly for daily classroom practice.

The MORE tests revealed how this wealth of information can be acquired, or – to paraphrase Joffe (1990) – how more can be obtained from written tests. What is characteristic of all the measures that can be taken in this context is that they usually work bilaterally. On the one hand, they often make the problems more accessible to the students, or – if you will – easier, because they contain built-in opportunities for solutions on various levels. On the other hand, the result of such measures is that the test pages can then reveal to the teacher a great deal about their students’ level of understanding and their applied strategies.

Although not exhaustive, the following sections will provide an overview of the measures that can be taken to make written (short-task) assessment more informative. Each measure will be illustrated with a few compelling examples of student work.

4.1.3a Scratch paper

An initial and very basic measure for making assessment problems more informative is to provide room on the test page for notating intermediate results or for making a sketch to support one’s solution (see Section 2.3.5). Aside from the support this can give students in their search for a solution, this ‘solidified student behavior’ (Ter Heege and Treffers, 1979) also reveals how the students set about their work and the different ways that problems can be solved. Examples of this can be found in Figure 2.14, Figures 3.3 and 3.5 and Figure 4.11b. There, the use of scratch paper is dealt with in detail. One aspect that should be mentioned here, however, is that the students must feel free to use the scratch paper and must not be given the impression that working without scratch paper is valued more highly.[15]

4.1.3b Asking about the strategy

As was mentioned earlier (see Section 2.3.5), the scratch paper in the MORE test problems was always – with an occasional exception – intended for optional use. This does not mean that no explicit questions could be asked about the applied strategies. A requirement, however, is that there is something that needs to be explained (cf. Lamon and Lesh, 1992). The assessment problems on percentage that were developed for the ‘Mathematics in Context’ project (see Van den Heuvel-Panhuizen, 1995a and Chapter 7) are examples of where explicit questions were posed about the applied strategy. In order to emphasize the fact that students could show in different ways how they solved a problem, not only was the formulation ‘explain your strategy’ used here, but, also, ‘show how you got your answer’. A drawing can namely also be used to explain how something was solved.

Questions on the applied strategy can be posed in a closed as well as an open manner. An example of the former is a series of multiple-choice problems (see Figure 4.2)[16] in which Dutch beginning fifth-grade students were asked how they would calculate three different arithmetic problems, if they were given little time to do so. In each problem, they could choose from three different methods: calculating on paper, mental arithmetic, or using a calculator. The strategies the students chose for each of the problems revealed a great deal about their insight into numbers and operations.

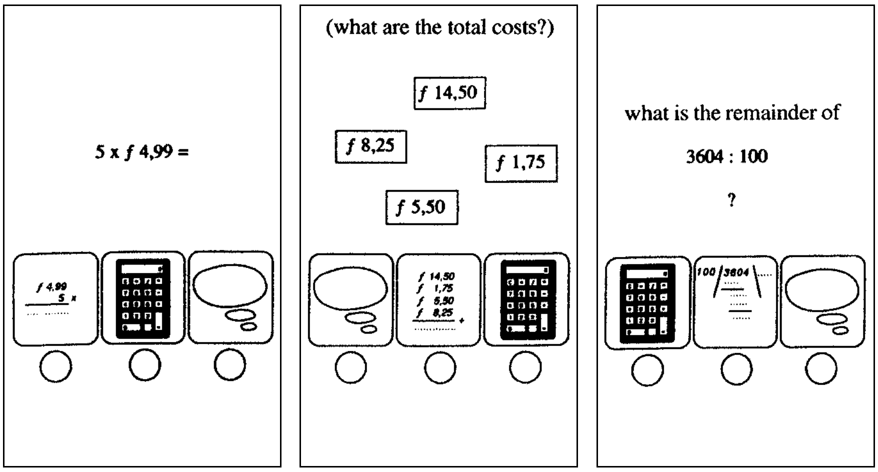


Figure 4.2: Three multiple-choice problems in which the strategy is explicitly requested

As can be seen in Table 4.1, it is clear that many of the students did not recognize the special properties of the numbers and operations in question, and, consequently, would have had difficulty with clever mental calculation. Only one-fourth of the students chose to calculate 5 × 4.99 mentally, and more than 80% of them clearly were not aware that 14.50 and 5.50, and 8.25 and 1.75 fit nicely together. Moreover, the problems revealed quite a bit about the students’ (lack of) familiarity with calculators. Consider, for instance, that 43% of the students chose to use the calculator for finding the remainder in the division problem. In a nutshell, these multiple-choice problems are quite eye-opening.

Table 4.1: Strategies selected by the fifth-grade students

grade 5, Nov (n = 376)	5 × *f* 4,99	*f* 14,50 + *f* 8,25 + *f* 5,50 + f 1,75	the remainder of 3604 ÷ 100
strategy	**% selected**	**% selected**	**% selected**
mental calculation	25	14	24
column caclulation	26	42	33
calculator	48	43	43
unknown	1	1	1

This example demonstrates quite plainly that the format itself is not the cause of bad assessment. Even the currently renounced multiple-choice format can elicit very useful information. What is needed, however, is a good match; in other words, the assessment format and the goal and content of the assessment should be well attuned to one another (see also Lamon and Lesh, 1992).

It is not always necessary, however, to explicitly ask about the applied strategy in order to discover something about the level of understanding and the approach taken. The Same Answer problem discussed earlier (see Section 3.2.9) makes this quite clear. This test problem was designed in such a way that, whether the scratch paper showed traces of calculations or was left blank (the students were, after all, not obligated to use it), conclusions could be drawn about the strategy followed. If a student did not use the scratch paper, then it could be assumed that the associative property had been used, in which case the problem was easy to solve mentally.

4.1.3c More than one correct answer

The assumption that mathematical problems always have only one correct solution is not only untrue, but it also cuts off a number of possibilities for assessment. By presenting them with problems where more than one correct answer is possible (a measure that has also been suggested by others, see Section 3.3.5c), the students not only have more latitude for thinking up solutions, but they will also reveal more information about their learning progress. An example of this is the Candle problem (see Figure 2.9), in which the students were asked to buy twelve candles, but could make their own choice of which boxes to buy – as long as the total was twelve. The solutions of two first-grade students (see Figure 4.3) show that this can expose relevant differences between students. The test problem was administered in February. The disparity that emerged may indicate a difference in degree of familiarity with the number structure. It would seem that one of the students already knew that twelve could be split into six and six, while the other student kept adding up until the total of twelve was reached. More information would be needed to be absolutely sure of this, however, as the choice made may, instead, have to do with a specific interpretation of the context, or may simply be coincidental.

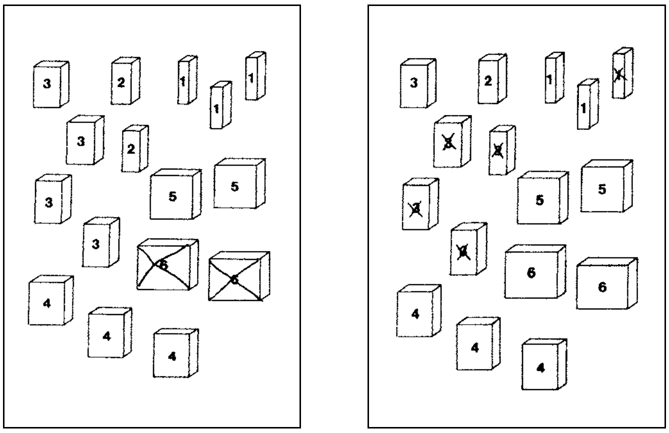


Figure 4.3: Candle problem: student work

In spite of these uncertainties, this kind of assessment problem – especially if the students are given a number of them – can still present an initial foothold for informed decision making.

4.1.3d Option problems

Options can be offered in terms of the question as well as the answer. This entails applying a measure that is quite customary in oral interviews, namely, that a student is asked a different question if an initial one is either too difficult or too easy. This can occur on a written test by letting the student decide which problem to do. Examples of this are the Shopping problems (see, for instance, Figure 2.8), in which the students themselves may decide what to buy. Because the prices of the things to be bought differ, the children actually control the degree of difficulty of the problem. Here, too, the children's abilities and strategies will be most accurately revealed if a few of such problems are included on a test.

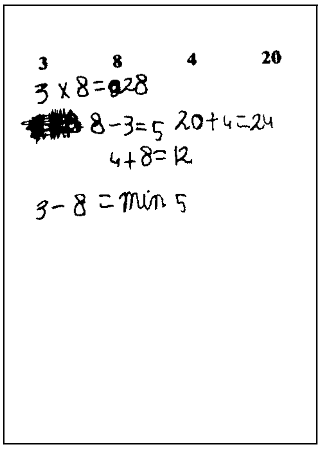
The idea of applying interview techniques to written tests was later elaborated upon in other ways as well (see Section 4.1.4).

4.1.3e Own productions

The 'own productions' – which have always been closely linked to RME (see Sections 1.2.3e and 1.2.5c)[17] – occupy a special place among the measures for making written assessment more informative. What is striking here is that the roles of teacher and student (or, when applicable, researcher or developer and student) have been reversed. The students are asked to think up and solve a problem or a task themselves, instead of doing problems thought up by others. One might say that problems requiring an 'own production' are the most open form of option problems. The opportunity given the students to make their own constructions not only offers them the chance to show what they can do, but simultaneously reveals a great deal about them and about the kind of education they have received.

As was demonstrated much earlier outside the RME circle (see, for instance, Grossman, 1975), own productions may sometimes hold absolute surprises.

An example of such a surprise in the MORE research is that of the second-grade student (see Figure 4.4) who, for an assignment in which the students were to make up their own problems out of given numbers, came up with a problem that resulted in a negative number.[18] This took place in the period April/May.



3 8 4 20

3 x 8 = 28

8 - 3 = 5 20 + 4 = 24

4 + 8 = 12

3 - 8 = min 5

Figure 4.4: Own production: 'Making up one's own problems using the numbers provided'

Other forms of own productions, which can reveal quite a bit about the breadth, in particular, of the students' skills, are those in which the students are asked to make up an easy problem and a difficult one. Figure 4.5 shows two third-grade students' work on this assignment from the MORE research. This took place in November.

A similar assignment, but involving percentage problems, was given to fifth-grade students in the 'Mathematics in Context' project (see Section 4.6d in Van den Heuvel-Panhuizen, 1995a; see also; Van den Heuvel-Panhuizen, Streefland, and Middleton, 1994; Van den Heuvel-Panhuizen, Middleton, and Streefland, 1995). Here, too, a splendid cross-section of student understanding at a particular moment was produced. Furthermore, however, it also produced a longitudinal section of the learning path that students will generally follow with respect to percentage.[19]

It is this last aspect that makes own productions an especially suitable source for deriving indications – and even educational material[20] – for further instruction. Another important feature of this type of problem is that not only the teachers, but the students, too, can become more conscious of where their limitations lie, and can improve their learning through using the tool of reflection that is elicited by own production tasks.

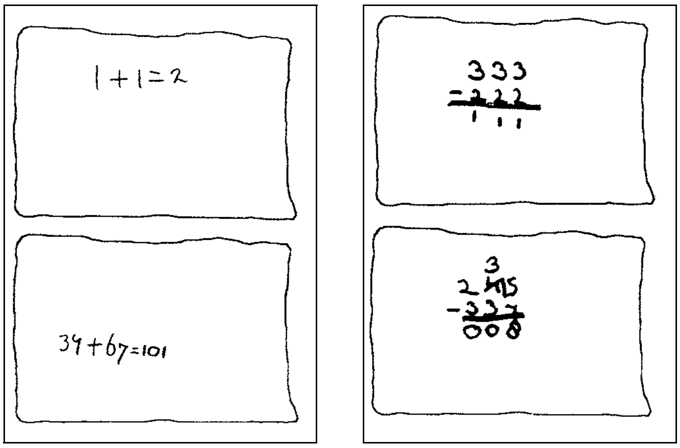


Figure 4.5: Own production: ‘An easy and a difficult problem’

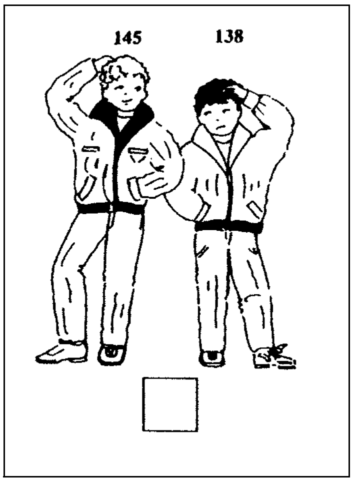
As can be seen from the examples cited above, own production tasks can be used on all educational levels. Furthermore, aside from the short tasks described here, own productions can also consist of extended tasks. An example of this is an alternative developed for secondary education, in which the students had to develop a test themselves (see Section 4.1.2a).

4.1.3f Alternations in presentation

Presenting one and the same problem in a variety of ways can also be extremely revealing. For instance, the MORE test problems disclosed large discrepancies between the context problem scores and the scores for the corresponding formula problems. An example of this, which was discussed earlier (see Section 2.3.4), involved a jar containing 47 beads, 43 of which were to be used to make a necklace. This Bead problem, which was given to second-grade students in November, was answered correctly by 60% of the students, while the corresponding formula problem (47 – 43 =), which was presented at the same time, was only answered correctly by 38%. Clements (1980), by the way, had already discovered similar discrepancies some years earlier (see Section 3.3.7d).

However spectacular these discovered discrepancies may be, it is what lies behind these differences that is important, rather than the differences themselves. Evidently, the context elicited a different strategy than did the formula problem. The strength of this measure lies in the fact that it provides students with the opportunity to solve certain problems by using informal strategies that are linked to contexts. As a result, instruction that has not yet been given can be anticipated. It is this opportunity for 'advance testing' that makes this measure so appropriate for providing indications for further instruction.

Another striking example of this aspect is the problem in which two boys compared their height (see Figure 4.6), which was first administered to a second-grade class in the period April/May.



Instructions to be read aloud:

"Evi and Onno are measuring how tall they are. Onno is 145 centimeters and Evi is 138 centimeters.
How great is the difference?
Write your answer in the empty box."

Figure 4.6: Comparing Height problem

Although, at the moment this problem was administered, problems involving bridging ten above one-hundred had not yet been handled, around 50% of the students were nonetheless able to calculate this difference. In contrast to the Bead problem, no corresponding formula problem (145 – 138 =) was simultaneously administered in this case. This was due to the wish not to frustrate the students with a type of problem for which they did not yet know the procedure. Moreover, such a problem would not have produced much information and would have revealed more about what the students were *not* able to do than what they *were* able to do.[21] In the context problem at hand, by contrast, it did become clear what the students could do. Compared with the bare subtraction problems involving bridging ten under one-hundred (33 – 25 = and 94 – 26 =), that were included on the test because the students had already dealt with them, the score for the Comparing Height problem was respectively 12% and 15% higher (see Table 4.2). Instead of using the often laborious subtraction procedures that are usually taught for formula problems, the students apparently also used the informal strategy of 'adding-up', that was elicited by the context.[22]

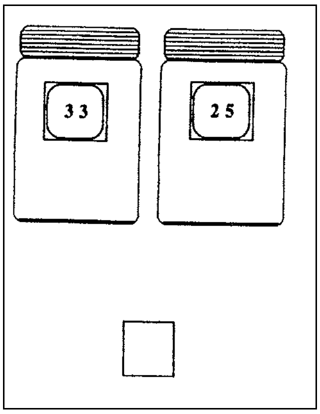
Aside from the actual lesson contained in these results with respect to introducing this type of formula problem, they also signify a confirmation of the RME educational theory that places informal application ahead of work on a formal level.

These results do not mean, however, that no further developmental research is necessary on this point. The assessment data (see Table 4.2) produced by the MORE research did, after all, also yield new questions as well as answers.

One of the most important questions that arose from this data concerned the relationship between context problems and formula problems in the course of the learning process. Another question was whether contexts are interchangeable. Take, for example, the difference in results between two context problems, both of which involved calculating 100 – 85, but where one entailed buying a marker and the other an ice cream (see Figure 4.8). Yet another question, that arose in connection with three different presentations of 6 × 25 on one and the same test (see Figure 4.10), was how quickly a given context problem can put students on the track of a clever strategy for a formula problem.

Table 4.2: Test results for context problem (C) and formula problem (F)

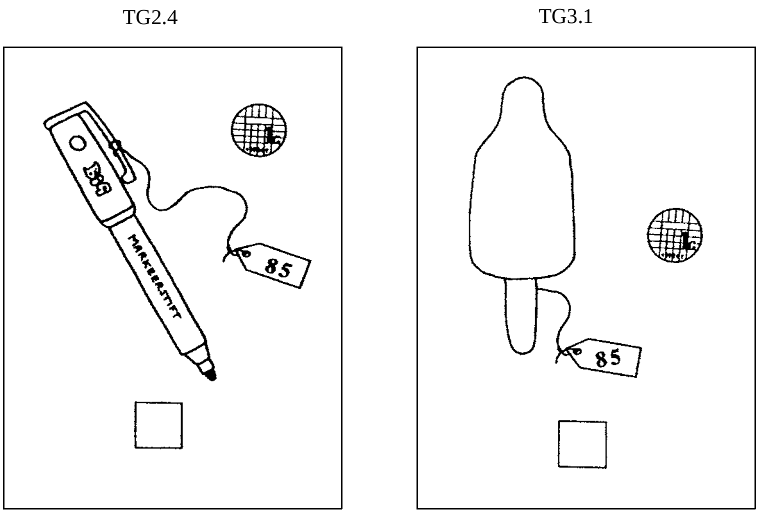
MORE test problems	% correct answers			
	TG2.2 (Grade 2, Nov) n = 427	TG2.4 (Grade 2, Apr/May) n = 432	TG3.1 (Grade 3, Sep) n = 425	TG3.2 (Grade 3, Nov) n = 416
C 47 – 43 (Fig. 2.12) F 47 – 43	60 38	- -	- -	- -
C 33 – 25 (Fig. 4.7) F 33 – 25	54 21	64 40	71 49	- -
C 100 – 85 (Fig. 4.8) F 100 – 85	- -	60 55	70 54	- -
C 145 – 138 (Fig. 4.6) F 94 – 29	- -	52 37	60 33	- -
C 67 + 23 (Fig. 4.9) F 67 + 23	- -	74 75	80 81	- -
C 25 × 6 (Fig. 4.10) C 6 × 25 (Fig. 4.10) F 25 × 6 (Fig. 4.10)	- - -	- - -	- - -	33 63 42



Instruction to be read aloud:

"Here are two jars containing beads. There are more beads in one of the jars than in the other.
Do you know how many more there are?
Write your answer in the empty box."

Figure 4.7: Context problem 33 – 25



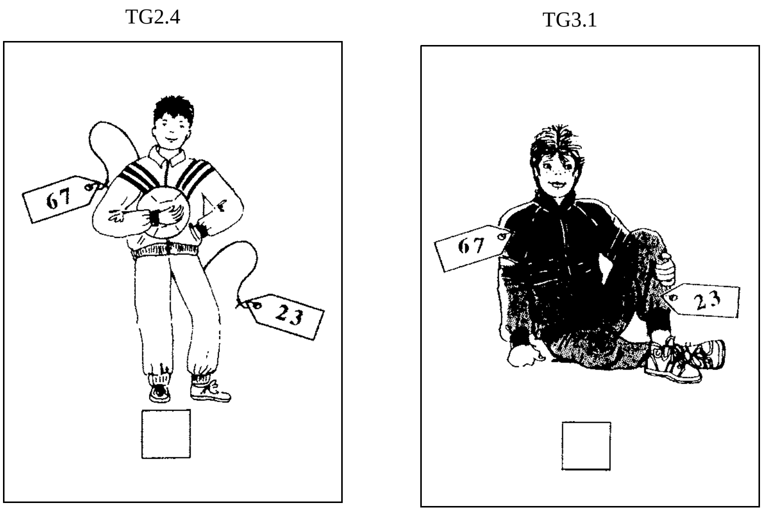
TG2.4

What will your change be if you pay with a one-guilder coin?

TG3.1

What will your change be if you pay with a one-guilder coin?

Figure 4.8: Context problems 100 – 85



TG2.4

How many guilders in all?

TG3.1

How many guilders in all?

Figure 4.9: Context problems 67 + 23

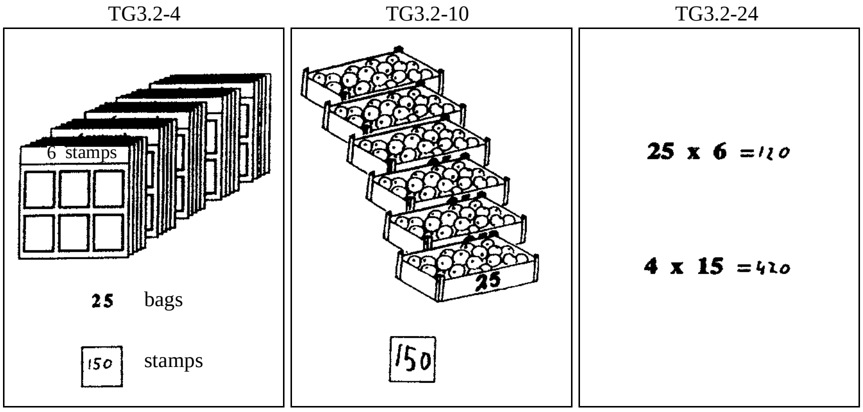


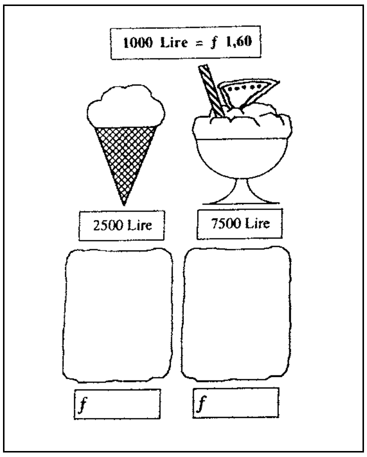
Figure 4.10: Three different presentations of 6 × 25

Aside from this particular alternation between context problem and formula problem, there are also other alternations in presentation that can be used to provide insight into students' thought processes. Take, for example, a presentation with or without support problems (an example of which can be found in Van den Heuvel-Panhuizen, 1990a, p. 69; see also Section 4.1.3i). Other such examples are, for instance, the presentation of problems both in numerical and non-numerical form (see the test on ratio discussed in Chapter 6), or with countable and non-countable objects (see the addition and subtraction problems from the first MORE test discussed in Chapter 5).

As stated earlier, most of the measures make test problems more informative because they contain built-in opportunities for solving the problems on different levels. This is particularly true of the following three measures.

4.1.3g Twin tasks

A twin task in fact comprises two linked problems, meaning that the answer to one part of the problem can be used to help find the answer to the other part. A crucial aspect here, of course, is whether the student takes advantage of this relationship. The Ice Cream problem (see Figure 4.11a) is an example of this.[23]



Instructions to be read aloud:

“If you buy an ice-cream cone in Italy it will cost you two-thousand-five-hundred Lire. Wow! And an ice-cream sundae will really make you gulp. That will cost you seven-thousand-five-hundred Lire! How much do you think that is in Dutch money?
Write your answers in the boxes at the bottom of the page, and use the scratch paper to show how you got your answers.”

Figure 4.11a: Ice Cream problem

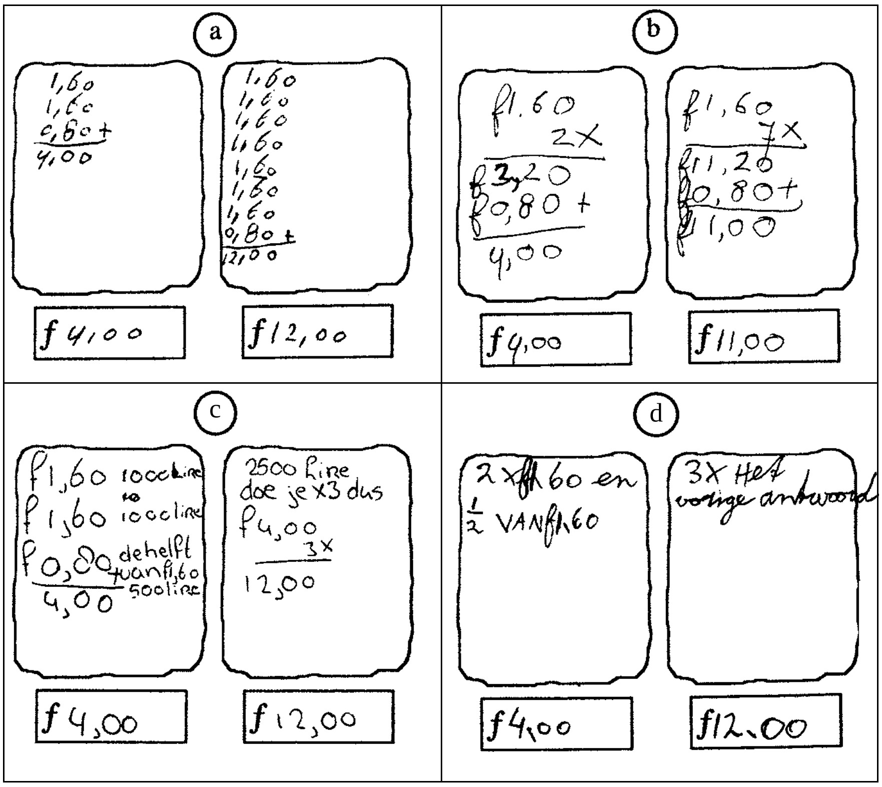
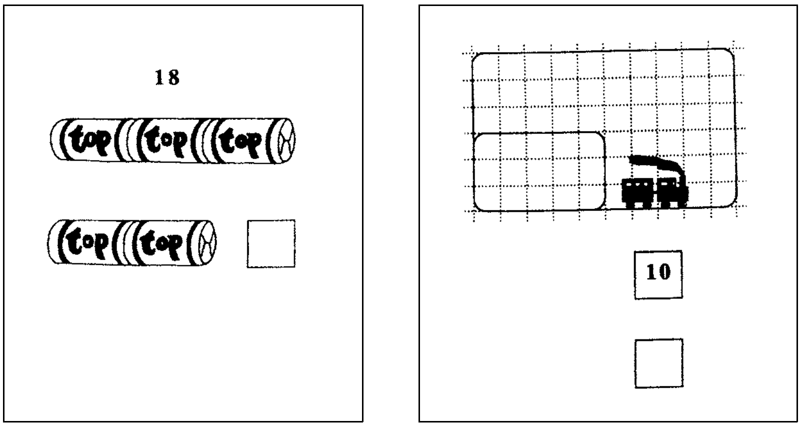

a
1,60
1,60
0,80 +
4,00
1,60
1,60
1,60
1,60
1,60
1,60
1,60
0,80 +
12,00
f 4,00
f 12,00
b
f1.60
2X
f3,20
f0,80 +
4,00
f1,60
7X
f11,20
f0,80+
f11,00
f 4,00
f 11,00
c
f1,60 1000 Lire
f 1,60 1000 lire
f 0,80 de helft van f1,60 500 lire
4,00
2500 Lire doe je x3 dus
f4,00
3x
12,00
f 4,00
f 12,00
d
2 x f1,60 en 1/2 VAN f1,60
3X Het vorige antwoord
f4,00
f12.00

Figure 4.11b: Student work on the Ice Cream problem

The examples of student work displayed in Figure 4.11b show but a small part of the range of solutions presented by students at the beginning of fifth grade.[24] Nevertheless, these examples clearly show how these problems – especially the second one – can be solved on different levels. Some students calculated each ice cream separately (see a and b), and some derived the price of the sundae directly from the ice-cream cone (see c and d). Even within these categories, however, different solution strategies can be distinguished.

4.1.3h Multi-level pictures

Another measure for providing assessment problems with a certain stratification is that of using pictures – or, if you will, context-drawings – that can be employed on different levels when solving the problems. Examples of this are the Tiled Floor problem (see Figures 2.15 and 2.16), which has already been discussed in Chapter 2, and the Candy problem and the Train problem (see Figure 4.12).



Instructions to be read aloud:

"There are 18 candies in the whole roll. How many candies do you think there are in the roll that is no longer whole?
Write your answer in the empty box."

Instructions to be read aloud:

"There's a train ride in the amusement park. The short ride takes 10 minutes. How long do you think the long ride takes?
Write your answer in the empty box."

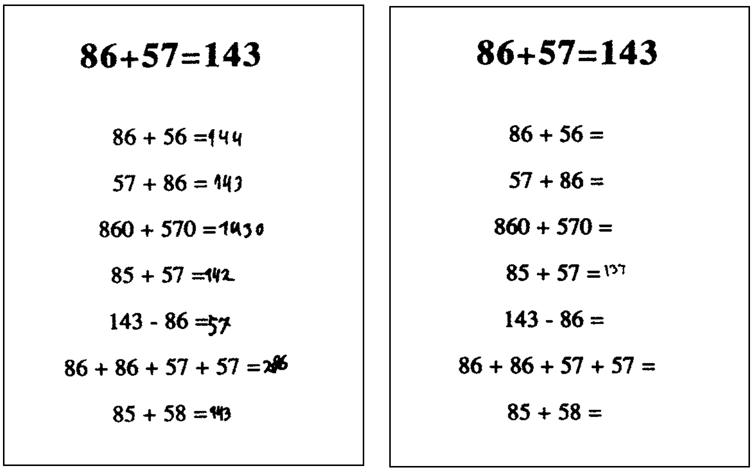
Figure 4.12: Multi-level pictures

The number of candies in the small roll, for instance, can be found by reasoning that the small roll contains two-thirds of eighteen candies, but also by actually dividing the large roll into three sections, drawing the number of candies in each section, and then counting the total number of candies in the small roll. In the Train problem, one can reason, for instance, that each horizontal and each vertical piece of the long ride is twice as long as the pieces in the short ride, so the long train ride will therefore take twice as long as the short ride. But those who still find this too difficult can mentally transfer the pieces of the short ride to the long ride, and then discover that the pieces of the short ride cover half of the long ride. Yet another way would be to count the number of ‘units’ in each ride.[25]

Sometimes, traces of these strategies can plainly be seen in the drawing – as is the case in the Tiled Floor problem (see Figure 2.16). In order not to be entirely dependent on fortuitous traces, however, one can explicitly ask the students about the strategies they applied.

4.1.3i Support problems

Another form of assistance which can make the problem more informative is the inclusion of support problems on the test page. This is namely a way of discovering whether the students have insight into the properties of the operations. The problems included on the test page were somewhat more advanced than the problems that had so far been handled in class. All the problems were related to one another in some way. The result to one of the problems (the ‘support problem’) had been provided, and the students were asked to find the results to the remaining problems. This may seem simple, but, in fact, is only so once one has acquired insight into the properties of the operations. It is clear that the two students whose work is displayed in Figure 4.13 had varying degrees of insight.



86+57=143

86 + 56 = 144

57 + 86 = 143

860 + 570 = 1430

85 + 57 = 142

143 - 86 = 57

86 + 86 + 57 + 57 = 286

85 + 58 = 143

86+57=143

86 + 56 =

57 + 86 =

860 + 570 =

85 + 57 = 137

143 - 86 =

86 + 86 + 57 + 57 =

85 + 58 =

Figure 4.13: Student work on task containing support problem

4.1.4 Applying interview techniques to written assessment

By using option problems and other conscious or unconscious measures that made the problems more dynamic and ‘elastic’ (see also Chapter 2), certain assessment problems increasingly began to display characteristics of oral interviews. The problems could be adapted in a certain sense to a student’s level, and the students could show in their own way how they had solved something. The more this interview-like nature of the problems was recognized and its advantages were felt, the more a need arose to apply certain interview techniques more explicitly, in order to make written assessment more informative. The initial ideas on this matter were put into effect in a test on percentage that was developed for the 'Mathematics in Context' project (see Van den Heuvel-Panhuizen, 1995a and Chapter 7).

4.1.4a The safety-net question

One of the first questions that may arise when correcting student work is whether the answer given does indeed accurately reflect the student's level of understanding. Did the student think the problem through? And was the problem understood as it was intended to be? These questions are even more pertinent when the problems are formulated in a very open manner and the students are not steered towards a particular answer. The need then arises – as may occur during an interview – to ask further questions[26], in order to acquire more certainty about the first answer that was given.

The first time this need was distinctly felt was with the Best Buys problem, in which the students had to compare two discount percentages. The purpose of this problem was to determine whether the students had insight into the relative nature of percentage (see Sections 3.3a, 3.3b, and 3.5c in Van den Heuvel-Panhuizen, 1995a). The issue that arose in response to the students' answers, was whether the students who had only compared the percentages absolutely were truly unaware that this really wouldn't work. In order to resolve this issue, an extra question was added, which, because of its function, was called a 'safety-net question'.[27] As became evident from the results, this extra question did, indeed, function as intended. Many students, who had only made an absolute comparison in response to the first question, demonstrated by their answer to the safety-net question that they did, indeed, have insight into the relative nature of percentage. Moreover, the safety-net question proved to have a converse function as well: as it turned out, other students, because of their correct answer to the first question, had been erroneously thought to have this insight, when, in fact, they did not.

A second application of the safety-net question in the case of the Parking-lots problem (see Section 4.5e in Van den Heuvel-Panhuizen, 1995a) confirmed its functionality. More detailed attention is devoted to the safety-net question in Chapter 7.

4.1.4b The second-chance question

Another interview technique that could be of service to written assessment, is that of eliciting reflection by confronting students with their own answer or strategy.[28] One could also call this inherent feedback that is acquired by means of a built-in control question. This is especially useful for questions in which the students can easily make a mistake. Such a control question would thereby offer the students an extra opportunity to arrive at a correct solution. An example of such an obviously tricky problem is the Binoculars problem (see Section 3.3 in Van den Heuvel-Panhuizen, 1995a). Here, the students had to determine the regular price of the binoculars on the basis of the sale price ($96) and the amount of discount (25%). Because there was a good chance that the students would mistakenly calculate 25% of $96 and add this to the sale price, the following 'second-chance question' was appended:

> "Now, check your answer by subtracting 25% of the regular price that you found. Are you satisfied by this result? If not, you can have a second chance. Try and correct your first answer."

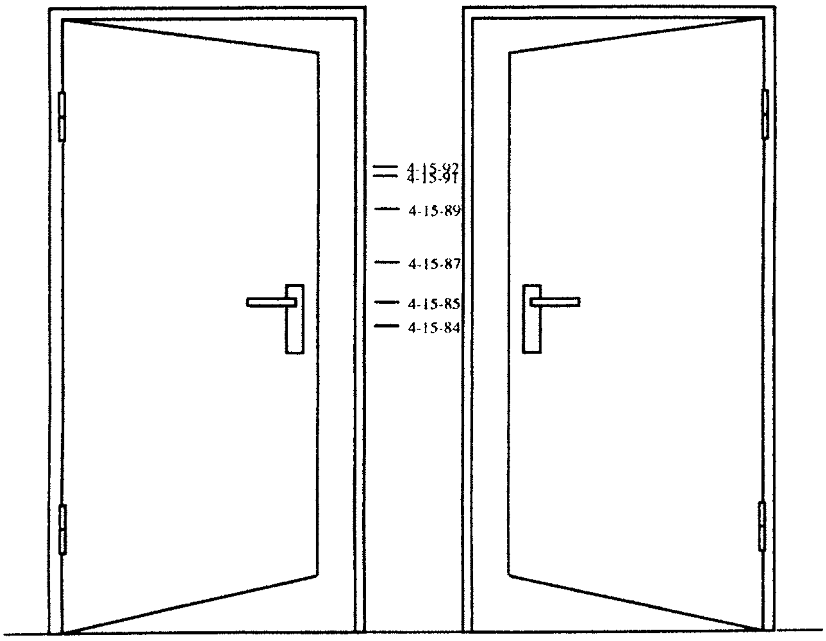
Although approximately one-third of the students whose first answer was incorrect did discover that fact, thanks to the extra question, only a few of them managed to correct their answer. Evidently, this particular second-chance question was inadequate for this purpose. The results suggest that students who have difficulty reasoning backwards with percentages need a more specific means of support, such as, for instance, a percentage bar. It may be, however, that this problem was simply too difficult for this group of students. A follow-up research involving the second-chance question has not yet been conducted, so many unanswered questions remain regarding the potential of the second-chance question. Nevertheless, this measure would seem to add perspective to written assessment.

4.1.4c The standby sheet

The last measure for making written assessment more informative to be discussed here was inspired by an interview technique in which assistance is offered the student, after which the student's reaction to this assistance is observed.[29] This idea of offering assistance during assessment (which dovetails with Vygotsky's (1978) concept of the 'zone of proximal development'), was already applied during the early years of RME as an enrichment of written assessment (see Ter Heege and Treffers, 1979 and Sections 1.2.3e and 1.2.5b). The assistance was offered by the teacher as he or she walked around the classroom while the students were taking a written test. The currently suggested measure involves effecting this assistance within the written test itself, in the form of a 'standby sheet'. This is a separate page containing a few extra questions that can guide and support the students as they answer the principal question. Use of this page is optional. The teacher takes note of which students request the standby sheet and thereby knows exactly which students answered the problem with assistance and which without. Furthermore, it can then be seen what the students did with this assistance.[30] In order to demonstrate what such a standby sheet might look like, an assessment problem with accompanying standby sheet was developed for the 'Line Graphs'[31] unit of the 'Mathematics in Context' project. The unit was intended as a sixth-grade introductory unit on graphs, and the purpose of the assessment problem (see Figures 4.14a and 4.14b) was to discover whether the students would be able to reproduce certain data on a graph.

The writing on the wall

Geert likes to keep track of how tall he is by writing on the wall. You can see here what the wall looks like. There's a date next to each mark. The first mark was made on his third birthday.

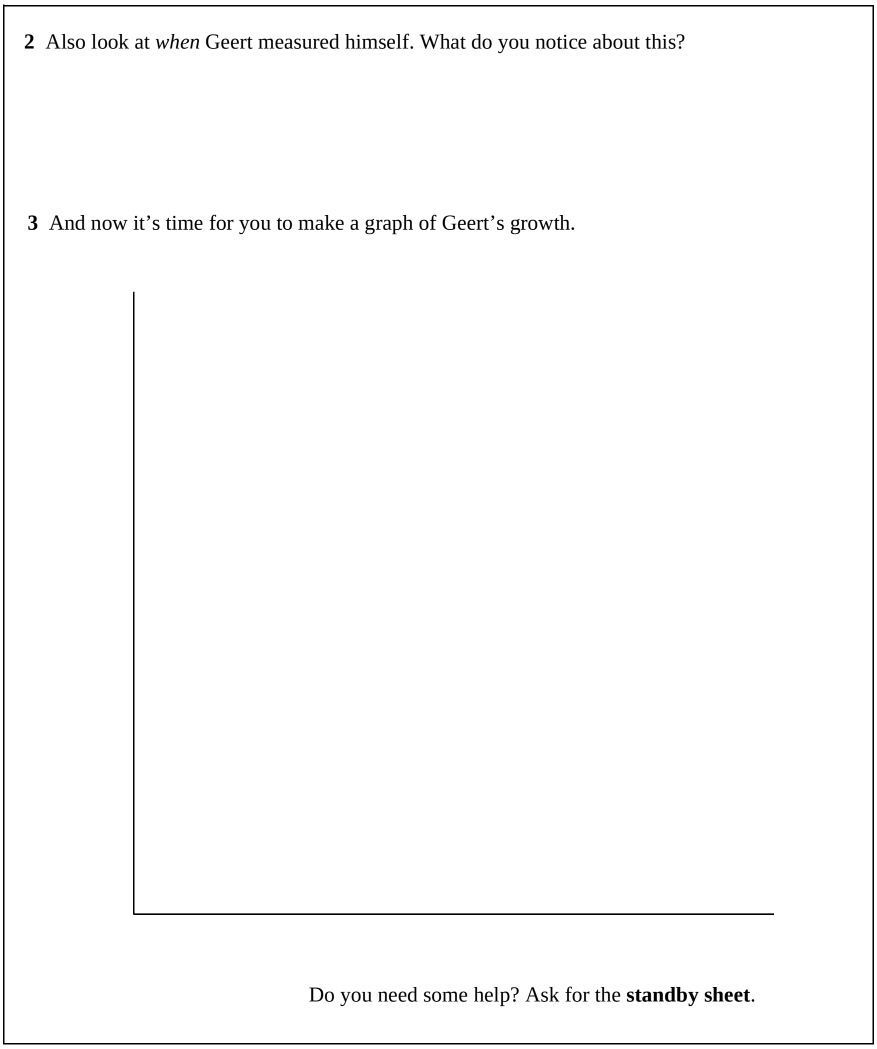


Well, Geert has just been learning about *graphs* at school. Now he thinks that making marks on the wall is kind of childish. So he's found a piece of paper and has made a graph. Let's see if you, too, can show Geert's growth on a graph.

1 Before you start, take a good look at the marks on the wall. How tall do you think Geert was when he started making these marks, and how tall was he when he made the last mark?

Figure 4.14a: Growth Graph problem, page 1

The assessment problem in question could be administered at the end of the unit. Compared with the contexts that appear in this unit, the context of this assessment problem was new. It was likely, however, that the students would already have come in contact with such growth graphs in everyday life. In order to assist the students, the data was even presented in such a way that it was already nearly a graph. The location was a bit of wall between two bedroom doors. Geert had often stood there in order to be measured, so the wall was full of marks.



2 Also look at *when* Geert measured himself. What do you notice about this?

3 And now it's time for you to make a graph of Geert's growth.

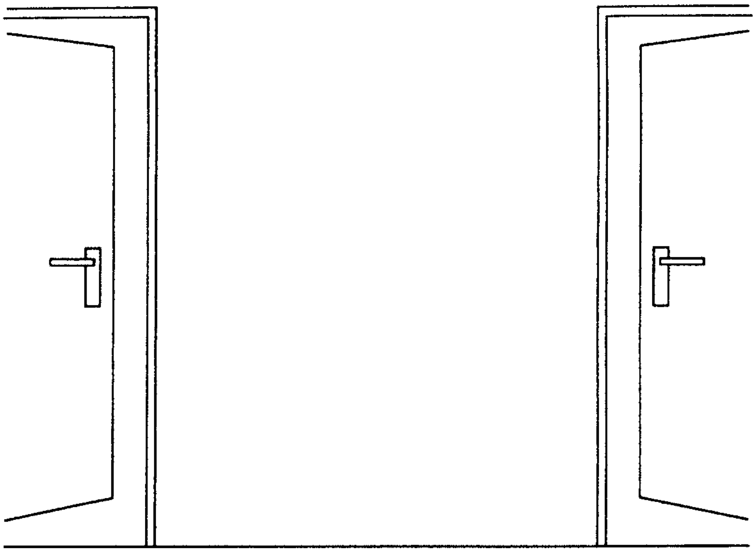
Do you need some help? Ask for the **standby sheet**.

Figure 4.14b: Growth Graph problem, page 2

The students were to use this data for making a graph. In a certain sense, the graph was there already, and only needed to be stretched out. No numerical information was provided about Geert's height. The students had to derive that themselves from the height of the doors. A complicating factor was that the intervals between the marks were not all the same. Hopefully, the students would take this into account when drawing the graph. Alongside the assistance given the students by the choice of context – which already contained the solution, as it were – additional assistance was also provided in the form of the standby sheet (see Figure 4.15).

Standby sheet

- Here you have the wall again, but now there's a lot more room to make the marks. If Geert had had this much space, he could have moved over a little each time. Then what would the wall have looked like? Go ahead and draw it.



- Now there are marks on the wall, but you still don't know how tall Geert is. Or do you? Imagine that you stood against this wall yourself, where would your mark be? Draw this mark and write your name next to it. How do you actually know that that's where your mark should be? About how tall are you? And the doors? So now do you know how tall Geert was when he measured himself?

- If you take a close look at the dates, you'll notice that Geert didn't measure himself every year. Sometimes he skipped a year. Keep this in mind as you draw the graph.

Now see if you can answer question number 3.

Figure 4.15 : Standby sheet for the Growth Graph problem

Students who had difficulty making the graph could ask the teacher for this extra page. At that point they would also be given a different colored pen, so that it would be clear what they had discovered on their own and where they needed help.

No trial run of this standby sheet has as yet taken place, so it remains to be seen whether the desired information can indeed be acquired in this way. Nevertheless, on the whole it would seem that this integration of oral and written assessment – which were previously regarded as two separate worlds – could have a beneficial effect on the further development of written assessment.[32]

4.1.5 Consequences for correcting student work

As a conclusion to these examples of RME alternatives to the traditional paper-and-pencil tests, the following sections will discuss the consequences of this different method of assessment for the way the students' answers are interpreted and analyzed. The issue of grading, however, will be mainly left aside. Not only is this an extremely complex matter, requiring much more in-depth consideration than is possible in the scope of this chapter[33], but there is another reason as well. In this new, alternative viewpoint on assessment, which involves much more concentration on collecting information to be directly used for instruction, grades are no longer the center of attention. This certainly does not imply that it is no longer appropriate to express appreciation for a particular approach to a problem in the form of a grade. Not at all. It simply means that more can be obtained from the students' answers than merely a grade, and that more can be said about these answers than simply whether they are right or wrong.

4.1.5a Right or wrong is not enough

Outside the RME circle, too, there is a fairly general consensus with regard to this last statement. Recent developments within the cognitive disciplines stress the importance of examining more than the correctness of the answer and point out that other aspects of the solution – such as representations, strategies and errors – must be considered as well (Resnick, 1988, cited by Magone et al., 1994).[34] Depending upon which strategies were used, one and the same answer may be based on different levels of understanding (Magone et al., 1994). In order to acquire footholds for further instruction, teachers must analyze their students' thought processes and may not merely check for right and wrong answers (Szetela and Nicol, 1992). This implies a drastic change in relation to a situation in which assigning grades has been the sole activity, and all further information – as Clarke (1993b) put it – has been discarded in the grading process. Such a change is particularly acute in the United States, where the preference for quantifiable information is deeply rooted (Roberts, 1994).

4.1.5b The richer the problems, the richer the answers

The fact that at times it is no longer possible to make a simple distinction between right and wrong answers is also due to the different types of assessment problems, as well as to the desire to obtain more information from the answers.[35] The answers are simply too complex to allow such a distinction. Take, for instance, students' answers to the Polar Bear problem (Figure 3.3). And, when it comes to extended tasks, such a distinction is even more difficult (see De Lange, 1987a). As it turns out, correctness may appear in several manifestations. The students' responses, which are elicited by the problems, do not allow one to confine oneself to the narrow criterion of correctness. Such a limitation would do an injustice to the richness of the answers and to the different aspects that are distinguishable in a result. One can better speak of the ‘reasonableness’ of a given answer, rather than of its correctness. In a sense, this implies putting oneself in the students’ shoes and asking oneself what they might have meant by their answer, or what their reasoning might have been. This will be discussed further in Section 4.1.5d.

Correcting the students’ work is also made more difficult because, as stated before, the aim is no longer merely to get results, but also to discover the thought processes and solution strategies applied by the students (see Freudenthal, 1981a; see also Section 1.2.3a). Compared with results, in the case of strategies it is often even more difficult to distinguish between right and wrong. A chosen strategy can, of course, either be applied correctly or incorrectly, and can be performed either *with* computational errors (see the strategy in Figure 3.3d) or *without* them. But other characteristics may be significant as well, making it impossible to maintain one given criterion of correctness. For instance, a strategy can provide *more* (see Figure 3.3f) or *less* perspective for future (more difficult) problems, be *more* (see Figure 3.3c) or *less* (see Figure 3.3a) clever in a given situation, and be a *more* or *less* assured way of arriving at the solution.

4.1.5c Various types of scoring rules

According to Lamon and Lesh (1992), in order to manage this rich information one needs a strong supporting framework that can provide a basis both for interpreting the students’ reasoning and for making decisions with regard to further instruction.[36] Scoring rules may also have the function of informing students about which points will be crucial when their work is corrected (see Bell, Burkhardt, and Swan, 1992b).

The scoring scales used for correcting open-ended tasks vary considerably, and may range from general scoring scales to task-specific ones (Wiliam, 1993).

- general scoring scales

Under general scoring scales may be understood categorizations in general levels of cognitive development – such as those of Piaget – or levels of problem solving ability. An example of the latter are the categories distinguished by the superitems, which are based on the SOLO taxonomy (see Section 3.3.5h). Other general forms of scoring are those labeled ‘analytic’ scoring and ‘holistic’ scoring (Lester and Lambdin Kroll, 1991). The first form involves a procedure whereby separate points are awarded for each aspect of the problem-solving process, i.e., understanding the problem, planning the solution, and getting an answer (see, for instance, Lambdin Kroll et al., 1992; Bell, Burkhardt, and Swan, 1992b). The holistic method of scoring focuses on the solution as an entity, rather than on its various components (see, for instance, Lester and Lambdin Kroll, 1991; Cross and Hynes, 1994). Both methods, however, usually involve a general description that in no way refers to the problems themselves (see Figure 4.16).[37] The advantage of these general scales is that they can be used for a wide range of problems (Szetela and Nicol, 1992).[38] The issue, however, is whether such general analyses actually provide sufficient footholds for further instruction.

ANALYTIC SCORING SCALE

Understanding the problem
0: Complete misunderstanding of the problem.
3: Part of the problem misunderstood or misinterpreted.
6: Complete understanding of the problem.

Planning a solution
0: No attempt, or totally inappropriate plan.
3: Partly correct plan ...
6: Plan could lead to a correct solution ...
Etcetera

HOLISTIC SCORING SCALE

0 points:
* Problem is not attempted or the answer sheet is blank.
* The data copied are erroneous and no attempt has been made to use that data.
* An incorrect answer is written and no work is shown.

1 point:
* The data in the problem are recopied but nothing is done.
* A correct strategy is indicated but not applied in the problem.
* The student tries to reach a subgoal but never does.

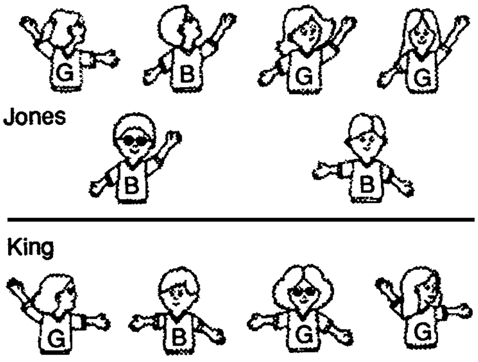
2 points:
* An inappropriate method is indicated and some work is done, but the correct answer is not reached.
* A correct strategy is followed but the student does not pursue the work sufficiently to get the solution.
* The correct answer is written but the work either is not intelligible or is not shown.

3 points:
* The student follows a correct strategy but commits a computational error in the middle ...
Etcetera

Figure 4.16: Parts of an analytic scoring scale (from Lambdin Kroll et al., 1992, p. 18) and a holistic scoring scale (from Cross and Hynes, 1994, p. 373)

- task-specific scoring scales

With respect to footholds for further instruction, more can be expected of the task-specific scoring scales, in which the categories of possible answers (often illustrated with student work) explicitly pertain to a specific problem. Examples of this can be found in Bell, Burkhardt, and Swan (1992a) and Lamon and Lesh (1992).[39] The example given by the latter authors has to do with ratio problems. In order to determine what sort of experiences are critical for understanding ratio and proportion, they interviewed children and analyzed many tasks involving these concepts. One of the essential aspects of the ratio concept that came to light was the ability to make a distinction between relative and absolute changes. So, in order to assess this ability, interview problems were created in which both additive and multiplicative reasoning are appropriate. An example of this can be seen in Figure 4.17. Both of these elements, each of which indicates a different solution level, can be found in the corresponding scoring rules (see Figure 4.18).



Here is a picture of the children in two families.
The Jones family has three girls and three boys and the King family has three girls and one boy. Which family has more girls?

Follow-up question: Which family has more girls compared to boys?

Figure 4.17: The Families problem (from Lamon and Lesh, 1992, p. 332)

0: The student reasons additively.

1: The student reasons multiplicatively in some situations when prompted to consider a relative comparison.

2: The student reasons multiplicatively in some situations without prompting.

3: The student's initial response uses relative thinking.

4: The student thinks relatively and explains his or her thinking by making connections to other pertinent material or by translating to an alternate form of representation.

Figure 4.18: Scoring categories pertaining to the Families problem
(from Lamon and Lesh, 1992, p. 334)

The nature of the scoring categories used by Lamon and Lesh is quite similar to how student work is viewed within RME. Moreover, the approach followed by Lamon and Lesh in designing these categories also strongly resembles the RME method, in which mathematical-didactical analyses and observing children are seen as the basis for assessment development (see Chapter 1).

- flexible scoring scales

The above-mentioned approach was also taken by Magone et al. (1994) when designing scoring categories for the assessment problems that were developed for the QUASAR project (see Section 3.3.6b). What is striking about these categories is that they are not only analytic (in the sense that they involve various aspects of the solution, such as the forms of representation, solution strategies, reasoning strategies, solution errors, mathematical arguments, quality of description), but also general and specific. The same aspects are not examined in each problem, and the categories are of a general nature for certain aspects (such as for the forms or the representations, where a distinction is made between explanations in words, pictures and symbols) and task-specific for others (such as the solution strategies). This flexible scoring method, in which the categories are dependent upon which problems have been presented to the students, is actually a logical result of the manner in which these categories arose, namely, through analyses of the task content and analyses of student solutions to the tasks. Here, too, there are many points of similarity to the approach within RME. The answer categories used in the test on percentage that was developed for the 'Mathematics in Context' project are a good illustration of this (see Van den Heuvel-Panhuizen, 1995a).

