

LARGE-SCALE ASSESSMENT OF CHANGE IN STUDENT ACHIEVEMENT: DUTCH
PRIMARY SCHOOL STUDENTS' RESULTS ON WRITTEN DIVISION IN 1997 AND 2004
AS AN EXAMPLE

MARJA VAN DEN HEUVEL-PANHUIZEN

FREUDENTHAL INSTITUTE FOR SCIENCE AND MATHEMATICS EDUCATION (FISME) AND
INSTITUTE FOR EDUCATIONAL PROGRESS, HUMBOLDT UNIVERSITY BERLIN

ALEXANDER ROBITZSCH

INSTITUTE FOR EDUCATIONAL PROGRESS, HUMBOLDT UNIVERSITY BERLIN

ADRI TREFFERS

FREUDENTHAL INSTITUTE FOR SCIENCE AND MATHEMATICS EDUCATION (FISME),
UTRECHT UNIVERSITY

OLAF KÖLLER

INSTITUTE FOR EDUCATIONAL PROGRESS, HUMBOLDT UNIVERSITY BERLIN

This article discusses large-scale assessment of change in student achievement and takes the study by Hickendorff, Heiser, Van Putten, and Verhelst (2009) as an example. This study compared the achievement of students in the Netherlands in 1997 and 2004 on written division problems. Based on this comparison, they claim that there is a performance decline in this subdomain of mathematics, and that there is a move from applying the digit-based long division algorithm to a less accurate way of working without writing down anything. In our discussion of this study, we address methodological challenges that come in when investigating long-term trends in student achievements, such as the need for adequate operationalizations, the influence of the time of measurement and the necessity of the comparability of assessments, the effect of the assessment format, and the importance of inclusion relevant covariates in item response models. All these issues matter when assessing change in student achievement.

Key words: large-scale assessment, primary school, achievement, change, written division.

1. Introduction

Investigating changes in educational outcomes is important for evaluating educational systems and the reform of these systems. Therefore, large comparative studies such as PISA, TIMSS, and NAEP put much effort in identifying long-term trends. This, however, involves a number of methodological challenges with respect to collecting data, using compatible test designs, and applying statistical analyses. This is especially true when these trends studies take place in a context of a change in educational policy and the implementation of educational reforms.

In the Netherlands, the five-yearly PPO (National Assessment of Educational Achievement) carried out by CITO (National Institute for Educational Measurement) is meant to study changes over time. For example, the most recent PPO has revealed that the level of mathematics achievement of Dutch primary school students has changed over the last 2 decades (cf. Janssen,

Requests for reprints should be sent to Marja van den Heuvel-Panhuizen, Freudenthal Institute, Utrecht University, Postbus 9432, 3506 GK Utrecht, The Netherlands. E-mail: m.vandenheuvel@fi.uu.nl

Van der Schoot, & Hemker, 2005; Van der Schoot, 2008). These changes were predominantly found in the whole number domain, whereas in the domains of rational numbers, measurement and geometry the scores were generally stable except for a few exceptions. Characteristic of the changes in the whole number domain is that achievement has strongly increased on number sense and estimation and to a lesser degree on mental calculation, especially with respect to mental addition and subtraction. However, at the same time, the opposite was the case in the domain of written calculations. Here, a strong decrease was found, especially in written multiplication and division.

This change in the competence profile of Dutch primary school students is in line with the reform proposal formulated some twenty years ago (Treffers & De Moor, 1984). The so-called “realistic” approach to mathematics education that is expressed in this document claimed the importance of mental calculation including smart calculation, estimation, and number sense and argued for spending less time on the mechanistic performance of algorithms. These ideas about the future direction of Dutch mathematics education in primary school were broadly supported by the educational community and by parents (Cadot & Vroegindeweij, 1986; Ahlers, 1987).

Now, we are 20 years further on and we have to conclude that the reform movement—which took place without any intervention by the government—has indeed accomplished the intended change. Surprisingly, however, this shift in the goals of mathematics education and the corresponding change in the competence profile of students is considered now—particularly in the public arena—to be a drop in mathematics achievement in general. Apparently, written arithmetic is identified more with mathematical competence than number sense, mental calculation and estimation.

Hickendorff, Heiser, Van Putten, and Verhelst (2009) also zoomed in on the “worryingly large decrease”—as they call it—of Dutch primary school student achievement on written arithmetic and compared the students’ results on written division in 1997 with those in 2004. The data for this analysis came from the PPON studies. In the classification used in PPON, these problems belong to a category labeled in Dutch as “Bewerkingen” which can be translated as “Operations.” The problems contrast with the problems for mental calculation in which the students are not allowed to write down their calculations. The problems in the category “Bewerkingen” are supposed to be solved in a written way, which can be the digit-based traditional algorithm or a whole-number-based prestage of it. To distinguish these problems from the problems meant for mental calculation, we call the problems that were the focus of the study of Hickendorff et al. (2009): *written division problems*.

We are well aware of the fact that Hickendorff et al. (2009) did not have the intention to evaluate the Dutch reform of mathematics education in primary school over a number of years, but purposely restricted themselves to examining the change in achievement in written arithmetic. However, the point is whether this focus on only one aspect of students’ mathematical competence makes sense when a rearrangement in educational goals and transformation of teaching methods took place during these years. Instructional time is limited. If one subcompetence gets more attention, this automatically means that another sub-competence will get less, and consequently that the first competence improves at the expense of the latter.

Although Hickendorff et al. (2009) only focused on one mathematical subskill, they enriched their study remarkably by not only looking at the correctness of the students’ answers, but also including the students’ strategies in their analysis. The reason for doing so was that the strategies the students applied might well explain the decline in achievement. Therefore, Hickendorff et al. (2009) compared the results of the students who used a realistic strategy—which is the reform-based strategy—with the results of those who used a traditional algorithm.

