The Math War between traditional and "realistic" mathematics education and its research. An analysis in institutional economics on research on education in arithmetic and algebra, with a focus on long term memory of pupils and using a causal model for valid testing on competence

Colignatus, Thomas

Samuel van Houten Genootschap

4 September 2018

The Math War between traditional and “realistic” mathematics education and its research. An analysis in institutional economics on research on education in arithmetic and algebra, with a focus on long term memory of pupils and using a causal model for valid testing on competence

Thomas Colignatus
September 4 2018

Journal of Economic Literature: JEL
P16 Political economy
I20 General education
D02 Institutions: Design, Formation, Operations, and Impact
O17 Formal and Informal Sectors • Shadow Economy • Institutional Arrangements

American Mathematical Society: MSC2010
00A35 Methodology and didactics
97-XX Mathematics education
97F02 Arithmetic, number theory ; Research exposition
97G70 Analytic geometry. Vector algebra
97H20 Elementary algebra

14,600 words

Abstract

Institutional economics investigates how institutions affect empirical events. The term “institution” can be taken widely, and may also represent engrained mental conceptions by organised groups of actors. There is a curious but counterproductive combination of three groups also at universities in Holland w.r.t. research on education in arithmetic and algebra: (1) adherents of "realistic" mathematics education, an ideology that compares to astrology or homeopathy, (2) traditional mathematicians, who have no expertise on the empirical science of didactics of mathematics either, (3) psychometricians, who look at statistical data but who have no expertise on the empirical science of didactics of mathematics either. This combination needs deconstruction and the present paper focuses on (3), though with influence from (1) and (2). Some psychometricians seem to have a sound dislike of both the ideologues from (1) and the discussion between (1) and (2), but they are less aware that (2) are ideologues too. Some psychometricians also throw away the child with the bathwater by disregarding (4) the proper science of didactics of mathematics. Measuring competence in arithmetic and algebra requires consideration of long term memory of students. What you learn in elementary school tends to stay with you for the rest of your life. What you learn in highschool has the property of “use it or lose it”. Algebra in highschool requires competence in the traditional algorithms of arithmetic, best learned in elementary school. “Realistic” mathematics education has reduced the competence of students at elementary school which affects them not only for algebra in highschool but also for the rest of their lives in both arithmetic and algebra. Inadequate testing by psychometricians allows this detrimental state to continue. The paper presents a causal model that identifies the engrained mental conceptions by psychometricians and where they would have to accept insights from didactics of mathematics. There is also a role for the Dutch Academy of Sciences KNAW that supported an inadequate report in 2009.

1. Introduction

1.1. The issue: education in arithmetic and algebra

Primary education has a window of opportunity.

- What you learn in elementary school tends to stay with you for the rest of your life.
- What you learn in highschool has the property of "use it or lose it".

We now look at arithmetic and algebra:

- Above two properties hold. Learning arithmetic in highschool comes with the property of "use it or lose it".
- Algebra at highschool requires pre-algebra training at elementary school on the algorithms of arithmetic. If you don't properly learn how to manipulate 1 / 2 + 1 / 3 or 2^H + 3^H at an early age then you will tend to fail on 1 / a + 1 / b or a^H + b^H at a later age (using H = -1, see [1]).
- If arithmetic at elementary school relies on the calculator or trial and error, then this will be your standard on arithmetic, while the window of opportunity on algebra closes. Highschool may try remedial teaching on arithmetic but your level of algebra will tend to remain low.
- For example: The teaching method of "equivalent ratios" using tables only [2] is called "pre-algebra" but might also be perused as "never-algebra". Proper didactics requires integration of text, formula, table and graph.

Let us look how these phenomena are dealt with by mathematics education research (MER) and policy making. I already discussed main aspects in *Elegance with Substance* (2009, 2015), also see its website, but now we look at the window of opportunity for arithmetic and algebra occurring in primary education. See the preface of *A child wants nice and no mean numbers* (2015, 2018) for my lack of expertise on primary education.

The situation in Holland could be interesting to the world (see the AAAS Project 2061 [3]). Holland is a middle sized country of 17 million people with data collection for the population of students (PPON) and not only samples (TIMSS). The population and education characteristics are not too heterogenous. Important is also that the "reform in mathematics education" in the whole world had a key impulse from Hans Freudenthal (1905-1990) from Utrecht University, to the extent that ICMI now features a Freudenthal Medal, see also Colignatus (2014, 2015).