- avoiding different interpretations

As with Magone et al. (1994), the results of this test on percentage were described by using examples of student work to illustrate the various categories. This method was also applied by, for instance, Bell, Burkhardt, and Swan (1992b) and is extremely important, according to Wiliam (1993), as a means of conveying information to the teachers. Wiliam's experience with various types of scoring scales has led to his belief that general descriptions too often lead to different interpretations, and that task-specific categories are not always recognized by the teachers; on the other hand, he has found the approach in which each task level is illustrated by student work to be very successful.

One method (used mainly in Great Britain) for helping teachers to agree on the scores is that of 'moderation' (Joffe, 1992; Bell, Burkhardt, and Swan, 1992c; Brown, 1993; Harlen, 1994). This may entail, for instance, having a moderator from the examining board mediate at a meeting in which teachers from different schools compare how they grade their students. Aside from moderation, however, there are also other ways to ensure some comparability in scoring. Baker, O'Neil, and Linn (1993), in addition to suggesting moderation and specification of the criteria, also mention procedures of monitoring and adjusting the scores of different raters, training, and procedures to double-check scoring.

4.1.5d Taking the student's standpoint

Better problems and more differentiated and flexible scoring rules will not on their own, however, provide a sufficient guarantee of better written assessment. A discrepancy can namely arise within the assessment itself, similar to the discrepancy identified by Galbraith (1993) between the educational philosophy and the assessment philosophy (see Section 3.3.1). The examples given by Cooper (1992) of realistic test problems that were not permitted to be answered realistically (see Section 3.3.7h) speak volumes on this matter.[40] Moreover, the eloquence of these examples is enhanced even further by the solution that was implied. In order to be able to give the exact answer that the test designer had in mind[41], the students just have to learn when they should excise reality from realistic problems and when not.

An entirely different, almost diametrically opposite approach is taken in RME.

Rather than having the student take the standpoint of the teacher, it is the person who is correcting the test (or otherwise assessing) who must try, as far as possible, to take the student's standpoint (see also Sections 1.2.3a, 1.2.4e, 1.2.4f, and 1.2.5d). Or, in other words, thinking along with the students is what matters in both mathematics instruction and assessment (Freudenthal, 1983b; Streefland, 1990b, 1992). It may then turn out that an incorrect answer is not necessarily wrong, but is instead the result of an incorrect question (Freudenthal, 1979c) or of the scoring rules being too narrow (Van den Brink, 1989). The distinction made by Cobb (1987) between an 'actor's perspective' and an 'observer's perspective' is thus a very useful concept for assessment, and fits the viewpoints of RME. Naturally, the designer of an assessment problem did have something specific in mind for that problem, but one must always be aware that students may interpret it differently and thus arrive at a different answer than was anticipated. Moreover, there are now sufficient indications that students will invent and use methods of their own to perform tasks for which they have actually been taught standard methods in school (Bell, 1993). Consequently, every scoring rule must be broad enough to do justice to this. Another possibility – as suggested by Wiggins (1992) – is to involve the students in designing the scoring system. Having criteria that are also clear and reasonable to the students, can – in addition to 'building ownership of the evaluation' – contribute as well to their rise to a higher solution level through the assessment process.

Although the intention may be to take the students' standpoint, the fact that students are not always able to communicate clearly (see Szetela and Nicol, 1992) can make understanding what they mean a problematic and delicate issue. A number of things may make students hard to follow – such as jumping from one idea to another, entirely new perspectives, a 'false start'[42], illegible handwriting, or a peculiar manner of making abbreviations and using symbols. The experiences gained in this area from the test on percentage showed that re-construction activities are also required in order to understand the students' responses (see Sections 4.5f, 4.7b, and 4.8 in Van den Heuvel-Panhuizen, 1995a; see also Scherer, in press). Correcting student work implies, as it were, putting oneself in the students' shoes.[43]

4.2 Revisiting written assessment within RME

Since the reform of mathematics education and the corresponding assessment reform, written assessment in particular has come under fire. In the light of the new education, existing written tests in the form of multiple-choice and short-answer tasks have proved inadequate on a number of points. The gravest objections concern the mismatch between what these tests measure and the altered goals and approach of the instruction, and, also, the lack of information provided by the existing tests about the students' thought processes and their applied strategies (see Section 4.1.1).

In the early stages of RME, it even looked as if written assessment might be entirely banished from education in favor of observation and individual oral tests. Only later was a real search undertaken for written assessment that would be appropriate in RME. What this search produced has been described in the first part of this chapter. The consequences for written assessment of the RME viewpoints on mathematics and on teaching and learning mathematics will be discussed in the part that now follows.

4.2.1 Shifting the boundaries of written assessment

The alternative forms of written assessment developed within RME and the experiences gained from these alternatives have meanwhile dispelled the resistance to written tests and have shown that such tests are not so bad after all. Written tests, too, are evidently able to do justice to the RME viewpoints on:

– the subject of mathematics (as a meaningful human activity) and the educational goals linked to this subject (in the breadth, the depth, and in terms of applicability);
– the manner of instruction (through contexts, models, students' own contributions, reflection, interaction and integration of learning strands);
– the learning processes (the students' own activities, their levels of understanding and of mathematization, and the discontinuous nature of learning processes).

It is necessary, however, that the existing canon of written test problems be abandoned. Such an abandonment, aside from leading to a number of changes in 'outward appearances'[44], also implies certain more fundamental alterations. In both cases, however, this means shifting the boundaries of the traditional method of written assessment.

The most obvious changes involve both task content and test format. The new tests no longer consist solely of bare, isolated problems, but also – and primarily – of realistic and meaningful problems. The emphasis lies on non-routine problems, so that the students can apply what they have learned, which is obviously not at all the same as copying what someone else has demonstrated. The answers, too, often consist of more than simply a number, and may even turn into an entire essay. Another external characteristic is the extensive use of drawings, graphs and tables. Furthermore, all kinds of hybrid forms of written and oral tests and of testing and instructing now exist. The conditions are now, therefore somewhat different from those that were previously customary in assessment situations. Take, for instance, the test-lessons, take-home tests, application of interview techniques (such as providing assistance and continued questioning), and having the students invent problems themselves, rather than just solving problems that were invented by others. What is astonishing, however, is that one can also fall back on traditional forms, such as multiple-choice tasks (see Section 4.1.3b). This shows that it is not the format in itself that determines the quality of the assessment. What is important is that the assessment method and the goal and content of the assessment be geared to one another.

As stated above, a number of more fundamental choices are linked to these 'outward appearances'. These ground-breaking choices – which have arisen from the RME educational theory – are what have made it possible to arrive at this other interpretation of written assessment.

4.2.1a From passive to active assessment

In a view of mathematics education where mathematics is seen as a human activity that can best be learned by doing (see Section 1.1.2a), a passive assessment in which the students merely choose or reproduce an answer will no longer suffice.[45] On the contrary, this view requires an active form of assessment[46], in which the students are given the opportunity to demonstrate that they are able to analyze, organize and solve a problem situation using mathematical means. If the goal of the education is learning mathematization, then this mathematization must also be assessed. Given that the ability to communicate is part of this goal, this usually implies that the students may present their answers in their own words. Assessment in RME is assessment with a broad 'zone of free construction' (see Section 3.2.4b and Section 3.3.3, Note 34).[47]

Consequently, the students' own productions occupy an important place within this assessment (see Section 4.1.3e). The multiple-choice form may nonetheless be quite useful at times, for instance, for determining the strategy when assessing clever calculation (see Section 4.1.3b), and when assessing the ability to estimate. Contrary to what one would expect, multiple-choice can be quite an appropriate form for assessing the latter. Because the possible answers are already present, the student need not calculate and is therefore encouraged to use estimation techniques. The most important criterion for a good assessment problem is that it elicits a certain thought process. This need not always mean, however, that the thought process also becomes visible (cf. Lamon and Lesh, 1992; see Section 3.3.5h).

Another aspect of the activity principle in assessment is that the students, themselves, must be actively involved as much as possible in the assessment. On the one hand, this means that the problems must be chosen in such a way that they contribute to this involvement: they must be meaningful and must provide the students with opportunities to 'own the problem' (see Sections 3.2.4a and 3.3.5b). The open-ended nature of the assessment problems is also important here, as this quality encourages students to assume greater responsibility for the response (Clarke, 1993a; see also Clarke, 1993b).[48] On the other hand, this active involvement may also mean that the students, through self-assessment, can participate both directly and indirectly in the assessment. By designing a test by themselves (see Section 4.1.2a and De Lange and Van Reeuwijk, 1993), they participate directly in the sense that they can provide important information about their knowledge and understanding. By designing an easy and a difficult problem (see Section 4.6d in Van den Heuvel-Panhuizen, 1995a; Van den Heuvel-Panhuizen, Streefland, and Middleton, 1994), they participate indirectly in the sense that, in this way, they are contributing to the development of assessment problems.

Lastly, the students' active involvement – which is so characteristic of this altered approach to instruction – also forms a manifest link with assessment in yet another way: when students are more active in instruction, this namely gives the teachers the opportunity to do on-line evaluation (Campione, 1989).[49]

4.2.1b From static to dynamic assessment

Teaching and learning mathematics is viewed in RME as a dynamic process, in which interaction and 'didactical' role-exchanges[50] elicit reflection and cognitive exchanges of perspective and the rise in levels that this generates (see Section 1.1.2c).[51] In just such a way is assessment also regarded in RME as a dynamic process. In contrast to a static notion of 'instruction as transfer of knowledge', in which the learning paths are virtually fixed and continuous, the standpoint in RME is that different students may follow different learning paths (see Freudenthal, 1979c), and that these learning paths may, moreover, contain discontinuities (see Freudenthal, 1978b). For this reason, RME has always preferred the much more flexible individual oral interviews to the static written tests.

This standpoint also characterizes the socio-constructivistic approach, in which assessment must provide information for dynamic instructional decision-making (Yackel, Cobb, and Wood, 1992). Many mathematics educators now recognize that, through teacher and student interaction, assessment can become a dynamic process and a means of collecting information for guiding instruction and enhancing educational experience (Mousley, Clements, and Ellerton, 1992; Webb and Briars, 1990).[52] According to Joffe (1990), the potential for interactive assessment must be given serious consideration. She describes various ways of accomplishing this, ranging from more structured approaches, in which the possible interventions (such as reporting back, clarification of problems and modification of problems) are standardized[53], to freer and more informal approaches, in which the intervention is left up to the individual teacher.

Another example of a more sensitive method of assessment are the 'probes' developed by Ginsburg, which were designed to be used after administering the TEMA Test[54] in the standard version (see Ginsburg et al., 1992; Ginsburg, Jacobs, and Lopez, 1993). The purpose of the 'probes' is to delve further into the thought processes that produced the observed performance, particularly in the case of errors.

A noteworthy phenomenon is that, alongside the endeavors of mathematics educators, interest in more flexible assessment methods is also increasing within the world of ability testing, due to the perceived limitations of standardized 'static' tests (see Campione, 1989; Campione and Brown, 1990; Burns et al., 1987).[55] The objections to the static tests are (i) that they do not provide information about what Vygotsky called ‘the zone of proximal development’ and hence do not offer specific instructions for dealing with students who are having trouble, (ii) that the results received by these tests are regarded as fixed and unlikely to change, and (iii) that these tests too strongly assume that all testees have had equivalent opportunities to acquire the knowledge and skills being evaluated.

In contrast to the static tests, where the tester is careful to avoid giving any information that might be helpful, in ‘dynamic assessment’[56] a learning environment is created within the assessment situation, in which the student’s reaction to a given learning task is observed. Comparative research conducted into static and dynamic assessment has shown that information about students that remains invisible in static assessment will come to light through dynamic assessment. Furthermore, it would appear that the weaker students in particular benefit the most from this assessment method (Burns et al., 1987; Campione and Brown, 1990).[57]

As can be seen from the previous examples, flexible questioning is generally associated with practical assignments and individual interviews, rather than with written tests (see also Foxman and Mitchell, 1983; Ginsburg et al., 1992; Joffe, 1990). The alternatives developed within RME demonstrate, however, that written tests can be created that break with their traditional counterparts. These traditional tests are characterized by a kind of one-way traffic, where students may submit answers but not ask questions, and where the teachers are not permitted either to substitute an easier question or to offer assistance. Examples of dynamic assessment, on the other hand, are the option problems (see Section 4.1.3d), the students’ own productions (see Sections 4.1.2a and 4.1.3e), the safety-net question (see Section 4.1.4a), the second-chance question (see Section 4.1.4b), and the standby sheet (see Section 4.1.4c). The last three examples are more explicitly taken from interview situations. The take-home test (see Section 4.1.2a), too, is clearly an exponent of this dynamic assessment. Lastly, the dynamic element can also be found in the concept of correcting assessment problems from the standpoint of the student, a notion that is particularly characteristic of RME (see Section 4.1.5d).

4.2.1c From objective to fair

Objectivity in the sense of ‘correcting as a machine would do’ is absolutely out of the question in this approach. Nonetheless, until recently, this was a requirement for assessment, at any rate where there was the assumption of a high-quality test. The test administration, too, was expected to proceed in a standardized manner and, if this did not occur, it was not regarded as true assessment. This might be considered a rather paradoxical situation, since, although non-standardized assessment was not accepted as a basis for important decisions – such as end-of-year evaluations and school-leaving certification – at the same time it was still deemed good enough for most of daily school practice (Joffe, 1990).

As has already been described in detail in Section 3.3.6, there has been a widespread shift in attitude, whereby fairness of assessment now prevails over objectivity and standardization. In RME, this choice was made at an early stage, and led, in the form of De Lange's (1987a) *fourth principle*, to fairness of assessment becoming one of the pillars upon which the RME alternatives to the traditional paper-and-pencil tests are based.

4.2.1d From limited certainty to rich uncertainty

Another issue, related to the last point, pertains to the certainty with which certain conclusions can be drawn. An attempt has long been made in education, by borrowing a concept of measuring from the natural sciences, to determine as precisely as possible what students know and can do.[58] According to Wiggins (1989b), even the root of the word 'assessment' refers to this precise measuring: it recalls the fact that an assessor should in some sense 'sit with' a learner, in order to be sure that the student's answer really does mean what it seems to mean. And yet, since the turn of the century, when Thorndike spurred on the development and use of achievement tests (Du Bois, 1970)[59], the acquisition of certainty has mainly occurred through written tests that have had little in common with the 'sit with' origin of the word assessment. Nevertheless, people have always been optimistic about the potential of written tests to precisely map out learning achievements. During the nineteen-sixties the conviction even prevailed that discoveries could be made through assessment that would rise far above common sense (Bloom and Foskay, 1967; see also Section 1.2.1). Although now weakened, this optimism has not yet disappeared. Even now, it is still stated that the quality of assessment:

> "...hinges upon its[!] ability to provide complete and appropriate information as needed to inform priorities in instructional decision making" (Lesh and Lamon (eds.), 1992, p. vi).

Everything, in fact, is still focused on controllability. This is attested to by the doubts that even proponents of the new assessment methods have recently cast on the open-ended problems, that is, on whether such problems are truly able to accurately portray students' learning (see Clarke, 1993b; see also Sections 3.3.5d and 3.3.5g).

But is there really any reason to strive for this certainty, even in the case of good problems (leaving aside for the present what these may be)? Aside from the fact that not everything a student learns necessarily rises to the surface – simply because not everything is written down or said – there are also a number of other reasons connected with learning that make this desire for certainty questionable at the very least.

- task-specificity

In the first place, there is the all but inevitable task-specificity. Learning is, after all, always a learning-in-context – that is, it is *about* something – and a particular ability always emerges within a particular (assessment) context. According to Joffe (1990, p. 144):

> "...we can say little about ability per se. The wide-ranging differences in achievement

> that can be seen, depending on what questions are asked, who is asking, what mode is considered acceptable for response, and so on, make it difficult to say with any degree of conviction whether achievement in a test reflects some underlying ability."

Wiggins has had the same experience:

> "What happens when we slightly vary the prompt or the context? One of the unnerving findings is: the students' score changes" (Wiggins in Brandt, 1992, p. 36).

Bodin (1993), too, subscribes to this opinion. He found considerable differences, for example, in scores between identical problems that had been included on different tests. Furthermore, there is also the research discussed earlier on the influence of context and wording on the test results (see Sections 3.3.7c and 3.3.7e, and Section 3.3.6b, Note 70).

- discontinuities

  The discontinuities that occur in learning processes (Freudenthal, 1978b, 1991; see Section 1.1.2c) are another reason why certainty is as good as unobtainable. Children sometimes pause in their development, only to leap forward later – a phenomenon that indeed makes any certainty about what one has measured quite precarious.

- no all-or-nothing phenomena

  There is another aspect, related to the discontinuities, that contributes to uncertainty in measuring. Education is now namely aimed at integrated knowledge, insights, and abilities (rather than at isolated skills and facts), whereby one can no longer speak of 'right or wrong' or 'present or absent'. According to Wilson (1992), learners have a variety of understandings and some of these understandings may be less complete than others. Hiebert and Wearne (1988), too, stress that changes in cognitive processes are not all-or-nothing phenomena. Some insights do not immediately emerge entirely and display only a few of their characteristics, while others turn out to be only temporary and disappear again after a while.

- ongoing learning

  Lastly, the measuring of learning results can also be greatly disturbed by the process of ongoing learning during the assessment. An assessment situation, aside from being a measuring situation, is also a situation in which new knowledge and insights are constructed. Learning does not stop during assessment (see Section 3.2.2a, Note 7). Certain forms of dynamic assessment (see Section 4.2.1b), particularly those intended for assessing the learning potential, make deliberate use of this fact. Instead of the assessment focusing on the products of previous learning processes, it is the learning process during the assessment that is, in this case, the actual object of research.

All in all, any certainty that can be obtained through assessment is relative, and some discretion and prudence are called for. Furthermore, the measures that can be taken to increase this certainty may also have the counterproductive effect – due to the choice of illusory certainty – of making one lose track of the task-specificity, the discontinuities and the ongoing learning. Because of the inextricable alliance between these aspects and RME, it is not surprising that this limited certainty has been exchanged in RME for a rich uncertainty; or, put another way, exchanged for the certainty of common sense. The students' responses to the Polar Bear problem (see Section 3.2.8), for instance, while inconclusive with respect to their precise developmental level, did offer footholds for further (assessment) questions and further instruction. Moreover, by not clinging at all cost to the pursuit of certainty, opportunities naturally arise for understanding the students better. An example of this is the safety-net question (see Section 4.1.4a and Chapter 7).

The choice made by RME to relinquish this pursuit of certainty corresponds with Moss' (1994) preference for the hermeneutic as opposed to the psychometric approach. Characteristic of the hermeneutic approach is the important role ascribed to human judgement and the attempt made to develop integrative interpretations based on all the relevant evidence. The goal is to construct a coherent interpretation of the collected performance, in which earlier interpretations can be constantly revised.[60] A salient feature of this approach is that:

> "...inconsistency in students' performance across tasks does not invalidate the assessment. Rather, it becomes an empirical puzzle to be solved by searching for a more comprehensive or elaborated interpretation that explains the inconsistency or articulates the need for additional evidence" (Moss, 1994, p. 8).

This is, furthermore, an approach that is also familiar to research in the natural sciences. The concept of measuring in physics, which has been so determinative for educational assessment, is, in fact, founded on a somewhat obsolete interpretation of measuring in physics. There is namely an inherent fundamental uncertainty in measuring in physics.[61] In other words, measuring in physics actually 'lies closer' to the RME views on educational measuring than had been thought. In order to make the hitherto rather implicit RME ideas on uncertainty in assessment more explicit, a 'mathematical-didactical uncertainty principle' has been formulated analogous to that of the natural sciences (see Streefland and Van den Heuvel-Panhuizen, 1994). The purpose of this uncertainty principle is to create space for further assessment development and for a practice of assessment on a human scale. Whether this principle can truly be regarded as an enrichment of the RME theory, however, remains to be seen.

4.2.1e From *problems* on different levels to *problems and answers* on different levels

One of the most important reasons why the existing written tests were no longer appropriate in the new mathematics education, was the absence of problems requiring higher-order thinking (see Sections 4.1.1a and 3.3.1). The paradoxical aspect of this situation, however, is that attempts had actually been made in the past to ensure that assessment would encompass all levels.

- assessment on different levels with the aid of taxonomies

Bloom's taxonomy, which was published in 1956, long served as a frame of reference for assessment on different levels. Along with goals on the level of knowledge, this taxonomy also distinguished complex higher goals, such as comprehension, application, analysis and synthesis. Many of the mathematics achievement tests which followed were based on 'content-by-process matrices', in which the content dimension was a particular classification of mathematical topics, and the process dimension was a version of Bloom's taxonomy (Kilpatrick, 1993).[62]

The SOLO taxonomy that was later developed by Biggs and Collis (see Collis, Romberg, and Jurdak, 1986) can be regarded as an improved version of this. The difference between Bloom's taxonomy and the SOLO taxonomy is that the levels in the latter were not determined a priori by educators and psychologists, but, rather, a posteriori, based on a large number of student responses. The criterion employed for determining these levels was the complexity of the information to be used – which ranged from an obvious piece of data to using abstract general principles. Afterwards, the levels were then used to develop the earlier mentioned 'superitems' (see Section 3.3.5h), which were designed to assess mathematical problem solving.[63]

- shortcomings of level classifications

Although the assessment results for the superitems did, indeed, confirm this level classification (Collis, Romberg, and Jurdak, 1986)[64], it is not the case that the classification could encompass all levels of mathematical understanding and thus be used for all types of problems, as it only involved the complexity of the information to be used. Moreover, it is questionable whether these findings were actually sufficient for regarding this classification as absolute – as did L.Wilson and Chavarria (1993).[65] M.Wilson (1992) was considerably more reticent on this matter and plainly expressed his doubts about whether all the assumed levels could, indeed, be realized for all problems.

An additional problem is that the inclusion of questions intended to elicit abstract responses can at times work counter-productively, as Collis, Romberg, and Jurdak (1986) were to discover. The students tended to spend a great deal of time on such questions without success, which lowered their motivation to proceed with subsequent problems.

Similarly, in other endeavors to construct a level classification, problems that should lie on the same level because of their structure, in fact display considerable differences in degree of difficulty (Wiliam, 1993; Ruthven, 1987). For Bodin (1993, p. 123) this was even enough reason to state that:

> "...the students' behavior respects neither the taxonomies of objectives nor the a priori analysis of difficulties."

Aside from the fact that the level classifications themselves are usually too simplistic and presume linear learning processes that proceed in the same manner for all students (Joffe, 1992), this absence of the assumed classification from the results has mainly to do, of course, with the way in which the levels are operationalized in test problems. Problems that were intended to measure synthesis (Bloom's highest level), for instance, often did not produce any higher-level outcome at all, simply because they involved nothing but remembering a formula and finding the correct answer through substitution (Wilson, 1992).

And yet, inadequate operationalizations are not the most important reason why level classifications can be unsuccessful. Even the most meticulously designed tests may in fact produce achievement profiles that fail to correspond with the conjectured hierarchies (Ruthven, 1987).

A more vital issue in designing problems on different levels is that a given problem does not necessarily 'include' the intended activity or level of cognitive process (Christiansen and Walther, 1986[66]; Linn, Baker, and Dunbar, 1991; Lesh et al., 1992; Magone et al., 1994). Moreover, the levels must not be regarded as absolute:

> "One of our most important findings of recent research on thinking is that the kinds of mental processes associated with thinking are not restricted to an advanced 'higher-order' stage of mental development. [...] Cognitive research on children's learning of basic skills reveals that [...] arithmetic [...] involves important components of inference, judgement, and active mental construction. The traditional view that the basics can be taught as routine skills with thinking and reasoning to follow later, can no longer guide our educational practice" (Resnick and Resnick, 1992, p. 39).

These last two reasons are precisely why Freudenthal (1978a) raised so many objections to taxonomies like Bloom's. According to Freudenthal, it is the *way* in which the answer to a question is found that determines the attainment of a given level, and not simply the ability to answer the question. He demonstrated, in fact, that the approach taken can even turn an assumed hierarchy upside down (see Section 1.2.4a). In other words, a teacher or researcher may view a given problem quite differently than a student may do (see also Section 3.3.5i).

> "From the conventional viewpoint of the teacher or researcher, the tasks may indeed be structurally identical. Equally, from the learners' perspective, the tasks may be amenable to quite different and distinct constructions, and thus produce very different patterns of response" (Ruthven, 1987, p. 247).

This view plainly breaks with the earlier consensus of opinion, which held that the level of response is primarily determined by the task rather than by the student (Bell, Burkhardt, and Swan, 1992a).

- an alternative

The issue that remains is whether this will mean the end of problems c.q questions on different levels. The answer is, no, definitely not. Even though the value of hierarchies of knowledge and skills is often limited in terms of determining an individual student's cognitive level and thereby obtaining specific information for further instruction, such hierarchies can still function as a global model for helping to understand developments in knowledge and skills. But one must not, according to Ruthven (1987), interpret these hierarchies as stereotypes.[67]

An additional point is that, although their responses do not always stick to the assumed level, the students must nevertheless be given the opportunity to respond on a particular level. And this requires appropriate problems. One should not expect to obtain any higher-order responses from simple, one-dimensional questions. Therefore, it is still very important that questions be asked on different levels. But it is equally important to ask questions that can be *solved* on different levels.

A comparison between the two superitems discussed earlier (see Section 3.3.5h) should provide sufficient illustration of this matter. While, in the Machine problem (see Figure 3.9), the final question goes no further than an all-or-nothing problem, the final question in the Area/Perimeter problem (see Figures 3.10 and 3.11) exposes various solution levels. In addition to producing much more differentiated information for further instruction, such a multi-level problem also avoids creating a situation where higher-level questions become counter-productive by demotivating the students.

This expansion of problems on different levels by including problems that can be solved on different levels signifies for RME a kind of integration of developments in assessment that, in a certain sense, originally occurred separately. The emphasis in secondary education, for instance, lay more on developing problems on different levels (see Section 4.1.2a), while, in primary education, more attention was devoted to inventing problems that could be solved on different levels. Although the latter pertained to nearly all the alternatives invented for primary education (see Sections 4.1.3 and 4.1.4), it was specifically true of the problems with more than one correct answer, the option problems, the problems with multi-level pictures and the students' own productions. Where own productions are concerned, by the way, primary and secondary education have already united.

4.2.2 Conclusion

In conclusion, it may be stated that RME requires a manner of written assessment in which the traditional boundaries of written assessment are shifted in the following ways:

- from passive to active
- from static to dynamic
- from objective to fair
- from certainty to an uncertainty that produces richer information
- from problems on different levels to both problems and answers on different levels.

These theoretical concepts regarding the nature of the assessment, together with the earlier-mentioned criteria that must be satisfied by RME assessment problems, and the practical execution in the form of measures to make written assessment more informative, form the essence of the further elaboration of assessment within the RME theory. This assessment was developed from within the RME educational theory, and cannot, therefore, be considered on its own. This is true for another reason as well, namely, that the influence also works in the other direction.

4.2.2a RME assessment as a source for further development of RME

It is precisely because students' thought processes are such an important breeding ground for RME, that the development of RME assessment can provide the impetus for further development of RME. The analogy to the integration of instruction and assessment in classroom practice can be plainly seen. Just as a teacher can adjust her or his approach according to the assessment findings, so can assessment appropriate to RME – and the assessment results thereby obtained – cause certain aspects of the RME theory to be adjusted, or at least reconsidered and submitted for further investigation. The following topics are certainly eligible for such further development.

- new strategies

The strategies applied by students in solving certain problems are an obvious topic in this context. Making the tests more informative will sometimes reveal very specific information about the students' chosen approach, which can in turn lead to a change in the instructional repertoire. An example of such a change took place during the correction of the first test on percentage, when the teacher in question used the strategy of one of the students to explain the problem to other students (see Section 3.5g in Van den Heuvel-Panhuizen, 1995a).

In a similar way, some of the students' approaches can also contribute on a more general level to an enrichment of RME teaching methods. After all, students have always played an important role in RME in the creation of strategies that would later become part of the established instructional repertoire. Examples of this are, among other things, the tables-symbol and the use of 'pseudonyms' for indicating equivalent fractions (Streefland, 1988; 1991), subtracting 'from the beginning' and 'from the end' (Veltman, 1993) and the informal 'put-it-aside strategy' of column subtraction (Boswinkel, 1995). These are all inventions that were picked up by developmental researchers, thanks to accurate observation and interpretation of what the students meant. On the other hand, had good questions not been posed or good problems presented, the students would not have come up with these inventions; nor would the inventions have become audible or visible in the absence of an approach that leaves some kind of a trace. In other words, good (assessment) problems can ensure that these student inventions will be produced and discovered.

- integration of mental and column arithmetic

The work on assessment appropriate to RME has also brought to light other points needing attention. One of these is the relation between mental and column arithmetic. As the scratch paper used in the MORE tests revealed (see Figure 4.19), students often use all kinds of hybrid forms of mental and column arithmetic, and yet the textbooks – including the RME textbooks – pay next to no attention to this phenomenon.

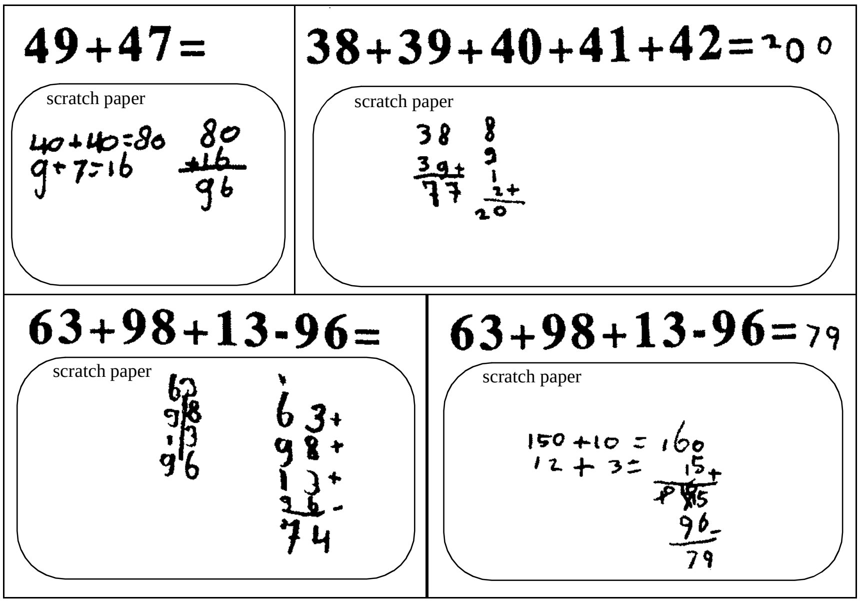


Figure 4.19: Examples of a hybrid form of mental and column arithmetic scratch paper[68]

- rediscovering the value of informal forms of notation

Another point needing attention, related to the above, is the informal forms of notation that are occasionally found on the scratch paper that accompanies test problems (see Sections 4.5f, 4.7b, and 4.8 in Van den Heuvel-Panhuizen, 1995a; see also Scherer, in press). If one takes the students' standpoint and gives them the benefit of the doubt, it becomes clear that they do usually write things that make sense – however imperfect and unconventional the manner of notation. These informal forms of notation, in addition to requiring re-construction activities during correction, also raise the issue of whether more should be done with this form of student expression, or whether the official language in math class should instead remain the language of the textbook.

- more opportunities for students' own contributions

Another point requiring attention involves the opportunities for students' own contributions to the teaching and learning process. This is one of the characteristics of RME and, as such, is nothing new. What is new, however, is that a better grasp of students' informal knowledge can be secured by using a different method of assessment. Consequently, one can proceed in advance of subsequent formal instruction (see Sections 1.1.2a and 1.2.3e) to a greater extent than was previously considered possible. In the MORE Entry-test, for instance, it became apparent that beginning first-graders' informal knowledge could even be traced by means of a written test (see Chapter 5). Other examples of written test problems that can provide insight into students' informal knowledge are the Bead problems (see Figures 2.12 and 4.7), and the Comparing Height problem (see Figure 4.6). Furthermore, the experience gained from the 'Mathematics in Context' unit on percentage demonstrated that, even on a formal topic like percentage, information on the students' informal knowledge can be gathered during class discussions (see Streefland and Van den Heuvel-Panhuizen, 1992). Taken all together, these findings indicate that there is every reason to provide a more explicit place in RME for 'advance testing' than has as yet occurred.

- reconsidering the role of contexts

Just as particular qualities of certain contexts may come to light during instruction, so is it also possible to discover the instructional quality of contexts in assessment problems. An assessment situation, in which the contexts must usually 'function' on their own without extensive explanation, is therefore pre-eminently suitable for determining the accessibility of a context, and for discovering to what extent the context does indeed suggest context-related strategies, which can be used to solve the problems on more than one level. This last point is important in that the context can be used to bring the students to a higher solution-level through interaction and reflection.

Consideration of suitable assessment problems, the role of contexts in such problems, and the way in which such contexts can be 'concretized' within the limited space of an assessment problem have all, in addition to the above-mentioned concrete information on particular contexts, been used to distinguish functions of contexts that use illustrations rather than text (see Section 3.2.6). These functions can also be valuable for instructional activities. Of particular interest in this sense, is the use of illustrations as 'model suppliers', or in other words, contexts that can elicit models. Some examples of this are the Twax problem (see Figures 6, 7, and 8 in Van den Heuvel-Panhuizen, 1995a), and the Father/Son problem (Figure 6.12) and the Bead Pattern problem (Figure 6.11), both of which are from the ratio test. These last two problems are particularly interesting due to the different levels of concreteness that emerge in the models. Another general point, moreover, is that 'assessment in context' has revealed the necessity of further considering and determining a standpoint with respect to the issue of when reality should and should not be abandoned (see Sections 3.3.7f, 3.3.7g, and 3.3.7h).

- further development of the micro-didactics

Another example of how assessment can contribute to the further development of RME is the impetus it can give to further development of the micro-didactics. Instruction, after all, consists to a great extent of asking questions and interpreting answers.[69] Crucial here is that one ask the right question and interpret the answer correctly. This is also precisely the point of assessment, and where instructing and assessing come together – regardless of their specific purposes.[70]

- emphasis on learning processes

  Although the three pillars of RME, namely, the viewpoints on the subject matter, the way in which this should be taught, and the way in which it is learned are also the pillars supporting RME assessment, it has recently become increasingly clear that the viewpoints on learning processes are what particularly influence assessment in RME.[71] Problems involving meaningful contexts and encompassing all goals both in breadth and depth still cannot, in themselves, produce an assessment that is appropriate for RME. For this, especially in the case of 'didactical' assessment, the problems must also correspond with "how one comes to know" (Romberg, 1993, p. 97). In RME, this means that:

  – the problems must contain opportunities for construction (because of the students' own activities)
  – opportunities must be available for a rise in level (because of the levels of understanding and mathematization)
  – the problems must have a certain elasticity or stratification (because of the discontinuities).

  Much of the shift in the boundaries of written assessment was inspired by these views on learning processes. In other words, the brackets that De Lange placed in 1992 around 'learning' (see Section 3.1.1) can now be deleted – something which Freudenthal actually did from the very beginning.

  Furthermore, this can also have a converse effect, meaning that RME assessment can contribute to an increase in attention to learning processes within RME itself. Reflection on and correction of assessment problems means, after all, that one must repeatedly reflect on and acquire insight into learning processes.

4.2.2b A further elaboration of the RME assessment theory?

In conclusion, the following is a brief reflection on the elaboration of assessment within the RME theory, as discussed in Chapters 3 and 4. An endeavor has been made to extend a number of lines from the early stages of RME and to combine these with the findings from subsequently conducted developmental research on assessment, consisting of theoretical ideas, research results and concrete assessment material. Although the emphasis here has been mainly on primary education, an attempt has been made to link up the often separate developments in primary and secondary education and to allow them to cross-fertilize as much as possible. In addition, an effort has been made to explore the entire subject more deeply and to provide it with more contrast by also including research results and theoretical ideas from outside the RME circle.

Since RME is continually being developed further, Chapters 3 and 4 should be regarded as a description of an interim state of affairs, based on current ideas and views on assessment within RME. No attempt at completeness has been attempted here, and much has been left undiscussed. Similarly, the further elaboration of the RME assessment theory is not complete in this sense, but is, instead a kind of theoretical infrastructure, in which a number of links are made both internally and externally, with the aid of concrete examples.

One terrain that has not been examined at all in this study, and that requires further investigation, is that of didactical assessment in actual classroom practice. A small step was taken in this direction in the framework of the test on percentage that was developed for the ‘Mathematics in Context’ project. This was done by interviewing teachers on their assessment practice and by having them develop assessment problems themselves (see Van den Heuvel-Panhuizen, 1995a). The research was conducted in the United States, however, and thus provides no indication of assessment practice in The Netherlands. No investigation has yet been conducted into which of the assessment alternatives developed by RME have actually found their way into the extensive Dutch supply of RME textbooks. Nor has there been any examination of the potential for enriching didactical assessment – in cases where these alternatives have been included.

A number of lacunae also remain when a comparison is made with Webb’s (1992) criteria for aspects that must receive attention in a theory of mathematics assessment. For instance, the theoretical foundation for RME assessment, as described in Chapters 3 and 4, is not responsive to the complete spectrum of purposes of assessment. Due to the emphasis on didactical assessment, things like large-scale assessments, for instance, have been left almost entirely out of the picture.

On the other hand, this theoretical foundation is useful for generating a wide range of assessment situations. And the general issues of validity and reliability are also dealt with, albeit not exhaustively. The same is true of the matter of generalizability, even though this was not dealt with in the way Webb implied. In his opinion, the theory must namely indicate how many assessment situations are necessary in order to perceive what a student knows. Within the elaboration of the RME theory for assessment described here, however, no pronouncements have at any rate been made on this matter. Moreover, it is questionable whether this can even be determined it its entirety. Attention has certainly been paid to the issue of asking questions, but without, however, specifically including a ‘sub-theory of questioning’, as was suggested by Webb.

A number of points also remain upon which it is not entirely clear whether or not Webb’s criteria have been satisfied. One may wonder, for instance, whether this elaboration of the RME theory can, as Webb suggests, help researchers decide whether certain assessment techniques are viable in helping to identify students’ knowledge of mathematics. While Chapters 3 and 4 do contain numerous suggestions for assessment, they do not include any concrete 'operating instructions' for choosing the correct approach in a given assessment situation.

Some of Webb's criteria are so specific, moreover, that it is questionable whether they even belong in every single assessment theory. Examples of this are, for instance: that a theory of mathematics assessment must provide a rationale for using different aggregation procedures; that it must describe and explain the differences in the assessment of mathematical knowledge with and without calculators; and that it must answer the question of why the type of administration of an assessment interacts with the results.

And then there are the criteria that cannot really be reconciled with RME and, therefore, can hardly be imposed on an elaboration of the RME theory for assessment. One of these is the criterion that a theory of mathematics assessment should address the relationship between the assessment of mathematical knowledge and the assessment of other content areas. The same may be said of the criterion that a theory should help make a distinction between domain-specific knowledge and general cognitive skills. Also, the criterion that a theory should address an explicit definition of the content to be assessed and that it would be beneficial to identify the advantages of one approach to content specification over another, is more reminiscent of a 'content-by-behavior' approach, whose focus is controllability, than of a 'mathematical-didactical' or 'didactic-phenomenological' analysis.

To sum up, the idea of a meta-theory for mathematics assessment as advocated by Webb is simply incompatible with RME. The RME views on subject matter, teaching and learning, which are so determinative for assessment, require a 'local' RME assessment theory. This does not mean a 'specific, but isolated' assessment theory, but, rather, one that is embedded in the RME theory in its entirety.

Finally, not all the questions that must (in Webb's opinion) be answered by a theory of mathematics assessment have indeed been answered in Chapters 3 and 4. Nevertheless, the substance of these two chapters can contribute to providing a structure for evaluating what has already been achieved in the area of assessment, and for organizing research into what remains to be accomplished. These two latter requirements were , as a matter of fact, also set by Webb.

Notes

[1] As will be shown in Chapter 6, this is not due to the paper-and-pencil format itself, but to the nature of the problems presented to the students.

[2] Similar recommendations can also be found in Szetela and Nicol (1992).They believe that new techniques must be invented for eliciting better communication of students' thinking.

[3] An example of just how counter-productive this can be is described in Section 3.3.5g.

[4] The RME alternatives described in this section only pertain to those which arose from projects and activities of the Freudenthal Institute and of its predecessor, OW&OC.

[5] 'Positive' testing means that students have the opportunity to show what they *can* do (see

Section 3.2.2d).

6 Besides developmental research on assessment for the HEWET project, this also includes the activities that were conducted in the Whitnall project (see De Lange and Van Reeuwijk, 1993).

7 This does not mean that these assessment methods were exclusively developed within RME. As has been repeatedly indicated in the previous chapter, development of new assessment methods has been taking place worldwide. Similar alternatives to the restricted-time written tests have therefore been developed elsewhere as well.
Assessment methods such as projects and portfolios, which are also mentioned by De Lange (1995), have not been included in this overview, however, as these formats did not originate specifically within RME.

8 De Lange (1987a) calls this an 'essay task'. Because, however, the test in its entirety is meant here, it has been decided to use the term 'test'. The same is true of the following two assessment methods.

9 Although the students may write down the answer to short-answer problems themselves, these are actually closed problems, as they require but one, specific, answer. De Lange (1995) calls them 'closed-open questions'.

10 The concept of problems on different levels can also be found in the superitems, which are based on the SOLO taxonomy (see Section 3.3.5h). The difference between the examples given here and the superitems is that, in the former case, the questions on the various levels need not refer to one single stem.

11 This last example is not mentioned in De Lange, 1995.

12 Among these is a test for fifth grade, that was developed in expectation of a sequel to the MORE research (one of its test problems is discussed in Section 3.2.9), a test on ratio that was developed for a small research project in special education (see Chapter 6), and a test on percentage that was developed for the 'Mathematics in Context' project (see Van den Heuvel-Panhuizen, 1995a and see also Chapter 7). Alongside these activities of the author, the MORE test problems are, among others, also being used as examples in an assessment development project in Sheboygan, Wisconsin (Doyle, 1993).

13 See also Van den Heuvel-Panhuizen (1991a).

14 The Polar Bear problem is an example in which this last point is true.

15 Szetela and Nicol (1992) also point out that student work often contains little process information because the students are proud of being able to make calculations without explanations.

16 These test problems were part of the test that was developed for fifth grade, in expectation of a sequel to the MORE research project (see Note 13 in Section 3.2.9). The text in parentheses was not printed on the test page, but was part of the instructions to be given out loud.

17 Selter (1993c) recently completed an extensive study into the role own productions can play in early arithmetic education that corresponds to these RME ideas on own productions.

18 The Dutch word 'min' means 'minus'. It is interesting to note that the student chose two different ways (symbolic and written) of expressing the two minuses (the operational sign and the verbal sign indicating the value of the number).

19 It was Treffers who, in connection with the results acquired, repeatedly emphasized the importance of these two types of sections, both of which can be produced by own productions. Examples of these sections can also be found in Streefland's (1988, 1991) research into fractions.

20 This can be done by, for example, making overhead sheets of the student work and discussing these in class.

21 This is even more true of procedures like long-division and column multiplication. Here, too, context problems offer more potential for advance testing.

22 It should be mentioned here, however, that the context may sometimes steer the student

towards a less clever strategy. An example of where this occurred is a test problem taken from the PPON research (see Section 1.4.2) on special education, which involves the age difference between Willemien, who is 9 years-old, and her grandmother, who is 63 (see Kraemer, Van der Schoot, and Veldhuizen, in press). Many children applied an 'adding-up' method that suited the situation. Because of the large amount needing to be bridged here, however, errors could easily be made when using this method, especially in the case of counting one by one.

23 This test problem was part of the test that was developed for fifth grade in expectation of a sequel to the MORE research project (see Note 13 in Section 3.2.9).

24 All four students whose work is displayed here were in the same class.

25 Considered in retrospect, the Train problem is actually a kind of pitfall. The visual deception it contains will chiefly entrap students who are not immediately aware that the issue here is one of linear ratios: if the length of the rectangular track has become twice as long, then this must also be true of the entire circumference and, therefore, of the time elapsed as well. Students who, unaware of this, go about figuring out the length of the longer ride bit by bit will be helped rather than hindered by the illustration. This is not the case, however, for the students who follow their first, visual, impression. Reactions to this problem indicate, in any case, that it still needs some tinkering in order to make it really suitable for everyone.

26 According to Sullivan and Clarke (1991), presenting students who have not answered, or incompletely answered a question with 'follow-up questions' is an important task for the teacher.

27 Looking back, this safety-net question closely resembles the 'follow-up questions' referred to by Sullivan and Clarke (1991) (see Note 26) and those posed by Lamon and Lesh (1992) during a number of interview problems. What is striking about the examples given by Lamon and Lesh is that these problems, too, involved absolute and relative comparisons (see Section 4.1.5c).

28 This resembles, to a certain extent, the reflection technique that was applied in the Kwantiwijzer instruments (see Section 1.3.1).

29 See, for instance, the Kwantiwijzer instruments (see Section 1.3.1).

30 Another, somewhat similar method, is the interview procedure developed by Zehavi, Bruckheimer, and Ben-Zvi (1988), whereby sequentially ordered hints were used, and note was made of the first effective hint and its result. In a sequel to this study, these hints were used in a kind of two-stage test. This test was corrected by the teacher and marked with hints that the students could use to improve their initial answers.

31 The unit itself was developed by Jan Auke de Jong and Nanda Querelle as the first version of the unit entitled 'Keeping on Track'.

32 It should be noted – although this aspect will not be discussed further here – that this standby sheet, as well as the other alternative methods of written assessment mentioned in this chapter, can provide significant opportunities for computer application. An example of this is the 'test-with-assistance' developed for research into test anxiety (Meijer and Elshout, 1994) . Some of the ninth-grade students who participated in this research were given hints and, if necessary, an extra explanation during a math test, while others received no hints or explanation. These hints and the extra explanation were displayed on the screen of a personal computer.

33 One must, after all, have norms in order to assign grades. And, in order to determine norms, decisions must first be made about which reference point is to be used: will it be what other students are able to do, what the student in question has already achieved, the educational goals that are to be achieved during a given stage, or a mixture of one or more of these reference points. Besides, assigning grades is also a matter of taste, and depends upon teachers' personal preferences and upon the prevailing standards with respect to this matter at a given school.

34 This standpoint goes back a long way with mathematics educators. Take, for instance,

Weaver (1955). He, too, urged that attention be given to more than just the results, because of the danger of arriving at erroneous conclusions.

35 See also Section 4.2.1d.

36 According to Lamon and Lesh (1992), this framework can also serve as a blueprint for creating new problems. Although it would be interesting to see how the various approaches to mathematics education regard the design philosophy of assessment problems, this, too, is beyond the scope of the present study. In Van den Heuvel-Panhuizen, 1995a, which contains an account of the development of assessment problems with respect to percentage, attention is paid to how this is viewed within RME. More on this topic can furthermore be found in Van den Heuvel-Panhuizen, 1993a.

37 This does not mean that no analytic and holistic scoring scales exist that do contain problem-specific descriptions. Lester and Lambdin Kroll (1991), for instance, give an example of a holistic scoring scale that, on certain points, focuses on a specific problem.

38 Moreover, Szetela and Nicol (1992) do not exclude the possibility of adapting these scales to specific problems.

39 Wiggins (1992), too, is not particularly impressed by general categories and their resulting general indications, such as ‘excellent’ or ‘fair’. Those who check the work must know where to focus their attention. This means that, when setting up a scoring system, it is important to know which are the most salient features of each level or quality of response, and which errors most justify lowering a score.

40 On this point, see also the results of Flener and Reedy (1990). Their research revealed that teachers often negatively evaluate solutions that they had not taught. Less than 25% of the teachers appeared to give credit for creative solutions to the problems.

41 One could regard this as a variant of ‘fishing’ during the lesson. Here, too, the students are continually asked leading questions in order to get them to follow the same train of thought as the teacher. The students must think along with the teacher instead of the teacher thinking along with the students (see, for instance, Gravemeijer, Van den Heuvel-Panhuizen, and Streefland, 1990). According to Wiliam (1993), where assessment is concerned – and especially in high-stake settings – the students are disciplined into adopting easily assessable, stereotyped responses and into playing the game of ‘Guess-what-teacher’s-thinking’.

42 This is one of the issues referred to in the report entitled ‘Measuring What Counts’ (MSEB, 1993b). Problem solving may namely mean that students first make a false start or follow a dead-end path. Although these attempts may not be successful in themselves, they do reflect important mathematical activities, which must be taken into account during the analysis and evaluation.

43 ‘Phonetical correction’, a term recently used by an English teacher to describe a way of discovering what students meant to say (reported in a Dutch daily newspaper: NRC, February 24, 1994), is a fine example taken from outside the area of mathematics education.

44 This is why they are sometimes referred to as ‘tests with a different face’ (Van den Heuvel-Panhuizen and Gravemeijer, 1993).

45 In certain assessment methods that focus on reproducing answers, the students must repeatedly ask themselves what answer the creator of the test problem had in mind. This makes them, in a sense, a kind of ‘second-hand thinkers’ (see De Lange, 1995)

46 Clarke (1986, p. 73), too, used the term ‘action-oriented assessment’. But this had another meaning for him, namely, that the information provided by assessment “should inform the actions of all participants in the learning situation”.

47 In order to emphasize this constructive element, Clarke (1993b) suggests using the term ‘constructive assessment’ rather than ‘authentic assessment’. A case can certainly be made for this, particularly considering the potential misinterpretations that can also be made of the word ‘realistic’ in RME (see Section 1.1.2b).

48 According to Clarke (1993a, p. 8), the “use of the term ‘constructive’ in the context of assessment is more than a rhetorical device.” It means “the progressive transferal of the lo-

cus of authority from teacher to pupil."

49 See also Freudenthal's (1978a) remark that, if instruction is given in a traditional manner, involving little of the students' own activities, then observation will not produce much information either (see Section 1.2.3a).

50 These are role exchanges that are introduced from an instructional perspective. Examples of these are, for instance, the ideas of 'reciprocal observation' (see Section 1.2.5d) and the 'student as arithmetic book author' (see Section 1.2.5c) developed by Van den Brink. Recently conducted class experiments by Streefland (1993; see also Elbers, Derks, and Streefland, 1995), in which the students play the role of researcher, may also be mentioned here.

51 "Learning processes are marked by a succession of changes of perspectives" (Freudenthal, 1991, p. 94).

52 Wiggins (1989b), too, points out that student understanding is a dynamic process requiring some kind of dialogue between the assessor and the assessed.

53 As an example of this, Joffe (1990) mentions an Australian adaptation of one of the APU tasks (see Foxman, 1993 and Section 3.3.2, Note 28). This was an extended task involving the planning of a class trip, in which separate cards were used to vary the complexity of the problem. If the teacher discovered, for instance, that using a timetable was too difficult for some students, this could be removed from the problem in a non-threatening manner and the students would be given a card with an adapted assignment.

54 TEMA is the acronym for Test of Early Mathematical Ability. This test was developed by Ginsburg and Baroody (1990) and is intended to be administered individually.

55 The same can be said of the increased interest in domain-specific abilities within this world. The reason for this interest is that, by focusing on school-like tasks, the leap to instruction can be minimized (see Campione, 1989).

56 In addition to this term, which was first coined by Feuerstein (1979), other terms are also used for this concept, such as 'assisted assessment', 'mediated assessment', and 'learning potential assessment' (Campione, 1989).

57 According to Burns et al. (1987), dynamic assessment may be an important tool for changing teachers' attitude about low-functioning children.

58 It is a pity, however, according to Resnick and Resnick (1992, p. 56), that this measuring by means of educational achievement tests cannot be as unobtrusive as measuring temperature. Forgetting that the measuring itself also takes energy and thus influences the physical system, they state that "we cannot place a 'test thermometer' in a classroom without expecting to change conditions in the classroom significantly." Elsewhere, reference has been made to other physical measuring instruments besides the thermometer, as a clarification of the purpose of the assessment. For example, Vaney, the head of the laboratory school for experimental pedagogy founded by Binet in 1905, was given the assignment by Binet to produce achievement tests – the first of which was an arithmetic test – that might serve as 'barometers' of instruction (Wolf, 1973, cited by Kilpatrick, 1993). Three decades later, the Dutch educator, Diels (1933) (who also wrote a widely used arithmetic textbook), spoke of tests as 'pedagogical measuring sticks'.

59 According to Du Bois (1970), it was Stone, a student of Thorndike, who published the first arithmetic test in 1908. This was actually an improved version of the arithmetic test designed by Rice in 1902, which had been used to test 6,000 children. It is interesting to discover the reason why Rice had designed such a test: "He was appalled by the rigid, mechanical, dehumanizing methods of instruction [...]. When the muckraking articles he wrote failed to ignite the fires of reform, he decided to gather some data to support his claims" (Kilpatrick, 1992, p. 12). In other words, in contrast to what is often the case, this test was not used by Rice to justify a kind of 'back to the basics' movement, but, in fact, to stimulate a discourse on less mechanistic education. (A similar motive was involved in the case of the research conducted in special education, which is described in Chapter 6.) Besides the tests designed by Stone and Rice, there is also the arithmetic test developed

by Vaney at the turn of the century (see Note 58).