The outcome of their analysis is that the students scored lower on the written division problems in 2004 than in 1997, but that

1. “[t]he effects of Realistic strategies and the Traditional algorithm did not differ significantly from each other in either 1997 or in 2004” (ibid.) and that
2. “[...] weak and strong students had as much success with the Traditional algorithm as with the Realistic strategies” (ibid.).

As far as we, the authors of this article, were involved in the design of the reform movement in the Netherlands or even responsible for the initiation of this movement, we can be contented with the result that using a realistic strategy instead of a traditional strategy does not influence the achievements of the students and that this is as true for the weak students as for the strong students in mathematics. However, as researchers with interest in the assessment of change in student achievement we are less satisfied, since Hickendorff et al. (2009) study has a number of substantive and methodological issues that question the validity of their findings.

In the following, we discuss large-scale assessment of change in student achievement and take Hickendorff et al. study (2009) as an example. First, we address the adequateness of the operationalization of constructs over time. In the case of the study of Hickendorff et al., this concerns the inclusion of all relevant item features of written division, the way the division problems are presented, and the classifications of strategies. Next, we discuss the crucial issue of time of assessment when assessing change in achievement. After that, we ask attention for the influence of the assessment format and covariates that might influence response behavior.

2. The Need of Adequate Operationalization of Constructs

2.1. *Generalizability of an Item Set Including Relevant Item Features*

In order to draw valid conclusions about students’ competences in a particular domain, it is necessary that researchers define a priori item features for the indicators which operationalize the construct. It has to be assured that these features are appropriately represented (with respect to the construct in mind) in the item set used. In assessing mathematics, i.e., when assessing mathematical operations such as division, it is important to include numbers of different sizes and different types of numbers (whole numbers and decimal numbers), problems with and without a remainder, as well as special anomalies in the algorithm, such as a zero in the quotient. Representativeness of item features is necessary to avoid that the intended construct is measured in a restricted way.

A first requirement when measuring a particular construct is that the items used evoke the mathematical operation that one intends to assess. Several of the division items included in the analysis by Hickendorff et al. (2009, see Table 1) are not typical problems, which today would be solved with written arithmetic. This point becomes a very serious problem when one has the intention to evaluate how the children’s ability in written arithmetic has changed over the years. For at least three out of the four items which were included in both the 1997 and the 2004 assessment, the answer would rather be found—at least nowadays—by mental calculation (or a strategy that is half-mental and half-written) rather than digit-based long division or a whole-number-based pre-stage of it. In the past, when children applied division algorithms more or less automatically without looking at the numbers involved, they would even carry out a long division for very simple numbers. Nowadays, the students are supposed to look critically at the numbers, which means that in the case of “easy” numbers they should switch to mental calculation. In Figure 1, it is shown how the items 7, 8, and 9 can be solved mentally.

As an example, we can take item 7 in which the students have to find the answer of: $872 \div 4$. This item is very suitable for mental calculation: $800 \div 4$ is 200, $40 \div 4 = 10$, and $32 \div 4 = 8$, all together 218. In 1997, 31% of the students still applied a long division method for this item, in 2004 only 8%. In the same period, the “no written work” strategy for this item increased from 41%

PSYCHOMETRIKA

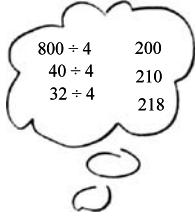
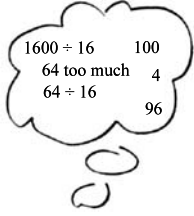
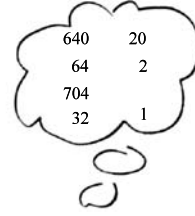
Item nr.	7	8	9
Division problem	$872 \div 4$	$1536 \div 16$	$736 \div 32$
Possible solution by mental calculation			

FIGURE 1.

How items used by Hickendorff et al. (2009) meant for assessing written arithmetic can be solved by mental calculation.

$ \begin{array}{r} 16 \overline{) 64800} \quad \quad 4 \\ \underline{64} \\ 0 \end{array} $	$ \begin{array}{r} 16 \overline{) 64800} \quad \quad 40 \\ \underline{64} \\ 08 \end{array} $	$ \begin{array}{r} 16 \overline{) 64800} \quad \quad 405 \\ \underline{64} \\ 080 \\ \underline{80} \\ 0 \end{array} $	$ \begin{array}{r} 16 \overline{) 64800} \quad \quad 4050 \\ \underline{64} \\ 080 \\ \underline{80} \\ 00 \\ \underline{0} \\ 0 \end{array} $	$ \begin{array}{r} 16 \overline{) 64800} \\ \underline{64000} \quad 4000x \\ 800 \\ \underline{800} \quad 50x \\ 0 \quad 4050 \end{array} $
--	--	---	---	---

FIGURE 2.

Division with a zero in the quotient solved through digit-based division (left) and whole-number-based division (right).

to 61%. If we assume that this means that more children applied a mental calculation strategy to solve this item, this would be completely in agreement with the expectations from the realistic approach to mathematics education. Moreover, we know that number sense over the years has increased and—to a lower degree—mental calculation and estimation as well. The increase in these subcompetences might have helped the students to decompose the numbers involved and recognize smart ways of mental calculation. The latter can explain why fewer children showed chunking on their scrap paper in 2004 than in 1997—Hickendorff et al. (2009) found a drop from 22% to 15% for the methods they called “realistic strategies.” More in general, the foregoing also clarifies why it is not a surprise that the “realistic strategy users” did not increase over time. With increased number sense, mental calculation and estimation, one does no longer need the written chunking strategy. The items that were initially intended for written calculation have now become items for mental calculation.