For the international context, a common reference is to Slavin & Lake (2008). Their p445: "More research is needed on all of these programs, but the evidence to date suggests a surprising conclusion that despite all the heated debates about the content of mathematics, there is limited high-quality evidence supporting differential effects of different math curricula."

1.2. Institutional economics and the math war in Holland

In economics, there is the branch of "institutional economics" that investigates how institutions can affect empirical events. The term "institution" can be taken widely, and may also represent engrained mental conceptions or ideologies. [4] Below we will mention some formal institutions that apply here but it appears that developments are more dominated by such engrained mental conceptions.

---

[1] https://doi.org/10.5281/zenodo.1251686
[2] https://www.khanacademy.org/math/pre-algebra/pre-algebra-ratios-rates/pre-algebra-visualize-ratios/e/solving-ratio-problems-with-tables
[3] https://www.aaas.org/program/project2061/about
[4] https://www.cambridge.org/core/journals/journal-of-institutional-economics/article/what-is-an-institution/3675101CE15BE2A7681CD5783C01F6D0

In Holland there is a "math war" [5] [6] that caused the Dutch Academy of Sciences KNAW to set up a committee, that produced the KNAW (2009) report. [7] This math war provides context to our issue, and it must be discussed to prevent confusion about what what this paper achieves.

1. The Dutch math war is between "traditional mathematics education" (TME) and "realistic mathematics education" (RME) a.k.a. "reform mathematics", as proposed by the Freudenthal Head in the Clouds Realistic Mathematics Institute (FHCRMI). [8]
2. My position is the third approach, consisting of scientific research, with re-engineering of mathematics education. [9]
3. The problem with TME and RME is that they derive from mathematicians trained on abstract thought who have little grasp of empirical research.
4. Both TME and RME have delegated empirical research to psychometricians, often at CITO. The subsequent problem is that psychometricians have no training in didactics of mathematics and mathematics education research (MER), whence such psychometric research runs the risk of invalidity. (See the present paper.)
5. In empirical science, when there are competing paradigms, then researchers set up a distinguishing experiment that shows which paradigm provides the best explanation. Adherents of TME and RME did not do so. However, a critical look at the available evidence would provide for such a decision, see Colignatus (2015c). [10] At issue is not which paradigm would be "right". There are useful ideas in both TME and RME, and it depends upon time and place what is most relevant, often decided by the teacher. At issue is to get rid of the blinding effect of ideology. If the distinguishing experiment shows that the TME (RME) textbook is best, then it can provide the baseline, and RME (TME) alternatives can be tested on a case by case basis.
6. Policy makers interfered and increased the chaos. Holland observed a reduction of competence in arithmetic and math, and TME claimed that this was being caused by RME. The minister of education imposed a separate test on arithmetic ("Rekentoets") as part of the highschool diploma. [11] While the problem was being caused at elementary school, requiring the re-training of 140,000 elementary school teachers, the minister approached the issue as end-of-pipe and put the burden on perhaps 12,000 math teachers in highschool. This neglected that incompetence in arithmetic at primary school would mentally maim students for algebra for highschool and the rest of their lives. (Students unable to do algebra are transferred to vocational schools, and would not be observed at pre-university schools.) (In 2018 the new state secretary adopted a new format for testing on competence, but I haven't had time to look into this.)
7. Because of the national debate a publisher created a textbook that uses "the best of TME and RME". They did so without scientific research to back this up, and without using a distinguishing experiment. This new mixture tends to make it more difficult to get such an experiment.

Thus we have two warring factions on the field, one playing soccer and the other playing American football, with an arbiter from basketball, and with the public throwing darts onto the field. My research hopes to help clean up the mess, while also hoping that others will be grateful for the clarity.

1.3. Formal and informal institutional setting

There is a large list of institutions for our issue. [12] [13] Key ones are:

[5] https://en.wikipedia.org/wiki/Math_wars (a portal and no source)
[6] https://www.theglobeandmail.com/opinion/article-in-the-ongoing-math-wars-both-sides-have-a-point/
[7] https://www.knaw.nl/nl/actueel/publicaties/rekenonderwijs-op-de-basisschool
[8] https://boycottholland.wordpress.com/2016/01/24/graphical-displays-about-the-math-war/
[9] https://zenodo.org/communities/re-engineering-math-ed/about/
[10] http://www.wiskundebrief.nl/721.htm#5
[11] http://www.wiskundebrief.nl/512.htm#1
[12] https://boycottholland.wordpress.com/2015/10/31/the-power-void-in-mathematics-education/
[13] See also the Presmeg chart at https://boycottholland.wordpress.com/2015/10/15/pierre-van-hiele-and-annie-selden/