60 The same ideas are also expressed by Broadfoot (1994, cited by Gipps, 1994, p. 288) who states: “Assessment is not an exact science.” According to Gipps (ibid., p. 288), there is a paradigm shift: “[...] the constructivist paradigm does not accept that reality is fixed and independent of the observer [...]. In this paradigm there is no such a thing as a ‘true score’. As Lauren Resnick has recently put it, maybe we should give up on ‘measurement’: what we are doing is making inferences from evidence [...].”

61 The reference here is to ‘Heisenberg’s uncertainty principle’, which states that the position and the speed of a particle cannot be measured simultaneously with unlimited accuracy.

62 The process dimension is also referred to as ‘behavior dimension’ (see Webb, 1992).

63 For the sake of completeness, it should be mentioned that Biggs and Collis also distinguished different levels of cognitive development that run parallel to the SOLO levels. This ‘Hypothetical Cognitive Structure’ (HCS) is comparable to Piaget’s classification, the difference being that the levels do not alternate but, instead, develop cumulatively; the later developing modes exist alongside earlier modes (Collis, 1992). Only the SOLO taxonomy was used, however, for developing the superitems.

64 The SOLO levels increased in degree of difficulty; there was a great deal of consistency in the levels recorded for each child as well as for children at the same grade level.

65 Wilson and Chavarria (1993) state that, if the results do not correspond to the level classification, then this probably has to do with a flaw in the wording or content of the problems. They do not discuss the level classification itself.

66 In another context, Christiansen and Walther (1986) call this an over-simplified conception of the relationship between task and learning.

67 Furthermore, according to Swan (1993), one must not make the error of automatically interpreting the empirical degree of difficulty of various problems as being determinative for the order in which something must be learned.

68 These examples are the work of four different students.

69 An intensive study conducted by Sullivan in ten primary school classrooms revealed that nearly 60% of the communications between teacher and students were in the form of questions and answers (Sullivan and Clarke, 1987).

70 Even in this matter, the differences are not always so clear. Take, for instance, De Lange’s (1987a) *first principle*.

71 As Romberg (1993) demonstrates, this is not only true of RME. Viewpoints on learning – from behaviorism to more contemporary beliefs about learning – have always been determinative for the method of assessment. In just this way does the contemporary ‘constructivistic’ view on learning, in which, according to Romberg, the emphasis is laid on the development of cognitive schemes, correspond to the choice of ‘authentic performance assessment’, which informs teachers about schema changes.

Part II

5 The MORE Entry-test – what a paper-and-pencil test can tell about the mathematical abilities of beginning first-graders

5.1 The MORE Entry-test

5.1.1 An unintended research project

It was not the original intention, as a matter of fact, to study first-graders' mathematical abilities at the beginning of the school year, nor to investigate the potential of paper-and-pencil tests for measuring these abilities. Only in retrospect did this take shape, and one may very well wonder how it occurred.

New developments in mathematics education in The Netherlands (see Chapter 1), and corresponding changes in the textbooks raised questions about the implementation and effects of this new, realistic approach. The MORE research project (see Chapter 2, and Gravemeijer et al., 1993) was designed and conducted to answer these questions. In order to discover the effects of the education, it was necessary to collect, among other data, information on the participating students' learning achievements. Since a great number of students were involved in the research, the only feasible way of gathering this information was through written tests. There was some resistance to this method within the research project, as paper-and-pencil tests were not considered a particularly suitable way of discovering students' abilities and strategies. The existing written tests, which chiefly consisted of rows of formula problems, were especially ill-suited for this purpose. Therefore, on behalf of the research, a new battery of paper-and-pencil tests (Van den Heuvel-Panhuizen and Gravemeijer, 1990a) were developed to evaluate students in grades 1 through 3. A comprehensive survey of this test battery and how it was developed can be found in Chapter 2. In addition, attention was paid to information that this research produced on designing problems, which formed the basis for further developments in written assessment. The present chapter will focus on the initial test of this battery – the 'entry-test' – and on the results it produced.

5.1.2 Some crucial decisions made regarding the entry-test

In order to answer as precisely as possible the question of what effect the type of mathematics education had on students' achievements, it was necessary to determine the students' skill level at the beginning of the school year. In other words, it was essential to discover what kind of numerical knowledge and skills the students already possessed when they first began to receive systematic instruction. In The Netherlands, not counting kindergarten preparatory activities, systematic mathematics instruction begins in first grade, when the children are six years-old.[1]

The assessment was to take place during the initial weeks of first grade. This meant that a test had to be developed that would be suitable for children who had not yet had any systematic mathematics instruction. The test therefore had to be designed in such a way that it could be taken by students who possessed no specific school-arithmetic knowledge. Consequently, a considerable effort was made to find links with various kinds of natural, everyday situations involving numbers and quantities.[2] Such links with real-life situations do not imply, however, the unsuitability of less realistic situations. What is important, is that the students can imagine something in the situation. The following sections will describe how this entry-test was developed, and what it produced in the way of data.

The choice of which mathematical knowledge and abilities to investigate stemmed from the traditional topics at the heart of the first-grade curriculum. The test was designed to respond to questions such as:

– have the students already mastered segments of the counting sequence?
– can they determine how many objects there are?
– can they already work with number symbols?
– can they already perform certain operations with quantities or numbers?

At first glance, this choice of topics would appear to be a fairly obvious one. Readers who are more familiar with early mathematics education, however, will notice the absence of all sorts of prerequisite abilities, such as classification, seriation and conservation. This absence was intentional, however, for two reasons. The first reason is closely connected to the nature of the MORE research, which was one of measuring effects and making comparisons. Because measuring effects involves, among other things, mathematical abilities, an initial measurement of these abilities will provide the best guarantee of a precise determination of potential effects. The second, more general, reason, has to do with the doubts that have arisen in recent years with respect to the relation between these so-called prerequisite abilities and arithmetic skills. Various studies have revealed that children are in fact capable of doing operations with small numbers while yet unable to solve conservation problems (see Gelman and Gallistel, 1978; Groen and Kieran, 1983; Hughes, 1986).

Administering the test at the beginning of the school year also meant that most of the first-grade students would not yet be able to read or write. The ideal choice would have been an individual, oral interview. This was not feasible, however, given the fact that more than 400 students would be participating in the research. So, in spite of the inherent drawbacks, there was no alternative but to administer a written test.[3] In order to circumvent the issue of reading and writing, it was decided to have the test instructions be given orally. In other words, the teacher would read the question and the instructions for each problem out loud. In order to prevent this becoming a test of the students' memory capacity, all relevant numerical information was printed on the test page.

It was also imperative, due to the size of the research group, that the assessment data be of a kind that could be processed by computer. This requirement, coupled with the inability of beginning first-graders to write out the answers themselves, led to the choice of a closed-problem format. That is to say, both what was asked and what could be answered were fixed. The students would answer by checking off one of the available possibilities. In order to acquire some insight into what the students were already able to do on their own, a number of more or less open problems were also included on the test. These problems involved student input by, for instance, allowing the students to choose which numbers to work with, or by presenting them with a choice of which problem to solve.[4] In order to rule out strokes of luck as far as possible, more than one problem was included for each topic to be assessed, and numerous answer possibilities were given. Because of the age of the students, the test could not be too long; the target length of the test was a half-hour.

The final version of the test – that is, the version that was used for the MORE research – can be found in the Appendix at the end of this chapter (see Section 5.4).

5.1.3 From trial version to final version

The test was developed in two stages. First, a trial version was designed, which was then administered to a kindergarten class at two different schools, neither of which participated in the actual research. The testing of the trial version took place at the end of the academic year. Only the kindergartners who would be in first grade that fall took part (see Section 5.1.2, Note 1). Based on the experiences with the pilot testing, certain aspects of the test were then adapted. The following description of these alterations anticipates, to some extent, the content of the final version. A separate description of the definitive content is given in Section 5.1.5.

The most conspicuous alteration involved replacing the sketches with professional drawings.[5] This was actually much more than merely a cosmetic change. The uniformity of the drawings helped give the test a sense of homogeneity. Although the contexts vary, the environment remains the same, and becomes recognizable. Moreover, the style of the drawings radiates a sense of tranquillity and friendliness.

A number of changes were also made in the problem content. Certain problems, for instance, were altered due to a lack of clarity, which could cause them to be interpreted differently than had been intended. An example of this is problem number 18 (see Figure 5.1), in which the students were asked to find the number of points scored in a pinball game.

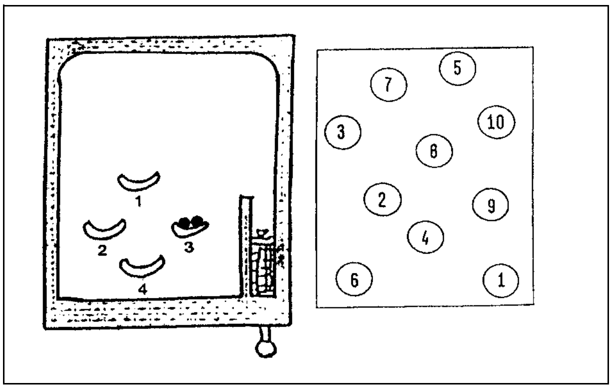


Figure 5.1: Problem 18, trial version

This was the second problem in this particular series. In the first problem, one pinball had been placed at ‘1’ and one at ‘3’. In the second problem, for fear that problems such as ‘3 and 4’ would be too difficult for the children, the decision was made to use ‘3 and 3’. After all, children master this kind of doubling quite early. During the trial run, however, many students crossed off ‘3’ as the answer (evidently in order to indicate that the pinballs were at the ‘3’). Apparently, for these students, the fact that the two pinballs were at ‘3’ masked what was really being asked, namely the total score. In order to avoid such an interpretation, in the final version of problem 18 (see Appendix) it was decided to place the two pinballs at different numbers. And, indeed, ‘3’ and ‘4’ were the numbers used after all. In the meantime, the trial run had revealed that the difficulty of such problems had been overestimated.

The most substantial alterations were made in the Shopping problems. These are problems in which the students have a certain amount of money with which to buy something, and are then asked to indicate how much money they have left. Problem number 24 (see Appendix) is an example of such a Shopping problem. The changes made in these problems chiefly involved the arrangement of the information on the test page. In the trial version, the money was presented as a bill of paper currency, and the different answers were printed on rectangles that looked like tickets (see Figure 5.2).

Later, because these tickets could be associated with paper currency, it was decided to print the possible answers on an open field (as was done in the other problems). The paper currency was also replaced by a coin-purse. The exact reason for this substitution can no longer be retraced, but the change had to do with increasing the students’ involvement: “You have 10 guilders in your coin-purse.” Before the decision was made to use a coin-purse, one suggestion had been to draw a piggy bank.

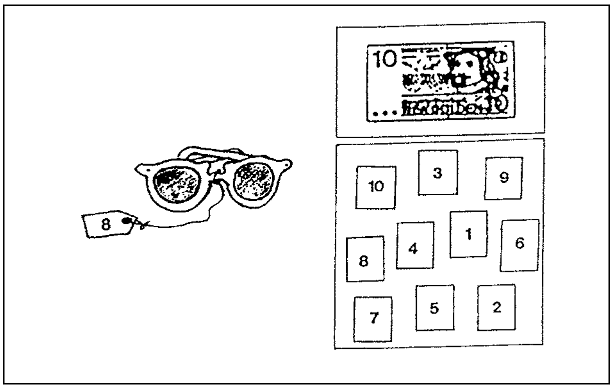


Figure 5.2: Problem 24, trial version

The test also contained two open Shopping problems in addition to the two closed ones. Here, too, the final version turned out somewhat different from the trial version. This had to do with the struggle between, on the one hand, making the problems open, and, on the other hand, not making them too open because of the insurmountable problems this would create for the data processing. This area of tension between open and closed was even apparent in the trial version. The starting point for designing the problems was clearly closed – because of the impending difficulties in processing the data. The problems were then gradually made more open, through the application of various possible choices.

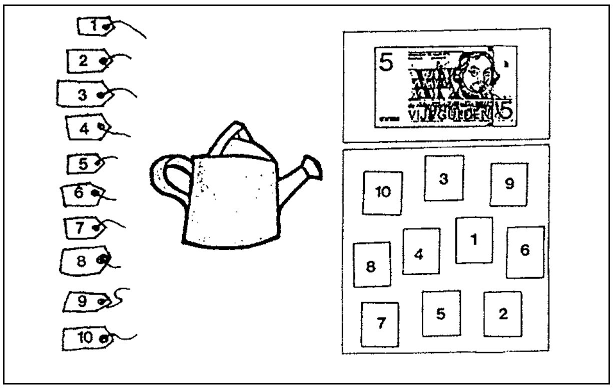


Figure 5.3: Problem 25, trial version

In the trial version of problem number 25 (see Figure 5.3), the students were first asked to determine a price for the watering can, then to pay for their purchase and, finally, to cross off the amount of money they would have left. When the test was administered, it was obvious that some children had lost their way entirely, because of the number of steps involved in this problem. It should not be forgotten, by the way, that the instructions were given to the class as a whole. Furthermore, the problem required the students to think ahead: if you only have 5 guilders, then you can't buy a watering can that costs 8 guilders. The events surrounding the trial version of this problem (and problem number 26) ultimately gave rise to the concept of the 'option problem', which was later to return on repeated occasions (see Section 4.1.3d).

In the final version of problem 25 (see Appendix), which took the form of an option problem, the students still had a choice. But the choice was now which item to buy, rather than its price. Although choosing which item to buy did involve choosing a price, this was a very different kind of choice than had to be made in the trial version. Aside from making the problem considerably less complicated, it was also seen as an advantage in the final version that no incorrect choices could be made. In other words, the prices now all fit the students' purse. This last argument, however, as was revealed later, could also be criticized as being somewhat shortsighted.

A great deal of revision work went into how the problems were worded. The object was to express the purpose as precisely as possible, but also to keep the text short, simple and natural. An example of this is problem number 13 (see Appendix), which was designed to ascertain whether the students could count backwards. In the end, the following instructions were given:

> "Here you see a rocket blasting off.
> At the bottom of the page, you can see where they're counting down.
> They're not quite done yet.
> Which number is next?
> Look at the cloud and cross off the number that will be next."

The most problematic issue was what to call the number to be crossed off.[6] In imitation of counting upwards, which had been assessed in the previous problems, the first phrase tried was "Which number will follow?" This was quickly replaced, however, by "Which number comes before?", because the problem was about counting backwards. But this could also cause confusion, as it was not clear which number was meant – up or down. During the trial run, in the presence of the children, the best wording just came up spontaneously: "Which number is next?" [7]

A spontaneous occurrence during the pilot testing was the count down – out loud – by the test administrator. Some of the children, however, couldn't contain themselves and gave away the answer by joining in. In order to prevent this, the counting out loud was removed from the final version. The empty cloud next to the large, number-filled cloud was removed as well (see Figure 5.4), because some of the children had written a number in this empty cloud, using it as the answer box. Not that there was anything specifically wrong with doing this, but due to a concern that it would lessen control of the assessment, it was removed. Unjustifiably, perhaps.

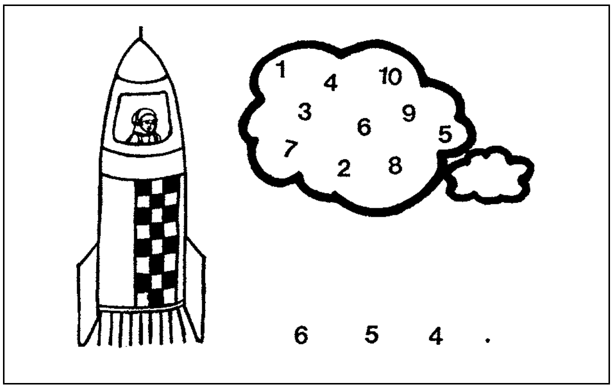


Figure 5.4: Problem 13, trial version

Since the duration of the test did, as estimated, turn out to be about a half-hour, no changes were made in the number of problems included. Nor did crossing off the answers cause any difficulties. The children used a pencil, and were told beforehand that if they made a mistake they should circle it. This was because erasing answers would take too much time and might make it ambiguous which answer was the intended one. The circles worked well. Some children did circle their answer instead of crossing it off, but this did not cause any confusion when the answers were corrected. After all, if there was only a circle, and nothing crossed off, then that had to be the intended answer.

5.1.4 Results of the trial version

The degree of difficulty was also left as it was, even though the two oldest groups of kindergartners that participated in the trial run did manage to answer a great number of problems correctly (see Figure 5.5). The highest score for the 28 problems was 27 correct, and the lowest was 8 correct. The average in both classes was 17 correct.

As stated, in spite of these high scores, the degree of difficulty was not increased for the final version. The reason behind this was the conjecture that perhaps these were two unusually bright kindergarten classes. Not that there was any reason to suspect this, aside from these test results. Nevertheless, to avoid scaring off the participants by presenting them with a too difficult test right at the start, it was decided not to make the test any more difficult. This did prove to be the right decision. Even though the level of difficulty was left unchanged, the teachers who participated in the study expressed their concern about the arduousness of the test when they saw the final version.

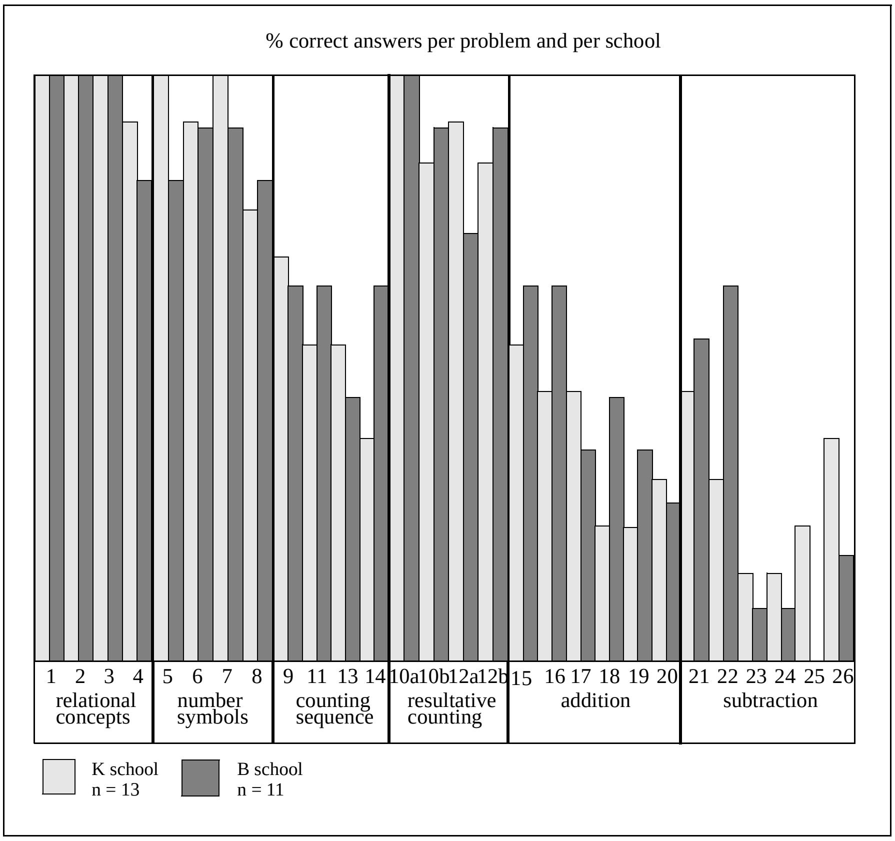


Figure 5.5: Results of the trial version of the MORE Entry-test

5.1.5 The content of the MORE Entry-test

The final version of the entry-test (which can be found in the Appendix), consisted of the following sections:

– relational concepts
– number symbols
– counting sequence
– resultative counting
– addition and subtraction in context

- relational concepts

In order to avoid assailing the students with ‘true’ arithmetic right away, four questions involving elementary relational concepts were included at the beginning of the test, namely: highest, smallest, thickest and most. These questions were quite similar to tasks found on ‘play-worksheets’, which are often used in kindergarten. In problem 1, for instance, the students were asked to cross off the highest building, in problem 2, the smallest ship, in problem 3 the thickest tree, and, in problem 4, the child that was holding the most balloons. This last problem offered the opportunity to discover whether the children understood the concept of number. Specifically, the point was to see whether their understanding of quantity was firm enough to no longer be distracted by the size of the balloons.

As mentioned above, these initial problems were kept as simple as possible. Only two or three choices were given in each problem, and the students were merely asked to cross off one of them. The drawback, however, was that crossing off the correct object did not automatically mean that the student had indeed understood the concept of number.

- number symbols

The four introductory problems were followed by four problems designed to ascertain the students' knowledge of number symbols. Specifically, could they recognize the numbers 3, 5, 10 and 14? All sorts of situations where children encounter numbers in everyday life were sought for these problems. It should be noted, however, just as with the other problems, that less realistic situations were not avoided either. In problem 5, the assignment was to cross off the number three on a license plate. In problem 6, the students were to cross off the hobbyhorse that had the number five. Problem 7 involved the numbers you pull when waiting to be helped at a bakery; in this case, the students were to cross off the number ten. And, in problem 8, the students were to cross off the soccer player who was wearing number fourteen.

In addition to determining whether the students were already familiar with number symbols, these problems also performed a kind of control function for the test itself. Written answers were necessary because the test was to be administered to an entire group, but, because of the age of the students, one could not assume that every child was already able to write numbers. For this reason, the students were to give their answers by crossing off the correct number. If, after administering the four problems on number symbols, it became clear that a given student was not yet able to read or recognize these symbols, then there would be no point in administering the rest of the test. Also, in order not to hamper the students beforehand in following the test instructions, drawings of little animals were used instead of page numbers.

- counting sequence

Here, too, the idea was to find a situation where children might encounter the counting sequence in a more or less natural fashion. A well-known Dutch board game, called the 'Goose Game' (which is similar to 'Candyland'), was the obvious choice. In the problems, only some of the numbers on the board were filled in. The children were asked: "Which number is next?" The object was to discover whether the students were able to find the next number in a given segment of counting sequence. In problem 9, the numbers one through four had already been filled in, and the students were to cross off the number five. Problem 11 was similar, only this time the numbers one through seven had already been filled in. These two problems caused a logistical difficulty: if they were printed next to one another in the test booklet, then the answer to problem 9 could be found from the game-board printed in problem 11. For this reason, the two board-game problems were presented alternately with problems involving a different skill, namely, that of resultative counting.

For counting backwards, however, it was more difficult to come up with a natural setting than for counting upwards. After all, when does one count backwards? Solutions are often found at unexpected moments. Suddenly, the thought of the televised NASA space shuttle lift-offs, and the preceding count-down came to mind. That gave the impetus to a new test problem. In order to avoid any association with weaponry, an astronaut was drawn looking out of the space-shuttle, to make it clear that this was a manned space flight. In contrast to the upward-counting problems, the counting backwards problems – numbers 13 and 14 – could be printed next to one another without there being any danger of copying, because they involved different segments of the counting sequence. It should be noted that these problems could only determine to a limited extent whether the students had mastered the counting sequence. In fact, the problems merely assessed whether the students knew one succeeding and one preceding number. Nevertheless, this choice was made due to the assumption that writing down part of the counting sequence would be too difficult for the students.

- resultative counting

  For the topic of resultative counting, the students were asked to color in a number of marbles. This is actually an indirect manner of resultative counting. Instead of determining the number of marbles, the students were to color a given number of marbles. In problem 10a/b, the students were first asked to color two marbles green, and then, on the same page, to color five marbles blue. Problem 12a/b was similar, only there the students had to color seven marbles yellow and then nine marbles red. The pattern of rows of five was purposefully chosen as an aid to the students. An entirely open problem had even been considered, in which the children themselves could draw seven marbles and nine blocks (or other objects). This would have supplied information on the degree and manner of structuring as well. But this idea was eventually rejected, due to the length of time all that drawing would have required. Moreover, the drawing activity might have diverted attention from the actual point of the problem, namely, resultative counting.[8]

- addition and subtraction in context

  This topic was used to ascertain the students' abilities in performing certain operations on numbers. It involved addition and subtraction up to ten. No formula problems were included. The students' abilities were assessed by using play situations that would be familiar to them from outside school as well.

The operations of addition and subtraction were assessed on two levels: with countable objects on the test page, and with number symbols substituting for the countable objects. Problems 15 and 16 involved the countable variant of addition. Again, the ‘Goose Game’ was used for this purpose. Two dice had been ‘rolled’ on the test page; the question was where the playing piece would end up.

Problems 17 and 18 involved the non-countable variant of addition. Here, addition was assessed with the aid of an illustrated pinball machine. Two pinballs had already been played, and the students were asked how many points had been scored in all.

Problems 19 and 20 were open tasks. This time, the students themselves could decide where the pinballs would lie. The purpose of this task was to determine whether the students were up to such an open problem and, if so, to see what their numerical choices would be. At first glance, such a task looks quite simple; after all, the students could choose easy numbers if they so wished. Viewed more closely, however, it becomes clear that the task is not, in fact, so simple. In order to take advantage of this (by choosing easy numbers), students must be able to discern this possibility, and must know which numbers are easy for them to add up. In any case, such a task requires input from the students, themselves.

Subtraction was conducted in the context of shopping. Again, countable objects were presented first. In problem 21, the students were asked how many fish had been sold, while, in problem 22, the question was how many balloons had been sold that day.

This was followed by the non-countable variant of subtraction, first in the closed form, and then more open. In problems 23 and 24, it had already been determined what would be bought. The students were only asked how much money they would have left in their coin-purse.

In the subsequent two problems, numbers 25 and 26, the students themselves could choose which item to buy. Which subtraction problem would then be done was created by their choice. They could choose to do an easy subtraction in problem 25, for instance, by buying the pencil. Or, they might think: “I like the doll better, I’m going to buy that.” As in the preceding open addition problems, the purpose was again to determine whether the students were able to deal with a relatively open task, and to see what their choices would be.

Aside from the difference in presentation, that is, countable objects versus number symbols, these problems also differ in another respect, in that they indicate two different underlying mathematical structures. In terms of how to calculate, it doesn’t make much difference. Both the problems with fish and balloons and the problems involving money can be solved using subtraction: 6 fish minus 1 fish, and 10 guilders minus 8 guilders.[9] However, the mathematical structure of the two types of problems is different. The structure of the problems involving fish or balloons, in which one must figure out what is being subtracted, is ‘6 – . = 1’ (missing subtrahend). The decision to use this mathematical structure in this concrete form of presentation was a fairly obvious one, because, thanks to this presentation, the disappearing second term was not problematic. In the problems involving the coin-purse, in which one must figure out what will be left over, the underlying structure is '10 – 8 = .' Here, in spite of the mathematical structure, there is not really a disappearing second term, thanks to the price context, which involves a subtraction that has yet to take place. Evidently, during problem development, the search for a suitable context in which the students could show their ability to solve subtractions was given a higher priority than the maintenance of one and the same mathematical structure.[10]

Another disparity – which is related to the difference in mathematical structure – is that the two types of subtraction problems refer, in a certain sense, to different manifestations of subtraction. The problems with fish and balloons involve subtraction as 'determining the difference' (between the initial amount and the amount left at the end of the day). The problems on the contents of the coin-purse, on the other hand, are clearly about subtraction as 'taking away'.

In this respect, the addition problems were more similar to one another. Both the board game and the pinball machine were suitable for either 'combining' or 'adding' as different manifestations of addition. For instance, if one die is rolled and the playing piece is moved to the requisite spot, whereupon the second die is rolled and the piece is moved further, then this is a case of adding. But if the dice are rolled simultaneously, this can be seen as combining the number of dots. The presentation of the addition problems was such that the students could, to a certain extent, choose the form that they found the most appealing. The test instructions did, as a matter of fact, lean more towards combining.

5.1.6 The administration of the MORE Entry-test

The test was administered to 22 first-grade classes, three weeks after the start of the 1987 school year. The participating schools were quite heterogeneous; there were urban and rural schools, schools with a high percentage of children whose first language was not Dutch and schools that were predominantly filled with native Dutch children, schools that used a realistic mathematics textbook and schools that followed a more traditional one. 441 first-grade students took the test. The teacher of each class administered the test according to an instruction booklet that prescribed every detail.

After the test was administered, members of the research staff collected the test booklets from the schools and corrected them.[11] A code book that described how different responses should be scored was used for this purpose. The research staff had discussed the code book beforehand. The responses were then converted into correct/incorrect scores. It had been explicitly agreed that responses that were not entirely in accordance with the instructions but that were nevertheless correct, should definitely be considered correct.

5.1.7 The results of the MORE Entry-test

As can be seen from Table 5.1 and Figure 5.6, the test scores were quite high.

Table 5.1: Psychometric data on the MORE Entry-test

Total number of problems	28	
Maximum score	28	(n = 31)
Minimum score	6	(n = 3)
Mean score	21	
Standard deviation	5.2	

An average of 21 problems (or 75%) of the 28 problems were answered correctly. There were even 31 students among the 441 who answered every single problem correctly.

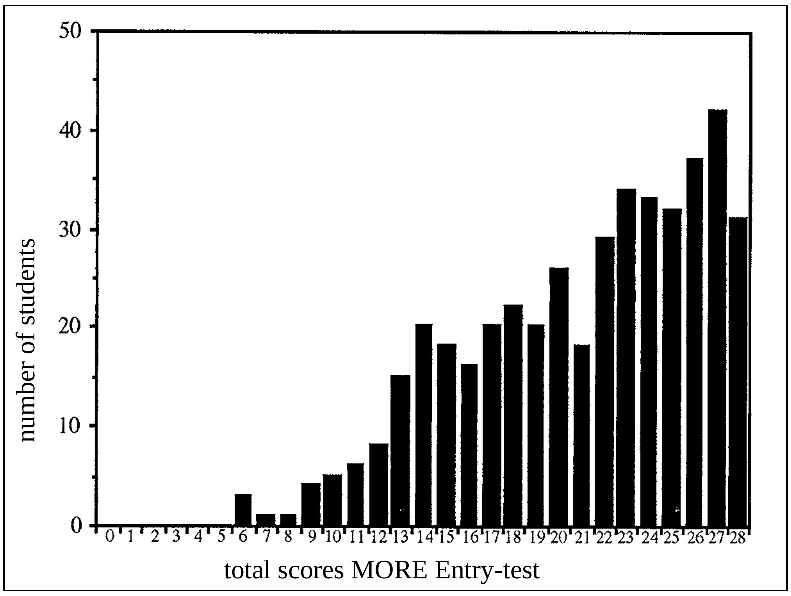


Figure 5.6: Distribution of the total scores from the MORE Entry-test

Certain supplementary data was collected by a detailed analysis of the scores for each problem. The p-values of the problems ranged from 0.99 (crossing off the thickest tree) to 0.39 (the final problem, in which the students could choose what to buy). The internal homogeneity (Cronbach's alpha) of the test was rather high (0.88). The most dominant problems in this respect were the four involving the pinball machine. Particularly high correlations were found in the problem-pairs that involved addition, for instance, between the two problems that used the board game.

The conclusion may be drawn from the p-values per problem (see Table 5.2) that children at the beginning of first grade can already do quite a bit in the area of arithmetic.

Table 5.2: P-values for problems on the MORE Entry-test (n = 441)

Test problems	% correct answers	Test problems	% correct answers
relational concepts		**addition and subtraction within a context**	
1 highest	97	15 board game (countable, 5 + 1)	80
2 smallest	98	16 board game (countable, 2 + 4)	78
3 thickest	99	17 pinball (non-countable, 1 + 3)	53
4 most	93	18 pinball (non-countable, 4 + 3)	43
number symbols		19 pinball (non-countable, open)	49
5 number 3	97	20 pinball (non-countable, open)	52
6 number 5	97	21 fish (countable, 6 – 1)	64
7 number 10	97	22 balloons (countable, 7 – 3)	55
8 number 14	81	23 coin-purse (non-countable, 5 – 2)	60
counting sequence		24 coin-purse (non-countable, 10 – 8)	44
9 after 4	86	25 coin-purse (non-countable, open, 7 – ..)	42
11 after 7	84	26 coin-purse (non-countable, open, 9 – ..)	39
13 before 4	59		
14 before 8	65		
resultative counting			
10a 2 marbles	99		
10b 5 marbles	97		
12a 7 marbles	95		
12b 9 marbles	84		

Nearly all of the students had mastered the relational concepts and were familiar with numbers up through ten. With respect to knowledge of the counting sequence, the great majority of the students were able to name the following number. This was not true of the problems involving a previous number, which were answered correctly by only slightly more than half the students. Nearly all of the students had also mastered resultative counting up to ten. In the topic comprising addition and subtraction in context form, the students proved to be quite good at adding small numbers, particularly when the tasks involved countable objects. But even the problems in which the numbers were solely indicated by number symbols were answered correctly by about half the students. In the subtraction problems, the scores were somewhat lower, and there was less of a definitive separation between problems with and without countable objects.

5.1.8 The estimated results

A number of experts were presented with the entry-test (but not the results) and asked their opinion on its degree of difficulty (see Van den Heuvel-Panhuizen, 1989a). This took place during the Seventh PANAMA Conference in Noordwijkerhout, The Netherlands in 1988. Four groups of experts were asked to estimate per problem the percentage of students that would respond correctly, given a class administration of the test at the beginning of first grade. Each group of experts consisted of four to five persons, all of whom were active either in school consulting, teacher education, educational research or educational development. Only a few of the experts were primary school teachers. The estimates of the four groups of experts are shown in Figure 5.7, along with the actual scores. The black column depicts the actual scores, while the four adjacent columns represent the estimates agreed upon by each of the four groups.

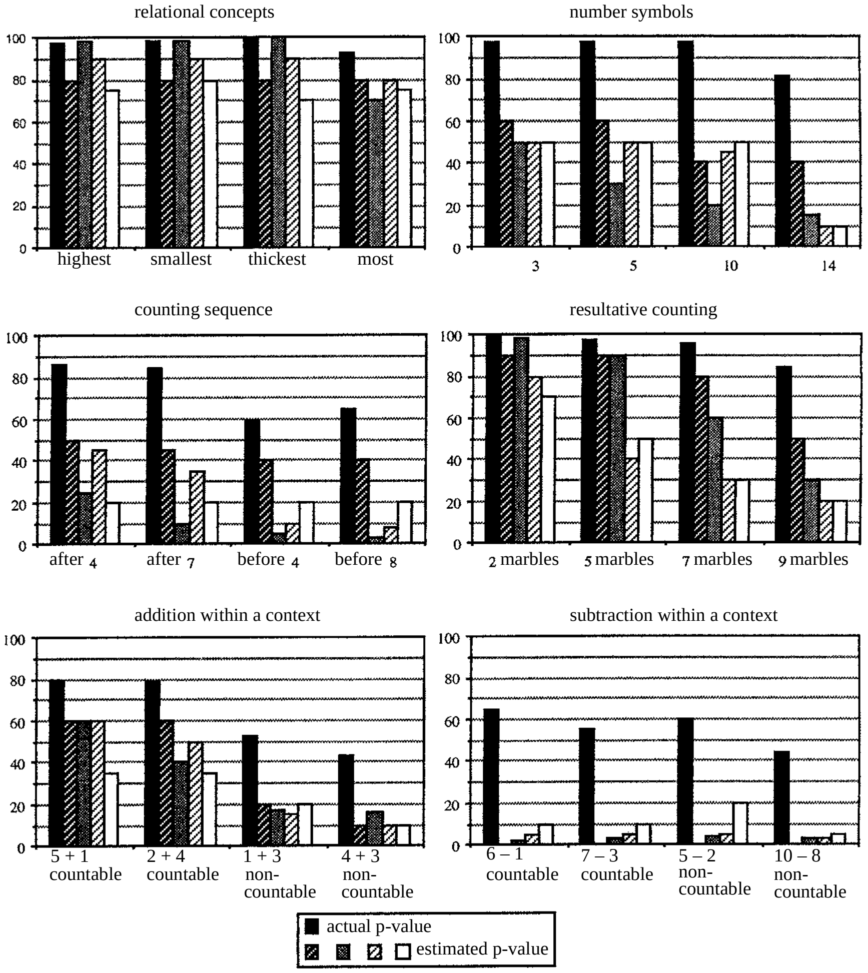


Figure 5.7: Actual and estimated p-values for the problems on the MORE Entry-test (n = 441)

The experts expected full or considerable mastery of the relational concepts, but their expectations with regard to the knowledge of number symbols were much lower. They thought that about half the students would know the numbers up to ten. The experts held even lower expectations with regard to the counting sequence; here, they only expected about one-fourth of the students to answer correctly. Approximately the same estimates were made for the resultative counting of seven and nine objects. The lowest expectations were held in the area of addition and subtraction. The experts thought that a fair percentage of the students would succeed with countable objects, but this estimated percentage dropped sharply for the non-countable variant. The problems involving subtraction received the very lowest estimate. On the whole, with the exception of the relational concepts, the experts underestimated the children's abilities enormously. Clearly, children at the start of first grade possess substantially more mathematical knowledge and abilities than these experts, at any rate, gave them credit for.

It will thus be no surprise to hear that the experts were astonished when confronted with the actual results. Nor were these experts the only ones to have underestimated such students' mathematical knowledge and abilities; one should keep in mind that the difficulty of the test was not increased after the results of the trial version were received. Although earlier researches had reported the considerable abilities of children of this age, these results were nonetheless entirely unexpected.[12]

5.2 Revealing, but nothing new

5.2.1 Earlier research

Research into children's mathematical knowledge and abilities before they have received formal schooling is nothing new. Ginsburg (1975) describes Binet's research in this area. In 1890, one of the things Binet determined through an experiment was that his four year-old daughter, Madeleine, could already compare amounts like 18 and 17, even though she was at that time only able to actually count three objects. Moreover, Binet had laid these numerous objects close to one another, and had distributed them irregularly.

Ginsburg (1975) also mentions the research of Smith, who, in 1924, reported on his investigation into first-grade children's everyday use of arithmetic. Five-hundred children were asked questions on the mornings of twenty-five successive school days. They were asked to recount everything they had done from the time they left school the afternoon before until returning to school that morning. Smith's study revealed that 30% were involved in shopping transactions, 14% in reading the roman numerals on clocks, 13% in reading arabic numerals and in finding certain pages in books, 6% in sharing food with playmates and pets, and 5% in depositing and withdrawing money from piggy banks.

In his own research, which he conducted together with Brush (Brush and Ginsburg, 1971, cited by Ginsburg, 1975), Ginsburg presented three different problems (involving the comparison of the number of marbles in two jars) to children 4 to 6-years-old. The most difficult problem, an 'addition and inequality' task, is described in Figure 5.8.

> *The researcher established an initial inequality, by placing 16 marbles in jar A and 12 marbles in jar B. After the child had acknowledged this, the jars were hidden, and the researcher then added one more marble to jar B. The child was then asked to determine again which jar had more.*

Figure 5.8: ‘Addition and inequality’ task

The results revealed that a majority of the children (19 out of 26) were successful at this task.[13] In a later study by Brush (1972, cited by Ginsburg, 1975), 50 out of 60 children answered this ‘addition and inequality’ task correctly.

Based on this and numerous other studies, and even including studies on ‘animal mathematics’, Ginsburg (1975) came to the conclusion that, prior to entering school, children already possess an informal knowledge of mathematics. Through spontaneous interaction with the environment, children develop various techniques for coping with quantitative problems. For Ginsburg, the educational implications of this discovery were that primary school education cannot proceed smoothly and effectively without devoting attention to children’s informal knowledge, because children use this informal mathematical knowledge as a frame of reference for assimilating the arithmetic taught in school.

Around the same time that Ginsburg was making these pronouncements in the United States, Koster (1975) was drawing similar conclusions in The Netherlands. These conclusions were based on the doctoral research of Westerveld (1972, cited by Koster, 1975). In this research, two groups of kindergartners (one group born in September and the other in October) (see Section 5.1.2, Note 1) were given a number of adding-up problems at two different instances (once in June, and then again in March of the following year). A comparison of the two measurements revealed that the group that had gone on to first grade had done no better after seven months of formal schooling than had the other group that had remained in kindergarten. The kindergartners, especially with respect to the problems that the first-graders had not yet learned ‘by heart’, were in fact better able to make use of the material presented them for adding-up. An example of such an adding-up problem is described in Figure 5.9.

> *There are two clearly separate groups of wooden dolls,*
> *one of which contains 7 dolls and the other 4.*
> *The interviewer asks:*
> *“Here are seven dolls.*
> *If I add these other dolls to them,*
> *how many will there be all together?”*

Figure 5.9: Adding-up problem

The results of this research raised doubts in Koster's mind about the contribution of education to the development of cognitive skills. Evidently, one year of first-grade education may sometimes have no more effect on children of virtually the same age than one year of kindergarten education.

This could be seen, according to Koster, as a Piagetian result, in which development is paramount to learning. But, in his opinion, another explanation might be that arithmetic instruction teaches no more to the students than what they have already learned in other situations. His impression was that first-grade arithmetic instruction spends too much time teaching things that most of the students have already mastered.

The well-known research of Carpenter and Moser (1984) into the acquisition of addition and subtraction concepts in first through third grades corroborated Koster's opinion. Carpenter and Moser followed a group of children for three years, during which the children were interviewed individually eight times. Eighty-eight children participated in all eight interviews. The first interview was conducted in October, 1978, when the children had just started first grade. At that time, they had received no formal instruction in addition and subtraction, but had only participated in the general readiness activities typical of kindergarten mathematics curricula.

The children were given several CGI-types of word problems[14]: 'join – result unknown', 'join – change unknown', 'separate – result unknown', and 'comparison' problems. An example of a 'join – change unknown' problem used in the interviews is shown in Figure 5.10.

Susan has 6 books.
How many more books does she need
if she wants to have 14 books all together?

Figure 5.10: Example of a 'join – change unknown' problem (Carpenter, 1993)

The word problems were read to the children by the interviewer. The study focused on the children's solution strategies and provided a reasonably detailed account of these strategies and of how they changed over time. Moreover, Carpenter and Moser discovered that the children were able to solve simple addition and subtraction word problems even before they had received formal arithmetic instruction. Figure 5.11 shows per type of problem the percentage of children that were able to solve these problems when interviewed in October of first grade.

The word problems involved larger number facts (the sum of the addends ranged from 11 to 16), and manipulatives were available.

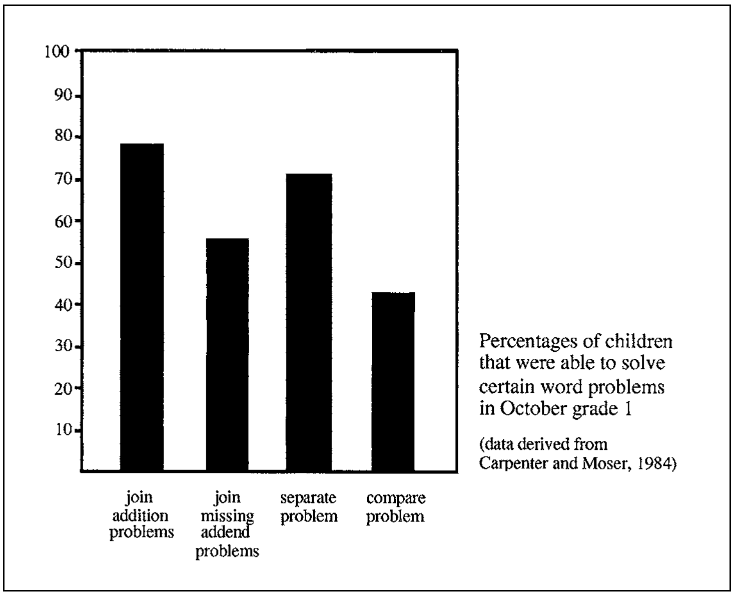


Figure 5.11: Performance on word problems at the beginning of first grade

Another, more recent, research project that also revealed the considerable extent of the arithmetic skills of three to five-year-olds was that of Hughes (1986) (see also Section 3.3.7c). Considering all these earlier findings, it is not surprising that Resnick's (1989) conclusion was that abundant evidence was now available on the ability of young children to develop and apply mathematical concepts before having attended school.

5.2.2 Implications for education

These findings hold significant implications for instruction. According to Carpenter and Moser (1984), the primary school mathematics curriculum of that time failed to capitalize on the rich, informal mathematics that children bring to instruction. The same conclusion was reached by Resnick (1989). And yet, now, some years later, the same comments are still being made (see, for instance, Urbanska, 1993). According to Leder and Forgasz (1992), it is characteristic of traditional mathematics education to simply ignore the rich store of children's spontaneous mathematical knowledge.

This rich, informal knowledge that children bring with them to school is also one of the pillars of RME. Freudenthal gave examples of this rich, informal knowledge in numerous publications (Freudenthal, 1975a, 1976c, 1979c). Not only for 'number', but even for topics like 'ratio' and 'geometry', he showed how deeply rooted these concepts have become before any formal schooling has taken place. Even though Freudenthal's views on learning processes and the corresponding realistic method of instruction (for which he, himself, together with the staff of the IOWO, laid the foundation in the nineteen-seventies) are now widely shared, there are still many textbook series in The Netherlands that start entirely from scratch in first grade. An example of this is the mechanistic textbook NZR (see Section 2.1.1), whose first-grade curriculum is summarized briefly in Figure 5.12.

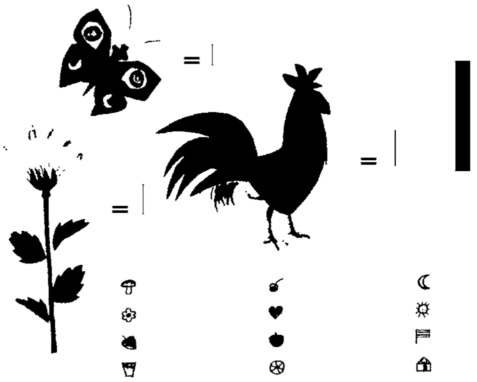
a	b
numbers and operations - numbers up to 20: the numbers are taught in the counting sequence: first 1, then 2, etc.; the numbers up to 10 are handled during the first half of first grade - problems up to 20: first the number 1 and the =-sign are taught (see part b of this figure), then the number 2 and +-sign, followed by addition problems; a few weeks later, after the number 5, the –-sign is taught and the students learn to do subtraction; the addition and subtraction problems up to 10 are handled during the first half of first grade	

Figure 5.12: NZR first-grade curriculum (a) and an illustration from the textbook (b)[15]

If one compares the above illustration with the findings of the MORE Entry-test, one will be struck by the incongruity between what the children are actually able to do and the content of this curriculum. Aside from the de-motivating element inherent in not encouraging students to use the knowledge and skills they already possess, such a curriculum can also sow the seeds of all sorts of breakdowns in communication between teacher and student. A child may be entirely in the dark about what the teacher is getting at exactly because he or she has already known something and been able to do it for a long time. The learning protocols that were collected and analyzed for the MORE research (see Streefland and Te Woerd, 1991) contain a multitude of examples of just such a situation.

According to Hiebert (1984), another pitfall that may result if links are not established between the children's prior knowledge and what they are taught, is that children will often abandon or ignore their own spontaneous strategies in order to conform to that of the 'formal' classroom.

In addition to ignoring all the numerical knowledge and skills that beginning first-graders already possess, curricula like NZR also commit another instructional error typical of the mechanistic approach, namely, that of introducing formalization too soon. Therefore, the results of the MORE Entry-test should certainly not be interpreted as a plea for more difficult problems. The analysis of second and third-grade scores from a later phase of the MORE research revealed that the score on the entry-test was one of the most significant explanatory factors for scores on later tests (see Gravemeijer et al., 1993). The students who had done well on the entry-test generally also did well later on. While these children may, indeed, have possessed higher cognitive skills or, if you will, have been more intelligent, they were at any rate better prepared for formal arithmetic thanks to the numerical knowledge and skills that they already possessed at the beginning of first grade. In other words, these findings from the MORE research could also be interpreted as an argument for ensuring that children experience all sorts of numerical activities, free from any formal arithmetic problems.[16]

5.2.3 Nothing new and yet revealing; why?

In spite of all the research described above, the high scores on the MORE Entry-test were nonetheless a surprise. In retrospect, it can be stated that both the test development and the score estimates took place in a frame of reference that, to a certain extent, was similar to the first-grade NZR curriculum; i.e., the assumption was made that one would have to start from scratch in first grade. Such an assumption, in the face of all the previous research, demands closer inspection. Why did these assorted research findings not lead much sooner to clearer instructional implications? The following sections describe three possible explanations.

5.2.3a Explanation number one

The first explanation is the assumed primacy of cognitive development. For a long time, the various numerical skills displayed by young children were not taken seriously. The prevailing opinion was that, if a child still foundered when doing conservation tasks, then the numerical concept was not yet present. It was Donaldson’s (1978; see also Section 3.3.7c) research in particular that showed unequivocally that children’s answers to such tasks are often responses to other questions than those intended by the adults and, furthermore, that the context is determinative for which answer a child gives. Freudenthal (1973, 1978b, 1979c) also pointed this out more than once. Since that time, the opinions on conservation as a prerequisite for doing arithmetic have become more differentiated (see Gelman and Gallistel, 1978; Groen and Kieran, 1983; Hughes, 1986; McClintic, 1988).

5.2.3b Explanation number two

A second explanation may be found in the tendency in education to teach children things that they in fact already know. A study by Desforges and Cockburn (1987), which followed seven experienced first-grade teachers for three weeks, revealed that, 50% of the time, the children already knew what to do before they were told.[17]

This was not only true of procedural learning, but of concept learning as well. In two out of every three tasks, the child had already grasped the underlying concept before doing the task. Take, for instance, Desforges and Cockburn's (ibid., p. 91) report on a student named Paul:

> "A week before his first lesson on subtraction, five-year-old Paul explained a problem to us thus, 'Six take away six is nothing. If I had six buns and someone ate them all up I'd have no buns left, would I?' "

Another example was that of six-year-old Joanna. She had been taught to trace rods in order to complete a series of addition problems, such as 4 + 3 = 7. She had great difficulty making the drawings because she could not hold the rods still. When this problem was discussed with her, she said that it was easier to do the addition in your head and that, if the number got bigger, you could always use your fingers.

Even though the students were often more advanced than what was being taught, Desforges and Cockburn did not feel that the teachers were wasting their time. They saw the growth of understanding of mathematical concept as a gradual process needing a considerable amount of practice and consolidation. They did, however, note the importance of emphasizing that consolidation is much more than over-learning.

An interesting phenomenon in this regard is that, in contrast to the underestimation that seems to occur at the beginning of primary school, by the end of primary school there is in fact a case of overestimation. The data collected in The Netherlands from the first PPON research project (see Section 1.4.2) revealed that both the sixth-grade teachers and the secondary education teachers estimated the level of students' mathematical skills at the end of sixth grade to be much higher than it actually was (Wijnstra (ed.), 1988). This disparity between the respective expectations at the beginning and at the end of primary school corresponds perfectly to a statement made by Wittmann, in 1991, namely:

> "...that pedagogy and didactics have always tended to *over*estimate the abilities of teachers and to *under*estimate the mental capacities of children" (Wittmann, 1991, p. 43, translation of original German text; see also Selter, 1993a).

A study by Bennett et al. (1984), in which Desforges and Cockburn also collaborated, clearly revealed both the complex nature of this matter and how the level of the students contributed to whether there was a question of over or underestimation. This study, which compared the intellectual demands of certain tasks with the students' levels of attainment, exposed a considerable degree of disparity between the two. Eleven classes were observed during the first four months of secondary education. The teachers involved were dedicated and conscientious people, and were rated as better than average by the school advisory service. Nevertheless, only 30% of the tasks matched the students abilities, whereas 37% were too easy and 30% too difficult. A breakdown according to the students' individual levels of attainment revealed that high achievers were underestimated on three-quarters of the tasks assigned to them, and that low achievers were severely overestimated; almost half the tasks given to the latter overestimated their abilities.

Even more astonishing, however, is the discovery of Bennett et al. (ibid.) that, while teachers were adept at recognizing a task that was proving to be too difficult, they were entirely oblivious of tasks that did not demand enough. According to Bennett et al., the responsibility for this lack of awareness lies, in the first place, with the conventional approach to teaching. Traditionally, the teacher usually remains at the front of the class and only occasionally traverses the classroom for brief observational glances. The second reason, in Bennett's opinion, is that, when teachers see children working hard, they interpret this as a validation of the appropriateness of the task. In other words, keeping busy is equated with appropriate demands. So the customary image of a class working cheerfully and industriously is accepted without question.

5.2.3c Explanation number three

The third explanation has to do with the tests themselves. All the above-mentioned research results, which revealed that children possess considerable mathematical knowledge and skills at the beginning of primary school, were based on individually conducted interviews. It may be that the data produced by these researches did not, therefore, make a sufficient impression. Or perhaps it was assumed that these results were linked to the specific assessment circumstances; in other words, there was a tacit suspicion that the children's high levels of achievement were due to having been aided by the interviewer in a one-on-one situation. Written tests are viewed differently; no one would ever assume that help was involved on such tests. Consequently, the MORE Entry-test results were quite a surprise. The most revelatory aspect was that these scores were achieved on a paper-and-pencil test that was administered to entire classes. This is even more astonishing if one keeps in mind that written tests are still considered a less appropriate means of assessing younger children (see NCTM/ASWG, 1994).