The better the set of problems includes all item features, the better the construct is measured, resulting in high construct validity. Among other things, this means that the item set contains division problems with a zero in the quotient. This is a very essential item feature and might have a large effect on the competence of students in relation with the strategy used. In the case of the study of Hickendorff et al. (2009), this item feature is underrepresented. Their set of items (see Hickendorff et al., 2009, Table 1) contains only one item ($64800 \div 16$) of that type. Problems like these are sensitive to mistakes, especially when they are solved through the long division algorithm, which is a digit-based division (see Figure 2). Consequently, not having those items

represented in the item set might result in an overestimate of the successfulness of the traditional algorithm.

As a matter of fact, the analysis by Hickendorff et al. (2009) could not deal with a sufficient variety of possible item characteristics because they only had 19 items. Analyzing item characteristics with explanatory IRT models like the linear logistic test models (or other model families) require a certain number of items to allow for generalization to an item universe. With 19 items, the validity might be challenged. Instead of investigating item characteristics, Hickendorff et al. (2009) summarized the strategy effect over all the items.

2.2. *Context Versus Bare Number Problems*

In mathematics, if not in any school subject, there is a difference between having achieved the “pure” ability and the application of this ability in a real situation. For example, being able to carry out a plain calculation is not the same as being able to use this skill to solve a problem in real life or to find an answer to a context problem in a test or textbook. Both abilities refer to a different construct. We should be aware of this difference when we operationalize a construct to be measured. If a construct is defined purely by bare number problems, then all items which are context problems measure a nuisance dimension. In other words, the aspect of converting the context problem into a bare number problem disturbs the construct. The reverse is the case when the intended construct has to cover application. Then bare number problems contaminate the measurement of this construct.

Instead of focusing either on a bare number construct or an application-like construct, it is also possible to define a multidimensional construct in which, for example, the ability to solve written division problems consisting of bare numbers is combined with the ability to solve written division problems which are presented as context problems. The resulting dimension is a weighted composite of these two subdimensions which weights are mainly defined by the number of items used in both subdimensions. For the purpose of evaluating educational outcomes, this can be a useful approach (Goldstein, 1979). In principle, this approach resemblances the definition of a (second order) formative construct (Edwards & Bagozzi, 2000) which means that the indicators are considered to form the construct.

These considerations refer especially to the definition of the construct in the study Hickendorff et al. (2009). They chose a formative construct. In itself, this choice is not a problem, but it does become one when the intention is to measure (the change in) applied strategies when solving written division problems. Context problems and bare number problems for assessing mathematical skills can trigger different strategies in students to solve these problems, even if they have the same mathematical structure. For example, solving 145–138 by adding on (when the problem is visually presented by means of two boys who are comparing their height) instead of by taking away 138 from 145 (when the problem is presented as a bare number problem) (Van den Heuvel-Panhuizen, 1996).

The set of 19 items used to assess the students’ ability in written division (see Hickendorff et al., 2009, see Table 1) is not very balanced with respect to item presentation. Most of the items are context problems; therefore, these items dominate the scale.¹ In total, there are only three bare number problems and in the four anchor items the bare number problems are not represented. The difficulty of having this preponderance of context problems is that it might influence strategy use. The contexts might have elicited a multiplying-on strategy or a repeated subtraction strategy rather than a long division strategy. Moreover, four of the context problems

¹This problem cannot be resolved with the use of latent variable models. Sijtsma (2006, p. 452) quotes: “I think that latent variables [...] are summaries of the data and nothing more...” In the same sense, Stenner, Burdick, and Stone (2008) claim that in the Rasch model, formative measures cannot be distinguished from reflective measures.

require that the children have to deal with a remainder in a context-dependent way. This, however, is a competence that should not be mixed up with the ability of carrying out a division calculation.

2.3. *Classifications and Interpretation of Strategies*

Especially in mathematics, where different strategies can both lead to a correct solution and indicate a different competence, level it makes sense to consider the applied strategy as a part of the construct intended to be assessed. Therefore, we think that it was a good idea of Hickendorff et al. (2009) to include strategies in their analyses to study substantial relations of correctness and strategy use.

Because any outcome of the statistical analyses depends on the classification of strategies, it is necessary to have categories for classifying strategies that are adequate with respect to the construct.

Unfortunately, the classification and interpretation of the strategies as used by Hickendorff et al. (2009) is somewhat doubtful. The category “traditional algorithm” is clear. It covers the digit-based algorithm for long division. The chunking and partitioning methods are labeled as “realistic strategies.” However, this label does not correspond to what should be named a realistic strategy. Indeed, it is true that the realistic approach to mathematics education uses a whole-number-based division strategy of chunking and repeated subtraction as an alternative for children who have difficulties with learning the most shortened way of long division, but if any particular strategy can be called a “realistic strategy,” then it is the flexible use of strategies that matches the problems involved. What is strongly emphasized in the realistic approach to mathematics education is that children adapt their strategy to the kind of calculation they have to do. So, in the case of $872 \div 4$, one may choose a mental division strategy, and in the case of $7839127 \div 12$, one may choose a written division strategy which can be the most shortened one (traditional long division) or a less shortened one (repeated subtraction with smaller or larger chunks). In other words, children who show this flexibility in strategy use can be called “realistic strategy users.” Therefore, the category of “no written working” is rather debatable. Assuming that a child immediately recognizes the 800, the 40 and the 32 in $872 \div 4$, and writes down 218, then this response would be classified as “no written working,” while the way the child solved the problem is completely in agreement with the realistic approach. In our view, it would have been better if Hickendorff et al. (2009) had used neutral terms such as “digit-based division,” “whole-number-based division,” “multiplying-on or repeated addition,” “just notating in-between steps,” and “no written traces” to classify the students’ responses. Such neutral terms would be more adequate, in particular, because Hickendorff et al. (2009) did not have the intention to evaluate the realistic reform in the Netherlands.