- Onderwijsraad (Education Council), an advisory body for the minister of education [14]
- Inspectie voor het Onderwijs (IvhO) (Inspectorate for Education) [15]
- CITO, that provides for tests at the end of primary education [16]
- National board for education research (NRO) [17]
- Association of education researchers (VOR), commonly from the universities [18]
- Teachers and educators of teachers, [19] association of teachers of mathematics (NVvW) [20] and association on arithmetic (NVORWO) [21]
- Publishers

While TME was the original standard in the 1960s, the takeover by RME was gradual. Dutch elementary school teachers started adopting RME and at some point the Inspectorate pushed for it. By 2009, all Dutch primary school textbooks used the RME method. (By comparison, the USA still has variety in TME and RME, see below.)

The distinguishing experiment between TME and RME has these aspects.

- The main aspect consists of pure logic. Preparation for algebra requires command of the traditional algorithms for arithmetic. Since RME spends much less attention on those (and aspires at their "guided reinvention" which is merely a hope and not proven), we can expect that RME performs less well on those algorithms. In Holland, this logic is not understood. TME has been singularly ineffective in bringing this logic into attention.
- The statistical aspect consists of the actual tests at the end of primary education, administered by CITO, with application of psychometric techniques and diagnostics. The focus of a group of education researchers has shifted to statistical testing.
- Let us use an analogy to compare logic and statistics. In 1950 there were no actual (statistical) observations about the other side of the Moon. A statistician could have hold that one can't infer the existence of this other side because statistical evidence was lacking. Hopefully such statistician would not defy the logic by physics. The relevance of statistics for TME and RME is only for the particular value of the effect size, graduated by the talents of pupils. Perhaps some students will never learn algebra and then are better served by a good command of the calculator.

The statistical aspect not only defies the first element of logic, but shows two other illogical phenomena.

(1) KNAW (2009) has documented that the shift from TME to RME has not been supported by statistical testing. Thus, the shift towards statistics did not come along with this notion of rigour. (The same happened in the USA, see below.) The CITO tests have shifted over time in favour of RME, with the use of the calculator or trial and error, but this shift itself was not corroborated by tests.
(2) KNAW (2009) supports teacher experience but does not investigate whether teacher experience was the cause for the shift from TME to RME, thus without such (perceived) need for statistical testing. The main cause may still be ideology (and the argument of authority of Hans Freudenthal).

I cannot avoid the conclusion that ideology has had a strong influence. Normally there would be strict rules on experimenting on humans. When there are two methods TME and RME, then you are supposed to develop a distinguishing experiment. When one method is shown to be superior then you abort the experiment and switch all subjects to the better method. (This holds per topic and may be extended to paradigms.) Curiously, TME were not able to convince the education community by merely pointing to the logic of the argument. Somehow,

[14] https://www.onderwijsraad.nl/english/item34
[15] https://www.onderwijsinspectie.nl/over-ons
[16] http://www.cito.nl/
[17] https://www.nwo.nl/over-nwo/organisatie/nwo-onderdelen/nro
[18] https://www.vorsite.nl/en/content/about-netherlands-educational-research-association
[19] http://www.lerarenopleider.nl/velon/wp-content/uploads/2016/12/K2_6KempenDietzeCoupe.pdf
[20] https://nvvw.nl/
[21] http://www.nvorwo.nl/

statistical testing by CITO started weighing in. The use of statistics, adopting some standard out of thin air, allowed a distinction between kids performing well and kids performing less, and who oh who was to argue that “well” wasn’t “enough”, or that the lesser performing kids could do better on the other method ?

Jan van de Craats is a professor of mathematics (now retired) without a degree for teaching at elementary school. Since about 2005 he started defending TME. He created the SGR foundation and got support also from some researchers who were later appointed in the Education Council (Onderwijsraad). Van de Craats was invited to participate in a committee that identified levels of competence (comparable to the US Common Core). [22] They identified fundamental and target levels (abbreviated as F and S). In this letter [23] Van de Craats states that the committee intended that the targets be adopted, while the Ministry of Education embraced only the fundamental levels. By this move, the lower level of competence became the new official level in Holland. There is now less need of criticism on the Ministry that the official level is not attained.