5.3 The MORE Entry-test abroad

5.3.1 Interest from abroad

The administration of an entry-test during the MORE research began as a more or less compulsory procedure, which was conducted in order to enable a subsequent determination of the effects of a particular textbook and of the instruction given with that textbook. But what began as a routine procedure led unintentionally to a renewed interest in the level of beginning first-grade students. Nor was this interest confined to The Netherlands. Freudenthal's efforts to disseminate the test and its results far and wide were not without effect. These efforts led, among other things, to an administration of part of the test in both Germany and Switzerland. Although these versions did differ to some extent from the original, on the whole they were all quite similar. Both in Germany and Switzerland, the test was again administered to beginning first-graders and, again, experts were asked beforehand to estimate the number of children that would answer the problems correctly.

The question that arose was to what extent the results found in The Netherlands would also apply to Germany and Switzerland, and whether the test could have an effect on education in those countries.

5.3.2 The German study

The research in Germany was conducted by Selter (1993a, 1993b), of the Institute for the Education of Mathematics at the University of Dortmund. A total of six problems were selected from the MORE Entry-test, in order to administer it to a greater number of students:

1 relational concepts: highest building (problem 1)
2 number symbols: number 3 (problem 5)
3 counting sequence: before 8 (problem 14)
4 resultative counting: number 9 (problem 12b)
5 addition in context, non-countable: 3+4 (problem 18)
6 subtraction in context, non-countable: 10-8 (problem 24)

As in The Netherlands, the test was administered to classes in their entirety and the instructions were given orally. Because, instead of the original test instructions, an abbreviated version was used that had been included in an article on the test (Van den Heuvel-Panhuizen, 1990a), these instructions were not entirely identical to those given in The Netherlands. Furthermore, in contrast to the original, the space-shuttle count-down in problem 14 was demonstrated out loud in the German version. There was also a minor alteration in problem 12b. Because the time needed to complete this problem varied considerably among the students, supplementary problems were presented to the quicker students, such as: “Write down the number 9”, “How many circles (marbles) have not yet been colored?”, “How many circles are there in all?” Supplementary information was also provided for the two context problems, as it was assumed that these problems would otherwise be difficult to understand. This supplementary information, however, was virtually identical to the instructions as originally written. One deviation from the Dutch test administration was particularly striking, however. The German students were namely told that many adults thought that children would not be able to do these problems, and that only with their help could the opposite be proven.

The test was administered in 1992, three weeks into the school year, at 14 schools in Northrhine-Westphalia. Even though they were not selected as a representative sample, the schools did, according to Selter, provide a good cross-section of schools in this area. In some of the schools, 45% of the students spoke a native language other than German, while, in other schools, all the students were native German speakers. Some were typical urban schools, others rural. All in all, they offered a representative picture of the area. The test was administered by staff and students from the Institute for the Education of Mathematics at the University of Dortmund, and was taken by a total of 893 students. In two of the classes, due to a misunderstanding, children who had limited knowledge of German did not participate in the assessment. The total number of participants therefore decreased to 881. The test administration lasted between 20 and 30 minutes. The test instructions were, when necessary, either paraphrased or repeated verbatim. When a child in the classroom did not understand the German sufficiently, another child was permitted to act as an interpreter.

Before the test was administered, a large group of experts – consisting of mathematics teachers and those who were studying to become mathematics teachers – was asked to make an estimate of the percentage of students that would answer each problem correctly. The experts were merely told that the purpose of the estimate was to determine whether the degree of difficulty of the test problems would need to be adapted. The group of experts consisted of a total of 426 persons. Of these, 51 were practicing primary school teachers, 130 were practicing primary school teachers in the second phase of their teacher education, and 245 were student teachers of primary school mathematics.

The results, on the whole, corresponded with those found in The Netherlands: there were high percentages of correct answers, and much lower prior estimates (see Table 5.3).

Table 5.3: The results of the German study

Test problems	% correct	
	actual n = 881	estimated n = 426
1 (problem 1)*	98	82
2 (problem 5)	95	65
3 (problem 14)	63	37
4 (problem 12b)	87	51
5 (problem 18)	66	29
6 (problem 24)	50	22

* The problem number in the MORE Entry-test is shown in parentheses

In Selter's opinion, despite certain weaknesses in the study, it was still possible to draw a number of conclusions. He summarized the potential implications for instruction as follows:

– The first-grade curriculum should correspond more closely to what the students already know and are able to do. A preparatory program that introduces the numbers up to twenty in stages may, in fact, have an inhibitory, rather than a stimulating effect on the students. In its place, Selter recommended a more holistic introduction to the numbers up to twenty.

– Not only should the results of the instruction be examined after the fact, but the extent of the children's knowledge of a particular topic, before they have received instruction, should also be determined beforehand. This could provide the teachers with a guideline for their instruction and should also receive a clearer place in the theory of mathematics education, in this case (as Selter calls it) the 'constructivistic' mathematics education.

5.3.3 Selter's supplementary analysis of the test results

In addition to the first two implications, Selter also mentioned a third that pertains to the test itself. A closer analysis of the test results namely revealed a number of aspects that needed improvement. This analysis – of the third problem (problem 14), the fifth problem (problem 18) and the sixth problem (problem 24) – reviewed frequently given answers, in order to determine the reasons behind these answers. The outcome was an even greater disparity between what people think children can do and the children's actual abilities, because of the high degree of rationality present in the 'incorrect' answers. Furthermore, instances of missing answers may have been caused by problematic wording of the text.

- the Space Shuttle problem (problem 14)

Nearly 10% of the students crossed off more than one number in this problem. In some of the answers, all the numbers from 10 down to 7 were crossed off, or all the numbers from 7 downwards. However, because only the single answer '7' was considered correct, these reasonable answers had to be marked wrong, even though one might certainly assume that these students could indeed count backwards. The single answers '9' and '10' were also given fairly often, in each case by 4% of the students. A possible explanation for this could be that the students had understood the word 'next' in an upward sense.[18] In the case of the 3% of the students who gave '1' as the answer, it is not inconceivable that they had first counted all the way down and then crossed off the lowest number.

- the Pinball problem (problem 18)

Here, too, 10% of the students crossed off more than one number. And, again, many of these answers were quite reasonable, such as the crossing off of the number symbols '3', '4', and '7'.

- the Glasses Shopping problem (problem 24)

In this problem, 6% of the students crossed off more than one number. Also, 4.5% had marked '10' and another 4.5% '8' as the answer. A case can be made for both these responses, as they are each an answer to a different question, respectively:
"How much money did you have?" and "How much do the glasses cost?" Another

explanation might be that the child had thought that the ‘10’ on the coin-purse referred to the situation after the purchase. It is also interesting to examine the answer ‘0’, given by 1.5% of the students. Maybe they had purchased the coin-purse for 10 guilders. After all, the children could not know that the ‘10’ on the coin-purse did not refer to a price.

- missing answers

  And then there were the children who gave no answer at all. For these three problems, the percentage of missing answers was, respectively, 10%, 10% and 12%. Unfortunately, one could not tell whether a child had attempted to do the problem but failed, or had not tried at all. According to Selter, there were two possible causes of these missing answers. Firstly, some of the students were insufficiently proficient in the German language. Secondly, these problems not only demanded certain arithmetic skills, but also required the students to understand the illustrations and the corresponding questions in a specific manner. Moreover, for each question, the students had to visualize a new and different context. All in all, some of the children may have been classified undeservedly as lacking the assessed skills. On the other hand, according to Selter, some of the children may indeed not have possessed these skills, but were able to give the correct answers by copying or because the answer was called out in class.

5.3.4 A closer look at Selter’s supplementary analysis

Selter’s supplementary analysis of the test data corresponds exactly to the RME ideas regarding developmental research on assessment. Characteristic of this analysis is that it takes the child’s standpoint (see Section 4.1.5d). Questions are posed such as, “What might the child have been thinking here?” and “Is this answer reasonable or not?” The importance of taking such a standpoint is apparent from what it produced.

It should be obvious that *the wording of the problems* was crucial. That was why so much trouble was taken with the text in the Dutch version, and why it received specific attention during the trial version. The German test instructions, on the other hand, were made from an abbreviated English version. Not only were these instructions shorter, but they had also been translated from Dutch into English, and then into German. This may very well have been an aggravating factor; translation, after all, must never be merely a precise transposition of words. It was of utmost importance that the test speak the children’s language. In the end, the purpose of this test was not to compare how children from different countries scored on an identical test, but to offer children the opportunity to show what they could do.

There was also the issue of some children’s lack of proficiency in German. Although the idea was to have the illustrations do most of the talking, some verbal explanation was nonetheless necessary. A separate version for these children in their own languages would certainly be desirable. Perhaps bilingual children could assist in developing such a version; they could function as interpreters, as indeed occurred in some of the German classrooms. And perhaps that should indeed have taken place during the development of the original, Dutch version. During the trial run, all that was determined was whether the test came across and was clear to everyone – an approach that is, in fact, both negative and passive. The children were not actively involved in the development process, and the test was only adapted when something did not come across well. It would certainly be preferable to actively involve the children in developing the test and, especially, the test instructions.

The repeated change of context, mentioned above, was, of course, due to the reduction of the test to only six problems. In the original test, one context was usually maintained for a few problems. The issue of changing contexts should be kept in mind, however, as it has been raised before (see Dekker, 1993). It is unclear whether and how changing contexts may be an inhibiting factor in demonstrating one's abilities. At any rate, it may be presumed that the nature of the contexts does play a role (see Section 3.3.7).

In contrast to the need for precision in how the questions are worded, *the students' answers* should not have to be precise. The student should not have to give one, specific, answer in order to have it be accepted. The answers of the children who crossed off the three numbers '3', '4', and '7' on the Pinball problem should, of course, have been marked correct. The open manner of assessment may in no way be circumscribed by strict rules of evaluation, or else the same mistake will be made as in Cooper's (1992) examples, where realistic test problems were not permitted to be answered realistically (see Section 3.3.7h). Not all answers were given their due, in fact, because of the traditional concept of objectivity. And the untenability of this objectivity was exposed thanks to the 'constructivistic' standpoint on mathematics education (see also Sections 3.3.6 and 4.2.1c). This standpoint should really have been taken when the tests were corrected. Why this did not occur probably had to do with the multiple-choice format of the answers. In an open-answer format, one is more inclined to investigate whether or not each answer is reasonable. As mentioned before, the assumption that beginning first-graders would not yet be ready for such an open format led to the decision to use multiple-choice. The results, however, revealed that the children had once again been underestimated. It is therefore quite plain what the next step will be in this developmental research on assessment: further investigation of the use of open problems.

Ultimately, *some uncertainty always remains* as to whether or not a child has mastered certain skills. This is a problem that will always be present to a greater or lesser degree. Even in the most optimal one-on-one interview situations, where additional questions can be asked, one cannot always be sure. There, too, a child may just not feel like participating and will simply say one thing or another in order to have done with it. Nor can this always be distinguished from truly not knowing the answer. Even if one is fairly certain whether a child can or cannot do a particular problem, it is still difficult to draw specific conclusions about the child's skills and insights. This is because, on the one hand, skills and insights are often context-related and thus depend on the specific problem, and, on the other hand, learning processes are full of discontinuities. In other words, who is to say that a child will demonstrate the identical skills and insights in another problem or at another moment (see also Section 4.2.1d).

5.3.5 The Swiss study

The research in Switzerland was conducted by Hengartner, of the Institute for Teacher Education in Zofingen, and by Röthlisberger, of the Institute for Education in Basel (Hengartner and Röthlisberger, 1993, 1995), in collaboration with student teachers from the two institutes. The Swiss study was intended as a kind of 'status quo' research. Its purpose was to ascertain both the extent of Swiss children's mathematical skills at the beginning of their schooling, and how teachers estimated these abilities. The occasion for the Swiss study – aside from the Dutch results to the MORE Entry-test and the results of Selter's study based on this test – were Spiegel's (1992) clinical interviews with first-graders. The Swiss study was also generated by teaching experiences gained from working with materials from the 'Project Mathe 2000' (Wittmann and Müller, 1990). During this work, it had been observed that first-graders during the first weeks of school were already able to perform and notate all kinds of operations with numbers up to 20 while playing various dice games, without yet having had this in class.

The Swiss study consisted of the administration of a group test containing problems from the MORE Entry-test, and of a supplementary individual interview consisting of problems involving time and money.[19]

In the Swiss study, as in the German study, the MORE Entry-test was not administered in its entirety but, instead, a number of problems were selected on the basis of an article on the MORE tests (Van den Heuvel-Panhuizen, 1990a). By mistake, in addition to problems from the entry-test, some problems were also selected from tests that were designed to be administered later on in first grade. The Swiss test consisted of a total of 13 problems. Seven of these were from the MORE Entry-test:

1 relational concepts: highest building (problem 1)
2 number symbols: number 3 (problem 5)
3 counting sequence: after 4 (problem 9)
4 counting sequence: before 8 (problem 14)
5 addition in context, countable: 2 + 4 (problem 16)

6 addition in context, non-countable: 3 + 4 (problem 18)
7 subtraction in context, non-countable: 10 – 8 (problem 24)

The test was administered to 6 classes in Basel (BS) and 5 classes in the area of Zofingen (Argau) (AG). A total of 198 first-grade students participated in the study. The test was administered three to four weeks after classes began, and was administered by student teachers from the two institutes mentioned above. Although these students did receive instructions beforehand on how to administer the test, small differences nevertheless arose in, for instance, the location of the children, the duration of the test, and the providing of assistance. Furthermore, it is unclear to what extent the oral instructions for this test diverged from the original Dutch instructions.[20]

After the test had been administered to the first-grade students, a group of 61 experts was asked to estimate for each problem the number of students in a class of 20 beginning first graders that would answer correctly. The group of experts consisted of primary school teachers. All of them had had experience teaching first grade, but none were teaching the participating students.

Once again, the first-grade students proved to do better than had been expected. In Table 5.4, one can see that the children's abilities in the problems taken from the MORE Entry-test were particularly underestimated in the areas of the knowledge of number symbols and the counting sequence.

Table 5.4: The Swiss results[21]

Test problems	% correct	
	actual n = 198	estimated n = 61
1 **(problem 1)***	**97**	**92**
2 (**problem 5**)	**96**	**73**
3 (problem 9)	82	56
4 (**problem 14**)	**68**	**46**
5 (problem 16)	77	62
6 (**problem 18**)	**52**	**43**
7 (**problem 24**)	**44**	**35**

* The problem number in the MORE Entry-test is shown in parentheses

A further analysis of the test results revealed striking differences between the individual students. For instance, the 20 students who scored the lowest answered only 2 or 3, and occasionally 4 questions (out of 13) correctly. On the other hand, the 20 students with the highest scores answered 11 or 12 questions correctly, and one of these students answered all 13 correctly. The teachers' estimates appeared to have focused more on the weaker students than on the stronger ones.

Furthermore, differences in the percentage of correct answers per problem were also found between the various first-grade classes. Figure 5.13, for example, shows how differently the classes scored on the Glasses Shopping problem (problem 24).[22]

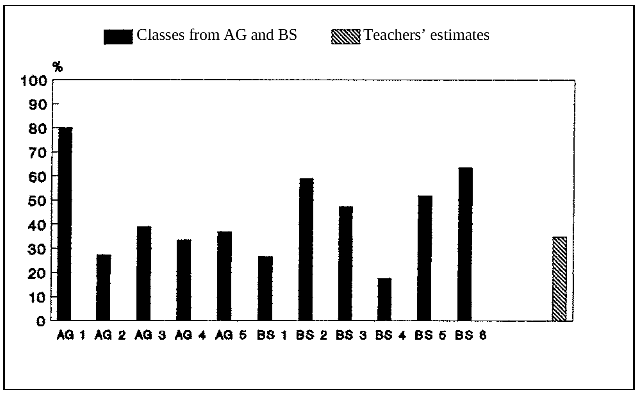


Figure 5.13: Problem 24, percentage of correct answers per class (from Hengartner and Röthlisberger, 1995, p. 73)

Lastly, differences in the scores were also found between boys and girls. On the initial problems, they were about even, but the boys clearly did better than the girls in addition and subtraction. The disparity here ranged from around 20% to nearly 50%.

The results of the time and money problems that were administered one to two months later were virtually the same. Here, too, the children demonstrated that they could do more than what had been taught in class. And here, again, the scores revealed a striking disparity with respect to gender.

In spite of these differences, Hengartner and Röthlisberger were convinced that the mathematical abilities of beginning first-grade students had been severely underestimated. In their opinion, this was caused more by the textbooks than by the teachers. Most of the Swiss textbooks assume that first-grade mathematics education must start from scratch. And, in a subject like mathematics, where the textbook plays a major role, it is not surprising that the teacher will tend to follow the textbook more than the students.

Hengartner and Röthlisberger drew a number of conclusions from the study and gave certain recommendations. These are in a nutshell, as follows:

– the students' abilities must be assessed more frequently
– the instruction must take into account what the students are already able to do
– one must not ingnore the fact that students also learn on their own.

According to Hengartner and Röthlisberger, teachers must not rely upon textbooks or long experience, but must determine each time anew what the students are actually able to do. In Hengartner and Röthlisberger's eyes, this is not the same as 'testing'. The point is not to establish a quantitative classification of the students, nor to make a prediction about their further development. The object is rather to gather an impression of the mathematical knowledge and skills that are available in the class, in order to better observe and listen to the students, and give them more appropriate tasks. In Hengartner and Röthlisberger's opinion, this research into what children are able to do in mathematics should be extended to other grades.

Furthermore, Hengartner and Röthlisberger believe that if children already know and can do so much, it makes no sense to start from scratch by presenting them with a preparatory program that prepares them for what many of them are already able to do. Nor should the numerical range be artificially restricted to the numbers under 6 if the children are already able to work with numbers up to 20 and above. Initial mathematics education should dovetail with what the children already know and are able to do, and the children should feel challenged by the activities, in order to develop further. Because the abilities of the various students are so disparate, there is no other choice but to make use of rich, complex problems that are both accessible to the weaker students and challenging to the stronger students. These are problems that offer the opportunity to engage in numerous activities on different levels.

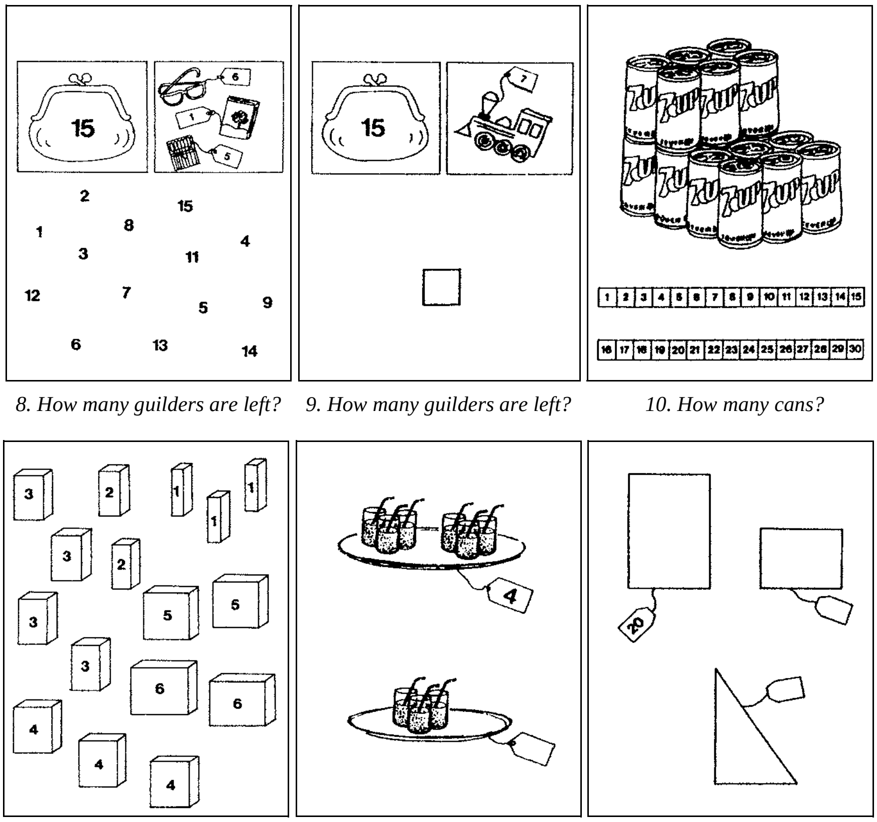
In Hengartner's and Röthlisberger's eyes, the results of their study reveal that children can acquire certain mathematical knowledge on their own, without any systematic instruction. This supports the view of learning as an active, meaningful, and constructive activity. A step-by-step approach, in which the students must again learn what they already knew on their own, can devalue and thereby block the students' own learning. This is particularly harmful to the weaker students. According to Hengartner and Röthlisberger, a different educational culture is needed, in which the focus is shifted from training and stepwise development to giving the children the opportunity to make their own discoveries. In this way, the teacher, too, is given the opportunity to better observe and listen to the children. As a result, it then also becomes easier to discover what children are able to do.

5.3.6 The surplus value of the Swiss study

The Swiss study, in addition to examining the differences between individual students, classes and genders, also provided extra information thanks to a fortuitous occurrence. Inadvertently, the following problems from other first-grade MORE tests (see Section 2.2.2, Figure 2.2) had been included on the test:

8 subtraction in context, non-countable: 15 – ... (TG1.2-18)
9 subtraction in context, non-countable: 15 – 7 (TG1.4-16)
10 geometry: number of stacked cans (TG1.2-4)
11 number: structuring the number 12 (TG1.3-17
12 ratio: deriving the price of 3 glasses from the price of 6 (TG1.4-11)
13 geometry/ratio: deriving the price of half a pizza (a/b) (TG1.4-13a/b).[23]

These problems, together with their (abbreviated) instructions, are shown in Figure 5.14.



8. How many guilders are left? *9. How many guilders are left?* *10. How many cans?*

11. Buy twelve candles *12. What do three glasses cost?* *13a/b. What do the pizzas cost?*

Figure 5.14: MORE problems used in the Swiss study
that were not taken from the MORE Entry-test

Table 5.5: The Swiss and Dutch results of problems from tests other than the entry-test

	Switzerland		The Netherlands		
Test problems	begin grade 1 % correct		TG1.2 Nov grade 1 % correct found n = 443	TG1.3 Feb grade 1 % correct found n = 440	TG1.4 Apr/May grade 1 % correct found n = 439
	found n = 198	estimated n = 61			
8	36	22	49	-	-
9	22	22	-	-	60
10	30	15	31	-	64
11	45	17	-	74	-
12	28	14	-	-	48
13a/b	28/16	6	-	-	37/23

The analysis of the answers revealed that approximately one-third of the students could already do these problems at the beginning of first-grade (see Table 5.5). Most textbooks only introduce the material covered by these problems halfway through or towards the end of first grade. A comparison with the Dutch results from later on in the school year reveals how clearly the roots of children’s abilities can be traced back to the beginning of the year.

Another remarkable finding was that nearly all the Swiss boys had scores approximately double those of the girls. An exception to this was problem 11, where the task was to choose boxes of candles so that one would buy a total of twelve candles. Here, the boys ‘only’ scored 25% higher than the girls. The gender difference was greater, furthermore, for problem 9 (15 – 7=) than for problem 8, where one could choose to buy an item that would not require bridging ten.

5.3.7 A closer examination of the Swiss results

The Swiss study, too, contributed to an enrichment of developmental research on assessment, albeit in a different manner than the German study. In the first place, the unintentional use of later test problems showed that the MORE Entry-test – while quite revelatory in itself – had by no means exposed the limits of what beginning first-graders can do.[24] This demonstrates, once again, how important it is, when developing tests, to dismiss all preconceptions of what children can and cannot do.

In the second place, by pointing out the sizable differences between individual students and between classes, the Swiss study demonstrated once more the importance of using appealing, elastic problems. Such problems are ones that are also accessible to the weaker students and that can be solved on different levels.

In the third place, the Swiss study pointed out the sizable difference between the scores of boys and girls. In addition to exposing this gap, the study simultaneously indicated where a solution to this problem might be found. The fact that the girls did not lag as far behind in problem 11 could indicate that girls are better able to show what they can do when there is more elasticity in the problem. This is also evident to a certain extent in the difference between the scores for problem 8 and problem 9, where the former, but not the latter, is an option problem. A characteristic of both problem 11 and problem 8 is that both can be solved on one’s own level. Some may raise the objection that this would then make the problems easier, which would raise the score without equalizing the differences in abilities. This objection is not completely valid, however. The girls must, in the first place, be given the opportunity to show what they can do, so that the teachers can become aware of this. If this is indeed less than what the boys can do, then so be it. The teachers will at any rate know where to start, which is certainly a precondition for resolving the situation. Aside from truly not being able to solve the problems, there may of course be other issues involved here, such as playing it safe, insecurity, conforming and suchlike. In such cases, too, problems with elasticity can provide a solution. This is not only of importance for assessment, but for instruction as well.[25]

5.3.8 The three studies combined

Even though the MORE Entry-test was not administered in all three studies in exactly the same way, the similarities are nonetheless sufficient to justify a comparison. Figure 5.15 displays the results of six of the problems in all three studies.[26]

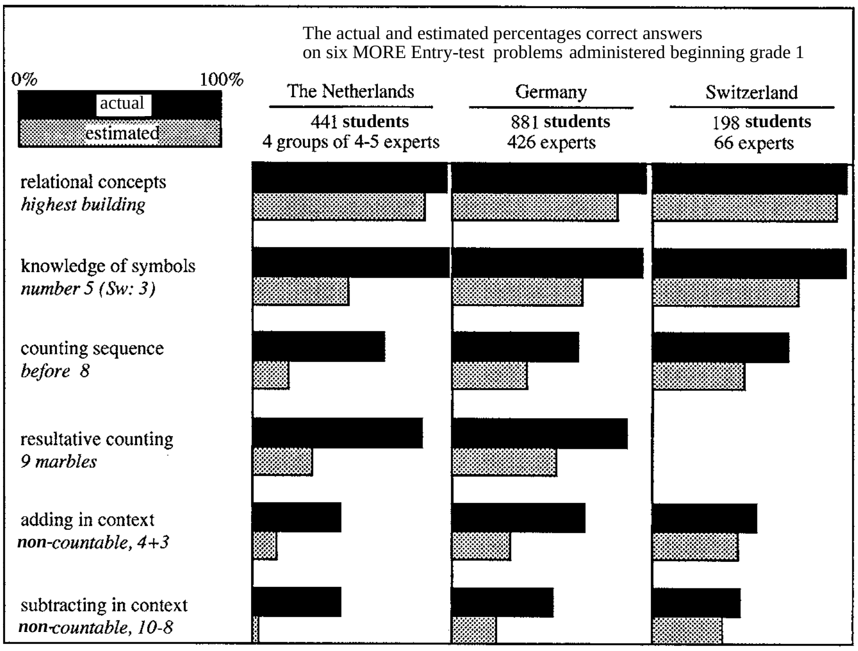


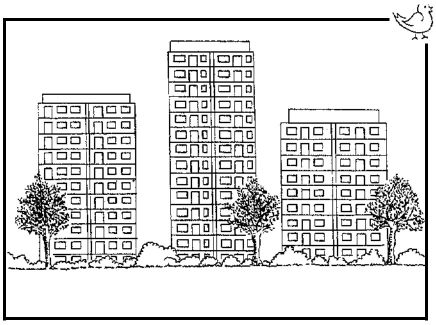
Figure 5.15: The results of the three studies

The data shows the same pattern across all three studies, both for the actual scores and for the estimates made by the experts. In all three countries, the children demonstrated significant abilities in the areas of number and operations, and, in all three countries, these abilities were severely underestimated. The disparity between the actual and expected results was the greatest in The Netherlands and the smallest in Switzerland. Besides certain procedural variations in gathering the estimates (such as requesting estimates from an individual versus a group, or requesting a percentage versus how many of the 20 students), the differences in composition of the groups of experts will certainly have contributed to this disparity. The more practical experience the experts had with first grade, the closer their estimates were to the actual results.[27] Hengartner and Röthlisberger have their own explanation for the smaller disparity in the Swiss study. In their opinion, this had to do with the involvement, in the Swiss study, of teachers who also trained student teachers.

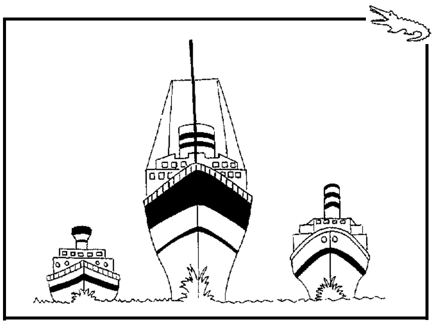
Both in Germany and Switzerland, the results of the studies are seen as a support for changes in the approach to education. The fact that children are already able to do so much[28] not only suggests that one should not start from scratch in first grade, but is also a confirmation of the view of education that regards learning as active discovery – which , in turn, requires a different kind of mathematics education.

It is paradoxical that this support for change in education emerged from a group written test. While much has been made of the widespread objections to traditional paper-and-pencil tests, such a simple paper-and-pencil test, administered at the beginning of first grade, has provided information that hitherto only surfaced through individual interviews.[29]

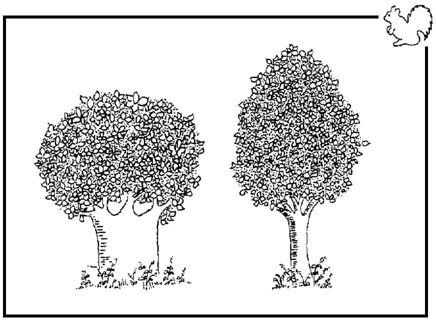
5.4 Appendix – The MORE Entry-test



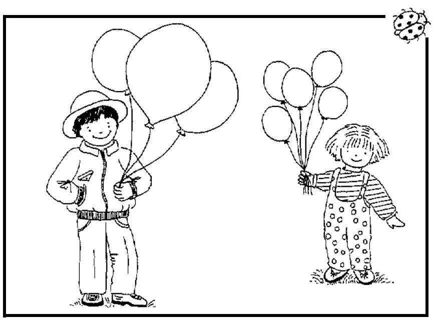
Problem 1



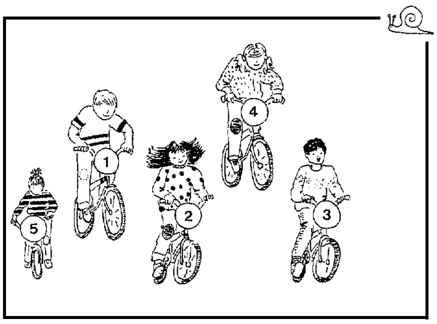
Problem 2



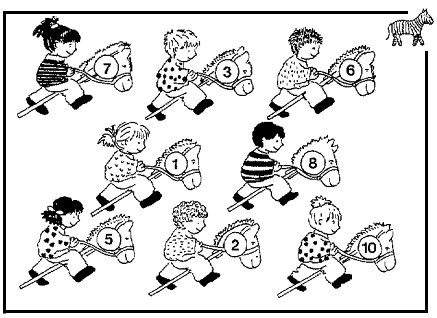
Problem 3



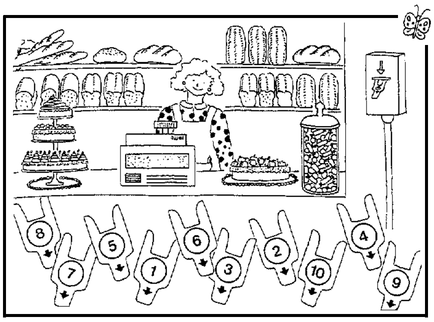
Problem 4



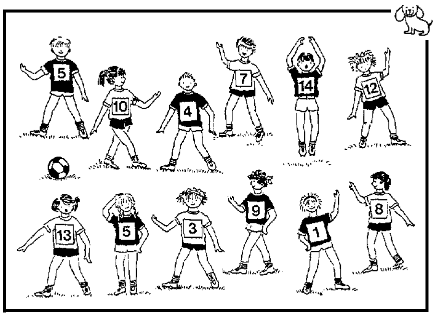
Problem 5



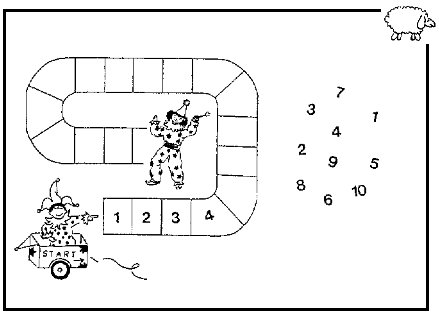
Problem 6



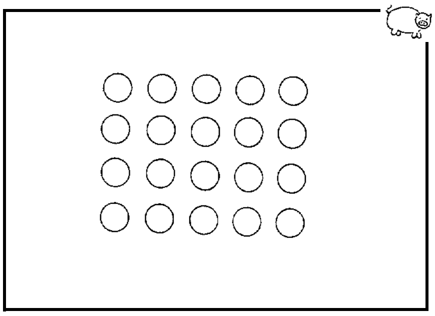
Problem 7



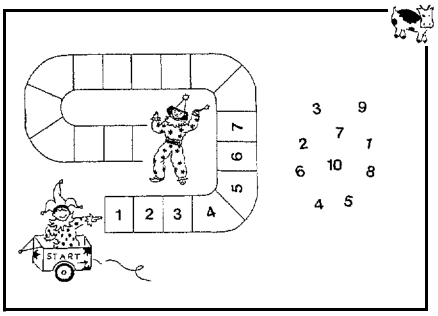
Problem 8



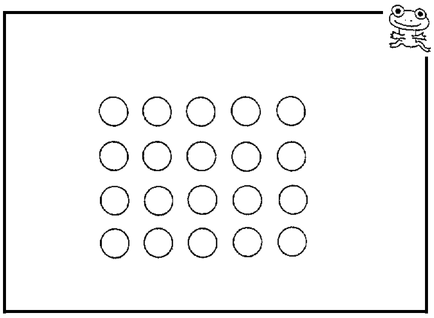
Problem 9



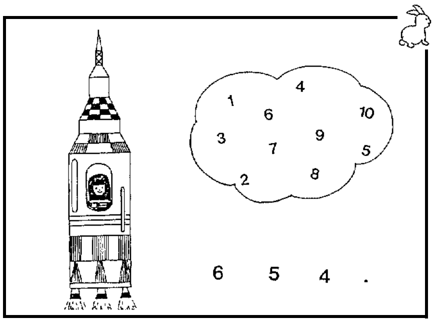
Problem 10a/b



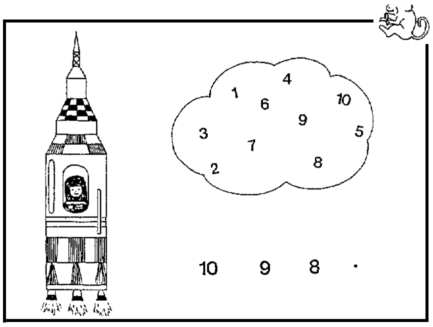
Problem 11



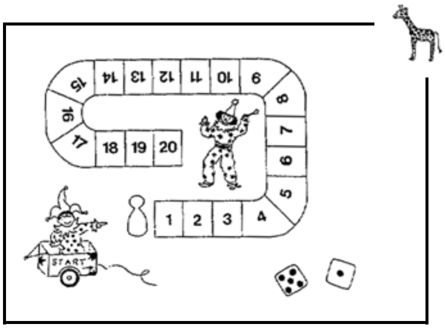
Problem 12a/b



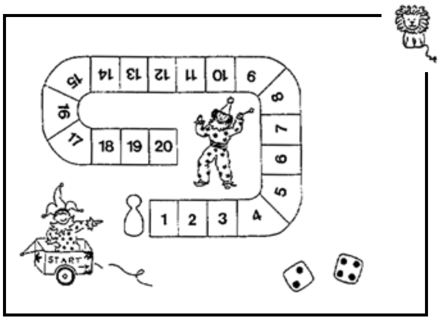
Problem 13



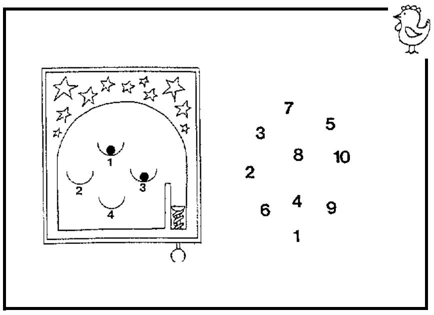
Problem 14



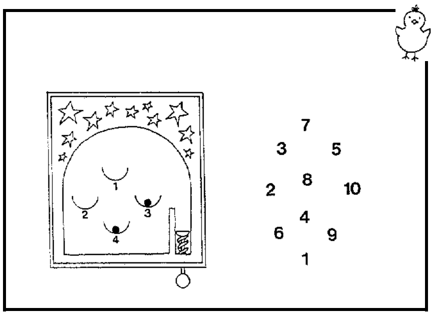
Problem 15



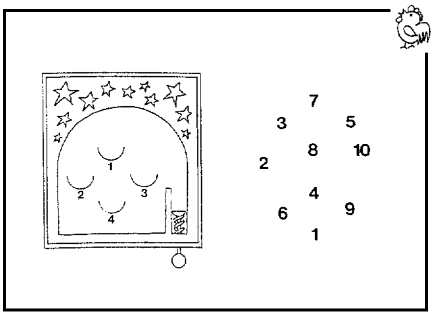
Problem 16



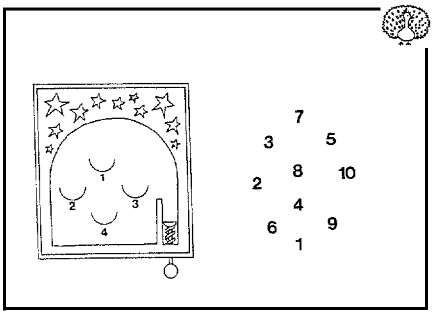
Problem 17



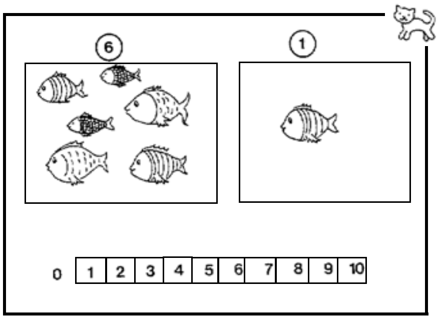
Problem 18



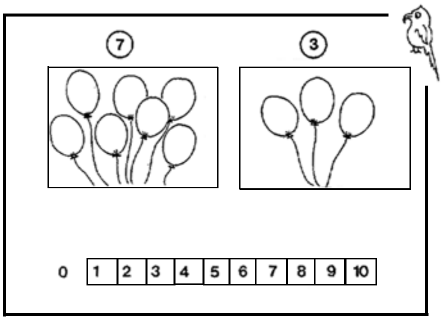
Problem 19



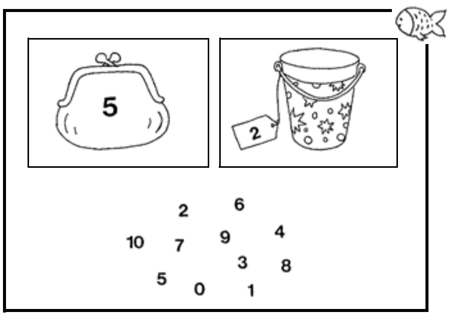
Problem 20



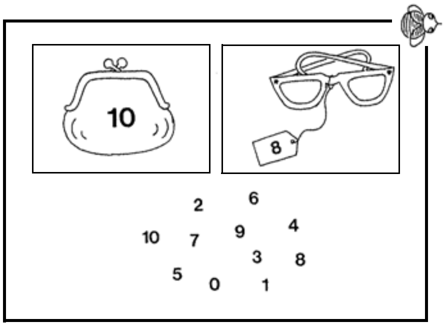
Problem 21



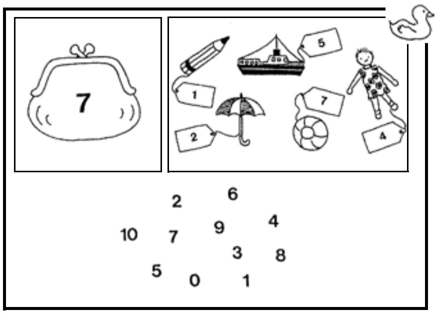
Problem 22



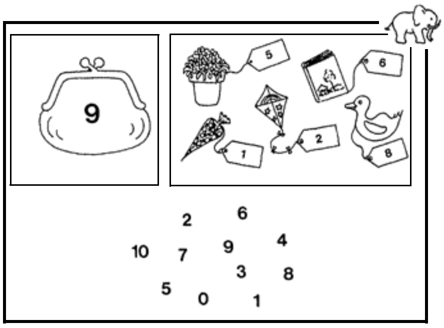
Problem 23



Problem 24



Problem 25



Problem 26

Notes

1 In The Netherlands, most children go to kindergarten for two years. To a certain extent, these two years are intended as learning preparation, and only the second year is compulsory. Students may enter kindergarten when they are 4-years-old. In order to start first grade, the student must turn 6 by October 1st of that academic year.

An exploratory research project conducted later by Harskamp and Willemsen (1991) revealed that 79% of the Dutch schools that were involved in their research had a program of preparatory arithmetic in kindergarten. However, only 67% of the teachers who taught the upper kindergarten class used such a program in their class. On the average, eight lessons from this program were given per year. The other material that was used consisted mainly of traditional learning-play activities for ordering, classifying, knowledge of number symbols, and comparing quantities. These were used three to four times per week. Materials involving games, such as simple board games and shopping were generally not used more than once a week. The use of certain materials did seem to influence the students' mathematical abilities. This, however, was not the case with the programs for preparatory arithmetic.

2 It is astonishing to observe how this approach, which was 'imposed' by the practical situation, corresponded with the ideas on mathematics education.

3 Some of the children, in addition to taking the group test, were also interviewed individually. See Section 2.2.2c.

4 This was, in fact, the beginning of the developmental research on assessment (see Chapter 2).

5 The drawings for the final version of the test were made by Lida Gravemeijer.

6 A similar experience with regard to the text occurred during the interviews on the issue of 'succeeding and preceding numbers'. Here, it became apparent that some children understood more clearly what was meant by the 'preceding number' if the question was posed in the past tense; for instance, "Which one came before 4?", rather than "Which one comes before 4?"

7 The subtle change in the Dutch version cannot be expressed in the English translation. The chief difficulties with the initially chosen wording, "Welk getal komt daarna?" ("Which number will follow?") and "Welk getal komt ervoor?" (Which number comes before?") are its directional aspect and the lack of clarity in indicating that the number desired is the one directly 'next to' the last-mentioned number. The text that was finally used, "Welk getal is nu aan de beurt?" literally means "Which number's turn is it now?"; this wording manages to avoid the directional aspect and makes the necessity of proximity clearer. Unlike the stilted sounding English translation, however, the Dutch phrase is the most common way of asking "Which number is next?"

8 Furthermore, this supplementary information on structuring could also be acquired from the individual interviews.

9 Depending on the magnitude of the second term, the students may either choose to subtract 'from behind' (count down), or to subtract 'from in front' (add up) (see also Veltman, 1993).

10 In the CGI project (see Section 3.3.2), the differences in mathematical structure were viewed as determinative for the degree of difficulty of the problems (see also Section 3.2.3, Note 10). In the MORE Entry-test, 'countable versus non-countable' was chosen as the distinguishing dimension, but this became obscured by the differences in mathematical structure.

11 This work was conducted by Gerarda van Donselaar, Wenny van der Hee, Tjako de Jong, and Nina Ruesink.

12 In the earlier-mentioned research of Harskamp and Willemsen (1991) (see Section 5.1.2, Note 1), the individually administered test developed for the research also proved to be on the easy side for most of the children. An average of 7 out of the 9 problems were answered correctly.

13 In addition, the children were given several other tasks, including Piagetian equivalence problems. It turned out that the children were more successful with the problems designed by Brush and Ginsburg than with the equivalence problems of Piaget (Ginsburg, 1975).

14 See also Section 3.3.2 and Note 10 in Section 3.2.3.

15 The illustration is a reduction of page 12 from the first booklet of the NZR textbook series (which was also lesson number 12), where the number 1 and the =-sign were taught.

16 This has already been implemented successfully in the Rightstart Program (see Griffin, Case, and Siegler, 1994). Children who do not possess the informal knowledge necessary for formal instruction can use the program to acquire this knowledge through all sorts of games. The program was developed to help close the gap caused by the vast differences in informal knowledge between first-graders from different socio-economic groups. One of the examples given by these authors was a problem in which you had four chocolate candies, and then were given three more. The question was how many candies you would have in all. 72% of the children from the higher socio-economic groups answered the problem correctly, 69% from the middle groups, and only 14% from the lower groups.

17 Similar results were found in another British study (Johnson, 1989, cited by Bell, 1993), which investigated the skills of 8 to 13 year-olds in the areas of measurement and fractions. In a nutshell, two of the six participating students had already understood beforehand, two failed to understand both before and after, and the remaining two did learn something – but not necessarily successfully.

18 In the German version, the word ‘nächtste’, which means ‘next’, was indeed used.

19 Individual interviews developed by Spiegel (1992) were also conducted, but these will not be discussed here.

20 Hengartner and Röthlisberger (1993) mention only the abbreviated version of their instructions.

21 The MORE Entry-test problems that were also administered in the German study are printed in bold type.

22 There is no data on the variability of scores between classes that also takes into account the variability within the classes.

23 The original MORE tests from which these problems were taken are indicated in parentheses. The test administered under the supervision of Hengartner and Röthlisberger contained two additional problems as well. The data on these problems has not been included here, in one case (TG1.1-22), because different instructions were given from what was intended and, in the second case (TG1.2-3), because of a badly printed illustration.

24 The discovered ceiling effect was already an indication in that direction.

25 For the importance of having some influence on the problem oneself, see Sections 3.2.4a and 4.1.3d.

26 The ‘resultative counting’ problem was not included on the Swiss test.

27 During the New Media project, the MORE Entry-test was presented to first-grade teachers who were enrolled in in-service education for mathematics. Their estimates lay much closer to the actual scores (NMP, 1989).

28 It should be noted, however, that this is not the case for all children. The differences found by Griffin, Case, and Siegler (1994) between children from various socio-economic groups are extremely important here (see Section 5.2.2, Note 16). This point is also strongly emphasized by Grassmann et al. (1995) (see the following note).

29 In addition to the three studies discussed in this chapter, there is now a fourth and fifth study as well. The fourth study is that of Grassmann, Mirwald, Klunter, and Veith (1995). In this study, the MORE Entry-test was administered to students in Berlin and Brandenburg. The findings from this fourth study corresponded on the whole to the preceding three (see table below). While Grassmann et al. also found that children’s knowledge at the beginning of first grade tends to be underestimated rather than overestimated, they emphatically stressed the danger of constructing a (new) ‘average student’ or ‘average class’ based on the average scores from this test. In their opinion, the differences between

individual children and between classes – even at one and the same school – are too great to warrant this. Furthermore, the pattern of differences between the classes was not the same for each problem. The percentage of correct answers was sometimes higher in one class, and sometimes in another. A new aspect in this fourth study is that the researchers also observed the students while they took the test, which enabled them to gather important information on the strategies applied. For instance, in Berlin, the children counted on their fingers much more often than they did in Brandenburg. And 8% of the total group of students colored the marbles in problem 12b from right to left – even though they were not left-handed.

Finally, the fifth study is that of Hošpesová, Kuřina, and Tichá (1995). In this study the problems of the MORE Entry-test were administrated in the Czech Republic and in Slovakia.

The table below contains all the results collected so far from six problems on the MORE Entry-test.

research site		**problem 1** highest building	**problem 6** number 5	**problem 14** before 8	**problem 12b** color 9 marbles	**problem 18** 4 + 3 in context	**problem 24** 10 – 8 in context
	n	**percentages correct answers**					
The Netherlands	441	98	98	66	83	44	44
Germany (Northrhine-Westphalia)	881	98	95	63	87	66	50
Switzerland	198	97	96	68	-	52	44
Berlin / Brandenburg	845	99	96	68	84	54	34
Czech Republic	661	99	98	44	88	59	48
Slovakia	325	92	92	41	81	54	53

6 A test on ratio – what a paper-and-pencil test can tell about the mathematical abilities of special education students[1]

6.1 Introduction

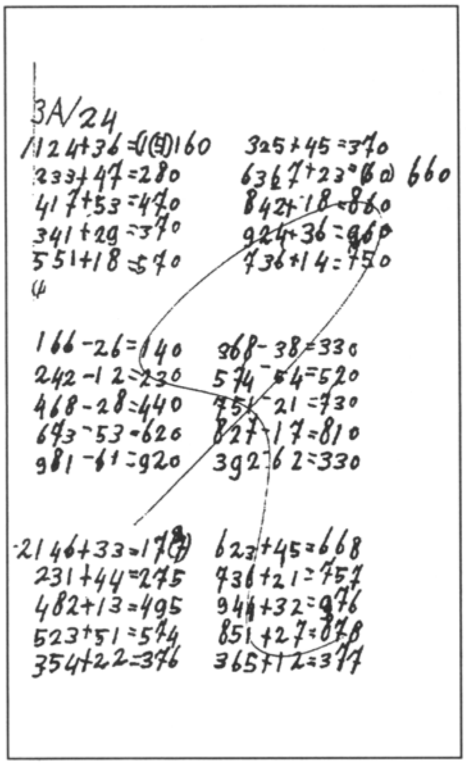
In The Netherlands, in addition to regular schools for primary education, there are also schools for special education. The system of special education comprises fourteen kinds of schools, which are attended by some 5% of primary school-age children.[2] Among these are schools for children with learning and behavioral difficulties, for mildly mentally retarded children, severely mentally retarded children, deaf children, the blind and visually handicapped, and for children with severe emotional problems.

Two kinds of schools, that is, for children with learning and behavioral difficulties[3] and for mildly mentally retarded children[4], account for the great majority of these children. Some three-quarters of the children in special education attend one of these two types of schools.

Although the students at these two types of schools do have much in common, there is a marked difference between them with respect to their level of ability. Whereas a child with learning and behavioral difficulties might possibly achieve the goals of regular primary education, this is virtually out of the question for a mildly mentally retarded child. At the end of primary special education the children's ability level in mathematics is assumed[5] to be as follows (see Van Luit (ed.) et al., 1989, and Damen, 1990): children with learning and behavioral difficulties eventually attain an ability level that lies somewhere between the middle of third grade and the end of sixth grade of regular primary education; mildly mentally retarded children attain an ability level that lies somewhere between the end of first or the beginning of second grade and the end of fourth grade. Sometimes, however, mildly mentally retarded children attain the level of the end of fifth grade of regular primary education.

Worksheets of two children are shown in Figures 6.1 and 6.2 as an illustration of the ability level in mathematics of mildly mentally retarded children at the end of primary special education. Both children were in sixth grade. The work was done halfway through the year.

The worksheet in Figure 6.1 was done by Martijn. He was then eleven years and ten months old and his mathematical ability was above the class average. Because he was fairly young, it was considered keeping him at the school for an additional year.



3A/24

124+36=1(5)160
233+47=280
417+53=470
341+29=370
551+18=570

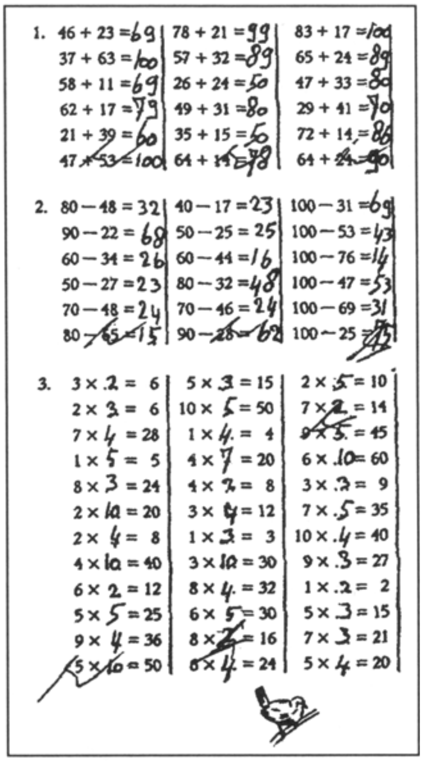
325+45=370
637+23=(6a) 660
842+18=860
924+36=960
736+14=750

166-26=140
242-12=230
468-28=440
643-53=626
981-61=920

368-38=330
574-54=520
751-21=730
827-17=810
392-62=330

2 146+33=179
231+44=275
482+13=495
523+51=574
354+22=376

623+45=668
736+21=757
944+32=976
851+27=878
365+12=377



1.	46 + 23 = 69	78 + 21 = 99	83 + 17 = 100
	37 + 63 = 100	57 + 32 = 89	65 + 24 = 89
	58 + 11 = 69	26 + 24 = 50	47 + 33 = 80
	62 + 17 = 79	49 + 31 = 80	29 + 41 = 70
	21 + 39 = 60	35 + 15 = 50	72 + 14 = 86
	47 + 53 = 100	64 + 14 = 78	64 + 24 = 90
2.	80 − 48 = 32	40 − 17 = 23	100 − 31 = 69
	90 − 22 = 68	50 − 25 = 25	100 − 53 = 43
	60 − 34 = 26	60 − 44 = 16	100 − 76 = 14
	50 − 27 = 23	80 − 32 = 48	100 − 47 = 53
	70 − 48 = 24	70 − 46 = 24	100 − 69 = 31
	80 − 65 = 15	90 − 28 = 62	100 − 25 = 75
3.	3 × 2 = 6	5 × 3 = 15	2 × 5 = 10
	2 × 3 = 6	10 × 5 = 50	7 × 2 = 14
	7 × 4 = 28	1 × 4 = 4	9 × 5 = 45
	1 × 5 = 5	4 × 7 = 20	6 × 10 = 60
	8 × 3 = 24	4 × 2 = 8	3 × 3 = 9
	2 × 10 = 20	3 × 4 = 12	7 × 5 = 35
	2 × 4 = 8	1 × 3 = 3	10 × 4 = 40
	4 × 10 = 40	3 × 10 = 30	9 × 3 = 27
	6 × 2 = 12	8 × 4 = 32	1 × 2 = 2
	5 × 5 = 25	6 × 5 = 30	5 × 3 = 15
	9 × 4 = 36	8 × 2 = 16	7 × 3 = 21
	5 × 10 = 50	6 × 4 = 24	5 × 4 = 20

Figure 6.1: Martijn’s worksheet

Figure 6.2: Harm’s worksheet

Harm, the boy whose worksheet is shown in Figure 6.2, was in the sixth grade at a different school for mildly mentally retarded children. He was then twelve years and nine months old. In terms of mathematics, he was one of the weakest students in his class. He would be leaving the school that year to attend a junior secondary vocational school.