3. Time of Measurement Matters

What a student knows is changing continuously, which means that the point of measurement influences his or her performance: the student’s score at a certain time point shows the current state of his or her development. The outcome of an assessment of a group is the result of the average of these “incidental” scores of the students. This average necessarily consists of a considerable amount of noise, which unavoidably makes the measurement somewhat uncertain. However, when it comes to assessing change, this noise may play a disturbing role. If the performance trajectories of the students are not homogeneous (say, they do not have the same development slope), then the time of measurement matters. Next, we elaborate the influence of time of measurement by first looking at differences in group scores at different time points and then zooming in on the intraindividual level and discussing differences in individual performance trajectories and their influence on inter-individual cross-sectional comparisons (for similar discussions, see Molenaar, 2004).

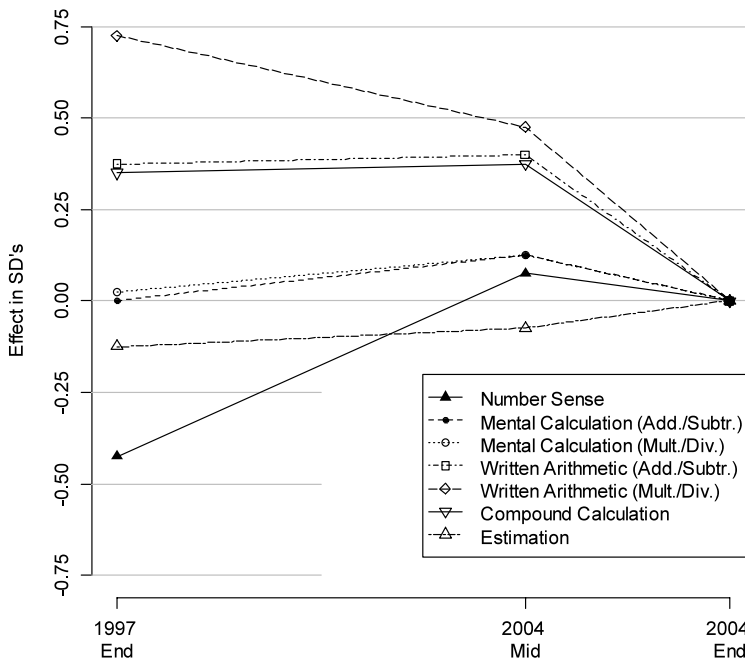


FIGURE 3.

Score change in mathematics achievement of Dutch primary school students from 1997 to 2004, including the midmeasurement.

3.1. Differences in Group Scores at Different Points in Time

Although one can conclude that in the Netherlands there is a decline of primary school students' performance on written arithmetic when the scores of 1997 are compared with those in 2004, a deeper look at available data shows that there is no steady downward movement. This deeper look is possible, because in the last PPON report (Janssen et al., 2005) CITO not only reported on the end of grade 6 scores, but also on the midyear scores for grade 6, and the scores of mid grade 5 and end grade 5.

These additional data show that when the 1997 scores on written arithmetic are compared with the mid grade 6 scores in 2004, the conclusion about the change in achievement is quite different (see Figure 3). For written multiplication and division (data taken from Janssen et al., 2005, p. 107), there is a drop of $\frac{3}{4}$ standard deviation from end grade 6 in 1997 to end grade 6 in 2004, while from end grade 6 in 1997 to mid grade 6 in 2004, there is just a small drop of $\frac{1}{4}$ standard deviation. For written addition and subtraction (ibid., p. 99), there is a decline of approximately 0.4 standard deviation between end 1997 and end 2004, whereas there is no difference between end grade 6 in 1997 and mid grade 6 in 2004. Almost the same trend is found for "compound calculation" (ibid., p. 115) including multistep problems in which the students have to combine addition, subtraction, multiplication, and division. For the subdomain of number sense (ibid., p. 55), the scores increased from end grade 6 in 1997 to mid grade 6 in 2004 by approximately $\frac{1}{2}$ standard deviation and remained more or less constant until end grade 6 in 2004. For mental calculations (ibid., p. 75 and p. 83), the mean student level seems to be constant between end grade 6 in 1997 to mid grade 6 in 2004, and also between the mid and the end grade 6 assessment in 2004. Estimation is more or less constant between end grade 6 in 1997 and mid grade 6 in 2004 and is the only competence that shows no decrease from the mid to the end assessment in 2004 (ibid., p. 91).

Unfortunately, we do not know what the difference would have been if the scores of mid grade 6 in 2004 had been compared with those of mid grade 6 in 1997, but making a firm statement about a “worryingly large decrease” from 1997 to 2004 is debatable. When one takes a point of measurement 4 months earlier, there is a much smaller effect. Furthermore, the data reported by Janssen et al. (2005) show that the decrease in the last part of the school year might not happen in every grade. For example, for written multiplication and division (ibid., 107), there is a decrease of about a $\frac{1}{2}$ standard deviation from mid grade 6 to end grade 6, while the opposite was found from mid grade 5 to end grade 5, where there was an increase of about 0.4 standard deviation.

What we can learn from the additional data from these other points of measurement (other than the measurement at the end grade 6) is that students do not develop in a monotone increase with respect to their mathematical competences. These so-called discontinuities in learning processes (Freudenthal, 1973, 1978a, 1978b) have consequences for *how* we assess the students—the problems must have a certain elasticity or stratification to cover variation in performance (Van den Heuvel-Panhuizen, 1996)—and have also consequences for *when* and *how often* the students are assessed.

A further point of attention is that there may even be differences in the learning pathway between subdomains: what is true for written arithmetic may not be so for number sense. The latter is more conceptual and may be less sensitive to practicing, while the latter is a skill and might be more influenced, for example, by doing less exercises in the last part of the final year in primary school. For monitoring and evaluating outcomes of education over several years, it is important to have a good image of how particular competences develop.