Van de Craats and his SGR have a strong argument because of the logic mentioned above, that should be sufficient to reject RME. However, in advocating TME they basically have a position in ideology because they lack expertise for education at primary school. [24] It should be qualified teachers and researchers who should make that decision. Van de Craats and his SGR supported the creation of a new TME textbook “Reken Zeker” (Noordhoff), that was introduced in 2010. Some 20 schools started with it. This textbook was written by elementary school teachers Piet Terpstra and Arjen de Vries, [25] and thus satisfied the criterion that it was backed by their degrees and experience. It still came without scientific support and testing that we would like to see nowadays. [26] [27] SGR claims: “Their [textbook “Reken zeker”] combines the best of the two worlds of traditional [arithmetic] and realistic [arithmetic]”, [28] without explaining how this “best” combination has been corroborated.

In 2015 I suggested the following idea to CITO [29] [30] and Dutch Parliament. [31] In 2016, the pupils taught with “Reken Zeker” would finish their primary education. Thus, if CITO would keep their tests apart, and perhaps test them additionally in comparison with a random selection of other kids that used the prevailing RME methods, then there would be a (natural) distinguishing experiment with adequate test results. Obviously, the school teams that opted for “Reken Zeker” would be motivated for TME and thus we should require a larger difference in success to warrant its claim on superiority.

In its reply, CITO shifted responsibility to the Inspectorate of Education (IvhO). [32] [33] In this reply, CITO abused the distinction between principal and agent. Indeed IvhO was the new

[22] http://www.steunpunttaalenrekenenvo.nl/sites/default/files/Over%20de%20drempels%20met%20rekenen.pdf
[23] https://staff.science.uva.nl/j.vandecraats/Mails_aan_Victor.pdf
[24] http://thomascool.eu/Papers/AardigeGetallen/2015-09-23-SGR-deugt-ook-al-niet.html
[25] http://www.goedrekenonderwijs.nl/wp-content/uploads/Interview-Reken-Zeker-1.pdf
[26] http://www.few.vu.nl/~jhulshof/2011bib2.pdf
[27] http://benwilbrink.nl/projecten/reken_zeker.htm
[28] Dutch: “Hun methode “Reken zeker” combineert het beste uit de twee werelden van traditioneel rekenen en realistisch rekenen." http://www.goedrekenonderwijs.nl/reken-zeker/ Google Translate tends to translate “rekenen” as “calculation” while “arithmetic” would be better here.
[29] http://thomascool.eu/Papers/AardigeGetallen/2015-10-18-Aan-CITO.html
[30] http://thomascool.eu/Papers/AardigeGetallen/2015-10-18-Tweede-brief-aan-CITO.html
[31] http://thomascool.eu/Papers/AardigeGetallen/2015-10-17-Aan-TK-commissie-OCW.html
[32] Google Translate 2018: “Through the PPON research we have provided insight in the past into the management of various domains within arithmetic. This research is no longer under the responsibility of Cito since 2015, but below that of the Inspectorate of Education. I want to refer you to that.”
Dutch: At 2015-10-15, Strijp wrote:
Geachte heer Cool,
Bedankt voor deze en uw eerdere uitgebreide email.
Doormiddel van het PPON onderzoek hebben wij in het verleden inzicht gegeven in de beheersing van verschillende domeinen binnen rekenen. Dit onderzoek valt sinds 2015 niet meer onder de verantwoordelijkheid van Cito maar onder die van de Inspectie van het onderwijs. Daar wil ik u dan ook naar verwijzen.
Ik hoop uw hiermee voldoende informatie te hebben gegeven.

principal: it may not do research itself but contracts external researchers or consortia. [34] Instead CITO should have taken research responsibility, by supporting my suggestion at IvhO. Nevertheless I wrote to IvhO, my letter officially recorded as number M0155149, [35] and also published my suggestion in a newsletter for teachers of mathematics, as Colignatus (2015c). [36] In its reply, the Inspectorate rejected its role and reponsibility in this, by interpreting my suggestion as if this would only concern a scientific experiment whether TME or RME would be right. [37] I protested that this was a false interpretation, and that the task of protecting children lies with IvhO (and not merely NRO). This protest at IvhO did not receive a reply. By January 2016 I looked deeper at the role of psychometricians, also at CITO. [38] I approached NRO only in 2016, with the practical point that my kind of research was excluded by their choice of criteria. [39] [40] I also informed the Dutch education researchers (VOR) about the paradigm shift w.r.t. mathematics education research (MER). [41] [42]

There may be reason to regard 2016 as a crucial year to test. The authors of "Reken Zeker" have retired and there was some rumour that the textbook might be stopped. Yet in 2018 the textbook is still available and it is unclear who took over. [43] Since the KNAW 2009 report, RME textbooks and teachers have started including more TME elements, though in unknown ways, making it less clear what "real" RME is, and making it more difficult to arrive at a distinction. Graduation year 2016 would be the least untainted one.