6.2 A disparity between two approaches to mathematics education

The above examples of written work not only illustrate the ability levels of mildly mentally retarded children in the uppermost grade of primary special education, but also indicate the kind of mathematics education typical of special education. Both schools follow a traditional, mechanistic approach, and use textbooks that can be characterized as mechanistic.[6]

In special education, the mathematics curriculum more often than not only covers the four main operations. These are supplemented by word problems, and by tasks dealing with measurement, money, time and the calendar. The teaching methods can be characterized as sparse, strict and step-by-step. Whenever possible, the children are presented with fixed solution procedures. No opportunity is allowed for different strategies, due to a concern that this would simply confuse the children.

In other words, the reform of mathematics education that has occurred – and is still occurring – in The Netherlands has had virtually no influence on special education. Developments in the direction of RME have taken place almost exclusively in the domain of regular education. As a consequence, there is a great disparity, with respect to mathematics education, between the instruction given in regular primary schools and in schools for special education.

Arguments in favor of an RME approach in special education (Van den Heuvel-Panhuizen, 1986; Ter Heege, 1988) have, up until now, fallen upon deaf ears. This is not surprising considering that, until recently, there has been little research data to substantiate these arguments.

Nonetheless, these arguments have not been entirely without effect. Various efforts are now being made in special education to move towards a realistic approach to mathematics. An example of this is a program intended for children who have difficulty learning mathematics, which was derived from a series of RME textbooks.[7] Another example can be found in an otherwise rather mechanistic textbook series, developed specifically for special education, to which a final, RME-like chapter was added.[8] A last, unmistakable, example is the recent endeavor to implement RME in the education of deaf children.[9]

On the whole, however, both teachers and psychologists in special education remain exceedingly reluctant to shift from the traditional approach to that of RME. Aside from their uncertainty about the feasibility of RME in special education due to a lack of research data, they raise many objections to the RME teaching methods. These objections pertain, in particular (see Van Luit, 1987, 1988; Damen, 1990), to (i) teaching an entire class at once, (ii) building on the students' informal knowledge, (iii) the variation in solution strategies (which is related to the previous point), and (iv) interaction in class; finally, they object to (v) the complicating factor of starting from contexts.[10]

What these objections actually boil down to is the concern that this kind of instruction would place too much of a burden on special education students. Time and again, school practice has validated these objections by pointing out the students' low ability level. In one way or another, each test the students take increases the distance from RME. The general conclusion is that, if the children are not capable of doing mathematics in the usual way, how could they possibly do it if contexts were added and they had to come up with solution strategies on their own?

When reflecting upon this conclusion, it is important to bear in mind that the children's abilities must not be regarded separately from the kind of instruction they receive. After all, perhaps another manner of mathematics education would lead to different learning results. Unfortunately, the classroom experiences described above have scarcely presented any incentives for moving towards RME.

6.3 Breaking the vicious circle

The following is an account of an attempt to break this vicious circle. Ideally, one should probably conduct a teaching experiment using both an experimental and a control group, whereby the first group would be taught according to the RME approach. In the research described here, however, the decision was made to take a different approach. This decision was chiefly made for practical reasons, but also due to the availability of a less complicated alternative that was expected to produce equally compelling results. The attempt to break the vicious circle was in fact carried out without involving any kind of RME – or, to put it more provocatively – without involving any education at all. In other words, an effort was made to prove the feasibility of RME by using the children's achievements.

At first glance, this might seem to contradict the above remarks on the children's limited achievements. There was one major difference however, namely, the manner in which the children's abilities were assessed. In this study, a type of assessment was employed that offered the children some assistance (see Van den Heuvel-Panhuizen, 1990a; Van den Heuvel-Panhuizen and Gravemeijer, 1990b, 1991a). Consequently, the children were better able to demonstrate their abilities. In order to achieve this, the tasks employed had to be very accessible. Tasks were therefore chosen whose intention the children would grasp immediately, and which would not require any prior knowledge of procedures or notations. In other words, these tasks made it possible to investigate the children's abilities without the hindrance caused by formal notation.

There is empirical evidence showing the revelatory nature of this manner of testing. This can be seen, for example, from the results of a test – administered to a first grade class after three weeks of instruction – that contained tasks with the features mentioned above. This test revealed that the children were capable of much more than had been assumed (Van den Heuvel-Panhuizen, 1990a; see also Chapter 5). The most remarkable aspect, however, was that these results – which had already been discovered through individual interview situations – came to light by means of a written test that was administered to an entire class.

In order to prove the feasibility of RME even more convincingly, the research in question focused on mildly mentally retarded children who, without doubt, are the weaker students within the total group of special education children – certainly when compared to children with learning and behavioral difficulties.

Moreover, the topic for the test – namely, that of ratio – was not one that is regularly included in the mathematics curriculum at schools for mildly mentally retarded children.

A salient feature of the traditional manner of teaching mathematics in special education is the way in which the subject matter is structured: small numbers are processed first, followed by larger ones; easy operations such as addition are dealt with before more difficult operations like subtraction; bare problems are performed before applications.

Because of this sequence, some students may not even have the opportunity to become acquainted with certain topics in the subject matter. This occurs not only because of the difficulty of these topics, but also because they are planned at the end of the curriculum. Some children, held back by an obstacle along the way, may therefore never even reach these topics. For instance, some children may never get a chance to do money problems because they had not succeeded in doing addition and subtraction problems up to one hundred. This flies in the face of the fact that one can certainly learn to calculate with guilders or dollars without being very skilled in arithmetic up to one hundred.

In addition to not being taught certain subject matter topics, the children may even end up missing entire areas of subject matter.

A study involving 82 learning disabled students and 78 mildly mentally retarded students from six schools for special education revealed that, for instance, by the end of sixth grade, neither the students with learning and behavioral difficulties nor the mildly mentally retarded students had even been introduced to the topic of ratio (Damen, 1990).[11]

The question is whether one can justify excluding the topic of ratio from the special education mathematics curriculum. In order to answer this question, a test on ratio was administered to a number of students in the upper two grades at two schools for mildly mentally retarded children.

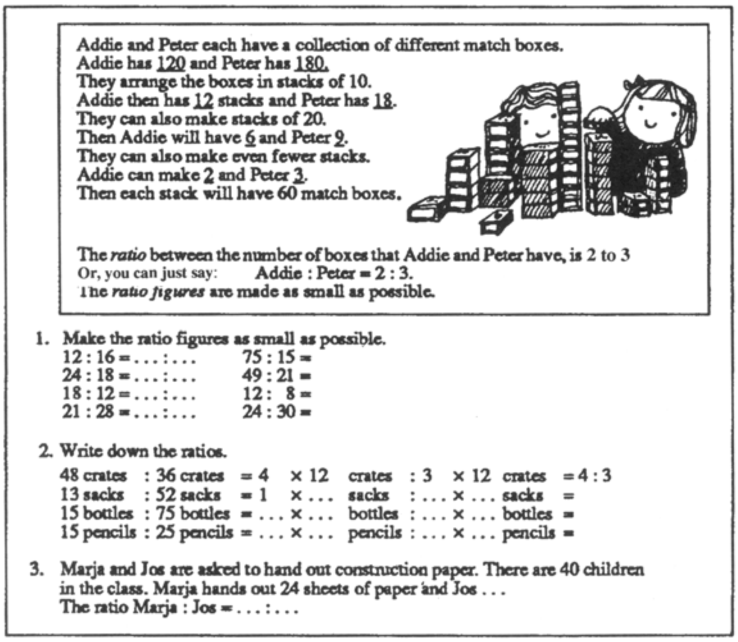
6.4 The topic of ratio

The topic of ratio involves performing operations on numbers that express a relation to one another. This relation may pertain to any and all measurable characteristics, such as number, length, area, volume, weight, duration, and price. These measurable characteristics – also called magnitudes – can be described in a relative manner by means of ratios.

There are several ways of expressing a relation by means of ratios. One way is to express the length of something in relation to the length of something else. One can also make a comparison within one object – for instance, by comparing its entire length to a part of its length or by comparing the length of something at different moments in time or in different situations.

In addition to comparing with respect to one magnitude – whether or not it involves one single object – the comparison can also incorporate different magnitudes. A relation can be expressed between the length of a certain route and the time it takes to cover this distance, or between the length of something and its price, or between the area of a country and the number of its inhabitants. As a matter of fact, relating different magnitudes to one another creates new compound magnitudes, such as velocity, price per meter, or density of population.

The ratio problems that students may be confronted with are different in nature due to the various mathematical structures behind the problems. Therefore, different kinds of ratio problems can be distinguished as follows: finding the ratio (? : ?), comparing ratios ($x : y$? $a : b$), producing equivalent ratios ($x : y$ = ? : ?) and, finally, finding the fourth proportional ($x : y = a$: ?).



Addie and Peter each have a collection of different match boxes.
Addie has 120 and Peter has 180.
They arrange the boxes in stacks of 10.
Addie then has 12 stacks and Peter has 18.
They can also make stacks of 20.
Then Addie will have 6 and Peter 9.
They can also make even fewer stacks.
Addie can make 2 and Peter 3.
Then each stack will have 60 match boxes.

The *ratio* between the number of boxes that Addie and Peter have, is 2 to 3
Or, you can just say: Addie : Peter = 2 : 3.
The *ratio figures* are made as small as possible.

1. Make the ratio figures as small as possible.

12 : 16 = . . . : . . .   75 : 15 =
24 : 18 = . . . : . . .   49 : 21 =
18 : 12 = . . . : . . .   12 : 8 =
21 : 28 = . . . : . . .   24 : 30 =

2. Write down the ratios.

48 crates : 36 crates = 4 × 12 crates : 3 × 12 crates = 4 : 3
13 sacks : 52 sacks = 1 × . . . sacks : . . . × . . . sacks =
15 bottles : 75 bottles = . . . × . . . bottles : . . . × . . . bottles =
15 pencils : 25 pencils = . . . × . . . pencils : . . . × . . . pencils =

3. Marja and Jos are asked to hand out construction paper. There are 40 children in the class. Marja hands out 24 sheets of paper and Jos . . .
The ratio Marja : Jos = . . . : . . .

Figure 6.3: An example of a mechanistic introduction to ratio[12]
(translation of original Dutch text)

It is not surprising that the topic of ratio has been excluded from the mathematics curriculum at schools for special education. Ratio is indeed a rather difficult topic, because of its departure from natural numbers that refer to concrete quantities. On the other hand, ratio has the special feature of being accessible in spite of this difficulty. The easy aspect of ratio is its strong, informal, roots, which are grounded in visual perception. Long before they have been introduced to a numerical approach or formal notation, children are already able to see ratios. A toy car looks the same as a real car, only on a smaller scale. Such a reduction can indeed be made on different scales, which is why toy cars can come in different sizes.

It should be mentioned here that this approach to ratio, which devotes attention to its non-numerical roots, is entirely absent from mechanistic textbooks. These textbooks introduce ratio on an exclusively numerical level, often by teaching the formal notation. An example of such an introduction is shown in Figure 6.3.

This mechanistic approach contrasts strikingly with the realistic introduction to ratio as shown in Figure 6.4.

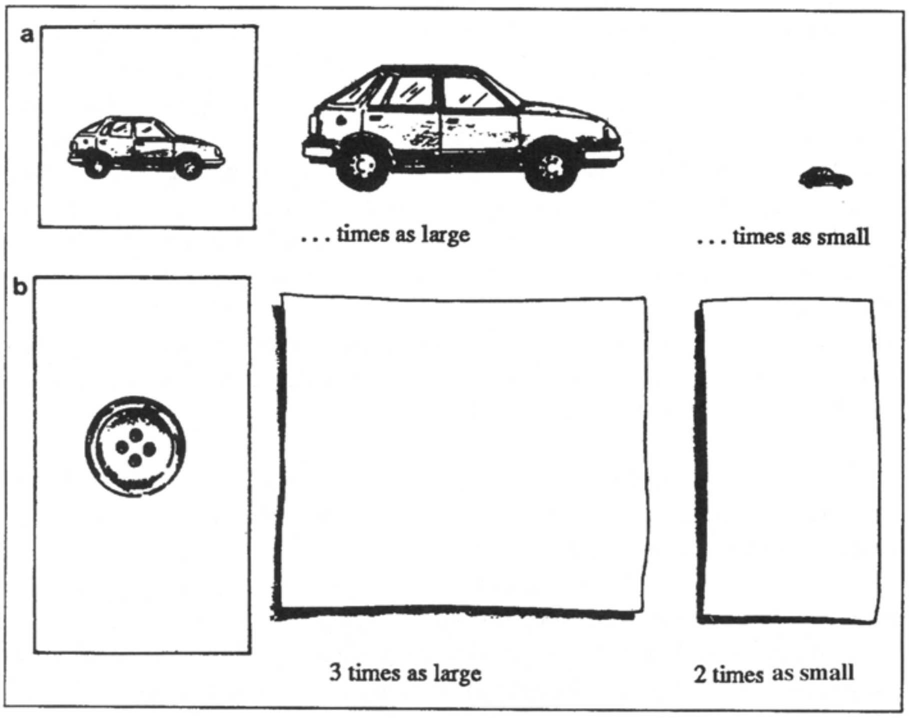


Figure 6.4: An example of a realistic introduction to ratio[13]
(translation of original Dutch text)

6.5 The test on ratio

The test on ratio developed for this study was devised in accordance with the principles of the MORE tests mentioned earlier (see Section 6.3 and Chapter 2). Consequently, a search was made for tasks that could be expressed by pictures which, wherever possible, were self-evident, referred to meaningful situations, and presented ideas for finding a solution. In order to prevent the test from becoming a test on reading comprehension rather than mathematics, the instructions were given orally and the test pages contained only the most essential textual information.

In terms of content, the intention was to design a test that would contain a variety of different situations in which children encounter ratio. These would be situations familiar to the children in one way or another through their experiences in everyday life (see Van den Brink and Streefland, 1979). Moreover, an effort was made to contrive tasks that would correspond with the different kinds of ratio problems as distinguished in the previous section: finding the ratio, comparing ratios, producing equivalent ratios and finding the fourth proportional.

In each kind of problem, some difference in level was introduced by incorporating non-numerical or qualitative tasks as well as numerical tasks (see Van den Brink and Streefland, 1979; Streefland, 1984; Van den Heuvel-Panhuizen, 1990b). Numerical tasks are those which contain numerical information, that is, in which the solutions can be found by means of a calculation. Non-numerical tasks, on the other hand, although they may involve numerical aspects, are tasks for which no numerical information is provided on the test page. It is reasonable to assume that the students would not solve these problems by calculating, but, rather, mainly by measuring and reasoning.

It should be noted that these tasks, each of which represented a particular kind of ratio problem, did not only differ with respect to the feature non-numerical/numerical. The framework of the test was not that strict.[14]

The entire ratio test consisted of sixteen problems.[15] The selected problems shown in Figure 6.5 give an impression of the entire test. The illustrations have been reduced in size for inclusion here.[16]

The text printed in italics is a summary of the oral instructions given to the students. These instructions were not printed on the test pages.

Unlike the reproductions shown in Figure 6.5, the actual test pages contained an illustration of a piece of scratch paper when the problem in question involved a numerical task. The children could use this scratch paper to write or draw something that would help them solve the problems.

The oral instructions did not mention measuring aids, meaning that the children were not encouraged to use them. They were, however, permitted to do so spontaneously.

Although the test was designed to be administered to an entire class, no time limit was set for each problem. Within reasonable limits, the children were allowed to work on the problems for as long as they wished. This meant that the faster students would have to wait a bit after most of the problems. The wait would not be very long, however, as the problems were not complex and did not require complicated calculations or reasoning.

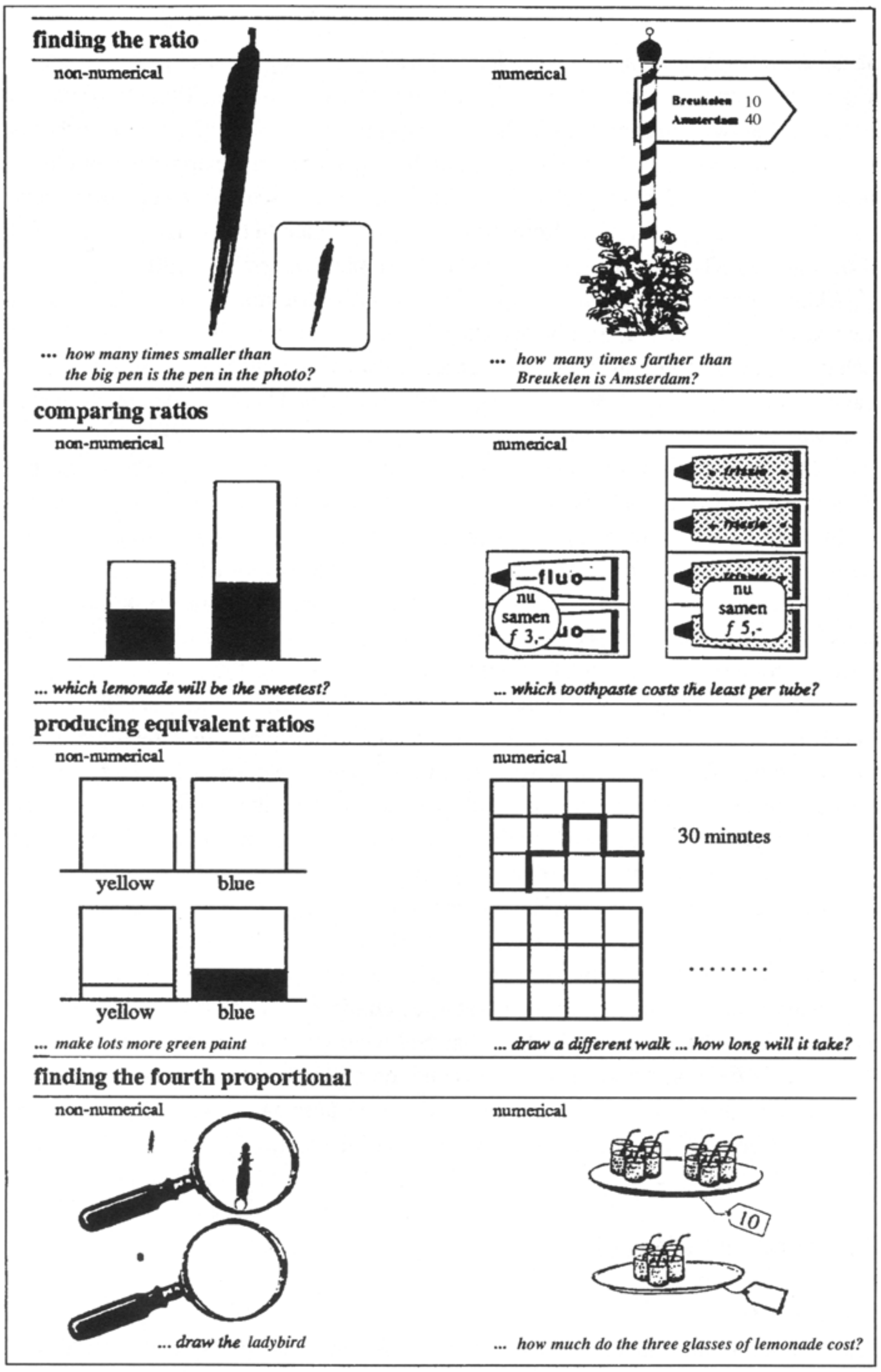

finding the ratio
non-numerical
... how many times smaller than the big pen is the pen in the photo?
numerical
Breukelen 10
Amsterdam 40
... how many times farther than Breukelen is Amsterdam?
comparing ratios
non-numerical
... which lemonade will be the sweetest?
numerical
fluo
nu samen f 3,-
nu samen f 5,-
... which toothpaste costs the least per tube?
producing equivalent ratios
non-numerical
yellow
blue
yellow
blue
... make lots more green paint
numerical
30 minutes
....
... draw a different walk ... how long will it take?
finding the fourth proportional
non-numerical
... draw the ladybird
numerical
10
... how much do the three glasses of lemonade cost?

Figure 6.5: Examples of test problems from the ratio test

6.6 Research design

Because the research was intended to be a pilot study for future research, the test group was restricted to two schools for mildly mentally retarded children. These two schools are located at two different sites in the southern part of the Netherlands. The schools were selected at random. The participants in the study were in the upper two grades at these two schools. 32 sixth-graders (16 children from each school), and 29 fifth-graders (14 from one school and 15 from the other) participated in the study, giving a total of 61 students. The test was administered in November/December 1990.

Along with the written test that was administered to determine whether the students were capable of solving ratio problems, other data on the students was also collected. This included information on their age and gender, on whether or not the student would be leaving school at the end of the year, and on mathematics level in class. Their mathematics level was determined by classifying the students in each class on a scale from good to poor. This classification was made by their teachers, whose point of departure was the progress that had been made by each student in the mathematics textbook used in class.

In each class, an inventory was also made of which mathematics topics had already been covered, either during that school year or previously.

In order to assess the opinions of the teachers of these four classes on the feasibility of teaching ratio in special education, they were asked beforehand to estimate per test problem the number of students that would answer the problem correctly. The teachers' estimates were made on the basis of the test booklet and the corresponding test instructions. The teachers were not explicitly told that the topic of the test was ratio. This information was intentionally withheld so as not to alarm the teachers in the event the mathematics they were teaching their students did not include the topic of ratio. They might namely have been concerned that something was going to be tested that had not yet been taught.

Two inspectors and two special education psychologists, in addition to the four teachers, were also asked to make estimates. Their task was to estimate the percentage of students that would prove capable of solving these test problems by the end of primary school for mildly mentally retarded children. These estimates were made on the basis of the same information that had been given to the teachers, the only difference being that the test instructions and the manner of scoring were discussed with the special education psychologists. However, here again, no information was supplied about the background and the purpose of the test.

6.7 Research results

6.7.1 The testing

Testing took approximately three-quarters of an hour. Hardly any explanation was necessary, as the children understood the tasks well. Now and again, some of the instructions were repeated. The questions the children asked not only pertained to the phrasing of the problems/tasks and the corresponding information, but also alluded to the solution strategies. This was especially true of the sixth-grade students. One of these students, for example, laughingly asked whether the walk (see Figure 6.5) could take just as long as the one already drawn. In short, the children's reactions were such that at times it was difficult to stick to the test instructions and not just start discussing the problems with them.

6.7.2 Scoring

Most of the test problems presented no ambiguity in terms of scoring because the answers were clearly either correct or incorrect. There were some problems, however, in which the difference between correct and incorrect was not altogether clear. These were primarily the problems in which the answer had to be drawn[17], such as the problem about paint (see Figure 6.5). In scoring this type of problem, the children's drawings were measured and a certain margin was allowed within which the 'answer' had to lie. For the paint problem, for instance, the ratio between yellow and blue paint (1 : 2) had to lie between $1 : 1\frac{2}{3}$ and $1 : 2\frac{1}{2}$.

The problems that posed the most difficulty in terms of psychometrics were those involving the comparison of ratios (see Figure 6.5), where a choice had to be made between only two options. Neither increasing the number of options per problem nor increasing the number of problems was a feasibility. The former might have made the problems too difficult and the latter would have made the test too long. An effort was made to reduce the chance factor in these problems in some other way, namely, by incorporating strong distracters that pointed to the incorrect answer. For instance, the lemonade in the glass with the most syrup would not be the sweetest, in spite of having more syrup.

6.7.3 Psychometric data

The following, prior to a discussion of the results, is some information about the research group and psychometric data on the test itself, based on its administration in this study. The age of the students who participated in the study ranged from ten-and-a-half to thirteen. Their average age at the time the test was administered was twelve years and one month. Considering the size of the research group, the test had a reasonable internal homogeneity. The alpha-value of the test was 0.61 and was not sensitive to omitting certain problems in the analysis. There is only a slight increase in the alpha-value if an analysis is made of the sixth-grade students alone; the alpha then becomes 0.64.

For the majority of the problems, there is evidence of a significant correlation with the total score. These correlations run from 0.22 to 0.57. No real clusters of related problems can be distinguished among the various problems. A few problem-pairs do show a significant correlation, but there is no evidence of a specific pattern.

The framework on which, in a certain sense, the development of the test was based – different types of problems and numerical versus non-numerical tasks for each problem type – was in no way reflected in the students' answers. Not that this was this really expected, however, as the problems differed from one another in many more aspects than just the type of ratio problem and the nature of the presentation (numerical or non-numerical). The frequency distribution of the students' total test scores revealed a fairly normal distribution, with neither a ceiling nor a bottom effect (see Figure 6.6).

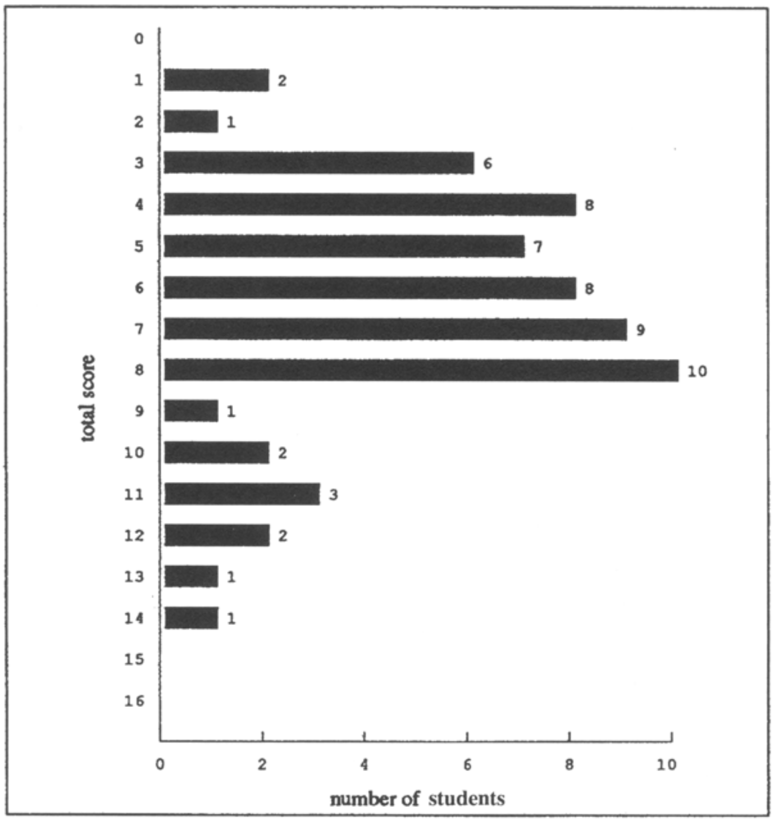


Figure 6.6: Frequency distribution of the students' total scores on the ratio test

6.7.4 Test results

As can be seen from the frequency distribution of the total scores, the lowest total score was 1 and the highest was 14. The average total score was 6.4 and the standard deviation 2.9. The percentages of correct answers (calculated for the entire research group) lay between 13% and 64%. Of the sixteen problems, six had a percentage correct answers of between 40% and 60%. Table 6.1 gives the percentage correct answers of each problem.

Table 6.1: Percentage correct answers per problem (total group tested)

	%		%
1. pen	39 (+10)	8. paint	43
2. paper clip	28 (+23)	9. coin dispencer	44
3. road sign	13 (+48)	10. walk	38
4. tree	57	11. ladybird	64
5. lemonade	54	12. string of beads	51
6. toothpaste	30	13. steps	38
7. swimming	44	14. glasses	64
		15. newspapers	26
		16. lpg	13

These percentages correct are elucidated briefly in the following section. In a few instances, certain incorrect answers are also discussed, as well as what the children wrote on the scratch paper.

Problems 1, 2 and 3 dealt with *finding the ratio* or, more precisely, determining the reduction or enlargement factor. Aside from the issue of discovering the size of this factor, these problems were also about formulation. The answers of a number of children suggested that an additive solution had been used instead of a multiplicative solution ('the pen is twice as small' instead of 'the pen is three times as small'). This phenomenon is familiar from other research (Hart, 1988; Küchemann, 1989), and will not be discussed further here. The percentage between brackets in Table 6.1 refers to the percentage of children that gave an 'additive' answer. If these answers had also been marked correct, the percentage correct would have risen considerably. This is particularly true of the problem about the road sign (see also Figure 6.5). In this problem, nearly half of the children described the difference in distance. However, the 'additive' answer given here differs somewhat from the ones mentioned above.

Problems 4 through 7 concerned the *comparison of ratios*. One should note here that the numerical problems (6 and 7) were not answered correctly as frequently as the non-numerical problems (4 and 5), and that the problem about toothpaste had fewer correct answers than the problem about the fastest swimmer. During the test, the children were also more animated while working on the latter problem. The toothpaste problem did not seem to appeal to them as much. Moreover, there was less implicit incentive to compare the price per tube of toothpaste.

Problems 8 through 10 involved *producing equivalent ratios*. Although, at first glance, the oral instructions for these problems seemed rather complicated, the problems created no difficulties and the percentage of correct scores was fairly high. Around 40% of the children solved these problems correctly. This was even true of the walk problem (problem 10). The work shown on the left in Figure 6.7 is Harm's solution, the student whose work was shown in Figure 6.2!

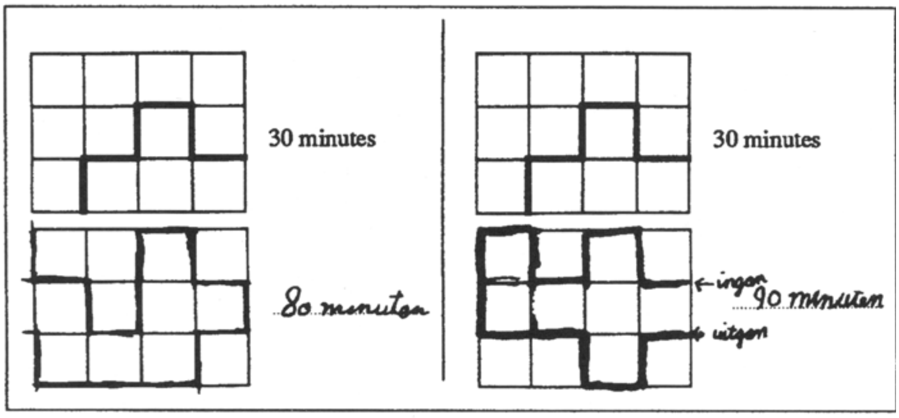


Figure 6.7: Two examples of answers to problem 10[18]

If problem 10 had been marked less strictly, the percentage correct would have been even higher. But, following the scoring rules, the solution on the right in Figure 6.7 was marked incorrect, even though it probably only contained a small error in calculating the time needed to complete one segment of the walk. In all, the answers of 8% of the children were based on a six-minute instead of a five-minute walk-segment. Another 16% of the children gave answers that implied the choice of a walk-segment ranging between four and six minutes (five minutes not included). In other words, these answers, although certainly realistic, were nevertheless marked incorrect.

In *problems 11 through 16*, the children had *to find the fourth proportional*. Of these problems, only number 11, the problem dealing with the ladybird, was non-numerical. The two problems that were answered correctly most frequently were this problem and the numerical problem on the price of three glasses of lemonade.

The most difficult problems (apart from the problem about the road sign) turned out to be numbers 15 and 16, which involved calculating the price of, respectively, twelve kilos of newspaper and forty liters of gasoline. This was not surprising, considering that these were students at schools for mildly mentally retarded children. Even so, some students did indeed solve these problems correctly.

Figure 6.8 shows two students' answers to problem 16. Notice that the answer on the right, although reasonably realistic, was marked incorrect according to the applied scoring rules. In all, four children gave similar answers. The two answers to this problem that appeared the most frequently, each given by thirteen children (= 21%), were 25 guilders and 16 guilders. These answers, too, indicate an additive solution: the number of liters had increased by ten or by one unit of ten so the price was also raised by ten or by one unit of one.

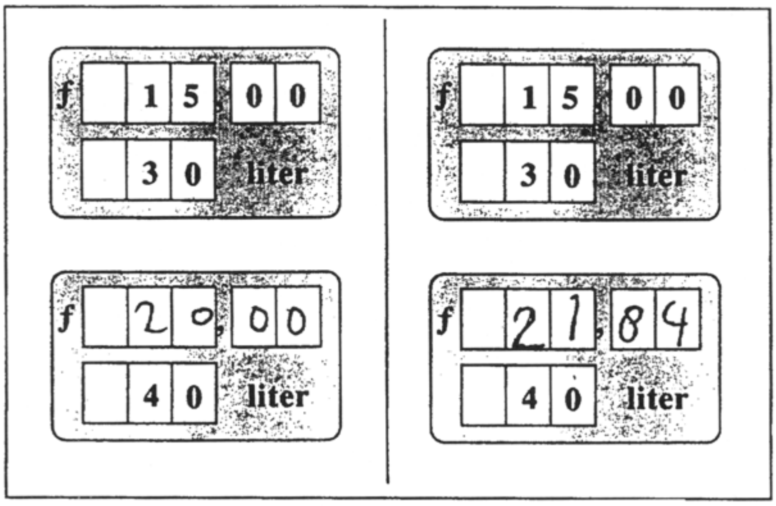


Figure 6.8: Two examples of student answers to problem 16

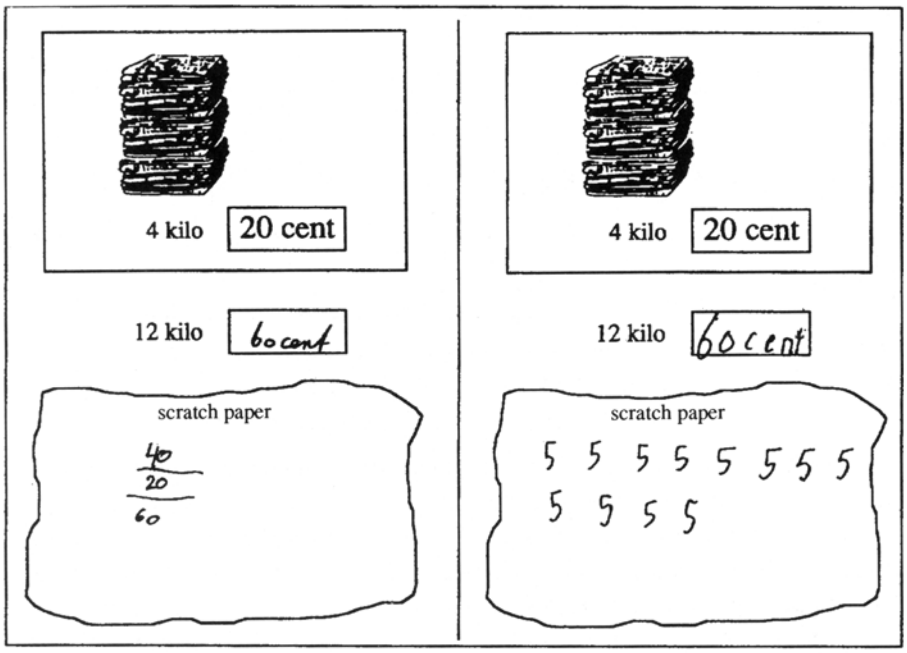


Figure 6.9: Two examples of scratch paper and solutions to problem 15

Figure 6.9, which refers to problem 15, demonstrates that the children were not only capable of solving the problem, but that they were also able to indicate how they had arrived at their answer. One child (see scratch paper on the left) arrived at the answer by calculating three times four kilos. The other (see scratch paper on the right) first calculated the price per kilo.

On the whole, the *scratch paper* was not used very frequently: 41% of the children used it once or more. It should be mentioned, however, that the students were not explicitly asked to work out their answers on the scratch paper. It was left entirely up to them whether they used it or not. Even so, some interesting scratch paper did turn up, from which it was clear that reflection on solution strategies is anything but impossible for students at schools for mildly mentally retarded children. Figures 6.10 through 6.12 show the scratch paper from three other test problems.

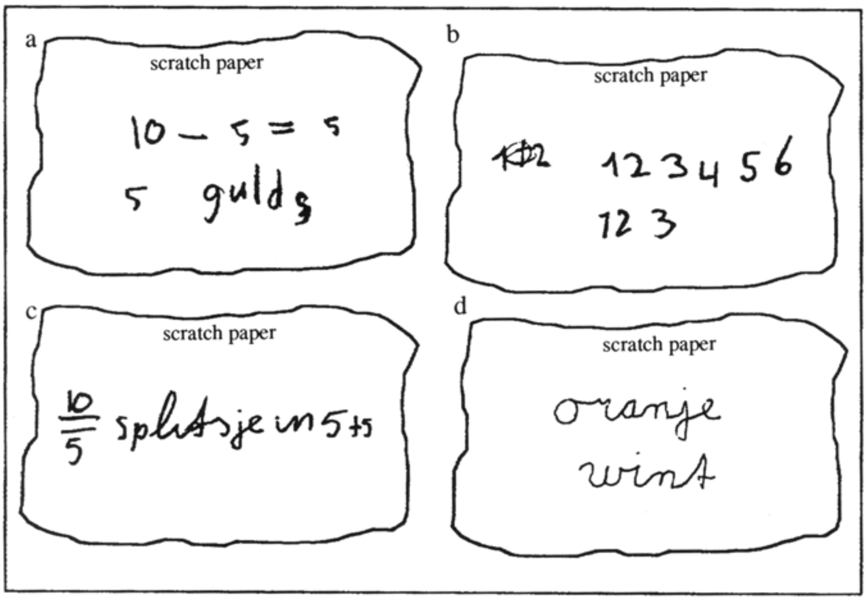


Figure 6.10: Four examples of scratch paper
with solution strategies pertaining to problem 14[19]

The four pieces of scratch paper in Figure 6.10 pertain to test problem 14, which involved determining the price of three glasses of lemonade (see Figure 6.5). In order to arrive at the solution to this problem, one must realize that the number of glasses has been reduced by half. Consequently, the price must also be halved. Scratch paper (c) shows this strategy the most directly: “You divide it into 5 + 5”. Scratch paper (a) shows a confirmation after the fact of the halving process, while scratch paper (b) reveals an earlier stage of the solution process.

This last child may have discovered the halving process through the two rows of numbers. As can be seen from the scratch paper on the lower right (d), not every piece of scratch paper provided all that much information on the applied strategy.[20]

In problem 12 (see Figure 6.11), the number of white beads had been given and the children were asked to figure out the number of black beads on the string. The difficult aspect was that not all the beads were visible in the illustration. The corresponding pieces of scratch paper show how three children used models on different levels to arrive at the solution.

The most concrete model is the one on the left (a), while the one on the right (c) is the most abstract. In the latter case, neither the number of beads, nor the specific pattern of two black beads followed by two white beads are important any longer. All that counts is the equivalent relationship between the black and the white beads.

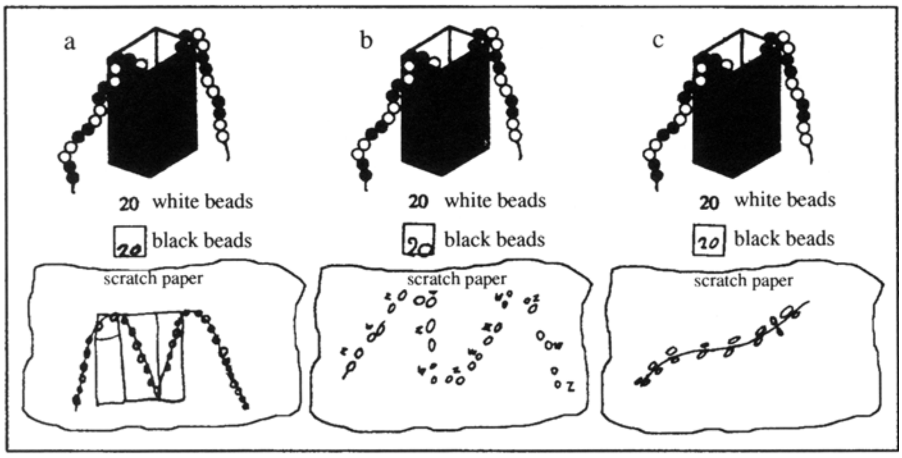


Figure 6.11: Three examples of scratch paper with solution strategies pertaining to problem 12

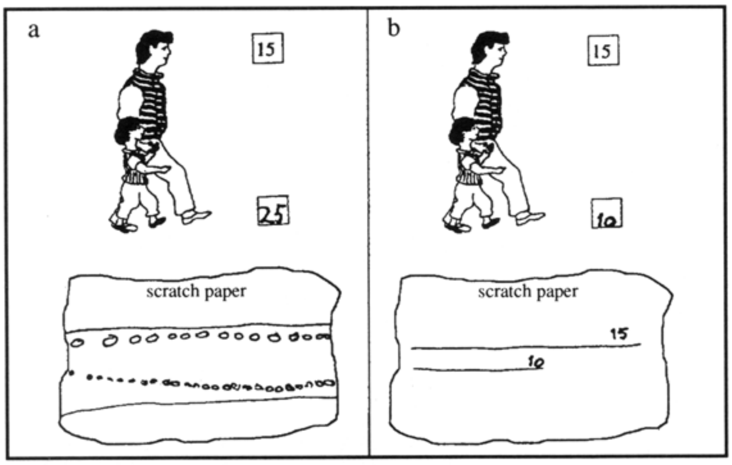


Figure 6.12: Two examples of scratch paper with solution strategies pertaining to test problem 13

In problem 13 (see Figure 6.12), a father and his son are measuring the length of their garden in footsteps. The father measures fifteen steps and the question was how many steps the son would have to take. The piece of scratch paper on the left (a) again shows how a concrete model enabled the student to find the answer by counting. Sometimes, the scratch paper will also show exactly where the student went wrong. An example of this can be seen in the scratch paper on the right (b). Instead of the number of steps, this student probably focused on the distance that would have been covered after a certain number of steps.

6.7.5 The implemented curriculum

How good, in fact, were the students' scores on this ratio test? Although this is difficult to ascertain without any reference data, the results certainly contrasted sharply with the findings that surfaced from the inventory of the implemented mathematics curriculum. As indicated in Table 6.2, none of the participating classes had ever been taught anything about ratio. The scores to the ratio test were therefore achieved in the absence of any explicit instruction in the area of ratio. It certainly makes one think!

Table 6.2: Inventory of subject matter components that had been dealt with either that school year or previously

	class 1.5[(1)]	**class 1.6**	**class 2.5**	**class 2.6**
(mental) arithmetic to 20	x	x	x	x
(mental) arithmetic to 100	x	x	x	x
column addition/subtraction	x	x	x	x
column multiplication	x[(2)]	x		
column division	x[(2)]	x		
fractions				x[(6)]
percentages				
decimal numbers				
ratio				
geometry				
measuring	x[(3)]	x[(5)]	x	x[(7)]
metric system	x[(3)]	x[(5)]		
arithmetic with money	x[(4)]	x[(5)]	x	x[(8)]
other				

1) school 1, fifth grade
2) not all children
3) m. cm, mm
4) money calculations up to 1 guilder
5) fourth-grade level at regular primary school

6) only the concepts $\frac{1}{2}$and $\frac{1}{4}$
7) m, cm, dm
8) assigning names to coins, assigning values to coins and comparing coin values up to two and a half guilders

6.7.6 Expectations

Even though the topic of ratio did not constitute part of the mathematics curriculum at the schools for mildly mentally retarded children, the estimates of the percentage of students that would find the correct answers were not particularly low. This may have been because those who made the estimates had not been told that the test was about ratio. Familiarity on the part of the special education psychologists with the background of this test and with the surprising differences that had been found on similar tests between the estimated and the resulting percentages of correct answers may also have contributed to the more accurate estimates.[21]

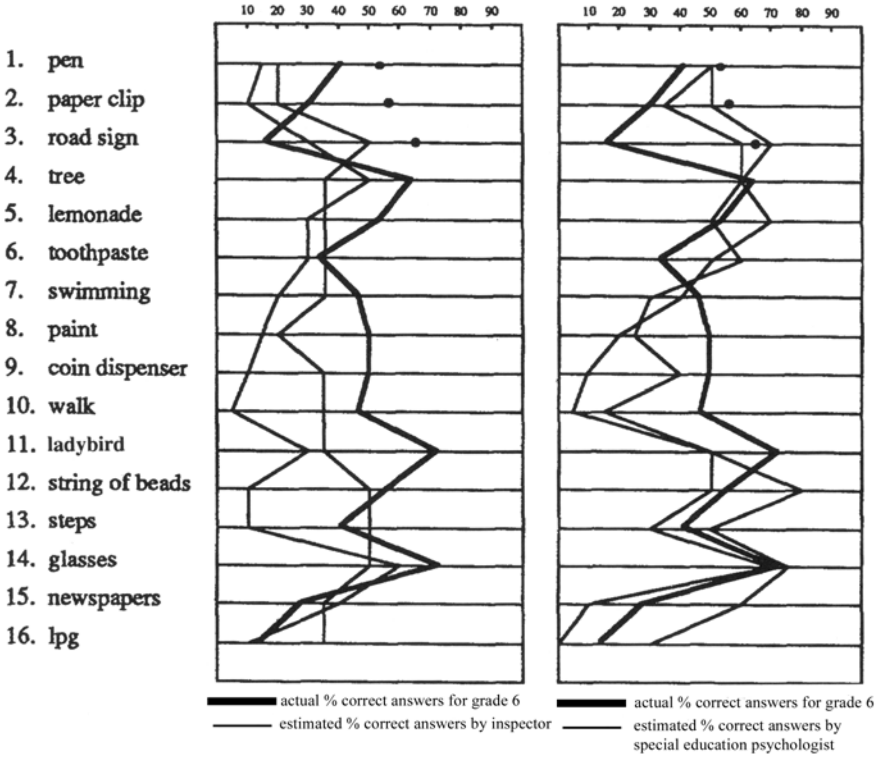


Figure 6.13: Actual sixth-grade percentages correct answers[22] and estimated percentages correct answers given by two inspectors and two special education psychologists

Nevertheless, it can be stated of the estimates on the whole, that the skills of students in the upper two grades of primary schools for mildly mentally retarded children were underestimated on a fair number of issues. The inspectors tended to have lower expectations than did the special education psychologists (see Figure 6.13). Also, the estimates of one of the two sixth-grade teachers (see Figure 6.14) were considerably lower than those made by the other, whose estimates generally corresponded with the actual percentages correct.

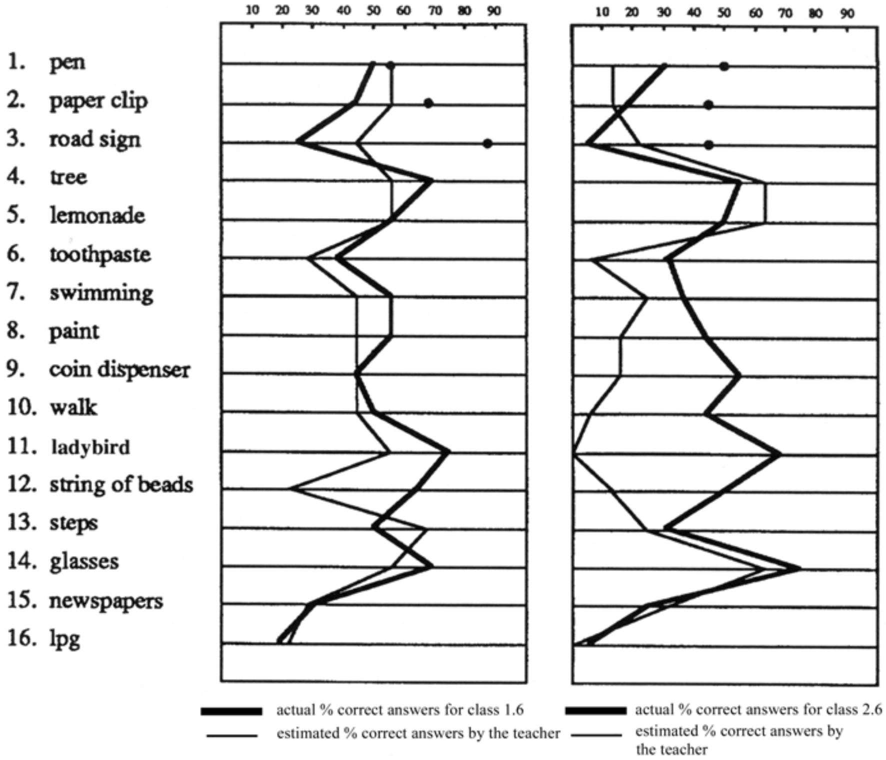


Figure 6.14: Actual percentages correct answers[22] of the two sixth-grade groups and estimated percentages correct answers given by the respective teachers

6.7.7 Relationship between test scores and certain student characteristics

Aside from the analysis of the test results, an investigation was also conducted into whether the total score was related to certain student characteristics.

By way of variance analysis, an examination was made of whether significant differences were present between the total scores of boys and girls, of fifth and sixth-graders, and of the two schools involved in the research. Two regression analyses were conducted to investigate the relationship between the age of the children and the test score, and between the class mathematics level and the test score. Of the five investigated relationships, only the relationship between the mathematics level in class and the total score appear to be significant (F (1,59) = 9.14; $p < 0.001$). The correlation between these two variables is 0.37 ($p < 0.01$) (see Figure 6.15).

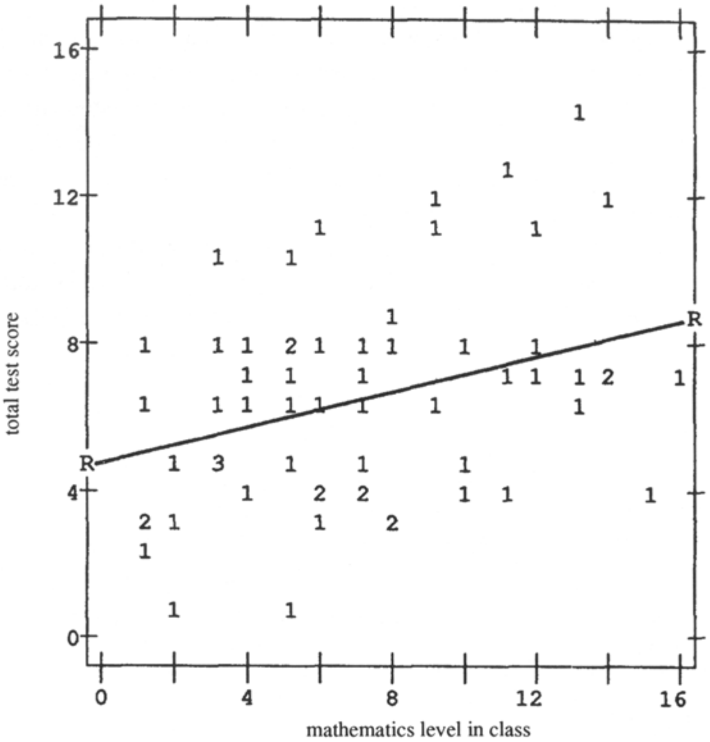


Figure 6.15: Relationship between the mathematics level in class and the total test score

6.8 Conclusions

Although no general conclusions can truly be made on the basis of such a limited study, the test results and the experiences gained from the testing do support the idea that the topic of ratio has undeservedly been omitted from the mathematics curriculum in special education.

Another conclusion, which must be regarded with equal caution, is that children in special education are indeed aware of the strategies they apply and are able to discuss them. The experiences gained from giving the test and the evidence on the scratch paper strongly point to this being the case. If the children are indeed spontaneously able to write down a strategy they have chosen themselves, then they may also be quite capable of talking about it.

The third, tentative conclusion concerns working with contexts. The test results revealed that this need not be the limiting factor it has so often been thought to be. Much depends, however, on the choice of contexts and on how these are presented.

It is essential that the contexts lead to student involvement and that they elicit strategies.