3.2. Differences in Individual Performance Trajectories

Every individual student has his or her own performance trajectory. This function is not necessarily increasing in a monotone way. If, for example, during a particular educational period the instruction in the competence that is measured was not optimal or had a lower impact on the student’s learning, then this function can be non-monotone. The performance trajectory can also be influenced by individual forgetting processes. This phenomenon has its consequences for determining an adequate time point for assessing students. In a large-scale assessment, every student is measured by a cross-sectional “snapshot” in time. If one wants to measure written arithmetic, it could be the case that the learning curve is increasing up to a particular time point, and after this point the curve decreases. Note that these points can differ from student to student. For illustrative purposes, such a situation is shown in Figure 4. Fictitious performance curves of three students belonging to two groups are depicted by either a solid or a dashed curve. The tops of the curves are marked with a circle or a triangle.

This picture shows that while the two groups of students have performance curves of the same shape, all are constantly shifted in time. The individual maximum performances do not differ in both groups. However, when we compare the scores at a particular measurement point, they do differ. When looking at the bold average curves of both groups, at measurement point 2, Group 1 (solid curves) performs strongly better than Group 2 (dashed curves). This is not the case for measurement point 1.

With respect to the results discussed in Section 3.1, Group 1 could be the 1997 cohort and Group 2 the 2004 cohort. Large differences at the end of grade 6 (measurement point 2) are reduced when one compares the groups at mid grade 6 (measurement point 1). To identify such individual performance curves, more than one time point for each student is necessary. Otherwise, when using several time points for measurement (say, three in a school year), a raw description of average performance curves and their changes over time can be made.

This is a useful approach to disentangle the vertical shifts (“true” score differences, which are the differences in the maximum values) and the horizontal shifts (differences in the locations

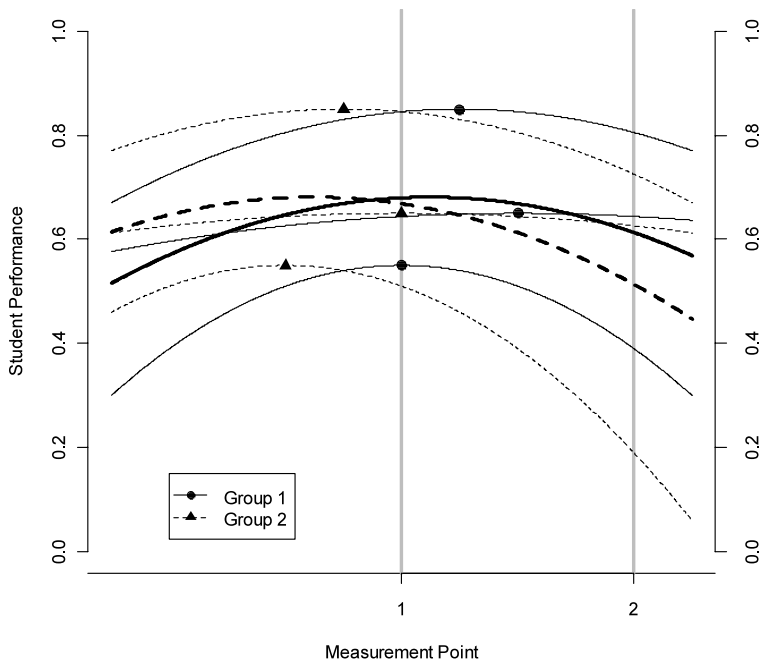


FIGURE 4.
Different performance trajectories (quadratic trend).

of the maximal values of average performance trajectories), whereas when there is only one cross-sectional assessment the observed score difference is confounded by these two factors. This approach of disentangling has been discussed in detail in the statistical analysis of functional data (Ramsay & Silverman, 2005).

An ideal situation is graphed in Figure 5. In both groups, the mean slopes of the linear performance remain the same over the time. Therefore, mean differences between the two groups are equal whatever measurement point is being used. However, it has to be argued whether such a situation is realistic in practice, especially when one compares different cohorts.

4. Format of Assessment Matters

Several studies have shown that the format of assessment matters (e.g., Danili & Reid, 2005; Caygill & Eley, 2001). When similar tasks are provided to students in a different format, remarkable differences in the students' responses show up. According to Danili and Reid (2005), this phenomenon raises questions about the validity of the formats of the assessment. One may question what different assessment formats are testing. A related question may be asked about how well a particular format is giving the right cue to elicit the behavior to be assessed.

In the case of the division problems from the PPO study which were used in the analysis of Hickendorff et al. (2009), one may wonder whether the test instruction "*In this arithmetic task, you can use the space next to each item for calculating the answer. You won't be needing scrap paper apart from this space.*" is strong enough to evoke written calculation and whether it is clear enough for the students that this test assesses whether they are able to carry out a written calculation. This question about how powerful this cue is to put the students on the track of written division, comes up when we compare the results found in the written format