In 2018 the Inspectorate started an evaluation targetted to show results in 2020/21. [44]

> Google Translate 2018: "An investigation will be conducted into the cause of the declining level of mathematics and mathematics education in the Netherlands. The research was announced in the annual work plan 2018 of the Education Inspectorate. This reports

---

Met vriendelijke groet,
Ineke Strijp

[33] https://www.onderwijsinspectie.nl/onderwerpen/peil-onderwijs

[34] Google Translate 2018: "Since 2014, the Inspectorate of Education has been in charge of the surveys under the name Peil.onderwijs. The surveys have been launched via the NRO since 2016." https://www.nro.nl/onderzoeksprojecten/peil-onderwijs/ For example, for 2017, IvhO / NRO contracted Marian Hickendorff at Leiden for another review study. https://www.nro.nl/nro-projecten-vinden/?projectid=405-17-920-rekenen%20op%20de%20basisschool

[35] http://thomascool.eu/Papers/AardigeGetallen/2015-11-23-Het-rekenexperiment-op-kinderen-moet-en-kan-stoppen.html

[36] http://www.wiskundebrief.nl/721.htm#5

[37] Google Translate 2018: "It is not the duty of the inspectorate to settle the discussion between supporters of the various [arithmetic] methods. Thank you for your interest. Perhaps there is interest in the National Education Research Foundation (NRO)"
Dutch: Date: Fri, 18 Dec 2015
From: Loket Onderwijsinspectie
To: Thomas Cool / Thomas Colignatus
Subject: M0155149 Memo: "Het reken-experiment op kinderen moet en kan stoppen"
Geachte heer Cool,
Het is niet de taak van de inspectie de discussie tussen aanhangers van de diverse rekenmethodieken te beslechten.
Dank voor uw interesse. Misschien is er belangstelling voor bij de Nationaal Regieorgaan Onderwijsonderzoek (NRO)
Met vriendelijke groet,
[XYZ]
Loket Onderwijsinspectie
.....
Inspectie van het Onderwijs
Ministerie van Onderwijs, Cultuur & Wetenschap
www.onderwijsinspectie.nl

[38] http://thomascool.eu/Papers/AardigeGetallen/2016-01-31-Enkele-emails-rekentoets-psychometrie.pdf

[39] http://thomascool.eu/Papers/Math/2016-04-15-Letter-to-NRO.pdf

[40] http://thomascool.eu/Papers/Math/2016-07-12-Second-Letter-to-NRO.pdf

[41] http://thomascool.eu/Papers/Math/2016-05-09-Letter-to-VOR-and-Trainers-of-teachers.pdf

[42] http://thomascool.eu/Papers/Math/2016-07-15-Second-letter-to-VOR-and-Trainers-of-teachers.pdf

[43] https://www.noordhoffuitgevers.nl/basisonderwijs

[44] https://www.onderwijsinspectie.nl/actueel/nieuws/2018/07/10/onderzoeken-rekenen-wiskunde-en-schrijfvaardigheid-voor-peil.onderwijs-gestart

> NU.nl. [...] international research (TIMSS) showed last year that 99 percent of primary school students in [grade 4] in the Netherlands master the basic skills in the field of arithmetic. The basic level is perfectly fine. At the same time, relatively few Dutch students achieve a higher level in comparison with other countries. The inspectorate wants to know how this is done and what can be done about it. [Arithmetic] education has had more narrative calculations since 2004. Some believe that this 'realistic [arithmetic]' is partly responsible for the decline of mathematical education. The Education Inspectorate thinks this is an outdated discussion: "You can see that in recent years the realistic [arithmetic] and the old form of arithmetic, ie [drilling] tables, are growing closer together."" [45]

The latter is a confused statement, as if the choice between RME and TME finds its proper answer in mixing those (in unknown ways), like the choice between astrology and homeopathy finds its answer in mixing those (in unknown ways), or like the choice between astrology and astronomy finds its answer in mixing those (in unknown ways). Basically the Inspectorate itself failed to take advantage of the “Reken Zeker” opportunity in 2016. Due to the current mixture we might be less able to deal with the ideologies.