In summary, although this study did not provide the certainty for special education that one would desire, there is, at the very least, cause for reflecting on the special education mathematics curriculum and its teaching methods. This study has exposed the feasibility of crossing the educational demarcation line between regular and special education and of reconsidering the presumed limitations of children who attend schools for special education.[23] This does not mean to imply that everything that is possible in regular education can also be realized in special education. It has been emphasized before (see Van den Heuvel-Panhuizen, 1987) that an RME approach to teaching requires modification when it is intended for students who are less than proficient in mathematics.

Notes

[1] This chapter was first published under the title of ‘Ratio in Special Education. A pilot study on the possibilities of shifting the boundaries’ (Van den Heuvel-Panhuizen, 1991b). The present chapter has been slightly modified.

[2] In The Netherlands, the percentage of children attending primary and secondary schools for special education varies from 3% to 8% (see Meijer, Pijl, and Kramer, 1989). The percentage depends on how the calculations were made. For six to thirteen-year-olds, this percentage is just under 5%.

[3] These schools are called LOM schools.

[4] These schools are called MLK schools.

[5] More precise data is not available. This will provided by the yet to be published report of the PPON research for special education (see Note 23).

[6] The school attended by Martijn uses the mechanistic textbook series ‘Niveaucursus Rekenen’ (see Note 12) with additional material from ‘Remelka’ and ‘Zo reken ik ook’ (see Notes 7 and 8). The school attended by Harm uses its own series of workbooks based on the mechanistic textbook ‘Steeds verder’. This series of workbooks contains sixty booklets which must be worked through successively. Each booklet covers a particular type of calculation. Harm’s work in Figure 6.2 is from booklet number thirteen, on column arithmetic. Thirty of the booklets are on this topic. Besides the booklets on column arithmetic, Harm had already completed some booklets on word problems, measurement, money, time, and the calendar.

[7] This is the ‘Remelka’ program for children who have difficulty doing mathematics. It is related to the realistic textbook series ‘De wereld in getallen’ (see Section 2.1.1, Note 3).

[8] This is the textbook series ‘Zo reken ik ook!’ (Pedologisch Instituut Rotterdam, 1987). Up until now, this textbook series on mathematics is the only one to have been specifically developed for special education.

[9] Since the academic year 1988/1989, the Institute for the Deaf in Sint-Michielsgestel has been involved in implementing RME. This implementation is based on the RME textbook series ‘Rekenen & Wiskunde’ (Gravemeijer et al., 1983).

[10] A recently completed study (Houben, 1995) has revealed that such viewpoints are not restricted to special education. In regular education, too, teachers tend to doubt the feasibility of RME for students who are weak in mathematics (see also Note 21 in Section 6.7.6).

[11] According to an outline of the content areas dealt with in special education by Thorton et al. (1983), this is the case in the United States as well as in The Netherlands.

[12] This example is from a sixth-grade unit in the textbook series ‘Niveaucursus Rekenen’ (Vossen et al., s. a.).

13 This example is from the Dutch educational television program ‘Pluspunt’ (Scholten and Ter Heege, 1983-1984; Ter Heege, Van den Heuvel-Panhuizen, and Scholten, 1983-1984). More about this program can be found in Section 1.3.2. This illustration is from the fifth-grade booklet.

14 Strictness was not necessary, due to the purpose of this test. However, such strictness would certainly be essential if the purpose were to investigate what exactly determines the level of difficulty of a ratio problem. Apart from whether each kind of ratio problem is non-numerical or numerical, many other features can be distinguished, such as:
- ratios within or between magnitudes;
- ratios involving one object or more than one;
- ratios involving simple magnitudes or compound magnitudes;
- ratios involving one, two or three-dimensional magnitudes;
- ratios which imply an increase or decrease in the value of something;
- ratio problems in which the standard of one is given or not, or can be calculated or not;
- ratio problems which can either be solved by internal or external comparison, or by both;
- ratio problems which do or do not require any numerical knowledge of everyday matters;
- ratio problems which do or do not use a formal ratio language;
- ratio problems in which something must be precisely calculated or which can be solved by estimating (also see Van den Heuvel-Panhuizen, 1990b).

15 All but one of the test problems involving finding the fourth proportional were taken from the MORE tests.

16 The actual size of the test pages was twelve by seventeen cm.

17 In addition to the test problems in which the answer had to be drawn, some other test problems, too, were constructed in such a way that a range of answers could be considered correct. This was the case, for instance, in the test problem about the pen (see Figure 6.5). In this problem, $3\frac{1}{2}$ could have also been marked correct; this was not done, however, in the current analysis. Another example was the problem about gasoline (see Figure 6.8), where, instead of ƒ20 being the sole correct answer, amounts in the vicinity could also be marked correct. Lastly, there was the problem about the father and son who were measuring the length of their garden (see Figure 6.12). Here any answer between 25 and 35 was accepted as correct.

18 The arrows on this piece of scratch paper indicate the entrance (‘ingan’) and the exit (‘uitgang’). The correct Dutch spelling of the former is ‘ingang’ and ‘uitgang’.

19 The translation of the Dutch text on the pieces of scratch paper is as follows: (a): ‘guilders’; (c): ‘you divide it into 5 + 5’; (d): ‘orange wins’.

20 In this instance, the scratch paper merely contained the words ‘orange wins’; ‘orange’ refers to the color of the Dutch national soccer team.

21 As a matter of fact, one of the special education psychologists, who remarked beforehand on the difficulty factor of using contexts, did not actually give a low estimate of the rate of success. The remark in question was: “Actually, the problems should also be administered in their bare form.” According to him, this would make the problems more accessible to the students. This remark demonstrates once again that, in special education, the assumption is that problems will be easier when presented in a bare form than within a context.

22 The dots in the diagram show the percentage correct for the first three test problems if both the multiplicative and the additive answers were considered correct.

23 Indications for this were also found in a pilot study on special education that was conducted for the PPON (see 1.4.2.). This study took place at about the same time as the study described in this chapter. Both with respect to the issue of the contexts and the issue of the application and awareness of strategies, the same conclusions as those stated in this chapter were drawn by Kraemer, Bokhove, and Janssen (1991).

7 The safety-net question – an example of developmental research on assessment[1]

7.1 Arguments and concerns regarding open-ended problems

One of the most striking characteristics of today's world-wide reform of mathematics assessment is the shift that has occurred from closed to open problems or constructed-response problems, in which the students must formulate the answers on their own.[2] Moreover, in most cases, these are problems that may have more than one correct answer.

An important feature of assessment within RME, too, is the use of this type of problem. Since the very outset of this movement to reform mathematics education, a strong preference for open questions has prevailed (see Sections 1.2.3e and 1.2.4d). This preference is inextricably bound to how mathematics is viewed within this approach, and to the goals pursued by this approach with respect to education. RME is based on the concept of mathematics as a human activity (Freudenthal, 1973), in which the main goal of mathematics education is that students learn to mathematize. Mathematization implies that students must be able to analyze and organize problem situations by using mathematical tools that, to a certain extent, they had developed themselves. Assessment adhering to this viewpoint must be designed in such a way that it will expose these mathematizing activities or their results as much as possible. The assessment, in other words, instead of merely providing a number of answers from which the correct one must be selected, should offer the students the opportunity to construct their own answer.

As mentioned before (see Section 3.3.5), RME is not alone in stressing the importance of open problems. Arguments in favor of such problems are being made internationally (see, among others, Sullivan and Clarke, 1987, 1991; Clarke, 1988; NCTM, 1989; Stenmark, 1989; Pandey, 1990; Romberg, Zarinnia, and Collis, 1990; Lamon and Lesh, 1992; Swan, 1993).

Recently, however, some concerns have also been raised about the use of open-ended problems. Such concerns can be found, for instance, in Clarke (1993b) and in Lamon and Lesh (1992); these concerns have already been discussed in Chapter 3 (see Sections 3.3.5d and 3.3.5g). Clarke, and Lamon and Lesh, mention the limitations inherent in open-ended problems, by which they mean that the teachers (or researchers) do not always manage to obtain the specific information required. The obvious reaction to this limitation is then to narrow the scope of the problem. In other words, 'improving' open-ended problems consists of a return to a more closed problem format.

This reaction offers a perfect illustration of the tension that can exist between openness and certainty. On the one hand, a constructed-response problem gives students the liberty to tackle it as they wish. This liberty makes the problem very informative by exposing the students' thought processes. On the other hand, however, such liberty means that one cannot always be certain of every aspect of these thought processes. This is especially true of those aspects that do not explicitly appear in the answers. In order to obtain more certainty about a particular aspect of the students' understanding, the problem must focus on that specific aspect. As a consequence, the problem might then be made less open. This, however, would in turn result in the problem being less informative, if one regards it in a broader perspective.

In the specific developmental research on assessment that is the focus of this chapter, this tension between openness and certainty was confronted in the context of written assessment.

7.2 The research context

The developmental research on assessment discussed in this chapter was conducted in the framework of the 'Mathematics in Context' project. This project commenced in 1991 and is expected to run through 1995. The aim of the project, which is being sponsored by the National Science Foundation, is to develop a new American middle school mathematics curriculum for grades 5 through 8 (Romberg (ed.), 1993). The project is being conducted by the National Center for Research in Mathematical Sciences Education at the University of Madison, in collaboration with the Freudenthal Institute of the University of Utrecht. The curriculum being developed must reflect the mathematical content and the teaching methods suggested by the Curriculum and Evaluation Standards for School Mathematics (NCTM, 1989). The philosophy behind this project is the belief that mathematics, like any other body of knowledge, is the product of human inventiveness and social activities. As such, this approach has much in common with RME.

One of the forty teaching units that has been developed for the 'Mathematics in Context' project is a fifth-grade unit entitled 'Per Sense' (Van den Heuvel-Panhuizen, Streefland, Meyer, Middleton, and Browne, in press). The goal of this unit is to help students make sense of percentage. In this unit, the students are confronted with problems requiring some form of standardization before comparisons between the different quantities can be made. Unlike the traditional approach to teaching percentage, the unit begins by exploring the students' informal knowledge. Moreover, the unit does not stress the learning of algorithms. Instead of learning all kinds of procedures, the students are introduced to a 'qualitative' approach to percentage, in which estimating and making a link to simple fractions and ratios play an important role. The percentage (or fraction) bar (which later becomes a double number line) and the ratio table are used as a support for this way of thinking.

The Per Sense unit contains several different kinds of assessment: (i) an initial assessment at the beginning of the unit, (ii) assessment activities at the end of each chapter, (iii) assessment activities during the unit, and (iv) a final, more formal, assessment at the end of the unit. The developmental research on assessment that was conducted in the design of this unit primarily focused on the final assessment. A two-stage research format was used in which assessment problems on percentage were designed and field tested, and then revised and field tested again. A detailed report of this research can be found in Van den Heuvel-Panhuizen, 1995a. The present chapter restricts itself to one test problem taken from this final assessment.

7.3 The first stage of the developmental research

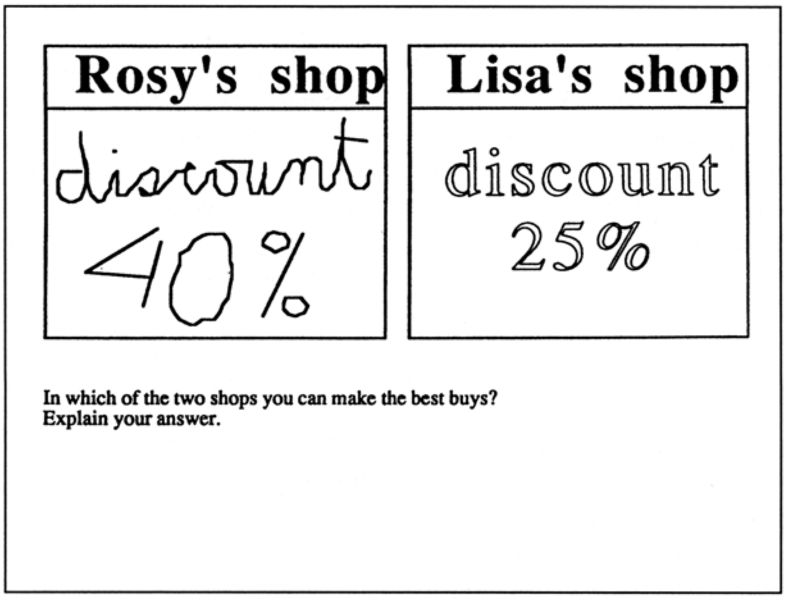
The initial ideas on how percentage should be assessed evolved more or less simultaneously with the development of the unit itself. The point of departure for both were mathematical-didactical issues such as:

- what is the quintessential feature of percentage
- what should one learn with respect to percentage
- in what kinds of situations do percentages arise (in what situations will students encounter them)
- how do percentages arise (what do they require from the students in a mathematical sense)
- how can the students' understanding of and skills with percentage become visible, and
- what kinds of appealing activities can students do with percentage?

Responses to these issues eventually produced the 'Per Sense Test' (Van den Heuvel-Panhuizen, 1992b), a test that was designed to cover at least some of the key concepts and skills involving percentage. It consists of a total of ten problems, including the Best Buys problem that lies at the heart of this chapter.[3]

7.3.1 First version of the Best Buys problem

A key feature of percentage – which must be understood in order to have insight into this concept – is that a percentage is a relation between two numbers or magnitudes that is expressed by a special ratio, namely, 'so-many-out-of-the-hundred'. It is not necessary, however, to explain percentage in this manner to fifth-grade students. On the contrary, it is inadvisable, unless one wishes schools to become temples of meaningless verbiage (Schoemaker, 1993). Students should develop insight into the meaning of percentages through using them, and not merely be handed definitions. They must develop a cognizance of the fact that percentages are always related to something and thus cannot be compared without taking this 'something' into account.



In which of the two shops you can make the best buys?
Explain your answer.

Figure 7.1: First version of the Best Buys problem

This awareness was assessed by means of the Best Buys problem (see Figure 7.1). The familiar situation of items on sale was chosen for assessing whether the students understood the above-mentioned relative property of percentages. The context of the problem was two stores, both of which were having a sale. The first store was offering a discount of 25%, the second a discount of 40%. A large poster advertising the sale had been placed in each shop window. The design of the two posters suggested that the quality of the wares in the two shops might be different. This hint was expressly given in order to encourage the students to consider what might actually be on sale. In other words: what were the percentages referring to?

7.3.2 Research issues

This problem (and the other test problems) were field tested in order to find out (i) what the problem exposed with respect to the students' understanding of and skills with percentage, (ii) whether the problem was informative for further instruction, and (iii) how the problem could be improved, if necessary. The focus in this chapter is on the final issue. The experiences gained from administering the problem, and the indications for improvement that were provided by the students' responses were crucial for answering this question.

7.3.3 Context of data collection

The ‘Per Sense unit was field tested in three seventh-grade classes[4] from two schools in a town near Madison, Wisconsin, in May, 1992. One of the seventh-grade classes was a special class for low achievers. Although this class did take the test, its scores are not included in the results discussed here. The two other classes, which contained a total of 39 students, can be classified as regular classes.

The Per Sense test was administered after the classes had completed the unit, which took just over three weeks. The test was administered by the respective teacher of each class, and the students worked on the test individually. There was no time limit. It was the first time that these students had taken a formal test that included problems like the Best Buys problem. The students’work as discussed in the following section was evaluated by this author.

7.3.4 Results of the first version of the Best Buys problem

The analysis of the responses (see Table 7.1) shows that at least half the students (20 out of 39) understood that one cannot compare percentages without taking into account what they refer to.

Table 7.1: The responses to the Best Buys problem[5]

Best Buys problem		
Response categories	**n**	**Examples**
a Taking the original price into account	15	– “It depends on the original price of the objects they are selling” – “Both, I mean how much do the items cost, nobody knows” – “Lisa’s, because if you buy something that’s already been used, you will have to fix it up or ...”
b Taking the same price as an example	3	– “Rosy’s, if something at both stores was $30.75. At Rosy’s it would be $12.33, at Lisa’s it would be $28.95”
c Taking the same price as an example; wrong conclusion	2	– “Lisa’s, because for example a shirt costs $50; 40% = $20 and 25% = $12.50; with Lisa’s deal you’re paying less”
d Comparing the percentages absolutely	18	– “Rosy’s, 40% is better than 25%” – “Rosy’s, because it is closer to one hundred percent, so there would be more off”
e No answer	1	

The majority of this group responded to the problem by indicating explicitly that the answer would depend on the original price. Three students did this more indirectly. They took as an example an item having the same original price in both shops and then performed some calculations. Two students proceeded in the same way but came to the wrong conclusion. This raised the issue of how to evaluate such a response. Even though they did not give the correct answer, these students obviously knew that a percentage is related to something. Because the assessment was focused on this awareness, and not on accurate reading and precise performance of the tasks, this was considered a reasonable response. Lastly, nearly half the students compared the two percentages absolutely.

Upon analyzing the results, however, it became clear that, although the responses showed a great variety in levels of understanding (which is very informative for further instruction), one could not always be sure that those students who had compared the percentages absolutely did, indeed, lack an understanding of the relative nature of percentage. Moreover, there was the issue of clarity in terms of how the problem was stated. What is meant by a best buy? Does it refer to the cheapest price or to the greatest discount in dollars? That this might confuse students can be seen from the response that reached the wrong conclusion. It is possible that these students switched their point of view while reasoning.

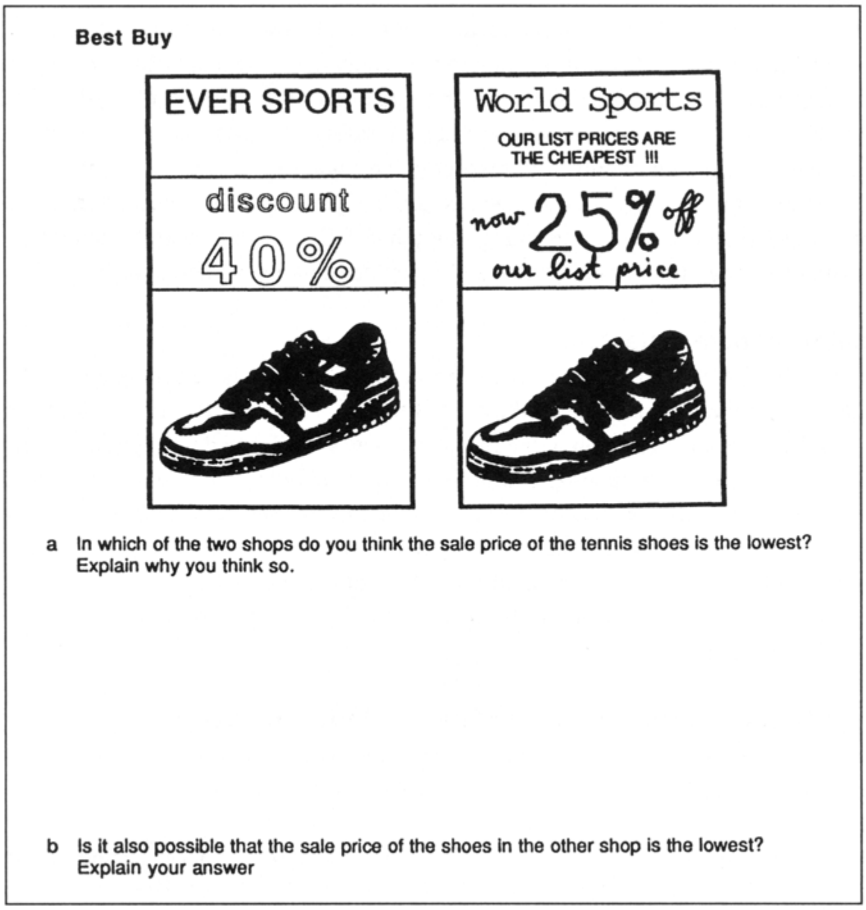
7.4 The second stage of the developmental research

The experiences with and results from the first stage of the developmental research formed the starting point for the second stage of the research. In this stage, a revised set of problems was developed that was made into a new version of the test. This new version was given the new title of 'Show-what-you-know Book on Percents' (Van den Heuvel-Panhuizen, 1993c). The test was developed through deliberation and reflection on the original problems and their responses by a number of the research staff.[6]

7.4.1 Second version of the Best Buys problem

The field test of the first version of the problem had revealed two issues needing resolution: the lack of clarity in what was meant by a best buy, and the uncertainty as to whether the students who compared the percentages absolutely did, indeed, lack an understanding of the relative nature of percentages.

In order to eliminate the confusion surrounding the interpretation of 'best buy', it was decided to make the question more specific and to ask the students which store would have the lowest price. The problem was also made more specific by having only one item be on sale, in this case, a pair of tennis shoes. During the deliberation among the research staff, it was also suggested that the price-tags on the shoes show the same list price, in order to indicate that the same tennis shoes were being sold by the two stores. This, however, would have taken the heart out of the problem. The students would then only have needed to compare the percentages absolutely. A better alternative was to show by the illustration in the advertisement that both stores were selling the same tennis shoes. As in the first version of the problem, different advertising styles were used to indicate that these same shoes might have different list prices. Because of the concern that this indication might not be strong enough, it was decided to stress the difference between the two stores by adding the slogan ‘Our list prices are the cheapest’ to the more slapdash advertisement (see Figure 7.2).



Best Buy

a In which of the two shops do you think the sale price of the tennis shoes is the lowest? Explain why you think so.

b Is it also possible that the sale price of the shoes in the other shop is the lowest? Explain your answer

Figure 7.2: Second version of the Best Buys problem with the safety-net question

To overcome the uncertainty involved in the openness of the problem, the decision was made to emulate what teachers do after having asked a question, when unsure of how to interpret the answers. In such situations, a teacher will often append an additional, more specific, question. Because the function of such an additional question, in this case, was to identify those students who understood the relative nature of percentage but needed some extra hints in order to be able to apply this understanding, this question was called a ‘safety-net question’. The problem, however, was how such a question should be stated. Asking whether the shoes could also be cheaper in the store offering a 25% discount would not be acceptable, because this would give away the assumed answer to the first question. It was therefore decided to append a more neutral question that could apply to either of the answers to the first question (see Figure 7.2).

7.4.2 Research issue

After the new test had been composed, another field test was organized. As far as the revised version of the Best Buys problem was concerned, the main objective of this field test was to provide information on the function of the safety-net question. The research issue to be addressed was whether the safety-net question had contributed to a higher degree of certainty with respect to the students' understanding of the relative nature of percentage. In other words, did the question really succeed in identifying those students who, even though they had failed to demonstrate this understanding initially, did, in fact, understand the relative nature of percentage.

7.4.3 Context of data collection

As with the first version of the test, the second version was also administered to the students after they had completed the Per Sense unit. This time, the unit – which had also undergone some revision in the meantime – was field tested in three fifth-grade classes. Again, the participating classes were from schools in a town nearby Madison, Wisconsin. Different teachers were involved in this second round of field testing. The ability levels within the classes in question ranged from mixed to homogeneous. Two of the three classes, involving a total of 44 students, did the new version of the test.[7] The students' work as discussed in the following section was again evaluated by this author.

7.4.4 Results of the second version of the Best Buys problem

The answers to the first part of the question in the second version were found to be about the same as the responses to the first version (see Table 7.2).

Table 7.2: Results in the two field tests; the results of the Best Buys problem compared with the results of the first question of the revised Best Buys problem

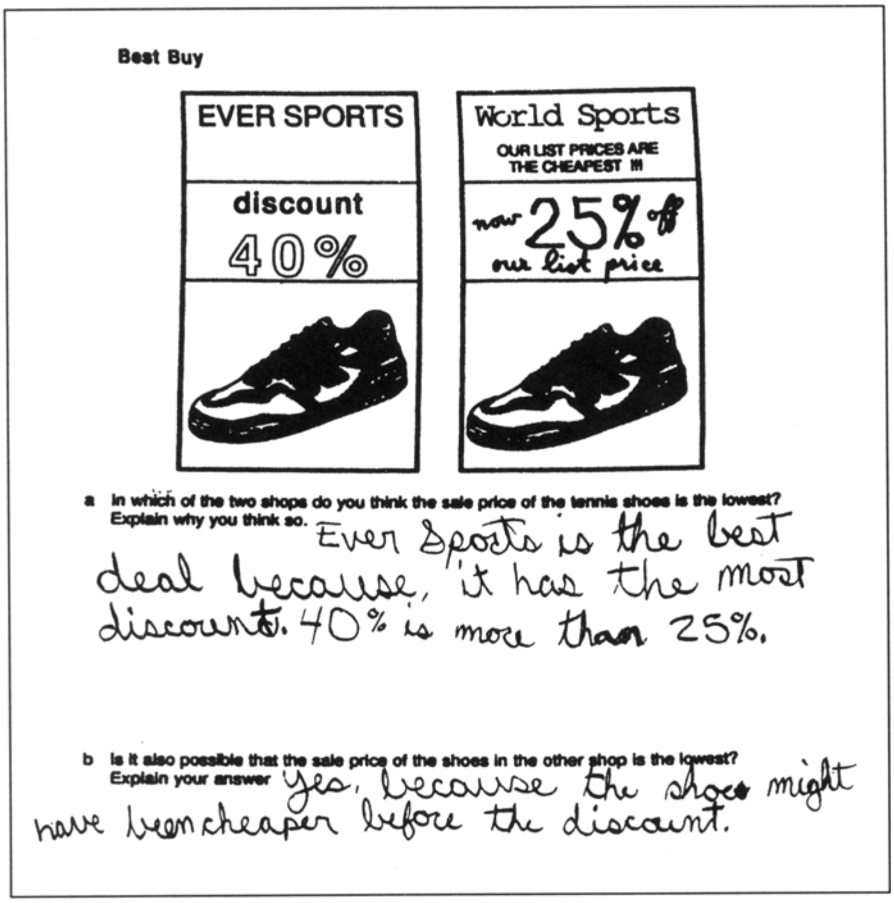
Best Buys problem	Per Sense Test grade 7 n = 39		Show-what-you-know Book (first question of the problem) grade 5 n = 44	
Response categories	**n**		**n**	
a Taking the original price into account	15	38%	15	34%
b Taking the same / a different price as an example	3/0	8%	3/1	9%
c Taking the same / a different price as an example; wrong conclusion	2/0	5%	0/1	2%
d Comparing the percentages absolutely	18	46%	21	48%
e No answer / unclear	1/0	3%	0/3	7%

Considering the differences one would expect between fifth and seventh grade classes, this lack of difference in the results was certainly surprising. Obviously, the revised problem that was presented to the fifth graders was stated more clearly. Maybe the fifth-grade classes were exceptionally bright or the seventh-grade classes unusually slow; or perhaps the unit had been dealt with more thoroughly in the fifth-grade classes.[8] Another explanation for the absence of any difference in results between the two grades might be that the problem assessed a qualitative understanding of percentage rather than a quantitative one. The capacity of a wider range of students to solve the problem correctly could be a characteristic of such qualitative assessment. The purpose of the study, however, was not to investigate this issue, but, rather, to examine the effect of the safety-net question.

Table 7.3: The results of both questions of the revised version of the Best Buys problem

Show-what-you-know Book **Best Buys problem**			**Safety-net question**				
			Response categories a Taking the original price into account	b Taking the same or a different price as an example	c Taking the same or a different price as	d Comparing the percentages absolutely	e No answer / unclear
First question	**Response categories**	**n**	**n**	**n**	**n**	**n**	**n**
	a Taking the original price into account	15	15				
	b Taking the same or a different price as an example	4	2			2	
	c Taking the same or a different price as an example; wrong conclusion	1	1				
	d Comparing the percentages absolutely	21	17			1	3
	e No answer or unclear	3	2				1
		44	37			3	4

As can be seen in Table 7.3, the results changed remarkably when the second question in the revised problem was taken into account. About 80% of the students (17 out of 21) who had at first compared the two percentages absolutely, showed a clear understanding of the relative nature of percentage in their response to the second question. Figure 7.3 is a typical example.



Best Buy

EVER SPORTS

discount

40%

World Sports

OUR LIST PRICES ARE THE CHEAPEST !!!

now 25% off our list price

a In which of the two shops do you think the sale price of the tennis shoes is the lowest? Explain why you think so. Ever Sports is the best deal because, it has the most discount. 40% is more than 25%.

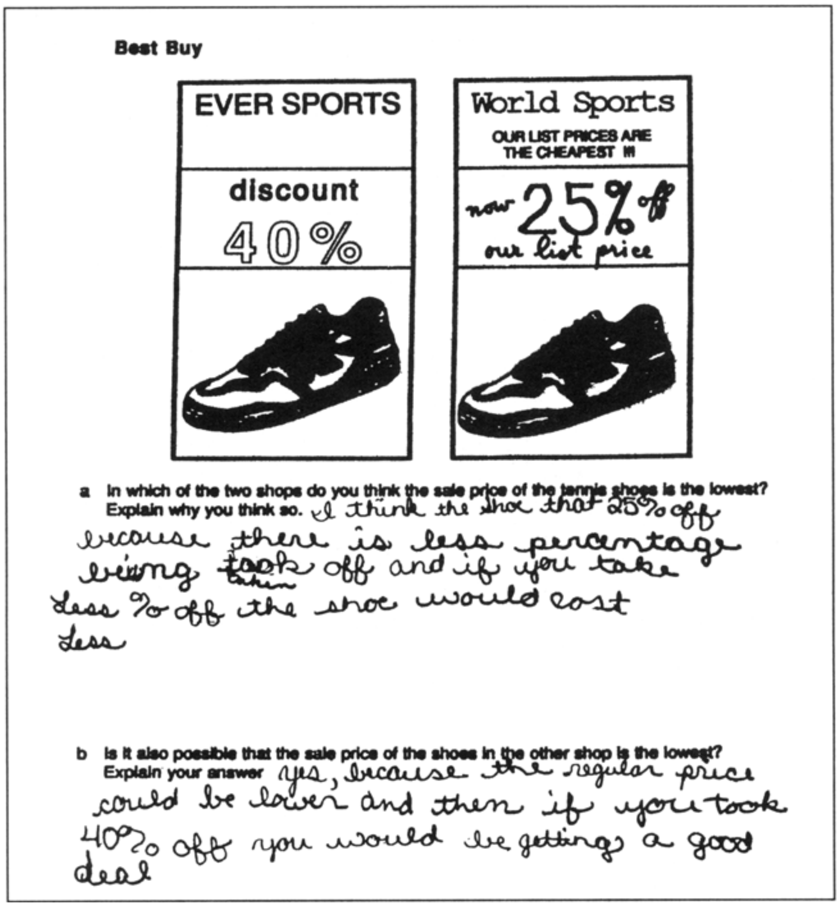
b Is it also possible that the sale price of the shoes in the other shop is the lowest? Explain your answer Yes, because the shoes might have been cheaper before the discount.

Figure 7.3: Example of student work

The advantage of following an open question with a safety-net question is that the students then truly have the liberty to answer the first question in their own way. As can be seen from the answer to the first question in Figure 7.4, this may not always exactly correspond with what the inventor of the problem had in mind.

At first glance, it is not very clear what the student, whose work is shown in Figure 7.4 meant by this answer. One might tend to conclude – particularly with respect to the statement,

"...if you take less % off the shoe would cost less"



Best Buy

EVER SPORTS

discount

40%

World Sports

OUR LIST PRICES ARE THE CHEAPEST !!!

now 25% off our list price

a In which of the two shops do you think the sale price of the tennis shoes is the lowest?
Explain why you think so. I think the shoe that 25% off because there is less percentage being ~~took~~ taken off and if you take less % off the shoe would cost less

b Is it also possible that the sale price of the shoes in the other shop is the lowest?
Explain your answer yes, because the regular price could be lower and then if you took 40% off you would be getting a good deal

Figure 7.4: Example of student work

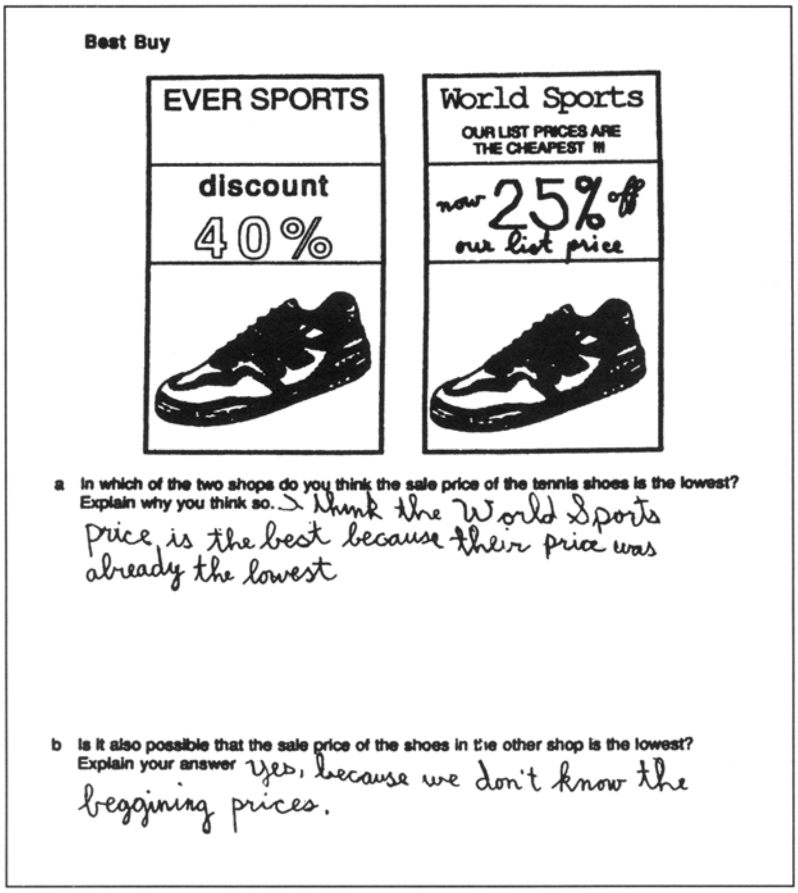
– that a true understanding of percentage is not yet present. But the student's answer to the safety-net question immediately removes all doubt on this matter. Moreover, it provides the reader the opportunity to regard the first question from this student's standpoint. Although formulated clumsily, the student obviously meant,

> "... if you take off less %, the shoe would [have] cost less [beforehand]".

The second question, by helping reconstruct the students' answers, can thereby function as a safety net in this way, too. More on reconstructing students' answers can be found in Van den Heuvel-Panhuizen, 1995a (see also Section 4.1.5d).

Furthermore, the safety-net question proved to function conversely as well. Two students, whose answers to the first question had been considered reasonable, revealed by their answers to the second question that they did not, in fact, have insight into the relative nature of percentage.

It is extremely important that students who display insight in their answers to the first question not become confused by the safety-net question. With respect to the revised version of the Best Buys problem (which is called Best Buy problem), most of the students who had answered the first question correctly actually provided further elucidation in their answer to the second question (see Figure 7.5).



EVER SPORTS
discount
40%
World Sports
OUR LIST PRICES ARE THE CHEAPEST !!!
now 25% off
our list price

Best Buy

a In which of the two shops do you think the sale price of the tennis shoes is the lowest? Explain why you think so. I think the World Sports price is the best because their price was already the lowest

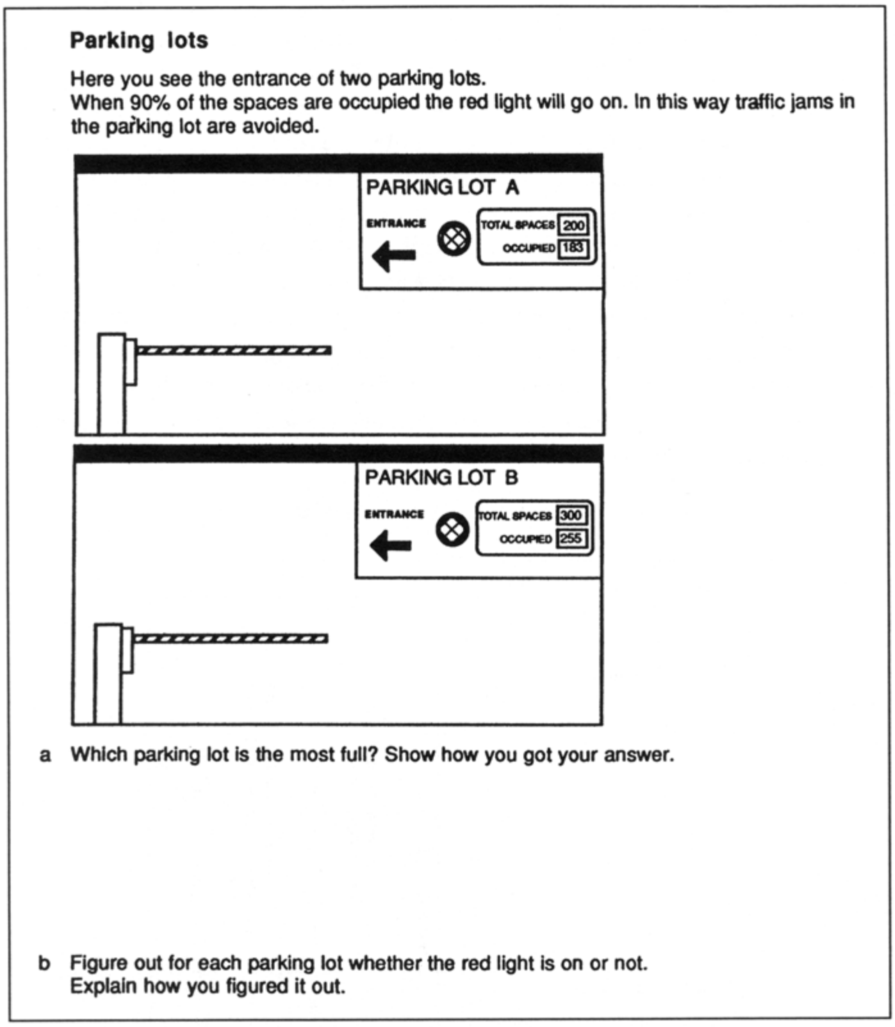
b Is it also possible that the sale price of the shoes in the other shop is the lowest? Explain your answer yes, because we don't know the beggining prices.

Figure 7.5: Example of student work

The safety-net question only proved to be redundant in one instance. This student did not answer the safety-net question, but instead drew arrows pointing to the first answer, which had already revealed an understanding of the relative nature of percentage.

7.4.5 A second examination of the safety-net question

In addition to the Best Buy problem, one other problem in the ‘Show-what-you-know Book on Percents’ also used the safety-net question. This was the Parking Lots problem (see Figure 7.6).



Parking lots

Here you see the entrance of two parking lots.
When 90% of the spaces are occupied the red light will go on. In this way traffic jams in the parking lot are avoided.

PARKING LOT A
ENTRANCE
TOTAL SPACES 200
OCCUPIED 183
PARKING LOT B
ENTRANCE
TOTAL SPACES 300
OCCUPIED 255

a Which parking lot is the most full? Show how you got your answer.

b Figure out for each parking lot whether the red light is on or not.
Explain how you figured it out.

Figure 7.6: Parking Lots problem

The purpose of the Parking Lots problem was to assess to what extent students would apply percentages spontaneously when comparing two 'how-many-out-of-the ...' situations which can be transposed to 'so-many-out-of-the-hundred'. Table 7.4 shows that about half the students (24 out of 44) solved the first part of the problem directly through relative reasoning, by taking into account the total number of spaces in each of the parking lots. A wide variety of strategies was used here: calculating the percentages, approximating by using the percentage bar, converting a fraction into a percentage, using the ratio table, and global relative reasoning. Nearly

40% of the students (16 out of 44), however, calculated the absolute difference, namely, the difference between the number of available spaces.

As was the case in the Best Buy problem, this latter strategy did not necessarily mean that these students were not able to solve the problem by using relative reasoning and applying percentages! Therefore, in order to avoid drawing the wrong conclusion regarding the abilities of these students, a safety-net question was added, which prompted the students more directly to demonstrate their abilities. Once again, the safety-net question functioned as intended. About half the students (9 out of 16) who had applied an absolute strategy when answering the first question, now applied a relative strategy (see Table 7.4). On the whole, it may be stated that a far clearer picture was obtained of the students' understanding, thanks to the safety-net question. This was particularly the case in one of the classes, where the great majority of those students who had first given an absolute answer, subsequently took into account the total number of spaces when answering the second question.

Table 7.4: The results of both questions of the Parking Lots problem

Show-what-you-know Book **Parking Lots problem**			**Safety-net question**		
		Response categories	a Relative answer, taking into account the magnitude of the parking lots	b Absolute answer, not taking into account the magnitude of the parking lots	c Questionable answer
First question	**Response categories**	**n**	**n**	**n**	**n**
	a Relative answer; taking into account the magnitude of the parking lots	24	22	1	1
	b Absolute answer; not taking into account the magnitude of the parking lots	16 (11)*	9 (8)	6 (2)	1 (1)
	c Questionable answer	4	3		1
		44	34	7	3

*The numbers between parentheses belong to one of the two classes

7.5 Final remarks

It has long been believed that individual interviews are the only possible means for obtaining true insight into students’ understanding, thought processes and strategies. Even the root of the verb ‘to assess’ – which means ‘to sit beside’ – refers to this. It implies that the assessor must ‘sit’ with a learner in order to be certain that the student’s answer really means what it seems to mean (Wiggins, 1989b). So it is not surprising that interviews are the first thing that come to mind if more certainty is required. See, for instance, the following remark by Clements (1980, p. 7):

> “It is obvious that any inferences about a child’s thinking drawn from his written response alone represent little more than guesswork on the part of the researcher. Written responses can suggest to a researcher, or teacher, reasons why a child is making errors, but structured interviews must be conducted with the child before consistent patterns of errors can be determined with any degree of certainty.”

This was indeed the viewpoint in RME as well. In the RME approach, a strong preference was expressed from the outset for observing and interviewing. In recent years, however, a new appreciation of written assessment has begun to emerge (see Chapters 3 and 4). A significant catalyst for this shift was the securing of the new RME-based secondary education curriculum by means of new written exams. Subsequently, new possibilities for written assessment were also developed on the primary school level.

The safety-net question discussed in this chapter was a result of further developmental research along these lines. The results presented here clearly demonstrate the capacity of the safety-net question to increase certainty with respect to students’ understanding without, however, making the problems more closed. By keeping the problems open, one avoids the disadvantage of losing important information on the students’ thought processes. Adding a more specific question to an open problem, by contrast, has the advantage of maintaining this information while also obtaining more certainty with regard to the students' understanding.

Another matter that was revealed by these findings is the feasibility of adopting a typical interview technique – like asking additional questions – in written assessment; moreover, this proved to be very informative. The integration of written assessment and interviewing, which up to now have been considered poles apart, has turned out to be extremely fruitful for written assessment. It opens a new avenue for improving written assessment, namely, the application of interview techniques. Other examples of such techniques are the ‘second-chance question’ and the ‘standby sheet’ (see Section 4.1.4).

Furthermore, this integration of interview techniques into written assessment is in accord with the shift from a static approach to written assessment to a more dynamic approach, which is characteristic of assessment both within RME and within other current reform movements in mathematics education and assessment (see also Section 4.2.1b).

But obtaining more certainty regarding students understanding by applying interview techniques is only one part of the story. One would be incorrect in assuming that the improvement of written assessment within RME is solely focused on increasing certainty. On the contrary, RME actually argues in favor of a rich uncertainty (see Section 4.2.1d). Indeed, this approach even advocates ending the pursuit of certainty that has traditionally dominated assessment, and thereby obstructed further developments in written assessment. The acceptance of a rich uncertainty can create room for further progress. As a matter of fact, the attempt to improve written assessment described in this chapter, while a result of such an acceptance, turned out in the end actually to provide more certainty.

Notes

[1] This chapter is a somewhat adapted version of a paper presented at the American Educational Research Association 1995 in San Francisco (see Van den Heuvel-Panhuizen, 1995c). Both the paper and this chapter are extracted from an extensive report on assessment research conducted in the framework of the 'Mathematics in Context' project. The research was linked to a teaching unit on percentage (see Van den Heuvel-Panhuizen, 1995a).

[2] See Section 3.3, and, especially, Note 47 in Section 3.3.5d.

[3] A more detailed description of the content and results of the first version of this test on percentage can be found both in the extensive report (see Van den Heuvel-Panhuizen, 1995a) and in Van den Heuvel-Panhuizen, 1994c.

[4] As no fifth-grade classes were available at the time the Per Sense unit was to be field tested, seventh-grade classes were used instead. Another reason for this switch was the assumption that the final draft version of the unit had turned out to be rather difficult for the fifth-grade level.

[5] The dotted line in the table indicates a possible cut-off point between reasonable and unreasonable answers.

[6] In addition to this author, the following people were involved in this deliberation: Koeno Gravemeijer, Jan de Lange, Meg Meyer, Jim Middleton, Martin van Reeuwijk, Leen Streefland and Adri Treffers.

[7] The other class took a test that was made by their teacher.

[8] The unit was improved before it was field tested the second time, and the teachers had received better preparation in how to present the unit.

Bibliography

Assessment Focus Group (AFG) (1991). *Issues and questions associated with research on alternative assessment.* A Report of the Assessment Focus Group, NCTM Standards Research Catalyst Conference II, December 6-8, 1991, Miami, Florida.

Australian Education Council (AEC) (1991). *A National Statement on Mathematics for Australian Schools*. Carlton, Victoria: Curriculum Corporation.

Baker, E.L., O'Neil Jr, H.F., and Linn, R.L. (1993). Policy and Validity Prospects for Performance-Based Assessment. *American Psychologist, 48* (12), 1210-1218.

Bell, A. (1993). Principles for the design of teaching. *Educational Studies in Mathematics, 24 (*1), 5-34.

Bell, A., Burkhardt, H., and Swan, M. (1992a). Balanced Assessment of Mathematical Performance. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 119-144). Washington: AAAS Press.

Bell, A., Burkhardt, H., and Swan, M. (1992b). Assessment of Extended Tasks. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 145-176). Washington: AAAS Press.

Bell, A., Burkhardt, H., and Swan, M. (1992c). Moving the System: The Contribution of Assessment. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 177-193). Washington: AAAS Press.

Bennett, N., Desforges, Ch., Cockburn, A., and Wilkinson, B. (1984). *The quality of pupil learning experiences*. London: Lawrence Erlbaum.

Bishop, A.J., and Goffree, F. (1986). Classroom organisation and dynamics. In B. Christiansen, A.G. Howson, and M. Otte (eds.), *Perspectives on Mathematics Education* (pp. 309-365). Dordrecht: Reidel Publishing Company.

Bloom, B.S., and Foskay, A.W. (1967). Formulation of hypotheses. In T. Husen (ed.), *International Study of Achievement in Mathematics. A comparison in twelve countries* (I, pp. 64-76). Stockholm: Almqvist and Wiskell.

Bloom, B.S., Hastings, J.Th., and Madaus, G.F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning.* New York: McGraw-Hill. (cited by Freudenthal, 1978a)

Bodin, A. (1993). What does to assess mean? The case of assessing mathematical knowledge. In M. Niss (ed.), *Investigations into Assessment in Mathematics Education* (pp. 113-141). Dordrecht: Kluwer Academic Publishers.

Boertien, H., and De Lange, J. (1994). *Model voor nationale optie TIMSS 1995* [Model for the national option TIMSS 1995]. Enschede: Universiteit Twente, OCTO. (internal paper)

Bokhove, J. (1987). Periodieke peiling onderwijsniveau (PPON) – organisatie en inhoudelijke bepaling [National Asesssment of Educational Achievement (PPON) – organisation and determination of content]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 18-22). Utrecht: OW&OC, Utrecht University.

Bokhove, J. and Moelands, F. (1982). Leerdoelgerichte toetsen kommagetallen [Criterion-referenced tests on decimals]. *Willem Bartjens, 2* (1), 29-35.

Borasi, R. (1986). On the nature of problems. *Educational Studies in Mathematics, 17* (2), 125-141.

Boswinkel, N. (1995). Interactie, een uitdaging. *Tijdschrift voor Onderzoek en Nascholing van het Reken-wiskundeonderwijs, 14*, 1, 4-11.

Brandt, R. (1992). On Performance Assessment: A Conversation with Grant Wiggins. *Educational Leadership, 49* (8), 35-37.

Broadfoot, P. (1994). *The myth of measurement.* Inaugural address, University of Bristol. (cited by Gipps, 1994)

Broekman, H., and Weterings, J. (1987). Invloed van toetsen en proefwerken op het wiskunde-onderwijs [Influence of tests and more informal written assessment methods on mathematics education]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 33-37). Utrecht: OW&OC, Utrecht University.

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics. In H.G. Steiner et al., *Theory of Mathematics Education* (pp. 110-119). Bielefeld: University of Bielefeld, Institut für Didaktik der Mathematik.

Brown, J.S., Collins, A., and Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher, 18* (1), 32-42.

Brown, M. (1993). Assessment in Mathematics Education: Developments in Philosophy and Practice in the United Kingdom. In M. Niss (ed.), *Cases of Assessment in Mathematics Education* (pp. 71-84). Dordrecht: Kluwer Academic Publishers.

Brush, L.R. (1972). *Children's Conceptions of Addition and Subtraction: The Relation of Formal and Informal Notations*. Unpublished Doctoral thesis, Cornell University. (cited by Ginsburg, 1975)

Brush, L.R., and Ginsburg, H. (1971). *Preschool Children's Understanding of Addition and Subtraction.* Unpublished manuscript, Cornell University. (cited by Ginsburg, 1975)

Burns, M.S., Vye, N.J., Bransford, J.D., Delclos, V., and Ogan, T. (1987). Static and Dynamic Measures of Learning in Young Handicapped Children. *Diagnostique, 12* (2), 59-73.

Buys, K., and Gravemeijer, K. (1987). Toetsen voor realistisch reken- en wiskunde-onderwijs [Tests for realistic mathematics education]. In: J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 38-43). Utrecht: OW&OC, Utrecht University.

Cain, R.W., Kenney, P.A., and Schloemer, C.G. (1994). Teachers as Assessors: A Professional Development Challenge. In D.B. Aichele (ed.), *Professional Development for Teachers of Mathematics*. 1993 Yearbook NCTM (pp. 93-101). Reston, VA: NCTM.

Campione, J.C. (1989). Assisted Assessment: A Taxonomy of Approaches and an Outline of Strengths and Weakness. *Journal of Learning Disabilities, 22* (3), 151-165.

Campione, J.C. and Brown, A.L. (1990). Guided Learning and Transfer: Implications for Approaches to Assessment. In N. Frederiksen, R. Glaser, A. Lesgold, and M.G. Shafto, (eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition.* Hillsdale, NJ: Erlbaum.

Carpenter, T. (1993). Personal communication.

Carpenter, T.P., and Fennema, E. (1988). Research and Cognitively Guided Instruction. In E. Fennema and Th.P. Carpenter (eds.), *Integrating Research on Teaching and Learning Mathematics* (pp. 2-17). Madison, WI: Wisconsin Center for Education Research, University of Wisconsin Madison.

Carpenter, T.P., and Moser, J.M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15* (3), 179-202.

Carraher, T.N. (1988). Street mathematics and school mathematics. In A. Borbás (ed.), *Proceedings of the 12th International Conference of the International Group for Psychology of Mathematics Education* (Vol. I, pp. 1-23). Veszprem, Hungary: OOK Printing House.

Carraher, T.N., Carraher, D.W., and Schliemann, A.D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology, 3,* 21-29.

Carter, R., Beranek, B., and Newman, D. (1990). *Some characteristics of a good mathematics problem.* A paper presented at the AERA 1990, April 16-20, Boston.

Chambers, D.L. (1993). Integrating Assessment and Instruction. In N.M. Webb (ed.), *Assessment in the Mathematics Classroom*. 1993 Yearbook NCTM (pp. 17-25). Reston, VA: NCTM.

Christiansen, B., and Walter, G. (1986). Task and Activity. In B. Christiansen, A.G. Howson, and M. Otte (eds.), *Perpectives on Mathematics Education* (pp. 243-307). Dordrecht: Reidel Publishing Company.

Clarke, D. (1986). Assessment Alternatives in Mathematics. In N.F. Ellerton (ed.), *Mathematics: Who Needs What?* (pp. 72-75). Parkville: The Mathematics Association of Victoria.

Clarke, D. (1988). *Assessment Alternatives in Mathematics*. Canberra: Curriculum Development Centre.

Clarke, D. (1992). The role of assessment in determining mathematics performance. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 145-168). Hawthorn, Victoria: Australian Council for Educational Research.

Clarke, D.J. (1993a). The language of assessment. In M. Stephens, A. Waywood, D. Clarke, and J. Izard (eds.), *Communicating mathematics perspectives from current research and classroom practice in Australia* (pp. 211-222). Hawthorn, Victoria: Australian Council for Educational Research / Australian Association of Mathematics Teachers.

Clarke, D.J. (1993b). *Open-ended tasks and assessment: the nettle or the rose*. A paper presented at the National Council of Teachers of Mathematics, Research Pre-session to the 71st Annual Meeting, Seattle, WA, March 29-30, 1993.

Clarke, D.J., Clarke, D.M., and Lovitt, Ch.J. (1990). Changes in Mathematics Teaching Call for Assessment Alternatives. In Th.J. Cooney and Chr.R. Hirsch (eds.), *Teaching and Learning Mathematics in the 1990s*. 1990 Yearbook NCTM. Reston, VA: NCTM.