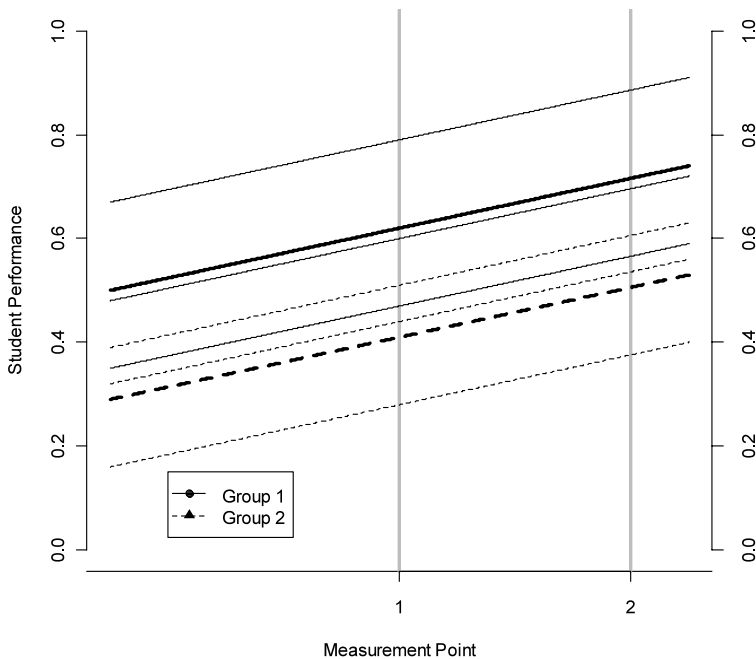


FIGURE 5.
Different performance trajectories (linear trend).

with those resulting from the individual interviews that were held parallel to the PPON 2004. These interviews were conducted with 140 students from 58 schools as part of the PPON study. This means that the composition and the mathematical level of the students and the moment of testing were the same as for the students who did the paper-and-pencil test (Van Putten & Hickendorff, 2006). The difference in results from the two test formats is remarkable. This is true for both the percentage of correct answers and the used strategies. Table 1 shows the findings with respect to two of the four anchor items. The data from the paper-and-pencil test were taken from Hickendorff et al. (2009) and those from the individual interviews from Janssen et al. (2005).

For item 9, the percentage of correct answers was 52% in the paper-and-pencil format, and 84% in the individual interviews. Item 10 moved from 29% to 60% correct answers. Although an individual administration makes problems in general less difficult for students, this difference of about 30% points is quite exceptional.

A closer look at the strategies shows that in the individual interviews, the students did not only use “realistic strategies” more often, but made more use of the “traditional algorithm” as well. Most noteworthy is that the category “no written working” is completely missing—or almost missing: Van Putten and Hickendorff (2006) report a frequency of 1%—in the individual interviews. The practical nonoccurrence of this response in this test format makes clear that the students are really able to write down their calculations and that this might have helped them to improve their performance. The finding that the “no written working” was minimized in the individual interviews can be seen as an indication that the prompt to write down the calculations is not the same in both assessment formats.

Although one may say that the paper-and-pencil format shows better what the students do spontaneously, this argument does not hold when one has the intention to assess the students’ competence in written arithmetic, which is not the same as assessing the students’ ability to find an answer—in one way or another—to division problems. Moreover, this argument is also not

TABLE 1.
Percentage strategy use and answers correct in two test formats.

Item 9	1997		2004	
736÷32 (in context)				
Strategy	Paper-and-pencil	Paper-and-pencil	Individual interview	
Traditional Algorithm	42	19	26	
Realistic ^a	24	33	71	
No Written Working	22	30	0	
Other	12	19	3	
Answer correct	71	52	84	
Item 10	1997		2004	
7849÷12 (in context) ^b				
Strategy	Paper-and-pencil	Paper-and-pencil	Individual interview	
Traditional Algorithm	41	19	27	
Realistic ^a	22	25	68	
No Written Working	17	35	0	
Other	20	21	5	
Answer correct	44	29	60	

^aThe “realistic” strategy as defined by Hickendorff et al. (2009) includes chunking and partitioning.

^bIn Table 1 of Hickendorff et al. (2009), the dummy version of this item is mentioned (9157÷14).

tenable if the problems used to assess written arithmetic are actually more suitable to be solved by mental arithmetic.

To make more valid conclusions about students’ competence in written division, additional data are needed about the two assessment formats of other problems in 2004. Moreover, additional data is necessary about the two assessment formats of the same problems in 1997.

The best way to assess whether the students can carry out a written division is to ask them explicitly to do this. Such approach is, for example, chosen the German mathematics test DEMAT (Gölitz, Roick, & Hasselhorn, 2006). In this test, the students are given a model of a long division followed by the instruction to solve the next problems in the same way. Another solution for getting more valid information about the students’ ability of written arithmetic and whether they master particular strategies, is to use Siegler’s and Lemaire’s (1997) Choice/No-Choice methodology. This solution is also suggested by Hickendorff et al. (2009).

Besides the specific strategy information about the two items for written division, the individual interviews connected to PPON revealed also important information about the change in strategies in general. These interviews have been conducted since 1987 and revealed that in many arithmetical subdomains there was an increase in the level of the strategies. According to Janssen et al. (2005), more advanced strategies were used in 2004 than in the earlier measurements.

5. The Necessity of Comparability of Assessments over Time

5.1. Test Design Issues

The comparability of two assessments becomes critical when test designs across assessments have been changed too much. In a recent critique of the PISA long-term design, Mazzeo and von Davier (2008) argue that stability in assessing trends is hampered by many factors; among other things, they mention design issues. They propose to use the same item clusters to avoid context

effects. Booklet effects can occur if items in particular booklets are more difficult than in others (in the situation that they are administered to the same population). Moreover, Mazzeo and von Davier (*ibid.*) argue that booklet effects decreases if focused designs instead of mixed designs are used. In a mixed design, a booklet contains items from different domains, while in a focused design only one domain is included. In most cases, these kinds of anomalies lead to violations of the usual assumptions of IRT models. Every adjustment to IRT models rests on additional assumptions that have to be defended when reporting about results (like when reporting about trends in PISA). According to Mazzeo and von Davier (*ibid.*), a better approach would be to use exact the same booklets in successive assessments. By doing this, construct irrelevant effects can be minimized.

Mazzeo et al.'s remarks also touch the data used by Hickendorff et al. (2009). As is explained by Janssen et al. (2005), in 2004 a different test design was used than in 1997. Whereas in 1997, a separate test booklet was used for every subdomain, in 2004, for part of the subdomains, the focused test design was changed into a mixed test design. This means that the problems on written division were distributed over a number of booklets containing several topics.