The PPON results come in two batches. Alongside the annual results on the population for the scores only, there are periodic samples that also collect data on textbooks used, social-economic-status (SES) and such. The PPON report on 2016 only gives the population. [46] An enquiry at CITO confirmed that 2016 had no collection of data on textbooks, SES and other factors. For a comparison of RME and TME such would have to be reconstructed from the school archives. The 2018 competition for research grants to do a periodic sample for 2018/2019 (but not 2016) was won by a consortium under leadership of psychometrician Marian Hickendorff. [47]

These formal institutions obviously have their role in these developments. My tendency is to think that agents at these formal institutions might be more influenced and motivated by their views on the role of science and the informal ideologies of TME and RME. Whatever this may be, there still is reason to look at the latter anyhow. This attention for the informal institutions brings us to the present paper.

1.4. Causal modeling

Didactics concerns the study of what an issue might be, what students might handle, how they might learn it, and how you would test this. Let me refer to Van de Grift (2010:16) (in Dutch) for the activities of a succesful teacher, and check that these activities are targeted at affecting learning behaviour. [48] Van de Grift and KNAW tend to refer to Hattie. It remains important to be aware of Slavin’s criticism w.r.t. Hattie’s approach. [49]

This issue on arithmetic and algebra also created some insights on causal modeling on didactics, psychology and student results. There is the distinction between instruction / direction (what teachers do) and learning behaviour (what students do). Some psychometricians seem to suggest that they study learning and that didactics studies teaching without studying learning. However, didactics obviously looks at learning.

In this discussion, there is the key point of grading exam questions. This obviously pertains to the psychometric measurement of test results. The point should be clear by itself, but apparently still contributed to confusion, and thus is best discussed.

This paper thus has the following structure: After clarifying grading and the effect measure, we summarise the Dutch situation, and then look at the causal modeling.

---

[45] https://blog.sbo.nl/onderwijs/onderzoek-dalende-niveau-reken-en-wiskundeonderwijs/
[46] https://www.onderwijsinspectie.nl/onderwerpen/peil-onderwijs/documenten/rapporten/2018/04/11/taal-en-rekenen-aan-het-einde-van-de-basisschool-2016-2017
[47] https://www.universiteitleiden.nl/nieuws/2018/05/subsidie-marian-hickendorff
[48] https://www.rug.nl/education/lerarenopleiding/onderwijs/oratie-van-de-grift.pdf
[49] https://robertslavinsblog.wordpress.com/2018/06/21/john-hattie-is-wrong/

2. *Grading and the effect measure*

Table 1 considers a problem and answers provided by students. How can we grade those answers ?

Table 1. A problem and its answers by two students

Problem: What is 100 / 4 ? Answer: Traditional or reform.

John: Traditional algorithm: long division	Susan: "Realistic" trial and error
4 / 100 \ 26 8 ---- - 20 20 ----- - 0	100 20 = 4 x 5 ------- - 80 20 = 4 x 5 ------- - 60 20 = 4 x 5 ------- - 40 = 4 x 10 3 x 5 + 10 = 25

There are at least two possible effect measures, or methods of scoring:

- A simple method is to "only check the final result": John gets a Fail. He used the advised algorithm but made a calculation error, with likely an oversight and lack of discipline to check up. Susan gets a Pass. She made various correct but simple calculations.
- Didactics of mathematics tends to advise: "also intermediate steps can show insight".

Suppose that you can earn 5 points on this sum.

- Then John might lose 2 points, for his answer is false, and the student should have checked the answer by multiplying 4 x 26. But otherwise the method is applied properly. For example, when John does another long division, and performs the algorithm again but then makes a calculation error at another point, then we verify that he knows the algorithm but should practice more on his tables of multiplication or rather his discipline on checking up. [50]
- Susan might earn 5 points simply because trial and error generated the right answer. I am not at home in “realistic” conventions (like astrology or homeopathy). I would find the steps too small, or the student should have recognised 80 as 4 x 20. Thus a score of 4 would make more sense. If the student would give a wrong answer, then I would find it hard to judge the trial and error process, since it might go anywhere.
- Thus the Pass / Fail method has scores 0 & 5 while didactics has scores 3 & 4.
- In general, didacticians have discussions about such grading steps, since it depends upon what students have been trained for and what they are being tested about.