Clements, M.A. (1980). Analyzing children’s errors on written mathematical tasks. *Educational Studies in Mathematics, 11,* 1-21.

Cobb, P. (1987). Information-processing psychology and mathematics education. *The Journal of Mathematical Behavior, 6* (1), 3-40.

Cobb, P., Wood, T., and Yackel, E. (1991). A constructivist approach to second grade mathematics. In E. von Glasersfeld (ed.), *Radical Constructivism in Mathematics Education* (pp. 157-176). Dordrecht: Kluwer Academic Publishers.

Cobb, P., Yackel, E., and Wood, T. (1992). A constructivist alternative to the representational view of mind in mathematics education. *Journal for Research in Mathematics Education, 23* (1), 2-33.

Cockcroft, W. H. (1982). *Mathematics Counts: Report of the Commission of Inquiry into the Teaching of Mathematics in Schools*. London: Her Majesty’s Stationary Office.

Collis, K.F. (1992). Curriculum and assessment: A basic cognitive model. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 24-45). Hawthorn, Victoria: Australian Council for Educational Research.

Collis, K.F., Romberg, Th.A., and Jurdak, M.E. (1986). A Technique for Assessing Mathematical Problem-solving Ability. *Journal for Research in Mathematics Education, 17* (3), 206-221.

Collison, J. (1992). Using Performance Assessment to Determine Mathematical Dispositions. *Arithmetic Teacher, 39* (6), 40-47.

Cooper, B. (1992). Testing National Curriculum mathematics: some critical comments on the treatment of ‘real’ contexts for mathematics. *The Curriculum Journal, 3* (3), 231-243.

Cross, L., and Hynes, M.C. (1994). Assessing Mathematics Learning for Students with Learning Differences. *Arithmetic Teacher, 41* (7), 371-377.

Damen, S. (1990). *Rekenen op lom- en mlk-scholen*. Ervaringen tijdens een stage bij het Instituut voor Toetsontwikkeling (CITO) [Arithmetic in schools for children with learning and behavioral difficulties and in schools for mildly mentally retarded children. Experiences during an internship at the National Institute for Test Development (CITO)]. Arnhem: Cito.

Davis, R.B. (1989). The culture of mathematics and the culture of schools. *Journal of Mathematical Behavior, 8* (2), 143-160.

De Block, A. (1975). *Taxonomie van Leerdoelen* [Taxonomy of instructional objectives]. Antwerpen. (cited by Freudenthal, 1981b)

De Jong, R. (1986). *Wiskobas in methoden* (Dissertatie) [Wiskobas in textbooks (Doctoral dissertation)]. Utrecht: OW&OC, Utrecht University.

De Jong, R. (ed.) (1977). *De abakas* [The abacus]. Utrecht: IOWO.

Dekker, A., Ter Heege, H., and Treffers, A. (1982). *Cijferend vermenigvuldigen en delen volgens Wiskobas* [Column multiplication and division according to Wiskobas]. Utrecht: OW&OC, Utrecht University.

Dekker, R. (1991). *Wiskunde leren in kleine heterogene groepen* (Dissertatie) [Learning mathematics in small heterogeneous groups (Doctoral dissertation]. De Lier: Academisch Boeken Centrum.

Dekker, T. (1993). *Checklist toetsen met contexten* [A checklist for tests with contexts]. Utrecht: Freudenthal Instituut, Utrecht University. (internal paper)

De Lange, J. (1979). Contextuele problemen [Contextual problems]. *Euclides, 55*, 50-60.

De Lange, J. (1985). Toetsen bij Wiskunde A [Tests for Mathematics A]. *Nieuwe Wiskrant, 4* (4), 12-16.

De Lange, J. (1987a). *Mathematics, Insight and Meaning* (Doctoral dissertation). Utrecht: OW&OC, Utrecht University.

De Lange, J. (1987b). Toetsen en HEWET-termen [Tests and Hewet targets]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 122-126). Utrecht: OW&OC, Utrecht University.

De Lange, J. (1992a). Critical factors for real changes in mathematics learning. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 305-329). Hawthorn, Victoria: Australian Council for Educational Research.

De Lange, J. (1992b). Higher order (un)teaching. In I. Wirszup and R. Streit (eds.), *Developments in School Mathematics Education Around the World. Volume Three* (pp. 49-72). Reston, VA: NCTM / Chicago, IL: University of Chicago School Mathematics Project.

De Lange, J. (1995). Assessment: No Change without Problems. In T.A. Romberg (ed.), *Reform in School Mathematics.* Albany, NY: SUNY Press.

De Lange, J. (ed.) (1985a). *Hewet & Toets 2* [Hewet & Test 2]. Utrecht: OW&OC, Utrecht University.

De Lange, J. (ed.) (1985b). *Hewet & Toets 3* [Hewet & Test 3]. Utrecht: OW&OC, Utrecht University.

De Lange, J., and Van Reeuwijk, M. (1993). The Test. In J. de Lange, G. Burrill, T. Romberg, and M. van Reeuwijk, *Learning and Testing Mathematics in Context – The Case: Data Visualization*. Madison, WI: National Center for Research in Mathematical Sciences Education.

De Lange, J., and Verhage, H. (1982). Kritisch kijken [A Critical View]. *Nieuwe Wiskrant, 1* (4), 19-28.

De Moor, E. (1984). *Pluspunt handboek* [Pluspunt handbook]. Hilversum: Nederlandse Onderwijstelevisie.

Department of Education and Science and the Welsh Office (DES / WO) (1989). *Mathematics in the National Curriculum*. London: Her Majesty's Stationary Office.

Desforges, Ch., and Cockburn, A. (1987). *Understanding the Mathematics Teacher. A Study of Practice in First Schools.* London: The Falmer Press.

Diels, P.A. (1933). *Op paedagogische verkenning. Studiën over moderne onderwijs-verschijnselen* [A pedagogical exploration. Studies on modern educational phenomena]. Groningen: J.B. Wolters Uitgeversmaatschappij.

Doig, B.A., and Masters, G.N. (1992). Through children's eyes: A constructivist approach to assessing mathematics learning. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 269-289). Hawthorn, Victoria: Australian Council for Educational Research.

Donaldson, M. (1978). *Children's Mind*. London: Fontana.

Doyle, J. (1993). *A New Way of Testing in Mathematics Classrooms*. Sheboygan, WI: Sheboygan Area School District. (internal publication)

Du Bois, P.H. (1970). *A History of Psychological Testing.* Boston: Allyn and Bacon, Inc.

Elbers, E. (1991a). The Development of Competence and Its Social Context. *Educational*

Psychology Review, 3 (2), 73-94.

Elbers, E. (1991b). Context, Culture, and Competence: Answers To Criticism. *Educational Psychology Review, 3* (2), 137-148.

Elbers, E. (1991c). Context en suggesties bij het ondervragen van jonge kinderen [Context and suggestions when questioning young children]. In J. Gerris (ed.), *Gezinsontwikkeling*. Amsterdam: Swets & Zeitlinger.

Elbers, E. (1992). Rekenen: de (sociale) regels van de kunst [Arithmetic: the (socially) approved manner]. In M. Dolk (ed.), *Panama Cursusboek 10. Rekenen onder en boven de tien*. Utrecht: HMN-FEO / Freudenthal Instituut, Utrecht University.

Elbers, E., Derks, A., and Streefland, L. (1995). *Learning in a Community of Inquiry: Teacher's Strategies and Children's Participation in the Construction of Mathematical Knowledge*. A paper presented at EARLI 1995, Nijmegen.

Elbers, E., and Kelderman, A. (1991). Verwachtingen en misverstanden bij het toetsen van kennis [Expectations and misconceptions in the assessment of knowledge]. *Pedagogische Studiën, 68* (4), 176-184.

Elsholz, R., and Elsholz, E. (1989). The Writing Process: A Model for Problem Solving. *Journal of Mathematical Behavior, 8* (2), 161-166.

Ernest, P. (1989). Developments in Assessing mathematics. In P. Ernest, *Mathematics Teaching. The State of the Art*. New York: The Falmer Press.

Feinberg, L. (1990). Multiple choice and its critics. *The College Board Review, No. 157*, Fall. (cited by Doig and Masters, 1992)

Feuerstein, R. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. Baltimore: University Park Press. (cited by Campione, 1989)

Flener, F.O., and Reedy, J. (1990). Can teachers evaluate problem solving ability? In G. Booker, P. Cobb, and T.N. de Mendicuti (eds.), *Proceedings of the Fourteenth PME Conference* (Vol. I, pp. 127-134). Mexico: Program Committee of the 14th PME Conference, Mexico.

Foxman, D. (1987). *Assessing Practical Mathematics in Secondary Schools*. London: Her Majesty's Stationary Office. (cited by Joffe, 1990)

Foxman, D. (1993). APU's monitoring survey. In M. Niss (ed.), *Investigations into Assessment in Mathematics Education. An ICMI Study* (pp. 217-228). Dordrecht: Kluwer Academic Publishers.

Foxman, D., and Mitchell, P. (1983). Assessing Mathematics 1. APU Framework and Modes of Assessment. *Mathematics in School, 12* (5), 1983, 2-5.

Foxman, D., Ruddock, G., Joffe, L., Mason, K., Mitchell, P., and Sexton, B. (1985). *Mathematical Development: A Review of Monitoring in Mathematics 1978 to 1982. Parts 1 and 2*. London: APU / Department of Education and Science. (cited by Joffe, 1990)

Frederiksen, N. (1984). The real test bias. *American Psychologist, 39* (3), 193-202. (cited by Romberg, Zarinnia, and Collis, 1990)

Freudenthal, H. (1968). Why to Teach Mathematics so as to Be Useful. *Educational Studies in Mathematics, 1*, 3-8.

Freudenthal, H. (1971). Geometry Between the Devil and the Deep Sea. *Educational Studies in Mathematics, 3*, 413-435.

Freudenthal, H. (1973). *Mathematics as an Educational Task*. Dordrecht: Reidel Publishing Company.

Freudenthal, H. (1975a). Wandelingen met Bastiaan [Walks with Bastiaan]. *Pedomorfose, 7* (25), 51-64.

Freudenthal, H. (1975b). Pupils' Achievements Internationally Compared – the IEA. *Educational Studies in Mathematics, 6*, 127-186.

Freudenthal, H. (1976a). De wereld van de toetsen [The world of tests]. *Rekenschap, 23*, 60-72.

Freudenthal, H. (1976b). *Toetsen, waarom, wat en hoe?* [Testing, why, what and how?]

Lecture presented on the NOT 76.

Freudenthal, H. (1976c). Wandelingen met Bastiaan [Walks with Bastiaan]. *Pedomorfose, 8* (30), 35-54.

Freudenthal, H. (1977). Antwoord door Prof. Dr. H. Freudenthal na het verlenen van het eredoctoraat [Answer by Prof. Dr. H. Freudenthal upon being granted an honorary doctorate]. *Euclides, 52*, 336-338.

Freudenthal, H. (1978a). *Weeding and Sowing. Preface to a Science of Mathematical Education.* Dordrecht: Reidel Publishing Company.

Freudenthal, H. (1978b). Cognitieve ontwikkeling – kinderen geobserveerd [Cognitive development – observing children]. In *Provinciaals Utrechts Genootschap, Jaarverslag 1977* (pp. 8-18).

Freudenthal, H. (1979a). Lessen van Sovjet rekenonderwijskunde [Lessons from Sovjet mathematics teaching]. *Pedagogische Studiën, 56* (1), 17-24.

Freudenthal, H. (1979b). Structuur der wiskunde en wiskundige structuren; een onderwijskundige analyse [Structure of mathematics and mathematical structures; an educational analysis]. *Pedagogische Studiën, 56* (2), 51-60.

Freudenthal, H. (1979c). *Learning processes*. Lecture at the Pre-session of the NCTM meeting Boston, 18 April 1979. (not published)

Freudenthal, H. (1979d). Konstruieren, Reflektieren, Beweisen in phänomenologischer Sicht [Constructing, Reflecting, Proving in a Phenomenological Perspective]. In W. Dörfler and R. Fisher (eds.), *Schriftenreihe Didaktik der Mathematik. Beweisen in Mathematikunterricht* (pp. 183-200). Klagenfurt / Wien.

Freudenthal, H. (1980). Wat is onderzoek van onderwijs? – een paradigma [What is research of education? – a paradigm]. In S. Pieters, *De achterkant van de Möbiusband* (pp. 11-15). Utrecht: IOWO.

Freudenthal, H. (1981a). Major problems of mathematics education. *Educational Studies in Mathematics, 12*, 133-150.

Freudenthal, H. (1981b). *Het toetsen van hiërarchieën en hiërarchieën om te toetsen* [The testing of hierarchies and hierarchies for testing]. Utrecht: OW&OC, Utrecht University. (internal publication)

Freudenthal, H. (1982a). Faibilité, validité et pertinence – critères de la recherche sur l'enseignement de la mathematique [Fallibility, validity and suitability – criteria for research on mathematics education]. *Educational Studies in Mathematics, 13*, 395-408.

Freudenthal, H. (1983a). *Didactical Phenomenology of Mathematical Structures*. Dordrecht: Reidel Publishing Company (the earlier version was written in 1976 / 1977)

Freudenthal, H. (1983b). Is heuristics a singular or a plural? In R. Hershkowitz (ed.). *Proceedings of the Seventh International Conference for the Psychology of Mathematics Education* (pp. 38-50). Rehovot, Israel: The Weizmann Institute of Science.

Freudenthal, H. (1984a). Onderzoek van onderwijs – Voorbeelden en voorwaarden [Educational research – Examples and conditions]. In P.G. Vos, K. Koster, and J. Kingma (eds.), *Rekenen, Balans van standpunten in theorievorming en empirisch onderzoek* (p. 7-20). Lisse: Swets & Zeitlinger.

Freudenthal, H. (1984b). Cito – Leerdoelgerichte toetsen [Cito – Criterion-referenced tests]. *Nieuwe Wiskrant, 4* (1), 15-23.

Freudenthal, H. (1984c). Toetsen, wie wordt er beter van? [Tests, who needs them?] *NRC-Handelsblad*, 29 november 1984.

Freudenthal, H. (1985). Wat wordt er niet getoetst, en kan dat wel? [That which is not tested, and is that acceptable?] *Euclides, 60* (8 / 9), 303-305.

Freudenthal, H. (1986). Didactical Principles in Mathematics Instruction. In J.A. Barroso, *Aspects of Mathematics and its Applications* (pp. 351-357). Amsterdam: Elsevier Science Publishers BV.

Freudenthal, H. (1987a). *Schrijf dat op, Hans. Knipsels uit een leven* [Write that down, Hans. Clippings from a life]. Amsterdam: Meulenhoff.

Freudenthal, H. (1987b). Theorievorming bij het wiskundeonderwijs. Geraamte en gereedschap [Theory of mathematics education. Framework and tools]. *Tijdschrift voor Nascholing en Onderzoek van het Reken-wiskundeonderwijs, 5* (3), 4-15.

Freudenthal, H. (1990). Wiskunde fenomenologisch (deel 3) [Mathematics phenomenological (part 3)]. *Tijdschrift voor Nascholing en Onderzoek van het Reken-wiskundeonderwijs, 8* (2), 51-61.

Freudenthal, H. (1991). *Revisiting Mathematics Education. China Lectures.* Dordrecht: Kluwer Academic Publishers.

Gagné, R.M. (1965). *The Conditions of Learning*. London: Holt, Rinehart, and Winston, Inc.

Galbraith, P. (1993). Paradigms, Problems and Assessment: Some Ideological Implications. In M. Niss (ed.), *Investigations into Assessment in Mathematics Education* (pp. 73-86). Dortrecht: Kluwer Academic Publishers.

Gelman, R., and C.R. Gallistel (1978). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

Ginsburg, H. (1975). Young Children's Informal Knowledge of Mathematics. *The Journal of Children's Mathematical Behavior, 1* (3), 3-156.

Ginsburg, H. (1981). The Clinical Interview in Psychological Research on Mathematical Thinking. *For the Learning of Mathematics, 1* (3), 4-11.

Ginsburg, H.P., and Baroody, A. (1990). *Test of Early Mathematics Ability. Second edition.* Austin, TX: Pro-ed.

Ginsburg, H.P., Jacobs, S.F., and Lopez, L.S. (1993). Assessing Mathematical Thinking and Learning Potential in Primary Grade Children. In: M. Niss (ed.), *Investigations into Assessment in Mathematics Education* (pp. 157-167). Dordrecht: Kluwer Academic Publishers.

Ginsburg, H.P., Kossan, N., Schwartz, R., and Swanson, D. (1983). Protocol Methods in Research on Mathematical Thinking. In H.P. Ginsburg (ed.), *The Development of Mathematical Thinking* (pp. 7-47). New York: Academic Press.

Ginsburg, H.P., Lopez, L.S., Mukhopadhyay, S., Yamamoto, T., Willis, M., and Kelly, M.S. (1992). Assessing Understanding of Arithmetic. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 265-289). Washington: AAAS Press.

Gipps, C. (1994). Developments in Educational Assessment: what makes a good test? *Assessment in Education: Principles, Policy & Practice, 1,* 3, 283-291.

Glaser, R. (1986). The integration of instruction and testing. In *The redesigning of Testing for the 21st Century. Proceedings from the 6th annual invitational conference of the Educational Testing Service,* 26 October 1985. (cited by Grouws and Meier, 1992)

Goffree, F. (1979). *Leren onderwijzen met Wiskobas* (Dissertatie) [Learning to teach with Wiskobas (Doctoral dissertation)]. Utrecht: IOWO.

Goffree, F. (1984). *Van redactiesom naar rijk probleem* [From word problem to rich problem]. Enschede: SLO.

Goffree, F. (1993). HF: Working on Mathematics Education. *Educational Studies in Mathematics, 25* (1-2), 21-58.

Goffree, F. (ed.) (1985). *Onderzoek en (leerplan)ontwikkeling* [Research and (curriculum) development]. Enschede: SLO.

Gould, S.J. (1981). *The Mismeasure of Man*. New York: W.W. Norton and Company.

Graf, R.G., and Riddell, J.C. (1972). Sex differences in problem solving as a funtion of problem context. *The Journal of Educational Research, 65* (10), 451-452. (cited by M.R. Meyer, 1992)

Grassmann, M., Mirwald, E., Klunter, M., and Veith, U. (1995). Arithmetische Kompetenz von Schulanfängern – Schluszfolgerungen für die Gestaltung des Anfangsunterrichtes – mehr Fragen als Antworten [The arithmetical compentence of beginning first graders – conclusions for education in the early grades – more questions than answers]. *Sachunterricht und Mathematik in der Primärstufe, 23* (7), 302-321.

Graue, M.E., and Smith S.Z (1992). *A Conceptual Framework for Instructional Assessment.* Madison, WI: National Center for Research in Mathematical Sciences Education.

Gravemeijer, K. (1982). Het gebruik van contexten [The use of contexts]. *Willem Bartjens, 1* (4), 192-197.

Gravemeijer, K. (1992). Onderwijsontwikkeling en ontwikkelingsonderzoek [Educational development and developmental research]. *Tijdschrift voor Nascholing en Onderzoek van het Reken-wiskundeonderwijs, 10* (3), 3-14.

Gravemeijer, K.P.E. (1994). *Developing Realistic Mathematics Education* (Doctoral dissertation). Utrecht: CD-ß Press / Freudenthal Institute.

Gravemeijer, K.P.E., Van den Heuvel-Panhuizen, M., and Streefland, L. (1990). MORE over zorgverbreding [MORE on the accommodating of special needs in regular schools]. In M. Dolk and E. Feijs (eds.), *Panama Cursusboek 8. Rekenen en Zorgverbreding.* Utrecht: HMN(FEO) / OW&OC.

Gravemeijer, K., Van den Heuvel-Panhuizen, M., Van Donselaar, G., Ruesink, N., Streefland, L., Vermeulen, W., Te Woerd, E., and Van der Ploeg, D. (1993). *Methoden in het reken-wiskundeonderwijs, een rijke context voor vergelijkend onderzoek* [Textbook series in mathematics education, a rich context for comparative research]. Utrecht: CD-ß Press / Freudenthal Institute, Utrecht University.

Gravemeijer, K., Van Galen, F., Kraemer, J.M., Meeuwisse, T., and Vermeulen, W. (1983). *Rekenen & Wiskunde* [Arithmetic & Mathematics]. Baarn: Bekadidact.

Greer, B. (1993). The Mathematical Modeling Perspectieve on Wor(l)d Problems. *Journal of Mathematical Behavior, 12*, 239-250.

Griffin, S., Case, R., and Siegler, R.S. (1994). Rightstart: Providing the Central Conceptual Prerequisites for First Formal Learning of Arithmetic to Students at Risk for School Failure. In K. McGilly (ed.), *Classroom Lessons: Integrating Cognitive Theory and Classroom Practice* (pp. 25-49). Cambridge, MA: USA Press.

Groen, G., and Kieran, C. (1983). In search of Piagetian Mathematics. In H. Ginsburg (ed.), *The Development of Mathematical Thinking* (pp. 351-375). New York: Academic Press.

Groenewegen, J.K.A., and Gravemeijer, K.P.E. (1988). *Het leren van de basisautomatismen voor optellen en aftrekken* [Learning the basic facts for addition and subtraction]. Rotterdam: OSM.

Gronlund, N.E. (1968). *Constructing Achievement Tests*. Englewood Cliffs: Prentice-Hall Inc.

Gronlund, N.E. (1970). *Stating behavioral objectives for classroom instruction*. New York: MacMillan.

Gronlund, N.E. (1991). *How to Construct Achievement Tests* (4th edition). Needham Heights, MA: Allyn and Bacon.

Grossen, M. (1988). L'Intersubjectivité En Situation De Test [Intersubjectivity in test situations]. Fribourg: Delval. (cited by Elbers, 1991b)

Grossman, R. (1975). Open-Ended Lessons Bring Unexpected Surprises. *Mathematics Teaching, 71*, 14-15.

Grouws, D.A., and Meier, S.L. (1992). Teaching and assessment relationships in mathematics instruction. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 83-106). Hawthorn, Victoria: Australian Council for Educational Research.

Hamers, J. (1991). A remark made during the opening address of the European Conference on Learning Potential Tests, Utrecht, November 28-29, 1991.

Harlen, W. (1994). Quality Assurance and Quality Control in Student Assessment: Purposes and Issues. In J. van den Akker and W. Kuiper (eds.), *Book of Summaries of the first European Research on Curriculum* (pp. 80-82). Enschede: Faculty of Educational Science and Technology, University of Twente.

Harskamp, E., and Suhre, C. (1986). *Vergelijking van rekenmethoden in het basisonderwijs* (Eindrapport*)* [Comparing textbook series in primary education (Final report)]. Groningen: RION.

Harskamp, E., and Willemsen, T. (1991). Programma's en speelleermaterialen voor voorbereidend rekenen in de basisschool [Programs and educational materials for preparatory arithmetic in primary school]. *Pedagogische Studiën, 68* (9), 404-414.

Hart, K. (1988). Ratio and Proportion. In J. Hiebert and M. Behr, *Research Agenda for Mathematics Education. Number Concepts and Operations in the Middle Grades, Volume 2*. Hillsdale, N.J.: Erlbaum / NCTM.

Hengartner, E., and Röthlisberger, H. (1993). *Rechenfähigkeit von Schulanfängern* [Students' arithmetical abilities at the beginning of primary school]. Zofingen / Basel: Höheren Pädagogische Lehranstalt / Pädagogischen Institut. (internal publication)

Hengartner, E., and Röthlisberger, H. (1995). Rechenfähigkeit von Schulanfängern [Students' arithmetical abilities at the beginning of primary school]. In H. Brügelmann, H. Balhorn, and I. Füssenich (eds.), *Am Rande der Schrift. Zwischen Sprachenfielfalt und Analphabetismus* (pp. 66-86). Lengwil am Bodensee: Libelle Verlag / Deutsche Gesellschaft für Lesen und Schreiben.

Herman, J.L., and Winters, L. (1994). Portfolio Research: A Slim Collection. *Educational Leadership, 52* (2), 48-55.

HEWET-team (1984). *Hewet & Toets* [Hewet & Test]. Utrecht: OW&OC, Utrecht University.

Hiebert, J. (1984). Children's mathematics learning: The struggle to link form and understanding. *The Elementary School Journal, 84* (5), 496-513.

Hiebert, J., and Wearne, D. (1988). Methodologies for Studying Learning to Inform Teaching. In E. Fennema, Th.P. Carpenter, and S.J. Lamon (eds.), *Integrating Research on Teaching and Learning Mathematics* (pp. 168-192). Madison, WI: Wisconsin Center for Education Research, University of Wisconsin-Madison.

Holland, J. (1981). Social class and changes in orientation to meaning. *Sociology, 15* (1), 1-18.

Hošpesová, A., Kuřina, F., and Tichá, M. (1995). Kennen wir die Kenntnisse unsere Schüler? [Do we know the knowledge of our students?] A paper presented at the 29th Bundestagung für Didaktik der Mathematik in Kassel.

Houben, S. (1995). De begeleiding van het realistisch rekenonderwijs (Doctoraalscriptie) [The counceling of realistic mathematics education (Master's thesis)]. Nijmegen: Katholieke Universiteit Nijmegen, Vakgroep Onderwijskunde.

Hughes, M. (1986). *Children and Number. Difficulties in Learning Mathematics*. Oxford: Basil Blackwell Ltd.

Huitema, S. (1988). We overvragen de basisschool [We ask too much of the primary school]. In J. Wijnstra (ed.), *Balans van het rekenonderwijs in de basisschool* (pp. 163-168). Arnhem: Cito.

Huitema, S., and Van der Klis, A. (1987). Evaluatie van het reken- en wiskunde-onderwijs in: De Wereld in Getallen [Evaluation of the mathematics education in: De Wereld in Getallen]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 90-92). Utrecht: OW&OC, Utrecht University.

Joffe, L.S. (1990). Evaluating assessment: examining alternatives. In S. Willis (ed.), *Being Numerate: What Counts?* (pp. 138-161). Hawthorn, Victoria: Australian Council for Educational Research.

Joffe, L.S. (1992). The English experience of a national curriculum and assessments. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 190-222). Hawthorn, Victoria: Australian Council for Educational Research.

Johnson, D.C. (1989). *Children's Mathematical Frameworks*. London: NFER / Nelson. (cited by Bell, 1993)

Kantowski, M.G. (1981). Problem solving. In E. Fennema (ed.), *Mathematics Education Research: Implications for the 80s* (pp. 111-126). Alexandria / Reston, VA: ASCD / NCTM. (cited by Borasi, 1986)

Kilpatrick, J. (1992). A History of Research in Mathematics Education. In D.A. Grouws (ed.), *Handbook of Research on Mathematics Teaching* (pp. 3-38). New York: NCTM / Macmillan.

Kilpatrick, J. (1993). The Chain and Arrow: From the History of Mathematics Assessment. In M. Niss (ed.), *Investigations into Assessment in Mathematics Education. An ICMI Study* (pp. 31-46). Dordrecht: Kluwer Academic Publishers.

Kindt, M. (ed.) (1986a). *Hewet & Toets 4* [Hewet & Test 4]. Utrecht: OW&OC, Utrecht University.

Kindt, M. (ed.) (1986b). *Hewet & Toets 5* [Hewet & Test 5]. Utrecht: OW&OC, Utrecht University.

Koster, K.B. (1975). *De ontwikkeling van het getalbegrip op de kleuterschool: een onderzoek naar de effecten van enkele trainingsprogramma's* [The development of number concept in kindergarten: research into the effects of certain training programs]. Groningen: Wolters-Noordhoff.

Kraemer, J.M., Bokhove, J., and Janssen, J. (1991). Vooronderzoek periodieke peiling onderwijsniveau in lom en mlk [Pilot study for the National Assessment of Educational Achievement in schools for children with learning and behavioral difficulties and in schools for mildly mentally retarded children]. *Tijdschrift voor Nascholing en Onderzoek van het Reken-wiskundeonderwijs, 9*, 3, 16-32.

Kraemer, J.M., Nelissen, J., Janssen, J., and Noteboom, A. (1995). *Rekenen-Wiskunde 1. Hulpboek* [Mathematics 1. Auxiliary book]. Arnhem: Cito.

Kraemer, J.M., Van der Schoot, F., and Veldhuizen, N. (in press). *Balans van het reken-wiskundeonderwijs aan het eind van het speciaal onderwijs. Uitkomsten van de eerste rekenpeiling einde speciaal onderwijs* [An account of mathematics education at the end of special education. Results of the first national assessment on mathematics at the end of special education]. Arnhem: Cito.

Küchemann, D. (1989). Learning and Teaching Ratio: A Look at Some Current Textbooks. In P. Ernst (ed.), *Mathematics Teaching: The State of the Art*. New York: The Falmer Press.

Kuipers, N., et al. (1978). *Naar zelfstandig rekenen* [Towards independent arithmetic]. Groningen: Wolters-Noordhoff.

Kulm, G. (1993). *A theory of Classroom Assessment and Teacher Practice in Mathematics*. A paper presented at the annual meeting of the AERA, Atlanta, April 15, 1993.

Labinowicz, E. (1985). *Learning from Children: New Beginnings for Teaching Numerical Thinking. A Piagetian Approach*. Menlo Park: Addison Wesley Publishing Company.

Lambdin Kroll, D., Masingila, J.O., and Tinsley Mau, S. (1992). Cooperative Problem Solving: But What About Grading? *Arithmetic Teacher, 39* (6), 17-23.

Lamon, S.J., and Lesh, R. (1992). Interpreting responses to problems with several levels and types of correct answers. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 319-342). Washington: AAAS Press.

Lankford, F.G. (1974). What can a teacher learn about a pupils thinking through oral interviews? *Arithmetic Teacher, 21* (5), 26-32.

Leder, G.C., and Forgasz, H.J. (1992). Perspectives on learning, teaching and assessment. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 1-23). Hawthorn, Victoria: Australian Council for Educational Research.

Lenné, H. (1969). *Analyse der Mathematik-didaktik* [Analysis of the didactics of mathematics]. Stuttgart: Klett. (cited by Christiansen and Walther, 1986)

Lesh, R., and Lamon, S.J. (1992a). Trends, Goals, and Priorities in Mathematics Assessment. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 3-15). Washington: AAAS Press.

Lesh, R., and Lamon, S.J. (1992b). Assessing Authentic Mathematical Performance. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 17-62). Washington: AAAS Press.

Lesh, R., and Lamon, S.J. (eds.) (1992). *Assessment of Authentic Performance in School*

Mathematics. Washington: AAAS Press.

Lesh, R., Lamon, S.J., Behr, M., and Lester, F. (1992). Future Directions for Mathematics Assessment. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 379-425). Washington: AAAS Press.

Lester Jr., F.K., and Lambdin Kroll, D. (1991). Evaluation: A new vision. *Mathematics Teacher, 84* (4), 276-284.

Linn, R.L., Baker, E., and Dunbar, S.B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher, 20* (8), 15-21.

Lit, S. (1988). *Verslag van het onderzoek naar de relatie tussen de instaptoetsen en opgavenseries. Kwantiwijzer, Memo 7* [Report on research into the relation between entry tests and the series of tasks]. Rotterdam: Erasmus University.

Loef, M.M., Carey, D.A., Carpenter, T.A., and Fennema, E. (1988). Integrating assessment and instruction. *Arithmetic Teacher, 36* (3), 53-56.

Long, M.J., and Ben-Hur, M. (1991). Informing Learning through the Clinical Interview. *Arithmetic Teacher, 38* (6), 44-46.

Lubinski, C.A., and Nesbitt Vacc, N. (1994). The Influence of Teachers Beliefs and Knowledge on Learning Environments. *Arithmetic Teacher, 41* (8), 476-479.

Madaus, G.F., Maxwell West, M., Harmon, M.C., Lomax, R.G., and Viator, K.A. (1992). *The Influence of Testing on Teaching Math and Science in Grades 4-12. Executive Summary.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation and Educational Policy.

Magone, M.E., Cai, J., Silver, E.A., and Wang, N. (1994). Validating the cognitive complexity and content validity of a mathematics performance assessment. *International Journal of Educational Research, 3* (21), 317-340.

Maher, C.A., Davis, R.B., and Alston, A. (1992). A Teacher’s Struggle to Assess Student Cognitive Growth. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 249-264). Washington: AAAS Press.

Marshall, S.P., and Thompson, A.G. (1994). Assessment: What’s New And Not So New – A Review of Six Recent Books. *Journal for Research in Mathematics Education*, *25* (2), 209-218.

Mathematical Sciences Education Board (MSEB) (1991). *For Good Measure: Principles and Goals for Mathematics Assessment.* Washington: National Academy Press.

Mathematical Sciences Education Board (MSEB) (1993a). *Measuring What Counts. A conceptual guide for mathematics assessment.* Washington: National Academy Press.

Mathematical Sciences Education Board (MSEB) (1993b). *Measuring What Counts. A conceptual guide for mathematics assessment. Executive Summary*. Washington: National Academy Press.

Mathematical Sciences Educational Board (MSEB) (1993c). *Measuring Up. Prototypes for Mathematics Assessment.* Washington: National Academy Press.

McClintic, S. (1988). Conservation – a meaningful Gauge for Assessment. *Arithmetic Teacher, 36* (6), 12-14.

McGarrigle, J., and Donaldson, M. (1974). Conservation accidents. *Cognition, 3*, 341-350.

McLean, L. (1990). Let’s call a halt to pointless testing. *Education Canada, 30* (3), 10-13.

Mehrens, W.A. (1992). Using Performance Assessment for Accountability Purposes. *Educational Measurement: Issues and Practice*, *11* (1), 3-9.

Meijer, C.J.W., Pijl, S.J., and Kramer, L.J.L.M. (1989). Rekenen met groei. Ontwikkelingen in deelname aan het (voortgezet) speciaal onderwijs (1972-1987) [Counting with growth. Developments in participation with (secondary) special education (1972-1987)]. *Tijdschrift voor Orthopedagogiek, 28* (2), 71-82.

Meijer, J., and Elshout, J.J. (1994). *Offering help and the relationship between test anxiety and mathematics performance.* A paper presented at the 15th International Conference of the Society for Stress and Anxiety Research, Madrid.

Meyer, C.A. (1992). What’s the Difference Between Authentic and Performance Assessment? *Educational Leadership, 49* (8), 39-40.

Meyer, M.R. (1992). Gender Differences in Test Taking: A Review. In T.A. Romberg (ed.), *Mathematics Assessment and Evaluation. Imperatives for Mathematics Educators* (pp. 169-183). Albany, NY: SUNY Press.

Moss, P.A. (1994). Can There Be Validity Without Reliability? *Educational Researcher, 23* (2), 5-12.

Mousley, J.A. (1990). Assessment in primary mathematics: the effects of item readability. In G. Booker, P. Cobb, and T.N. de Mendicuti (eds.), *Proceedings of the Fourteenth PME Conference* (Vol. III, pp. 273-280). Mexico: Program Committee of the 14th PME Conference, Mexico.

Mousley, J.A., Clements, M.A., and Ellerton, N.F. (1992). Teachers interpretations of their roles in mathematics classrooms. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 107-144). Hawthorn, Victoria: Australian Council for Educational Research.

Mumme, J. (1990). *Portfolio Assessment in Mathematics*. Santa Barbara: University of California.

Nationaal Informatiecentrum Leermiddelen (NICL) (1992). *Toetsen en Tests basisonderwijs* [Tests for primary eudcation]. Enschede: NICL.

Nationaal Informatiecentrum Leermiddelen (NICL) (1993). *Toetsen en Tests basisonderwijs, aanvulling (incl. leerlingvolgsystemen)* [Tests for primary education, supplement (incl. student monitoring systems]. Enschede: NICL.

National Commission on Excellence in Education (NCEE) (1983). *A Nation at Risk: The Imperative for Educational Reform*. Washington: US Government Printing Office.

National Council of Teachers of Mathematics (NCTM) (1987). *Curriculum and Evaluation Standards for School Mathematics* (draft). Reston, VA: NCTM. (cited by Doig and Masters, 1992)

National Council of Teachers of Mathematics (NCTM) (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: NCTM.

National Council of Teachers of Mathematics (NCTM) (1991). *Professional Standards for Teaching Mathematics*. Reston, VA: NCTM.

National Council of Teachers of Mathematics / Assessment Standards Working Group (NCTM / ASWG) (1994). *Assessment Standards for School Mathematics* (draft). Reston, VA: NCTM. (the final version of this report was published in 1995)

Nelissen, J., and Post, M. (1987). Toetsen in Rekenwerk [Tests in Rekenwerk]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 127-130). Utrecht: OW&OC, Utrecht University.

Nesbitt Vacc, N. (1993). Questioning in the Mathematics Classroom. *Arithmetic Teacher, 41* (2), 88-91.

Nieuwe Media Project (NMP) (1989). *Observatieverslag Nieuwe Media*, 6 oktober 1989 [Observation report New Media, 6 October 1989]. (internal publication)

O'Brien, T.C., and Richard, J.V. (1971). Interviews to Assess Number Knowledge. *Arithmetic Teacher, 18* (May), 322-326.

O'Neil, J. (1992). Putting Performance Assessment to the Test. *Educational Leadership, 49* (8), 14-19.

Oonk, J. (1984). Toetsen en methoden [Tests and textbook series]. In E. de Moor (ed.), *Panama Cursusboek 2. Methoden en reken/wiskundeonderwijs* (pp. 115-120). Utrecht: SOL / OW&OC.

Pedologisch Instituut Rotterdam (1987). Zo reken ik ook! [That's how I do math, too!] Gorinchem: De Ruiter.

Pandey, T. (1990). Power Items and the Alignment of Curriculum and Assessment. In G. Kulm, *Assessing Higher Order Thinking in Mathematics* (pp. 39-51). Washington: AAAS Press.

Perret-Clermont, A.N., and Schubauer-Leoni, M.L. (1981). Conflict and cooperation as opportunities for learning. In W.P. Robinson (ed.), *Communication in Development* (pp. 203-233). London: Academic Press. (cited by Elbers, 1991b)

Polya, G. (1981). *Mathematical Discovery: On Understanding, Learning and Teaching Problem Solving*. New York: Wiley. (cited by Borasi, 1986)

Popham, W.J. (1975). *Educational Evaluation*. Englewood Cliffs, NJ: Prentice Hall.

Popham, W.J., and Baker, E.L. (1970). *Establishing instructional goals*. Englewood Cliffs, NJ: Prentice Hall.

Quellmalz, E.S. (1985). Needed: Better Methods for Testing Higher-Order Thinking Skills. *Educational Leadership, 43* (2), 29-35.

Querelle, W.M.G. (1987). Mavo-examens en opvattingen over wiskunde-onderwijs [Mavo-exams and viewpoints on mathematics education]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 131-135). Utrecht: OW&OC, Utrecht University.

Resnick, L.B. (1988). Treating mathematics as an ill-structured discipline. In R.I. Charles, and E.A. Silver (eds.), *The teaching and assessing of mathematical problem solving* (pp. 32-60). Reston, VA: NCTM / Erlbaum. (cited by Magone et al., 1994)

Resnick, L.B. (1989). Developing mathematical knowledge. *American Psychologist, 44* (2), 162-169.

Resnick, L.B., and Resnick, D.P. (1992). Assessing the Thinking Curriculum: New Tools for Educational Reform. In B.R. Gifford and M.C. O'Connor (eds.), *Changing Assessments: Alternative Views of Attitude, Achievement and Instruction* (pp. 37-75). Norwell, MA: Kluwer.

Roberts, E.S. (1994). Integrating Assessment and Instruction? The tip of the Iceberg. *Arithmetic Teacher, 41* (7), 350-351.

Romberg, T.A. (1992). Overview of the Book. In T.A. Romberg, *Mathematics Assessment and Evaluation. Imperatives for Mathematics Educators* (pp. 1-9). Albany, NY: SUNY Press.

Romberg, T.A. (1993). How one comes to know: models and theories of the learning of mathematics. In M. Niss (ed.), *Investigations into Assessment in Mathematics Education* (pp. 97-111). Dordrecht: Kluwer Academic Publishers.

Romberg, T.A. (1995). Personal communication.

Romberg, T.A. (ed.) (1993). *A Blueprint for Maths in Context: A connected curriculum for Grades 5-8*. Madison, WI: National Center for Research in Mathematical Sciences Education.

Romberg, T.A., and Wilson, L.D. (1992). Alignment of Tests with the Standards. *Arithmetic Teacher, 40* (1), 18-22.

Romberg, T.A., and Wilson, L.D. (1995). Issues Related to the Development of an Authentic Assessment System for School Mathematics. In T.A. Romberg (ed.), *Reform in School Mathematics*. Albany, NY: SUNY Press.

Romberg, T.A., Zarinnia, E.A., and Collis, K.F. (1990). A New World View of Assessment in Mathematics. In G. Kulm, *Assessing Higher Order Thinking in Mathematics* (pp. 21-38). Washington, DC: AAAS.

Romberg, T.A., Zarinnia, E.A., and Williams, S. (1989). *The Influence of Mandated Testing on Mathematics Instruction: Grade 8 Teachers Perceptions*. Madison, WI: NCRMSE.

Rudnitsky, A.N., Drickamer, P., and Handy, R. (1981). Talking Mathematics with Children. *Arithmetic Teacher, 28* (8), 14-17.

Ruthven, K. (1987). Ability stereotyping in mathematics. *Educational Studies in Mathematics, 18*, 243-253.

Säljö, R. (1991). Learning and mediation: Fitting reality into a table. *Learning and Instruction, 1*, 261-272.

Scheer, J.K. (1980). The Etiquette of Diagnosis. *Arithmetic Teacher, 27* (9), 18-19.

Scherer, P. (in press). “Zeig', was Du weiszt” – Ergebnisse eines Tests zur Prozentrechnung [“Show what you know” – Results of a test on percentage].

Schoemaker, G. (1993). Is Mathematics meant to be understood? In P. Bero (ed.), *Proceedings of the 3rd Bratislava International Symposium on Mathematical Education* (pp. 119-124). Bratislava: Comenius University.

Schoen, H.L. (1979). Using the Individual Interview to Asses Mathematical Learning. *Arithmetic Teacher, 27* (3) (November), 34-37.

Schoenfeld, A.H. (ed.) (1987). *Cognitive science and mathematics education*. Hilsdale, NJ: Lawrence Erlbaum. (cited by Galbraith, 1993)

Scholten, P., and Ter Heege, H. (1983/1984). *Pluspunt. Werkbladen klas 2, 3, 4 en 5* [Pluspunt. Worksheets grades 2, 3, 4, and 5]. Hilversum: Nederlandse Onderwijs Televisie.

Schwarz, J.L. (1992). The Intellectual Prices of Secrecy in Mathematics Assessment. In R. Lesh and S.J. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 427-437). Washington: AAAS Press.

Selter, Chr. (1993a). *Die Kluft zwischen den arithmetischen Kompetenzen von Erstkläszlern und dem Pessimismus der Experten* [The gap between the arithmetical competence of first-graders and the pessimism of experts]. Dortmund: Universität Dortmund. (internal publication)

Selter, Chr. (1993b). Die Kluft zwischen den arithmetischen Kompetenzen von Erstkläszlern und dem Pessimismus der Experten [The gap between the arithmetical competence of first-graders and the pessimism of experts]. In *Beitrage zum Mathematikunterricht 1993* (pp. 350-353). Hildesheim: Franzbecker.

Selter, Chr. (1993c). Eigen Produktionen im Arithmetikunterricht der Primarstufe [Students' own productions in primary school arithmetic education]. Wiesbaden: Deutscher Universitäts Verlag.

Selter, Chr. (1995). Zur Fiktivität der 'Stunde Null' im arithmetischen Anfangsunterricht [The fiction of starting from scratch at the beginning of arithmetic education]. *Mathematische Unterrichtspraxis, 16* (2), 11-20.

Silverman, F.L., Winograd, K., and Strohauer, D. (1992). Student Generated Story Problems. *Arithmetic Teacher, 39* (8), 6-12.

Smaling, A. (1990). Enige aspecten van kwalitatief onderzoek en het klinisch interview [Certain facets of qualitative research and the clinical interview]. *Tijdschrift voor Onderzoek en Nascholing van het Reken-wiskundeonderwijs, 8* (3), 4-10.

Spiegel, H. (1992). Rechenfähigkeiten von Schulanfängern im Bereich von Additionen und Subtraktionen [Arithmetical abilities of beginning first-graders with respect to addition and subtraction]. In *Beitrage zum Mathematikunterricht 1992* (pp. 447-450). Bad Salzdetfurth: Franzbecker.

Stenmark, J.K. (1989). *Assessment Alternatives in Mathematics*. Berkeley, CA: EQUALS, University of California / California Mathematics Council.

Stenmark, J.K. (ed.) (1991). *Mathematics Assessment. Myths, Models, Good Questions, and Practical Suggestions*. Reston, VA: NCTM.

Stephens, M., Clarke, D., and Pavlou, M. (1994). Policy to practice: high stakes assessment as a catalyst for classroom change. A paper presented at the 17th Annual Conference of the Mathematics Education Research Group of Australia (MERGA), Southern Cross University, Lismore, Australia, July 5-8, 1994.

Streefland, L. (1978). Some observational results concerning the mental constitution of the concept of fraction. *Educational Studies in Mathematics, 9*, 51-73.

Streefland, L. (1979). Davydov, Piaget en de breuken [Davydov, Piaget and fractions]. *Pedagogische Studiën, 56* (7 / 8), 289-307.

Streefland, L. (1980). Kijk op wiskunde doen [A look at doing mathematics]. In S. Pieters (ed.), *De achterkant van de Möbiusband* (pp. 92-95). Utrecht: IOWO.

Streefland, L. (1981a). Van Erathostenes tot Cito-toets [From Erathostenes to Cito-test]. *Nieuwe Wiskrant, 1* (1), 34-40.

Streefland, L. (1981b). Cito's kommagetallen leerdoelgericht getoetst (1) [Cito's decimals tested in a criterion-referenced way (1)]. *Willem Bartjens, 1* (1), 34-44.

Streefland, L. (1982). Cito’s kommagetallen leerdoelgericht getoetst (2) [Cito’s decimals tested in a criterion-referenced way (2)]. *Willem Bartjens, 1* (2), 92-97.

Streefland, L. (1984). Search for the Roots of Ratio: Some Thoughts on the Long Term Learning Process (Towards? A Theory). Part I. *Educational Studies in Mathematics, 15*, 327-348.

Streefland, L. (1985a). Vorgreifendes Lernen zum Steuern Langfristiger Lernprozesse [Anticipatory learning to steer long-term learning processes]. In W. Dörfler and R. Fischer (eds.), *Empirische Untersuchungen zum Lehren und Lernen von Mathematik. Beiträge zum 4. Internationalen Symposium für Didaktik der Mathematik in Klagenfurt in 1984* (pp. 271-285). Wien: Hölder-Pichler-Tempsky.

Streefland, L. (1985b). Wiskunde als activiteit en realiteit als bron [Mathematics as an activity and reality as a source]. *Nieuwe Wiskrant, 5* (1), 60-67.

Streefland, L. (1987). Free productions of fraction monographs. In J.C. Bergeron, N. Herscovics, and C. Kieran (eds.), *Proceedings of the Eleventh Annual Meeting of the International Group for the Psychology of Mathematics Education* (Vol. I, pp. 405-410). Montreal.

Streefland, L. (1988). *Realistisch breukenonderwijs* (Dissertatie) [Realistic instruction of fractions (Doctoral dissertation)]. Utrecht: OW&OC, Utrecht University.

Streefland, L. (1990a). Free Productions in Teaching and Learning Mathematics. In K. Gravemeijer, M. van den Heuvel-Panhuizen, and L. Streefland, *Contexts, Free Productions, Tests and Geometry in Realistic Mathematics Education* (pp. 33-52). Utrecht: OW&OC, Utrecht University.

Streefland, L. (1990b). Developmental research and tests – sine functions as a paradigm. In *2nd Bratislava International Symposium on Mathematics Education* (pp. 78-98), August 23-25, 1990, Bratislava.

Streefland, L. (1991). *Fractions in Realistic Mathematics Education. A Paradigm of Developmental Research*. Dordrecht: Kluwer Academic Publishers.

Streefland, L. (1992). Thinking Strategies in Mathematics Instruction: How is Testing Possible? In R. Lesh and S. Lamon (eds.), *Assessment of Authentic Performance in School Mathematics* (pp. 215-246). Washington: AAAS Press.

Streefland, L. (1993). Theorievorming door ontwikkelingsonderzoek [Theory development through developmental research]. *Tijdschrift voor Onderzoek en Nascholing van het Reken-wiskundeonderwijs, 12* (2), 20-24.

Streefland, L., Hartings, T., and Veldhuis, E. (1979). *Leeronderzoek redactiesommen* [Pilot research on word problems]. Utrecht: IPAW, Utrecht University.

Streefland, L., and Te Woerd, E. (1991). *Protocolanalyse kwalitatief* [A qualitative protocol analysis]. Utrecht: OW&OC / ISOR, Utrecht University. (internal publication)

Streefland, L., and Van den Heuvel-Panhuizen, M. (1992). *Evoking Pupils’ Informal Knowledge on Percents. Proceedings of the Sixteenth PME Conference* (Vol. III, pp. 51-57). Durham, NH: University of New Hampshire.

Streefland, L., and Van den Heuvel-Panhuizen, M. (1994). Het mathematisch-didactisch onzekerheidsbeginsel – een mogelijke verrijking van de realistische theorie [The mathematical-didactical principle of uncertainty – a potential enrichment of the realistic theory]. *Tijdschrift voor Onderzoek en Nascholing van het Reken-wiskundeonderwijs, 12* (4), 19-21.

Sullivan, P., and Clarke, D. (1987). Good teaching and good questions. In W. Caughey (ed.), *From now to the future*. Parkville: The Mathematics Association of Victoria.

Sullivan, P., and Clarke, D.J. (1991). Catering to all abilities through the use of good questions. *Arithmetic Teacher, 39* (2), 14-21.

Swan, M. (1993). Improving the design and balance of mathematical assessment. In M. Niss (ed.), *Investigations into Assessment in Mathematics Education* (pp. 195-216). Dordrecht: Kluwer Academic Publishers.

Szetela, W., and Nicol, C. (1992). Evaluating Problem Solving in Mathematics. *Educational*

Leadership, 49 (8), 42-45.

Telese, J.A. (1993). *Effects of Alternative Assessment from the Students View*. A paper presented at the annual meeting of the American Educational Research Association, Atlanta, 15 April, 1993.

Ter Heege, H. (1978). Testing the maturity for learning the algorithm of multiplication. *Educational Studies in Mathematics, 9,* 75-83.

Ter Heege, H. (1980). Vermenigvuldigen met een afhaker [Doing multiplication with a quitter]. In S. Pieters (ed.), *De achterkant van de Möbiusband* (pp. 77-83). Utrecht: IOWO.

Ter Heege, H. (1981-1982). Het rekenen van Gijsbert [Gijsbert's arithmetic]. *Willem Bartjens, 1* (1), 25-26; (2), 67-68; (3), 109-111; (4), 172-177.

Ter Heege, H. (1988). Realistisch reken-wiskundeonderwijs. Op weg naar vernieuwing in het speciaal onderwijs [Realistic mathematics education. En route to a reform of special education]. *School & Begeleiding, 5* (17), 12-14.

Ter Heege, H., and Goffree, F. (1981). Hoofdrekenen getoetst [Mental calculation being tested]. *Willem Bartjens, 1* (1), 45-52.

Ter Heege, H., and Treffers, A. (1979). Peiling [Gauge]. In A. Treffers (ed.), *Cijferend vermenigvuldigen en delen (1). Overzicht en achtergronden. Leerplanpublikatie 10* (pp. 107-130). Utrecht: IOWO.

Ter Heege, H., Van den Heuvel-Panhuizen, M., and Scholten, P. (1983-1984). *Pluspunt. Handleiding klas 2, 3, 4, en 5* [Pluspunt. Teacher's guide grades 2, 3, 4 and 5]. Hilversum: Nederlandse Onderwijs Televisie.

Tessel, C.M. (1987). Mavo-examens en opvattingen over wiskunde-onderwijs [Mavo exams and viewpoints on mathematics education]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 142-145). Utrecht: OW&OC, Utrecht University.

Teunissen, F. (1988). Een hoge norm [A high standard]. In J. Wijnstra (ed.), *Balans van het rekenonderwijs in de basisschool* (pp. 169-180). Arnhem: Cito.

Thornton, A., Tucker, B.F., Dossey, J.A., and Bazik, E.F. (1983). *Teaching Mathematics to Children with Special Needs.* Menlo Park, CA: Addison Wesley Publishing Company.

Tissink, J., Hamers, J.H.M., and Van Luit, J.E.H. (1993). Learning Potential Tests with Domain-general and Domain-specific Tasks. In J.H.M. Hamers, K. Sijtsma, and A.J.J.M. Ruijssenaars (eds.), *Learning Potential Assessment Theoretical, Methodological and Practical Issues* (pp. 243-266). Amsterdam / Lisse: Swets & Zeitlinger.

Treffers, A. (1978). *Wiskobas doelgericht* (Dissertatie) [Wiskobas goal-directed (Doctoral dissertation)]. Utrecht: IOWO.