5.2. *Linking Issues*

The precision of trend results depends on how stable the assessment of a trend can be. In theory, using one item as a link item is sufficient. In successive assessments (especially in different contexts) items will change their difficulty which leads to item parameter variation (DIF or item drift). Then a linking error for the comparison is introduced. This linking error is reduced if a large number of items occur in both assessments.

Because of statistical and validity issues it is necessary to administer a minimum number of items (say, 10 or 20 items—it depends how “broad” the measured construct is) in both assessments as anchor items. These items should be placed in the same contexts to avoid artificially changing item difficulties so that item drift is as minimal as possible. In the study of Hickendorff et al. (2009), 19 items on written division problems were used in 1997 and 2004, but only four items are anchor items, e.g., were used in both studies. From a linking perspective, these few items can lead to high linking errors. In addition, any proposition about change on one scale rests on the appropriateness of these anchor items. If there are only four anchor items, the generalizability of the findings can be questioned. Of course, the link between 1997 and 2004 could be seen stronger than found, if we assume that there are item clones (i.e., parallel variants of items with equal difficulty at one assessment point), but this assumption seems to be more difficult to realize for context embedded items. In addition, differential item functioning (DIF) between assessments (item drift) can occur for “wrongly selected” linking items. If all anchor items show the same DIF direction, estimated mean scale changes are prone to bias in this direction. Moreover, differences in opportunity-to-learn, due to using different textbook or training on published example items, can lead to DIF.

In addition, using only a few items in a few booklets as anchor items, factors that affect item difficulty like position effects, context effects (other items surrounding the item under study) or booklet effects can have a high impact. The situation becomes more complicated in the study of Hickendorff et al. (2009), because statements about students' competence in written division can only be done based on context items, which—because of other item ingredients—might measure other competences as well and cannot be seen as “pure items” for written arithmetic. As a consequence, a change for pure written arithmetic could be confounded with a change in other competences such as problem solving in contexts. We do not claim that this study is affected by all the problems mentioned here, but the probability that these factors can come into play increases with a weakly linked design.

6. Covariates in Item Response Models

When assessing change in student achievement, it is very important to include covariates (item covariates, person covariates and items x person covariates) in the item response models that can explain differences in achievement. In this respect, one may think of a change in variables, such as students' reading ability, their attitude to mathematics and their degree of mathematics anxiety, and their opportunity to learn ("OTL", Husén, 1967) particular subject matter content and processes. We take OTL as an example. Several studies have shown that there is a strong correlation between what is taught to students and their achievement (e.g., Floden, 2002; Haggarty & Pepin, 2002; Törnroos, 2005).

Of course, we realize that an investigation into the implemented curriculum in the Netherlands would go beyond the study carried out by Hickendorff et al. (2009). Nevertheless, we think that for a good understanding of the results of their study the issue of what the students have been taught should be considered. Roughly speaking there are three main methods to measure opportunity-to-learn: using teacher reports, analysis of curriculum documents and textbook series, and classroom observations. In the Netherlands, only sparse examples from the first two are available to inform about what is taught in primary school mathematics education in the domain of written arithmetic, in particular written division.

A recent example of the latter is a textbook analysis carried out by Treffers (2008) in which the two textbook series which have the largest market share (respectively, 25% and 40%) have been compared with respect to how they have outlined the teaching of written arithmetic. This analysis revealed that one textbook series has a very clear learning-teaching trajectory reflecting a progressive shortening of the written calculation procedures toward the most curtailed ways of written arithmetic, while in the other textbook series, the students are stimulated all along to choose their own method. The difference between these two textbook series manifests itself most sharply when, for example, a comparison is made of the amount of problems that both textbook series devote to the most simple form of written—i.e., algorithmic digit-based—multiplication. These are the multiplication problems with a one-digit multiplier and a multi-digit multiplicand. The first textbooks series includes around 500 to 600 of these exercises, while the second textbook series contains barely 100. This means that the students using these two textbook series are completely differently prepared when they come to written division in which the basic ability of written multiplication plays a crucial role.

It will be clear that in our view the above means that just quoting the statement of Janssen et al. (2005) that "Dutch primary schools have almost uniformly adopted mathematics textbooks based on the principles of RME" (Hickendorff et al., 2009) is not sufficient if one takes the importance of the opportunity-to-learn seriously.

Further information about the opportunity-to-learn can also be derived from the results of the teacher survey included in the PPON report (Janssen et al., 2005). This survey among a large sample of teachers disclosed that 17% of the teachers teach digit-based long division, 58% whole-number-based division and 24% do both. We suppose that it was not possible for Hickendorff et al. (2009) to relate the strategies the students used to what strategy their teachers taught them, as it would have been very interesting to know the result of this analysis.

Another pressing question that still is waiting for an answer is what other factors may cause change in mathematics achievement, as the PPON researchers were puzzled that for the domain of number, including written arithmetic, there is a negative year effect when the results of 1992 and 1997 are compared with those of 2004, while at the same time the (RME-based) textbooks series have a positive effect on the achievements of the students (Janssen et al., 2005).

7. Concluding Remarks

Large-scale assessment of change in student achievement over a number of years is an incredibly difficult job to do. The paper by Hickendorff et al. (2009) provides suitable approaches on how to deal with many of the problems that typically occur in this field. They took into account different solution strategies, identified those strategies by means of latent class analysis and presented convincing IRT-analyses predicting achievement change with different strategies. However, from our point of view, many additional variables are at play at different levels of influence. Moreover, an additional complicating factor comes up when the change that one wants to investigate, has occurred within a reform context where its implementation—for instance, because the freedom of education is regulated by law—is neither institutionalized, nor supported by a compulsory professional training of in-service teachers. As a consequence, such a reform may develop in a number of different directions, and may result in a far from unified approach to teaching. Moreover, many misconceptions of what such a reform means can arise.