These issues are fundamental for didactics and psychometrics. The definitions and observations are closely connected. There isn't just measurement but this depends upon the definitions. In the 100 / 4 example it seems as if the algorithms might be well defined. But when you don't score the steps properly, then it might still be trial and error. See **Table 2** on a contrived case that might only occur seldomly but that highlights the aspects. A simple

[50] Referring to grading on scale of 0-10, Henk Boonstra thinks that the grades 7+ are more indicative of discipline in execution rather than understanding of principle. Current testing is deficient in making this distinction and giving pupils the proper feedback. Boonstra also calls attention to the fact that students are heterogeneous, in primary and secondary education alike. See https://henkboonstra.blogspot.com/2010/01/de-ongelijkheid-van-kansen-in-het.html

scoring method would only look at the right answer 25 and give Jack a Pass. If Jack found the right answer by trial and error, but also has learnt that the teacher is only happy when shown a semblance of a long division, then he might mimic this. Categorising him as following the traditional method could be wrong. The categorisation is less relevant because the relevant measure of using the method of long division requires that you also grade the intermediate steps. In this case Jack might get 1 out of 5 points because 2 times 4 is 8, while the right answer of 25 is judged as deriving from trial and error and not from following the algorithm.

Table 2. Why grading steps tends to be advisable

Jack: "Traditional algorithm: long division"	
4 / 100 \ 25 8 ---- 180 180 ----- 0	

This discussion only exemplifies the key importance of defining your measurements. **Table 3** gives an overview of the possible combinations. Psychometricians Van Putten & Hickendorff (VPH) (2009) classify answers by students on strategies but they still score on outcomes only. It is not clear to me whether they would still categorise the semblance of long division in **Table 2** as that Jack would have really worked in traditional manner. Nevertheless, when they score on outcomes only, then they don't really score on strategies, because they do not assign points per step. A (vertical) categorisation on semblance on strategy is not the same as (horizontal) assigning points for the various steps. A vertical comparison is at risk of invalid conclusions because the strategies are not scored properly. A conceptual base for algebra is not merely arithmetic success of getting the right outcomes, but requires command of the traditional algorithms. These considerations are so blatantly obvious to teachers and researchers on testing that it is almost painful to restate them here again, but surprisingly there is confusion about them in Holland.

Table 3. Categorising and scoring

Scoring	*Outcome only*	*Score on steps*
Traditional algorithm	VPH (2009)	Valid
RME, trial and error or context allows calculator	VPH (2009)	No standardised steps
No distinction	CITO	-

By 2009 all textbooks used in Dutch elementary schools were of the RME kind. The VPH (2009) categorisation only concerned a distinction by *technique* and not by *didactics*. All pupils using the traditional algorithm had received their training in an environment of RME. Thus, a conclusion, upon this classification, that there was no real difference in performance cannot be translated into a conclusion about RME and TME. The VPH (2009) classification thus doesn't allow for a test on didactics (and thus the link to later algebra and the TME claim that it are precisely the less talented pupils who would benefit from a training on the algorithms without distraction from other solution techniques).

3. The math war in Holland and the KNAW 2009 report

3.1. Declining competence in arithmetic

The competence of students in arithmetic has been deteriorating over the years, with CITO duly recording this, as they provide official tests at the end of primary education. We can be grateful to CITO, because they actually monitor this, while the ideologues of "realistic" mathematics education don't do so, and while the traditional mathematicians actually don't do so either (for they are trained to think abstractly and they don't like empirical methods) – with the exception of A.D. de Groot (1914-2006) [51] [52] who with a BSc in mathematics switched to psychology and was key in founding CITO. However, if CITO had been measuring with the proper effect measure (highlighting the preparation for algebra) then we should have seen the deterioration much earlier and much larger. (Obviously, this is a counterfactual based upon logic without statistical evidence.)

Van der Plas (2009:210-211) [53] explains that the shift to "context questions" has obscured the lack of algebraic competence, i.e. the arithmetic competence of methods that are also relevant for algebra.