Treffers, A. (1979). Overzicht van praktische aanwijzingen [Overview of practical guidelines]. In Treffers, A. (ed.), *Cijferend vermenigvuldigen en delen (1). Overzicht en achtergronden. Leerplanpublikatie 10* (pp. 154-171). Utrecht: IOWO.

Treffers, A. (1980a). Cito item dito – over leerdoelgerichte toetsen meten [Cito idem dito – on the criterion-referenced testing of measuremenet]. *Wiskobas-Bulletin, 9* (6), 81-99.

Treffers, A. (1980b). Sorry Cito – over ondeugdelijke leerdoelgerichte toetsen [Sorry Cito – on unsound criterion-referenced tests]. In Pieters, S. *De achterkant van de Möbiusband* (pp. 145-154), Utrecht: IOWO.

Treffers, A. (1983). Leerdoelen en Cito-toetsen [Instructional objectives and Cito tests]. *Willem Bartjens, 3* (1), 55-58.

Treffers, A. (1985). Op weg naar een nationaal plan voor het reken / wiskundeonderwijs [En route to a national plan for mathematics education]. In F. Goffree (ed.), *Onderzoek en (leerplan)ontwikkeling* (pp. 99-100). Enschede: SLO.

Treffers, A. (1987a). *Three Dimensions. A Model of Goal and Theory Description in Mathematics Instruction – the Wiskobas Project*. Dordrecht: Reidel Publishing Company.

Treffers, A. (1987b). Beschrijvingen van eindtermen [Descriptions of attainment targets]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 146-

150). Utrecht: OW&OC, Utrecht University.

Treffers, A. (1988). Over de merkbare invloed van onderwijsmethoden op leerprestaties [On the noticeable influence of textbook series on educational achievements]. In J. Wijnstra (ed.), *Balans van het rekenonderwijs in de basisschool* (pp. 181-189). Arnhem: Cito.

Treffers, A. (1990). *Het voorkomen van ongecijferdheid op de basisschool* (Oratie) [The occurrrence / prevention of innumeracy at primary school (Inaugural lecture)]. Utrecht: VOU / OW&OC, Utrecht University.

Treffers, A. (1991a). Realistic mathematics education in The Netherlands 1980-1990. In L. Streefland (ed.), *Realistic Mathematics Education in Primary School.* Utrecht: CD-β Press / Freudenthal Institute, Utrecht University.

Treffers, A. (1991b). Meeting Innumeracy at Primary School. *Educational Studies in Mathematics, 22*, 333-352.

Treffers, A. (1992). Terug naar de toekomst. Reken-wiskundeonderwijs voor de basisschool 1972-2002 [Back to the future. Mathematics education for primary school 1972-2002]. In A. Treffers, F. Goffree, and J. de Lange, *Rekenen anno 2002 toekomstverwachtingen van het reken-wiskundeonderwijs* (pp. 11-34). Utrecht: NVORWO.

Treffers, A. (1993a). Wiskobas and Freudenthal: Realistic Mathematics Education. *Educational Studies in Mathematics, 25* (1-2), 89-108.

Treffers, A. (1993b). Ontwikkelingsonderzoek in eerste aanzet [Developmental research in its first stage]. In R. de Jong, and M. Wijers, *Ontwikkelingsonderzoek, theorie en praktijk* (pp. 35-58). Utrecht: NVORWO / Freudenthal Institute.

Treffers, A. (1995). Personal communication.

Treffers, A. (ed.) (1979). *Cijferend vermenigvuldigen en delen (1). Overzicht en achtergronden. Leerplanpublikatie 10* [Column multiplication and division (1). Overview and backgrounds. Publication on curriculum 10]. Utrecht: IOWO.

Treffers, A., and De Moor, E. (1984). *10 voor de basisvorming rekenen/wiskunde* [An A for basic education in mathematics]. Utrecht: OW&OC.

Treffers, A., and De Moor, E. (1990). *Proeve van een nationaal programma voor het reken-wiskundeonderwijs op de basisschool. Deel II. Basisvaardigheden en cijferen* [Design of a National Curriculum for mathematics education at primary school. Part II. Basic abilities and column arithmetic]. Tilburg: Zwijsen.

Treffers, A., De Moor, E., and Feijs, E. (1989). *Proeve van een nationaal programma voor het reken-wiskundeonderwijs op de basisschool. Deel I. Overzicht einddoelen* [Design of a National Curriculum for mathematics education at primary school. Part I. Overview of goals]. Tilburg: Zwijsen.

Treffers, A., and Goffree, F. (1982). Inzicht in BOVO-toetsen voor rekenen [Insight into BOVO tests for arithmetic]. *Nieuwe Wiskrant, 2* (1), 42-48.

Treffers, A., and Goffree, F. (1985). Rational analysis of realistic mathematics education – the Wiskobas program. In L. Streefland (ed.), *Proceedings of the Ninth International Conference for the Psychology of Mathematics Education* (Vol. II, pp. 97-121). Utrecht: OW&OC, Utrecht University.

Urbanska, A. (1993). On the numerical competence of six-years-old children. *Educational Studies in Mathematics, 24* (3), 265-275.

Van de Molengraaf, G.W.J., et al. (1981). *De wereld in getallen* [The world in numbers]. Den Bosch: Malmberg.

Van den Berg, W., and Van Eerde, D. (1983a). De Kwantiwijzer: diagnostisch instrumentarium voor het reken/wiskundeonderwijs [The Kwantiwijzer: diagnostic instruments for mathematics education]. In E. de Moor (ed.), *Panama Cursusboek 1. Reken/wiskundeonderwijs voor het jonge kind (4-8 jaar)* (pp. 39-49). Utrecht: OW&OC / SOL.

Van den Berg, W., and Van Eerde, D. (1983b). Ontcijfering van het rekenen: diagnostiek met de Kwantiwijzer [Deciphering of arithmetic: diagnosing with Kwantiwijzer]. *Tijdschrift voor Orthopedagogiek, 22* (2), 44-67.

Van den Berg, W., and Van Eerde, D. (1985). *Kwantiwijzer*. Rotterdam: Erasmus University.

Van den Brink, J. (1973a). Bijna noemen [Almost mention it]. *Wiskobas-Bulletin 3*, 129-131.

Van den Brink, J. (1973b). Laat ze voor je uit lopen [Let them walk ahead of you]. *Wiskobas-Bulletin 3*, 229-233.

Van den Brink, J.F. (1980). IOWO material tested. In R. Karplus (ed.), *Proceedings of the Fourth International Conference for the Psychology of Mathematics Education* (pp. 361-369). Berkeley: Lawrence Hall of Science University of California.

Van den Brink, J. (1981a). Intensiever observeren van jonge kinderen [Closer observation of young children]. *Het Jonge Kind, 8* (9), 228, 237.

Van den Brink, J. (1981b). Mutual Observation. *For the Learning of Mathematics, 2* (2), 29-30.

Van den Brink, J. (1987). Children as arithmetic book authors. *For the Learning of mathematics, 7*, 44-48.

Van den Brink, J.F. (1989). *Realistisch rekenonderwijs aan jonge kinderen* (Dissertatie) [Realistic arithmetic education to young children (Doctoral dissertation)]. Utrecht: OW&OC, Utrecht University.

Van den Brink, J., and Streefland, L. (1979). Young Children (6-8) – Ratio and Proportion. *Educational Studies in Mathematics, 10*, 403-420.

Van den Heuvel-Panhuizen, M. (1986). Het rekenonderwijs op de lom-school opnieuw ter discussie [Arithmetic education at schools for children with learning and behavioral difficulties again up for debate]. *Tijdschrift voor Orthopedagogiek, 25* (3), 137-145.

Van den Heuvel-Panhuizen, M. (1987). *Handle with care. Een theoretische en praktische terreinverkenning op het gebied van zorgverbreding bij reken-wiskundeonderwijs* [Handle with care. A theoretical and practical exploration into the accommodating of special needs in regular school mathematics education]. Enschede: SLO.

Van den Heuvel-Panhuizen, M. (1989a). De eerste uitkomsten. De eerste MORE-gegevens over het rekenonderwijs in groep 3 van de basisschool [The initial results. The initial MORE results on first-grade arithmetic education]. In E. de Moor (ed.), *Panama Cursusboek 7. Rekenen-wiskunde. Periodieke peiling onderwijsniveau, beredeneerde eindtermen, proeve van een nationaal programma* (pp. 59-68). Utrecht: HMN / SOL and OW&OC.

Van den Heuvel-Panhuizen, M. (1989b). Realistic Arithmetic/Mathematics Instruction and Tests. In C.A. Maher, G.A. Goldin, and R.B. Davis (eds.), *Proceedings of the Eleventh Annual Meeting of the PME-NA* (Vol. 2, Plenary Lectures and Symposia, pp. 143-147). New Brunswick, NJ: Rutgers, State University of New Jersey.

Van den Heuvel-Panhuizen, M. (1990a). Realistic Arithmetic/Mathematics Instruction and Tests. In K. Gravemeijer, M. van den Heuvel, and L. Streefland, *Contexts, Free Productions, Tests and Geometry in Realistic Mathematics Education.* Utrecht: OW&OC.

Van den Heuvel-Panhuizen, M. (1990b). Lijn in verhoudingen [Structure in ratio]. *Tijdschrift voor Nascholing en Onderzoek van het Reken-wiskundeonderwijs, 9* (2), 21-26.

Van den Heuvel-Panhuizen, M. (1991a). Three Taboos. Handout for workshop in Sheboygan, Wisconsin, 15 February 1991.

Van den Heuvel-Panhuizen, M. (1991b). Ratio in Special Education. A pilot study on the possibilities of shifting the bounderies. In L. Streefland (ed.), *Realistic Mathematics Education in Primary School* (pp. 157-181). Utrecht: CD-ß Press / Freudenthal Institute.

Van den Heuvel-Panhuizen, M. (1992a). Onderzoek van reken-wiskundeonderwijs: gaan we weer gouden tijden tegemoet? [Research on mathematics education: are we approaching a new Golden Age?]. *Tijdschrift voor Nascholing en Onderzoek van het Reken-wiskundeonderwijs, 11* (1), 33-43.

Van den Heuvel-Panhuizen, M. (1992b). *Per Sense Test*. Utrecht: Freudenthal Institute, Utrecht University. (internal publication – draft version)

Van den Heuvel-Panhuizen, M. (1993a). Toetsontwikkelingsonderzoek [Developmental research on assessment]. In R. de Jong and M. Wijers (eds.), *Ontwikkelingsonderzoek, theorie en praktijk* (pp. 85-110). Utrecht: NVORWO / Freudenthal Institute.

Van den Heuvel-Panhuizen, M. (1993b). New forms of assessment, but don't forget the problems. In *Proceedings of the Seventeenth International Conference for the Psychology of Mathematics Education (*Vol. III, pp. 186-193). Tsukuba, Japan: University of Tsukuba.

Van den Heuvel-Panhuizen, M. (1993c). *Show-What-You-Know Book on Percents*. Utrecht: Freudenthal Institute, University of Utrecht. (internal publication – draft version)

Van den Heuvel-Panhuizen, M. (1994a). New chances for paper-and-pencil tests. In L. Grugnetti (ed.), *Assessment focussed on the student*. Proceedings of the 45th CIEAEM Meeting (pp. 213-221). Cagliari, Italy: CIEAEM.

Van den Heuvel-Panhuizen, M. (1994b). New chances for paper-and-pencil tests in mathematics education. In J.E.H. van Luit (ed.), *Research on learning and instruction of mathematics in kindergarten and primary school*. Doetinchem / Rapallo: Graviant Publishing Company.

Van den Heuvel-Panhuizen, M. (1994c). Improvement of (didactical) assessment by improvement of the problems: An attempt with respect to percentage. *Educational Studies in Mathematics, 27* (4), 341-372.

Van den Heuvel-Panhuizen, M. (1995a). *Developing assessment problems on percentage. An example of developmental research on assessment conducted within the MiC project along the lines of Realistic Mathematics Education.* Utrecht: Freudenthal Institute, Utrecht University. (internal publication)

Van den Heuvel-Panhuizen, M. (1995b). *A representational model in a long term learning process – the didactical use of models in Realistic Mathematics Education.* A paper presented at AERA 1995, San Franscisco.

Van den Heuvel-Panhuizen, M. (1995c). *The Tension between Openness and Certainty: An Example of Developmental Research on Assessment*. A paper presented at the American Educational Research Association 1995, San Franscisco.

Van den Heuvel-Panhuizen, M. (1995d). Toetsen bij reken-wiskundeonderwijs [Assessment in mathematics education]. In L. Verschaffel and E. De Corte (eds.), *Naar een nieuwe reken/wiskundedidactiek voor de basisschool en de basiseducatie* (Leerinhoud 6, pp. 196-247). Brussel/Leuven: Studiecentrum Open Hoger Onderwijs / ACCO.

Van den Heuvel-Panhuizen, M., and Goffree, F. (1986). *Zo rekent Nederland* [This is the way The Netherlands does arithmetic]. Enschede: SLO.

Van den Heuvel-Panhuizen, M., and Gravemeijer, K. (1990a). *Reken-wiskunde Toetsen* [Mathematics Tests]. Utrecht: OW&OC / ISOR, Utrecht University. (internal publication)

Van den Heuvel-Panhuizen, M., and Gravemeijer, K.P.E. (1990b). Toetsen zijn zo slecht nog niet [Tests are not that bad]. *Didaktief, 20* (10), 13-15.

Van den Heuvel-Panhuizen, M., and Gravemeijer, K. (1991a). Tests are not all bad. An attempt to change the appearance of written tests in mathematics instruction at primary school level. In L. Streefland (ed.), *Realistic Mathematics Education in Primary School* (pp. 139-155). Utrecht: CD-β Press / Freudenthal Institute.

Van den Heuvel-Panhuizen, M., and Gravemeijer, K. (1991b). *Toets 7.1* [Test 7.1]. Utrecht: OW&OC, Utrecht University. (internal publication)

Van den Heuvel-Panhuizen, M., and Gravemeijer, K. (1993). Tests aren't all bad. An Attempt to Change the Face of Written Tests in Primary School Mathematics. In N.L. Webb and A.F. Coxford (eds.), *Assessment in the Mathematics Classroom, 1993 Yearbook* (pp. 54-64). Reston, VA: NCTM.

Van den Heuvel-Panhuizen, M., Middleton, J.A., and Streefland, L. (1995). Student-Generated Problems: Easy and Difficult Problems on Percentage. *For the Learning of Mathematics, 15* (3), pp. 21-27.

Van den Heuvel-Panhuizen, M., Streefland, L., Meyer, M., Middleton, J.A. and Browne, J.

(in press). Per Sense. In T.A. Romberg (ed.), *Mathematics in Contexts: A Connected Curriculum for Grades 5-8*. Chicago, IL: Encyclopaedia Britannica Educational Corporation.

Van den Heuvel-Panhuizen, M., Streefland, L., and Middleton, J.A. (1994). Students' own Productions as a Source for Developing Assessment. In J. van den Akker and W. Kuiper (eds.), *European Research on Curriculum. Book of Summaries of the first European Conference on Curriculum* (pp. 28-29). Enschede: Faculty of Educational Science and Technolgy, University of Twente.

Van der Blij, F. (1987). Toetsen, eindtermen en onderwijsontwikkeling [Tests, attainment targets and development of education]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 10-17). Utrecht: OW&OC, Utrecht University.

Van der Kooij, H. (1987). Opvattingen en toetsen [Viewpoints and tests]. In J. De Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwi*js (pp. 109-118). Utrecht: OW&OC, Utrecht University.

Van der Linden, W.J., and Zwarts, M.A. (1995). De opbrengsten van het basisonderwijs: een repliek [The output of primary education: a rebuttal]. *Tijdschrift voor Onderwijsresearch 20* (1), 34-41.

Van der Veer, R., and Valsiner, J. (1991). *Understanding Vygotsky. A quest for synthesis.* Oxford: Blackwell. (cited by Elbers, 1992)

Van Eerde, D., Lit, S., and Van den Berg, W. (1992). *Kwantiwijzer voor leerkrachten* [Kwantiwijzer for teachers]. Tilburg: Zwijsen.

Van Eerde, D., and Van den Berg, W. (1984). Kleuters en de Kwantiwijzer [Kindergarteners and the Kwantiwijzer]. *Willem Bartjens, 4* (1), 23-26.

Van Galen, F., Gravemeijer, K., Kraemer, J.M., Meeuwisse, A., and Vermeulen, W. (1985). *Rekenen in een tweede taal* [Arithmetic in a second language]. Enschede: SLO.

Van Galen, F., and Meeuwisse, A. (1986). Anderstalige leerlingen en rekenonderwijs [Non-Dutch speaking students and arithmetic education]. In E. Feijs and E. de Moor (eds.), *Panama Cursusboek 4* (pp. 128-134). Utrecht: SOL / OW&OC, Utrecht University.

Van Hoorn, M.C. (1987). Kritische kanttekeningen bij 'Problemen met het construeren van examens' [Critical remarks to 'Problems in constructing exams']. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 74-89). Utrecht: OW&OC, Utrecht University.

Van Luit, J.E.H. (1987). *Rekenproblemen in het speciaal onderwijs* (Dissertatie) [Learning difficulties with respect to arithmetic in special education (Doctoral dissertation)]. Nijmegen: Katholieke Universiteit Nijmegen.

Van Luit, J.E.H. (1988). Realistisch reken-wiskundeonderwijs in het speciaal onderwijs? [Realistic mathematics education in special education?]. *School & Begeleiding, 5* (17), 15-18.

Van Luit, J.E.H. (ed.), Kaskens, J.M.M., Van Zessen, T., and Timmermans, C.A. (1989). *Bronnenboek methoden rekenen/wiskunde. Invoeringsprogramma Speciaal Onderwijs* [Source book on mathematics textbook series. Implementation program for Special Education]. Utrecht: Faculteit Sociale Wetenschappen, Utrecht University.

Van Parreren, C.F. (1981). Leerproblemen bij kleuters, gezien vanuit de handelings- en leerpsychologie [Learning difficulties of kindergarteners, considered from the viewpoint of activity psychology and learning psychology]. *Tijdschrift voor Orthopedagogiek, 20,* 4-26.

Van Reeuwijk, M. (1995). Students' knowledge of algebra. In L. Meira, and D. Carraher (eds.), *Proceedings of the 19th PME Conference* (Vol. 1, pp. 135-150). Recife, Brazil: Universidade Federal de Pernambuco.

Van 't Riet, A. (1987). Toetsen in wiskundemethoden in het voortgezet onderwijs [Tests in mathematics textbook series in secondary education]. In J. De Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 136-141). Utrecht: OW&OC, Utrecht University.

Veltman, A. (1993). Van het begin en van het eind – ontwikkelingsonderzoek naar het rekenen tot honderd op de (lege) getallenlijn [From the beginning and from the end – developmental research into doing arithmetic up to one hundred on the (empty) number line]. *Tijdschrift voor Nascholing en Onderzoek van het Reken-wiskundeonderwijs, 11* (4), 7-13.

Verschaffel, L., De Corte, E., and Lasure, S. (1994). Realistic considerations in mathematical modeling of school arithmetic word problems. *Learning and Instruction, 4* (4), 273-294.

Victorian Curriculum and Assessment Board (VCAB) (1990). *Mathematics Study Design*. Melbourne: VCAB.

Vossen, H.M.M., et al. (s. a.). *Niveaucursus Rekenen* [A course arithmetic in levels]. Den Bosch: Malmberg.

Vygotsky, L.S. (1978). *Mind in Society. The Development of Higher Psychological Processes.* Cambridge, MA: Harvard University Press.

Weaver, J.F. (1955). Big Dividends from little Interviews. *Arithmetic Teacher,* April 1955, 40-47.

Webb, N.L. (1992). Assessment of Students' Knowledge of Mathematics: Steps Toward a Theory. In D.A. Grouws (ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 661-683). New York: NCTM / Macmillan.

Webb, N.L. (1993). Visualizing a Theory of the Assessment of Students Knowledge of Mathematics. In M. Niss (ed.), *Investigations into Assessment in Mathematics Education. An ICMI Study* (pp. 31-46). Dordrecht: Kluwer Academic Publishers.

Webb, N., and Briars, D. (1990). Assessment in Mathematics Classroom. In T.J. Cooney and C.R. Hirsch (eds.), *Teaching and Learning mathematics in the 1990s. 1993 Yearbook NCTM* (pp. 108-117). Reston, VA: NCTM.

Westerveld, M. (1972). *Het leren bijtellen in de eerste klas van de basisschool* (Doctoraalscriptie) [The learning to add on in the first grade of primary education (Master's thesis)]. Utrecht: Rijksuniversiteit Utrecht, IPAW. (cited by Koster, 1975)

Wiggins, G. (1989a). Teaching to the (Authentic) Test. *Educational Leadership, 46* (7), 41-47.

Wiggins, G. (1989b). A True Test: Towards More Authentic and Equitable Assessment. *Phi Delta Kappan, 70*, May, 703-713.

Wiggins, G. (1992). Creating Tests Worth Taking. *Educational Leadership, 49* (8), 26-33.

Wijdeveld, E. (1980). Zich realiseren [To realize]. In: S. Pieters (ed.), *De achterkant van de Möbiusband* (pp. 23-26). Utrecht: IOWO.

Wijffels, E. (1993). *Laat eens horen wat er allemaal in je hoofd rommelt* (Doctoraalscriptie) [Tell us what's on your mind (Master's thesis)]. Utrecht: Universiteit Utrecht.

Wijnstra, J.M. (1995). De opbrengsten van het basisonderwijs volgens de CEB: enkele kanttekeningen bij de gevolgde procedures [The output of primary education according to the CEB: some remarks on the applied procedures]. *Tijdschrift voor Onderwijsresearch, 20* (1), 28-33.

Wijnstra, J. (ed.) (1988). *Balans van het rekenonderwijs in de basisschool* [An account of mathematics education in primary school]. Arnhem: Cito.

Wiliam, D. (1993). *Assessing authentic tasks: norms, criteria, and other referents*. A paper presented at the Nordic Symposium Research on Assessment in Mathematics Education, University of Göteborg, November 5, 1993.

Wilson, L.D., and Chavarria, S. (1993). Superitem Tests as a Classroom Assessment Tool. In N.M. Webb (ed.), *Assessment in the Mathematics Classroom. 1993 Yearbook NCTM* (pp. 135-142). Reston, VA: NCTM.

Wilson, M. (1992). Measuring levels of understanding. In T.A. Romberg, *Mathematics Assessment and Evaluation. Imperatives for Mathematics Educators* (pp. 213-241). Albany, NY: SUNY Press.

Winograd, K. (1992). What fifth graders learn when they write their own math problems. *Educational Leadership, 49* (7), 64-67.

Wittmann, E. (1991). Die weitere Entwicklung des Mathematikunterrichts in der Grundschule – was musz sich bewegen? [The further development of mathematics education in primary school – what must change?]. In *Beiträge zum Mathematikunterricht 1991* (pp. 41-48). Bad Salzdetfurth: Franzbecker.

Wittmann, E., and Müller, G.N. (1990). *Handbuch produktiver Rechenübungen.* [Handbook on productive arithmetic exercises]. Stuttgart/Düsseldorf: Klett-Schulbuchverlag.

Wolf, D.P. (1989). Portfolio Assessment: Sampling Student Work. *Educational Leadership, 46* (7), 35-39.

Wolf, T.H. (1973). *Alfred Binet*. Chicago: University of Chicago Press. (cited by Kilpatrick, 1993)

Wood, D. (1988). *How children think and lear*n. Oxford: Blackwell. (cited by Elbers, 1992)

Woodward, H. (1993). *Negotiated Evaluation*. Newtown, Australia: Primary English Teaching Association.

Yackel, E. (1992). *The Evolution of Second Grade Children's Understanding of What Constitutes an Explanation in a Mathematics Class.* A paper presented at ICMI-7, Quebec.

Yackel, E., Cobb, P., and Wood, T. (1992). Instructional development and assessment from a socio-constructivist perspective. In G.C. Leder (ed.), *Assessment and Learning of Mathematics* (pp. 63-82). Hawthorn, Victoria: Australian Council for Educational Research.

Zehavi, N., Bruckheimer, M., and Ben-Zvi, R. (1988). Effect of assignment projects on students mathematical activity. *Journal for Research in Mathematics Education, 19* (5), 421-438.

Zwaneveld, B. (1987). Kort-antwoordvragen als alternatief voor meerkeuzevragen [Short-answer questions as an alternative to multiple-choice questions]. In J. de Lange (ed.), *Toetsen, eindtermen en opvattingen over wiskunde-onderwijs* (pp. 171-175). Utrecht: OW&OC, Utrecht University.

Samenvatting

In het ontwikkelingsonderzoek dat aan dit proefschrift ten grondslag ligt, is nagegaan wat de implicaties zijn van de realistische onderwijstheorie voor het toetsen.

In de vijfentwintig jaar die inmiddels in Nederland is gewerkt aan de ontwikkeling en implementatie van realistisch reken-wiskundeonderwijs, is de bijbehorende specifieke manier van toetsen lang op de achtergrond gebleven. Zonder afbreuk te doen aan al het werk dat in het verleden hiervoor is gedaan, kan gesteld worden dat deze studie wat het basisonderwijs betreft in zekere zin de eerste is die speciaal aan dit onderwerp is gewijd.

Voor het voortgezet onderwijs ligt dit duidelijk anders. Om de gewenste veranderingen in realistische richting veilig te stellen, is men daar in het begin van de jaren tachtig, tegelijk met de vernieuwing van het wiskundecurriculum van het vwo, begonnen met de ontwikkeling van alternatieve examens.

Later is ook ten behoeve van het basisonderwijs naar alternatieven gezocht voor de bestaande manier van toetsen. Het MethodenOnderzoek REkenen-wiskunde (in het kort aangeduid met MORE-onderzoek), dat in feite een onderzoeksproject was naar de implementatie en effecten van wiskundemethoden, heeft hierbij een cruciale rol gespeeld. De toetsontwikkeling die nodig was om de leerresultaten te kunnen vergelijken, is gaandeweg verbreed en zich steeds meer gaan richten op de consequenties en mogelijkheden van toetsen bij realistisch reken-wiskundeonderwijs. In dit opzicht vormde het MORE-onderzoek dan ook een proeftuin voor realistisch toetsen. De ideeën en bevindingen die hierbij naar voren zijn gekomen, hebben uiteindelijk geleid tot de onderhavige studie. Het doel hiervan is om op basis van toetsontwikkelingswerk, literatuurstudie en reflectie, via aanscherping van theoretische noties en met behulp van concrete voorbeelden een bijdrage te leveren aan de ontwikkeling van het toetsen bij realistisch reken-wiskundeonderwijs.

Dit proefschrift geeft de huidige stand van zaken weer met betrekking tot het realistisch toetsen. Dit houdt echter geenszins in dat alle aspecten van het toetsen worden bestreken.

Zo gaat het op de eerste plaats om het *toetsen van rekenen-wiskunde in het basisonderwijs*.

Een tweede inperking is het *didactisch toetsen*, bedoeld ter ondersteuning van het onderwijsleerproces. Dit is het toetsen dat dicht staat bij het onderwijs en in principe deel uitmaakt van de dagelijkse onderwijspraktijk. Als zodanig onderscheidt het zich van het selecteren van leerlingen en het evalueren van onderwijs. Overigens betekent deze keuze niet dat deze andere oogmerken van toetsen zouden worden afgewezen – in tegendeel.

De derde inperking betreft de toespitsing op *schriftelijk toetsen*. Het accent ligt hierbij op korte vragen. Hoewel een dergelijke invulling van toetsen op het eerste gezicht in tegenspraak lijkt met de realistische opvattingen, werd in de loop van de studie duidelijk dat schriftelijk toetsen met behulp van korte vragen zeer wel verenigbaar is met realistisch reken-wiskundeonderwijs. Sterker, het kan zelfs een inspiratiebron vormen voor de verdere ontwikkeling ervan.

Toen in Nederland aan het eind van de jaren tachtig de tijd rijp bleek voor een nieuwe doordenking van de consequenties voor het toetsen op het niveau van het basisonderwijs, was er op dit gebied ook internationaal een vernieuwing op gang gekomen. Deze gelijktijdige internationale ontwikkelingen hebben echter nauwelijks invloed gehad op het hier uitgevoerde toetsontwikkelingswerk. Toch zijn in de onderhavige studie, waar mogelijk, verbindingen gelegd met deze nieuwe internationale inzichten.

Het proefschrift bestaat uit twee delen en omvat in totaal zeven hoofdstukken. Het eerste deel met vier hoofdstukken vormt de kern van het boek. Hierin komen achtereenvolgens aan bod:
1 de rol van het toetsen in de beginperiode van realistisch reken-wiskundeonderwijs,
2 de toetsontwikkeling binnen het MORE-onderzoek,
3 de huidige stand van zaken met betrekking tot het toetsen bij realistisch reken-wiskundeonderwijs, en
4 de mogelijkheden van schriftelijke toetsen hierbij.
In het tweede deel zijn drie hoofdstukken opgenomen die in feite als bijlagen moeten worden beschouwd. Ze doen verslag van deelonderzoeken en hebben achtereenvolgens betrekking op:
5 een rekentoets voor het begin van groep 3,
6 een toets over verhoudingen die in het speciaal onderwijs is afgenomen, en
7 een deel van een toets over procenten.

Hoofdstuk 1 beschrijft de beginjaren van de Nederlandse vernieuwingsbeweging. Op basis van literatuuronderzoek wordt een overzicht geboden van de toenmalige denkbeelden over toetsen. Het hoofdstuk start met een korte uiteenzetting van de hoofdlijnen van realistisch reken-wiskundeonderwijs. Hierbij worden met name die kenmerken naar voren gehaald welke van beslissende betekenis zijn voor het toetsen:
- de eigen activiteit en inbreng van de leerlingen,
- de koppeling aan de realiteit en de toepassingsgerichtheid, en
- de verschillende niveaus van begrijpen.

Daarna wordt ingegaan op de plaats van het toetsen. De beschrijving ervan strekt zich uit over de periode die loopt van de jaren zestig, toen de ontwikkeling van dit reken-wiskundeonderwijs begon, tot het jaar 1987.

In de beginjaren was de houding ten opzichte van toetsen vooral gericht tégen toetsen. Tenminste, die indruk kan gemakkelijk ontstaan vanwege de vaak heftige strijd die destijds tegen de toenmalige toetsen werd gevoerd. Men had grote bezwaren tegen de doelen, de doelbeschrijvingen en de taxonomieën die werden gebruikt bij het construeren van toetsen, en tegen de eenzijdig psychometrische benadering hierbij. Voorts had men ook bedenkingen tegen de sterk geformaliseerde vorm van de toetsen en tegen de valkuilen in toetsvragen. Ook was er kritiek op de wijze waarop antwoorden van leerlingen werden beoordeeld.

Bij een nadere beschouwing van de publikaties wordt echter duidelijk dat, behalve opvattingen over hoe het toetsen niet moest, er toen ook heel duidelijke opvattingen leefden over hoe het wel diende te gebeuren. Zo werd hoge prioriteit toegekend aan observeren en kreeg het continue en geïntegreerde karakter van toetsen veel nadruk. Aan de leerkracht werd een centrale rol toebedacht en verder was men van mening dat naast cognitieve ook sociaal-emotionele aspecten getoetst moesten worden. Tevens had men een sterke voorkeur voor open toetsvormen en streefde men naar echte toepassingsproblemen in plaats van aangeklede redactieopgaven.

Behalve algemene opvattingen over hoe het toetsen wél moest en over hoe de bestaande toetsen verbeterd konden worden – waarbij zowel het standpunt van de vakinhoud werd ingenomen als dat van het lerende kind – zijn destijds ook concrete alternatieven ontwikkeld. De zogenoemde toetsles is daar een duidelijk voorbeeld van. Het toetsen is hierbij ingebed in een lessituatie met de hele groep. Daarnaast werd gezocht naar geschikte observatie- en interviewtechnieken.

Het zoeken naar alternatieve manieren van toetsen bleef overigens niet beperkt tot het IOWO en het latere OW&OC (de voorlopers van het Freudenthal Instituut). Ook buiten deze kring werd in die tijd gewerkt aan een wijze van toetsen die meer in overeenstemming is met de realistische opvattingen over reken-wiskundeonderwijs. Een voorbeeld hiervan is het Kwantiwijzer-instrumentarium.

In de geschiedenis van het toetsen binnen realistisch reken-wiskundeonderwijs neemt 1987 een bijzondere plaats in. In dat jaar verscheen namelijk de dissertatie van De Lange over nieuwe toetsvormen voor het voortgezet onderwijs, vond de eerste afname plaats van de door het Cito uitgevoerde Periodieke Peiling van het Onderwijsniveau voor het vak rekenen-wiskunde en werd door OW&OC een toetsconferentie georganiseerd. Deze gebeurtenissen kunnen beschouwd worden als een afsluiting van een periode waarin het fundament werd gelegd voor de uitbouw van het theoretisch toetskader.

In *hoofdstuk 2* staat het MORE-onderzoek centraal. Dit in 1987 gestart onderzoek behoefde alleen de ontwikkeling van evaluatieve toetsen. De ervaringen hiermee leidden echter tot een nieuwe doordenking van het didactisch toetsen en wel speciaal het schriftelijk toetsen.

De volgende cruciale momenten in het ontwikkelingsproces gaven de aanzet hiertoe. De eerste impuls kwam voort uit de onverwachte uitkomsten op de toets van begin groep 3. Bij jonge kinderen bleek een schriftelijke toets meer te kunnen onthullen over hun reken-wiskundevaardigheden dan tot nu toe werd aangenomen. Ook scoorden de leerlingen op bepaalde onderdelen veel hoger dan de deskundigen voorspelden. Een andere ervaring was dat de toetsing aanzienlijk informatiever kan worden indien de leerlingen zelf de getallen in een opgave mogen kiezen. Voorts opende een bij toeval gemaakte fout in een toetsopgave de ogen voor het feit dat schriftelijk toetsen niet per se éénrichtingsverkeer hoeft te zijn. Evenzo leerde het feit dat van bepaalde opgaven zowel de contextvorm als de kale versie was opgenomen, dat presentatiewisselingen belangrijke informatie kunnen opleveren. Ook leidden de sporen van oplossingsstrategieën op de toetsbladen ertoe, dat er steeds explicieter werd gezocht naar allerlei manieren om strategieën bloot te leggen. Het afbeelden van kladblaadjes op de toetsopgaven is hier een voorbeeld van. De meest ingrijpende ontdekking was echter dat door te werken met steunbiedende contexten en modellen van situaties er een zekere gelaagdheid in opgaven kan worden aangebracht. Dit idee van ‘toetsopgaven met rek’ betekende een nieuw element in het denken over toetsen – in ieder geval binnen het MORE-project.

Achteraf moet geconstateerd worden dat de inzichten in de mogelijkheden van schriftelijke toetsen niet los gezien kunnen worden van de context waarin de MORE-toetsen ontwikkeld zijn. Zo bepaalde de grootte van de onderzoeksgroep dat er schriftelijke toetsen ontwikkeld moesten worden, terwijl de voorkeur eigenlijk uitging naar observaties en mondelinge interviews. Dat de plaatjes bij de betreffende toetsopgaven zo dominant zijn, heeft weer alles te maken met het gegeven dat begonnen moest worden in groep 3. Ook het feit dat er twee totaal verschillende reken-wiskundemethoden bij het onderzoek betrokken waren, heeft duidelijk zijn stempel op de toetsontwikkeling gezet. Voorkomen moest worden dat een bepaalde groep werd geconfronteerd met opgaven die ze ‘nog niet hadden gehad’. Dit opende echter tegelijkertijd de weg naar het zogenoemde vooruit-toetsen. Op dezelfde manier vroeg het longitudinale karakter van het onderzoek om opgaven met een grote reikwijdte. Bovendien bood de duur van het onderzoek ruimschoots de mogelijkheid om opgaven bij te stellen en opnieuw uit te proberen. Verder heeft het gelijktijdig afnemen van een mondelinge variant van de toetsen ertoe bijgedragen dat het dynamisch schriftelijk toetsen in beeld kwam. Bij dit alles is ten slotte ook nog van grote invloed geweest dat het toetsen, noch door de leerlingen noch door leerkrachten, als beoordelend werd ervaren. Deze onbelaste onderzoekscontext verschafte de nodige experimenteerruimte.

Hoofdstuk 3 biedt een algemene plaatsbepaling van de huidige stand van zaken. Tevens wordt in dit hoofdstuk een aanzet geleverd tot een verdere uitwerking van de realistische onderwijstheorie voor het toetsen.

Toetsen bij realistisch reken-wiskundeonderwijs is vooral didactisch toetsen. Dit is in alle aspecten ervan herkenbaar. Doel, inhoud, procedures en te gebruiken instrumenten zijn alle nauw verbonden met het onderwijs. Een ander kenmerk van realistisch toetsen is de spilfunctie van de problemen. Belangrijker dan de vorm waarin iets wordt gevraagd, is *wat* er wordt gevraagd. Mathematisch-didactische analyses zijn hiervoor onmisbaar. Daarbij komt naar voren welke inzichten en vaardig-heden van belang zijn en in welke situaties toepasbaar. Afgezien van inhoudspecifieke eisen zijn er ook twee algemene criteria waaraan problemen dienen te voldoen: ze moeten zinvol en informatief zijn. Het eerste criterium houdt in dat ze zowel vanuit het vak als vanuit de leerlingen gezien zinvol moeten zijn. Hierin onderscheiden toetsproblemen zich overigens niet van de andere problemen die bij realistisch reken-wiskundeonderwijs worden gebruikt. Het tweede criterium daarentegen is meer specifiek voor toetsproblemen.

Met name door het gebruik van contexten kan aan deze vereisten tegemoet worden gekomen. Op de eerste plaats kunnen ze bijdragen aan de toegankelijkheid van toetsopgaven. Verder bieden ze de mogelijkheid om oplossingen op verschillende niveaus te geven. Zo komt er als het ware meer rek in toetsopgaven en wordt het toetsen daarmee tegelijkertijd doorzichtiger. Ten slotte bieden contexten vaak aangrijpingspunten voor verschillende oplossingsstrategieën. Dat aan contexten een belangrijke betekenis wordt toegekend, houdt echter niet in dat kale sommen niet bij realistisch reken-wiskundeonderwijs zouden passen. Contexten kunnen immers ook refereren aan puur wiskundige structuren. Kenmerkend voor de aanduiding ‘realistisch’ is in dit verband niet alleen de relatie met de objectieve werkelijkheid maar ook met de subjectieve werkelijkheid. Contexten moeten betekenisvol en voorstelbaar zijn voor de leerlingen. Bovendien moeten beide soorten contexten zich lenen voor mathematisering. Hetgeen betekent dat contextproblemen met wiskundige methoden en modellen opgelost kunnen worden.

In het tweede deel van dit hoofdstuk worden de kenmerken van realistisch toetsen gespiegeld aan de ideeën die buiten de kring van realistisch reken-wiskundeonderwijs ontwikkeld zijn, met name aan de internationale toetsvernieuwing. Veel van de realistische kenmerken van het toetsen zijn ook aanwijsbaar in de ontwikkelingen die in Amerika, Engeland en Australië op gang zijn gekomen. Ook hier blijkt duidelijk sprake van een didactische gerichtheid van het toetsen. Een apart punt van aandacht, dat eveneens aansluit bij de realistische opvattingen, vormt het zogenoemde toetscontract. Dit houdt onder meer in dat de leerlingen precies op de hoogte moeten zijn van de bedoeling van het toetsen.

Naast overeenkomsten zijn er ook verschilpunten: in het begin van de toetsvernieuwing lag met name in Amerika het accent vooral op vorm- en organisatieaspecten. De laatste tijd is er echter ook aandacht voor de toetsopgaven zelf. Tegelijkertijd is deze gerichtheid op de inhoud van de problemen ook een bron van mogelijke verschillen. Als er immers sprake is van een andere didactiek, zal het didactisch toetsen ook een andere inhoud krijgen.

Inventarisatie van algemene eisen waaraan toetsproblemen dienen te voldoen laat een grote mate van consensus zien. Zo moeten goede toetsproblemen een wiskundig relevante inhoud bevatten, de moeite waard zijn om op te lossen, meer dan één antwoord opleveren of op meerdere manieren zijn op te lossen en het oplossingsproces zichtbaar maken. Goede toetsproblemen kunnen verder verschillende verschijningsvormen hebben. Bepalend is wat men wil toetsen en met welke bedoelingen. In bepaalde gevallen kan dit betekenen, dat een multiple-choice opgave heel geschikt is. Aan de andere kant leidt het open maken van gesloten problemen niet automatisch tot verbetering. Een belangrijk gegeven dat bij het zoeken naar goede toetsproblemen niet uit het oog verloren mag worden, is dat iedere toetsopgave door de leerlingen op eigen wijze wordt geïnterpreteerd. Het is dus moeilijk om over *de* toetsopgave te spreken.

Ook kenmerkend voor de nieuwe ideeën over toetsen is, dat er een andere invulling wordt gegeven aan de traditionele psychometrische kwaliteitseisen. Zo wordt het criterium van objectiviteit steeds meer vervangen door ‘fairness’, recht doen aan de leerlingen. Voorts vindt er een accentverschuiving plaats van betrouwbaardheid naar validiteit. Dit is precies waarvoor binnen realistisch reken-wiskundeonderwijs ook altijd is gepleit.

Een laatste punt van overeenkomst tussen realistisch toetsen en de internationale toetsvernieuwing vormt het belang dat men hecht aan het realiteitsgehalte van het toetsen en meer in het bijzonder aan de rol van de context. De vele onderzoeksgegevens die van buiten de realistische kring hierover beschikbaar zijn, vormen onmiskenbaar een rijke aanvulling op het realistische gedachtengoed. Toch is het niet zo dat wat ‘authentic assessment’ heet, gelijk gesteld mag worden met realistisch toetsen. Behalve dat contexten binnen realistisch rekenwiskundeonderwijs ruimer worden opgevat dan authentieke situaties, hebben ze ook een bredere functie: in realistisch reken-wiskundeonderwijs zijn contexten zowel doel als bron.

In *hoofdstuk 4* wordt als aanvulling op de voorgaande algemene plaatsbepaling de aandacht verlegd naar de toetsinstrumenten. Hierbij gaat het om korte, schriftelijke toetsopgaven en de verrijking ervan vanuit de realistische onderwijstheorie.

Met de verandering van het reken-wiskundeonderwijs zijn met name de bestaande schriftelijke toetsen – zowel nationaal als internationaal – onder druk komen te staan. Het sterkst geldt dit voor multiple-choice toetsen. De bezwaren tegen deze en andere schriftelijke toetsen zijn, dat ze niet passen bij de veranderde doelen en de aanpak van het onderwijs en dat ze bovendien geen informatie geven over de toegepaste strategieën.

Dit hoofdstuk is voor het grootste deel gewijd aan alternatieve vormen van schriftelijk toetsen die binnen realistisch reken-wiskundeonderwijs voor het basisonderwijs zijn ontwikkeld. De centrale vraag hierbij was, hoe korte schriftelijke opgaven informatiever gemaakt kunnen worden. Voorwaarde om dit te bereiken is, dat er eerst gebroken wordt met de veronderstelling dat opgaven niet op verschillende manieren opgelost kunnen worden, niet meerdere antwoorden kunnen hebben en dat het altijd duidelijk moet zijn wat het goede antwoord is. Pas daarna kunnen maatregelen in beeld komen die van schriftelijke toetsen een rijk instrument kunnen maken, zoals het aanbieden van:

- kladpapier op het toetsblad,
- expliciete vragen naar de strategie,
- opgaven waarbij de leerlingen verschillende goede antwoorden kunnen geven,
- keuze-opgaven waarbij de leerlingen zelf de moeilijkheidsgraad van de opgave in de hand hebben,
- eigen produkties,
- opgavenparen met verschillende presentatievormen,
- opgavenparen waarbij de uitkomst van de ene opgave kan worden gebruikt voor het oplossen van de andere,
- opgaven met illustraties die de mogelijkheid bieden van verschillende niveaus van oplossen,
- opgaven met hulpsommen.

Kenmerkend voor een aantal van deze maatregelen is, dat er meer dynamiek en rek in de opgaven komt. Later zijn deze aan interviewtechnieken verwante maatregelen ook toegepast in de vorm van toetsopgaven met een vangnet-vraag, een tweedekans-vraag of een hulpblad met hints.

Realistisch toetsen heeft niet alleen consequenties voor de vorm en de inhoud van de toetsopgaven, maar ook voor de manier waarop de antwoorden van de leerlingen worden geïnterpreteerd en geanalyseerd. Buiten de kring van realistisch reken-wiskundeonderwijs wordt hierop eveneens gewezen. Afgezien van andere scoringscategorieën vraagt het nakijken echter ook dat zoveel mogelijk het standpunt van de leerlingen wordt ingenomen.

De kern van het hoofdstuk wordt gevormd door de terugblik op de consequenties van de realistische opvattingen voor het schriftelijk toetsen. Hierbij komt naar voren dat binnen realistisch reken-wiskundeonderwijs de grenzen van de traditionele schriftelijke toetsen op een aantal punten zijn verlegd:

- van passief naar actief toetsen,
- van statisch naar dynamisch toetsen,
- van een op zekerheid gericht toetsen naar een toetsen waarbij meer onzeker is, maar wel rijkere informatie wordt verkregen,
- van toetsen met opgaven op verschillende niveaus naar een toetsen met vooral opgaven die op verschillende niveaus zijn op te lossen.

Deze noties omtrent de aard van het toetsen vormen, samen met de eerder besproken criteria en de praktische uitwerkingen, de kern van de uitbouw van de realistische toetstheorie. Behalve dat de realistische onderwijstheorie hiervoor het vertrekpunt vormt, is er ook sprake van een beïnvloeding in de omgekeerde richting. Juist omdat het denken van leerlingen zo'n belangrijke voedingsbron is voor realistisch reken-wiskundeonderwijs, kan de verdere ontwikkeling van realistisch toetsen ook impulsen geven aan de verdere ontwikkeling van de realistische onderwijstheorie.

In het tweede deel worden drie onderzoeken beschreven die als basis hebben gediend van het voorgaande.

In *hoofdstuk 5* wordt de ontstaansgeschiedenis geschetst van de eerste MORE-toets, bedoeld voor de beginmeting in groep 3. Vanwege het grote aantal leerlingen dat bij het onderzoek betrokken was, moest noodgedwongen gekozen worden voor een schriftelijke toets. Nadat in het hoofdstuk eerst enige achtergrondinformatie is gegeven over de hiervoor ontwikkelde toets, wordt vervolgens uitgebreid stilgestaan bij de onverwachte uitkomsten ervan. De leerlingen bleken namelijk al over meer wiskundige kennis en vaardigheden te beschikken dan door deskundigen werd voorspeld. Het zijn, zoals eerder gezegd, met name deze verrassende bevindingen geweest die aanleiding waren om de mogelijkheden van schriftelijke toetsen verder te onderzoeken.

Daarnaast heeft het onderzoek ook in het buitenland geleid tot een hernieuwde belangstelling voor het beginniveau van kinderen die voor het eerst geconfronteerd worden met systematisch reken-wiskundeonderwijs. Zo zijn bepaalde onderdelen van de toets onder andere afgenomen in Duitsland en Zwitserland. Ook hier bleken de kinderen het vaak beter te doen dan verwacht. Bovendien gaven deze onderzoeken nog aanwijzingen voor de verbetering van de toetsen.

Hoofdstuk 6 doet verslag van een onderzoek naar de mogelijkheden van realistisch reken-wiskundeonderwijs in het speciaal onderwijs. De aanleiding tot dit onderzoek vormde de kloof die er in Nederland bestaat tussen het reguliere onderwijs en het speciaal onderwijs. De vernieuwing van het reken-wiskundeonderwijs in realistische richting is in het speciaal onderwijs eigenlijk betrekkelijk lang uitgebleven. De voornaamste reden hiervoor is de veronderstelling dat de realistische aanpak (de nadruk op contexten, het starten vanuit informele kennis, de variatie in oplossingswijzen en de reflectie daarop) te hoge eisen aan de leerlingen zou stellen. Het doel van de ondernomen studie was om de juistheid van deze opvatttingen te toetsen. Dit is gedaan met behulp van een op de realistische principes gebaseerde toets over verhoudingen. Dit onderwerp hoort niet tot het vigerende onderwijsprogramma van het speciaal onderwijs. De toets is afgenomen in de twee hoogste groepen van twee mlk-scholen. De resultaten laten zien, dat de leerlingen heel behoorlijk met de contextproblemen over verhoudingen overweg kunnen. Bovendien zijn ze in staat via kladblaadjes te laten zien hoe ze aan hun oplossingen zijn gekomen. Deze uitkomsten staan in schril contrast met de mogelijkheden die doorgaans aan deze leerlingen worden toegedicht. Er is dus alle reden om het curriculum en de gangbare didactiek van het speciaal onderwijs aan een herbezinning te onderwerpen.

Hoofdstuk 7 handelt over een onderzoek dat is uitgevoerd in het kader van het ‘Mathematics in Context’-project. Dit is een Amerikaans curriculumproject, gericht op de ontwikkeling van een nieuw wiskundeprogramma voor de hoogste twee jaren van het basisonderwijs en de eerste twee jaren van het voortgezet onderwijs. Doel van het betreffende toetsontwikkelingsonderzoek was de kwaliteit van korte schriftelijke toetsopgaven te verbeteren. Hierbij is onder andere gezocht naar open toetsopgaven die de leerlingen de mogelijkheid bieden om te laten zien wat ze kunnen. In twee onderzoekscycli zijn een serie toetsopgaven ontwikkeld, uitgeprobeerd, gereviseerd en nogmaals beproefd. De toetsopgaven hadden betrekking op het onderdeel procenten. Dit hoofdstuk spitst zich toe op één van deze toetsopgaven. De opgave is bedoeld om te meten of leerlingen inzicht hebben in het relatieve karakter van procenten. De resultaten van het eerste deel van het onderzoek toonden aan dat open opgaven niet altijd voldoende informatie geven om te kunnen besluiten of de leerlingen wel of niet een bepaald niveau van begrip hebben. Het tweede deel van het onderzoek wijst uit dat het toevoegen van een zogenoemde ‘vangnet-vraag’ een geschikt middel is om deze problemen te voorkomen. Via zo’n vraag kan meer duidelijkheid verkregen worden over wat leerlingen kunnen, zonder dat men de toetsopgave meer gesloten hoeft te maken. Deze oplossing in de vorm van een vangnet-vraag staat niet op zich, maar hangt samen met andere toepassingen van interviewtechnieken bij schriftelijk toetsen, waarmee het schriftelijk toetsen kan worden verrijkt.

Bovenstaande onderzoeken hebben voor verschillende leerlingpopulaties de mogelijkheden van schriftelijke toetsen blootgelegd. Schriftelijke toetsen blijken meer te kunnen opleveren dan alleen uitkomsten in de vorm van antwoorden. Ze kunnen ook houvast geven voor verder onderwijs. Daarmee zijn het didactische instrumenten geworden en hebben ze een onvervreemdbare plaats binnen de domein-specifiek realistische theorie voor reken-wiskundeonderwijs verworven – en mogelijk in de toekomst ook in de praktijk van het onderwijs.

Curriculum vitae

Marja van den Heuvel-Panhuizen was born in 1950 in Gemert, The Netherlands. After completing secondary education (ULO-B) in Gemert she studied at the Kweekschool in Veghel to become a teacher. She taught in primary education and special education for the next eleven years. While teaching, she studied Pedagogy at the Katholieke Leergangen in Tilburg, and later enrolled in a graduation program in Pedagogy, with a speciality in Educational Science, at Utrecht University. She received her doctorandus degree in 1983 and became involved in the research and development of mathematics education at a number of institutes and organizations.

She collaborated on a program for Dutch Educational Television (NOT) called ‘Pluspunt’, whose purpose was to introduce teachers to realistic mathematics education. Subsequently, at the Pedagogical Institute (PI) in Rotterdam, she participated in the development of ‘Zo reken ik ook!’, a mathematics textbook series for children with learning disabilities. This was followed by a study – commissioned by the National Institute for Curriculum Development (SLO) – on the practice of mathematics education, which resulted in the publication of the report entitled ‘Zo rekent Nederland’. Thereafter, as a member of the ‘Zorgverbreding’ project at the SLO, she conducted a study into the potential of accommodating a wider range of educational needs in mathematics education.

Since 1987, she has been employed by the Freudenthal Institute of Utrecht University, where she has been involved in a variety of research and development projects. She participated in the MORE project, an extensive study involving research into the use and effects of mathematics textbooks in primary education, which was conducted in collaboration with the Educational Science Department (VOU) at Utrecht University. The development of tests that took place during this research, and the experiences gained from developing these tests, provided the stimulus for further research in this area and, in turn, led to the present dissertation on assessment in realistic mathematics education. She was also involved in the large-scale ‘Mathematics in Context’ project. This project was a collaboration with the University of Wisconsin at Madison, aimed at developing a new American mathematics curriculum for grades 5 through 8. Her primary activities in this project focused on number.

Presently, a significant part of her work at the Freudenthal Institute is the development of an in-service training course for primary school teachers. She is also engaged in research into the differences in mathematics achievements between girls and boys. Furthermore, she is involved in a number of (international) assessment projects, including a project in the United States on developing a multi-media assessment tool for teachers.