In the case of the realistic approach, one such fallacy is the idea this approach is against digit-based long division. Furthermore, it is thought that this division method is completely different from whole-number-based division. These and other erroneous beliefs that might guide teachers' way of teaching makes it extremely difficult to make valid conclusions about the change in student achievement in written division. Large-scale assessment of change in student achievement has to find ways to deal with all these complicating factors that operate on what we see as the final result. Moreover, all the methodological problems that have been raised in the context of PISA (e.g., Mazzeo & von Davier, 2008) clearly demonstrate that we are still in the beginning of understanding all requirements that have to be fulfilled so that we can draw valid conclusions about changes in achievement from large-scale assessment.

More in general, assessment of change in student achievement has to disentangle the multifaceted learning processes that take place within complex educational settings, the latitude of educational policy, and societal forces. A good understanding of all these ingredients, as well as dealing with all the methodological issues, form absolute requirements for undertaking such an assessment. Therefore, we think such a job requires a joint research enterprise of didacticians and psychometricians.

References

- Ahlers, J. (1987). Grote eensgezindheid over basisonderwijs. Onderzoek onder leraren en ouders [Large consensus about primary education. A survey among teachers and parents]. *School*, 15(4), 5–10.
- Cadot, J., & Vroegindewij, D. (1986). *10 voor de basisvorming onderzocht [Ten points for basic education in mathematics investigated]*. Utrecht University, OW & OC: Utrecht.
- Caygill, R., & Eley, L. (2001). *Evidence about the effects of assessment task format on student achievement*. Paper presented at the Annual Conference of the British Educational Research Association, University of Leeds, England, September 13–15, 2001. Retrieved from <http://www.leeds.ac.uk/educol/documents/00001841.htm>.
- Danili, E., & Reid, N. (2005). Assessment formats: do they make a difference? *Chemistry Education Research and Practice*, 6(4), 204–212.
- Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and their measures. *Psychological Methods*, 5(2), 155–174.
- Floden, R. E. (2002). The measurement of opportunity to learn. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 231–266). Washington: National Academy Press.
- Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht: Reidel.
- Freudenthal, H. (1978a). *Weeding and sowing. Preface to a science of mathematical education*. Dordrecht: Reidel.
- Freudenthal, H. (1978b). Cognitieve ontwikkeling—kinderen geobserveerd [Cognitive development—observing children]. In *Provinciaals Utrechts Genootschap, Jaarverslag 1977* (pp. 8–18)
- Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 5, 211–220.
- Gölitz, D., Roick, T., & Hasselhorn, M. (2006). *DEMAT 4: Deutscher Mathematiktest für vierte Klassen [DEMAT 4: German mathematics test for grade 4]*. Göttingen: Hogrefe.
- Haggarty, L., & Pepin, B. (2002). An investigation of mathematics textbooks and their use in English, French and German classrooms: Who gets an opportunity to learn what?. *British Educational Research Journal*, 28(4), 567–590.

- Hickendorff, M., Heiser, W. J., Van Putten, C. M., & Verhelst, N. D. (2009). Solution strategies and achievement in Dutch complex arithmetic: Latent variable modeling of change. *Psychometrika*, *74*(2), doi:10.1007/s11336-008-9074-z
- Husén, T. (1967). *International study of achievement in mathematics: A comparison of twelve countries* (Vol. II). New York: Wiley.
- Janssen, J., Van der Schoot, F., & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 4* [Fourth assessment of mathematics education at the end of primary school]. Arnhem: CITO.
- Mazzeo, J., & von Davier, M. (2008). *Review of the programme for international student assessment (PISA) test design: Recommendations for fostering stability in assessment results* (OECD Education Working Papers) (EDU/PISA/GB(2008)28). Paris: OECD
- Molenaar, P. C. M. (2004). A manifesto on Psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218.
- Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*. New York: Springer.
- Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General*, *126*(1), 71–92.
- Sijtsma, K. (2006). Psychometrics in psychological research: Role model or partner in science. *Psychometrika*, *71*, 451–455.
- Stenner, A. J., Burdick, D. S., & Stone, M. H. (2008). Formative and reflective models: Can a Rasch analysis tell the difference?. *Rasch Measurement Transactions*, *22*, 1152–1153.
- Törnroos, J. (2005). Mathematical textbooks, opportunity to learn and student achievement. *Studies in Educational Evaluation*, *31*(4), 315–327.
- Treffers, A., & De Moor, E. (1984). *10 voor de basisvorming rekenen/wiskunde* [Ten points for basic education in mathematics]. Utrecht: Utrecht University, OW&OC.
- Treffers, A. (2008). *Comparing WIG's en PLUSPUNT's teaching of written arithmetic* (Unpublished manuscript). Utrecht: Utrecht University, Freudenthal Institute for Science and Mathematics Education.
- Van der Schoot, F. (2008). *Onderwijs op peil? Een samenvattend overzicht van 20 jaar PPON* [A summary overview of 20 years of national assessments of the level of education]. Arnhem: CITO.
- Van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education*. Utrecht: CD-β Press/Freudenthal Institute, Utrecht University.
- Van Putten, C. M., & Hickendorff, M. (2006). Strategieën van leerlingen bij het beantwoorden van deelopgaven in de periodieke peilingen aan het eind van de basisschool van 2004 en 1997 [Students' strategies when solving division problems in the PPON test end primary school 2004 and 1997]. *Reken-wiskundeonderwijs: onderzoek, ontwikkeling, praktijk*, *25*(2), 16–25.

Manuscript Received: 29 SEP 2008

Final Version Received: 23 DEC 2008