It is an innovation by Kees van Putten that he looked at student strategies, which CITO neglected, as it only looked at the outcome of sums. If I understand this correctly, it was for this project that Marian Hickendorff was recruited for, for her Ph.D. thesis. Van Putten to Jan van de Craats 2008-01-28: [54]

> Google Translate 2018: “In 2006, Marian Hickendorff and I, together with six students, looked at almost 10,000 multiplication assignments of over 1,500 pupils in the PPON testbooks that were made available by the Cito at Leiden University. These are the first results and my AIO Marian will soon start with more detailed analyzes. The traditional multiplication 'under each other' (as the grandfather did) is still very common (in contrast to the tail division [long division]), but it is declining in 2004 compared to 1997. I have specifically zoomed in on the task '99 × 99 = ? ' because it lends itself so well to the so-called realistic approach. I inspected a large number of testbooks with this assignment from PPON 2004 one night and slept exceptionally badly that night: as long as the students counted 'according to grandpa', it usually went well, but realistic approaches via for example 100 × 99 or 100 × 100 provided a battlefield with erroneous effects and answers. It already started with errors in 100 × 99 or 100 × 100 (with errors like 990 and 1000 or 100 000 respectively), and then the problem how much to subtract (compensate) with errors like 1 or 2, 100 or 200 off. In fact, only the traditional approach was successful here and only the strong calculators (best 33%) could afford a realistic approach; all other combinations had no chance.”

[51] https://en.wikipedia.org/wiki/Adriaan_de_Groot
[52] https://boycottholland.wordpress.com/2015/11/24/a-general-theory-of-knowledge/
[53] http://www.liesbethvanderplas.nl/rekenvaardigheid-in-relatie-tot-wiskunde
[54] http://www.onderwijskrant.be/kranten/ok146.pdf page 22: “In 2006 hebben Marian Hickendorff en ik samen met zes studenten bijna 10 000 vermenigvuldigopgaven van ruim 1500 leerlingen bekeken in de PPON-toetsboekjes die door het Cito aan de Universiteit Leiden ter beschikking zijn gesteld. Dit zijn de eerste resultaten en mijn AIO Marian gaat binnenkort beginnen met gedetailleerdere analyses. De traditionele vermenigvuldiging ‘onder elkaar’ (zoals opa het deed) komt nog steeds veel voor (in tegenstelling tot de staartdeling), maar is wel aan het teruglopen in 2004 vergeleken met 1997. k heb speciaal ingezoomd op de opgave ‘99 × 99 = ?’ omdat deze zich zo goed leent voor de zogenaamde realistische aanpak. Ik heb een groot deel van testboekjes met deze opgave uit PPON 2004 op een avond zelf nagekeken en heb die nacht bijzonder slecht geslapen: zolang de leerlingen maar ‘volgens opa’ rekenden, ging het meestal goed, maar realistische aanpakken via bijvoorbeeld 100 × 99 of 100 × 100 leverden een slagveld aan foutieve uitwerkingen en antwoorden op. Het begon al met fouten in 100 × 99 of 100 × 100 (met fouten als 990 respectievelijk 1000 of 100 000), en vervolgens het probleem hoeveel daarvan af te trekken (compenseren) met fouten als 1 of 2, 100 of 200 eraf. Eigenlijk was alleen de traditionele aanpak hier succesvol en konden alleen de sterke rekenaars (beste 33 %) zich een realistische aanpak veroorloven; alle andere combinaties waren kansloos.”

However, Van Putten should also have realised the key notion of measurement (psychometrics), i.e. that definitions matter about what you observe. Categorising strategies into either traditional or "realistic" or both or none, is one step, but it matters whether one measures (scores) the intermediate steps, to indicate to what extent such strategies are actually pursued (for it might also be just trial and error).

An observation by Van Putten and Hickendorff (VPH) was that students who used pen and paper did better than students who did not (relying on mental calculations only). In itself teachers know this already, but one still needs to check what it actually means. Perhaps students who did not write much were mostly deficient anyway (excepting those who got the right answer). (I did not check this part of their analysis.) But, if you grade intermediate steps, then students are aware that they should also record intermediate steps, and then there is an automatic reward for recording these steps. Thus the very way of measurement would affect whether students actually perform better or worse.

3.2. The main claim of lack of evidence on a difference

The “math war” caused the Dutch Academy of Sciences KNAW to set up a committee, that produced a KNAW (2009) report. The KNAW committee consisted of mathematicians and psychometricians. Key researchers were psychometricians Kees van Putten and (non-member and Ph.D. student at the time) Marian Hickendorff. The Hickendorff (2011) thesis [55] partly refers to her research for the KNAW report. KNAW (2009:10) gives a summary in English and its mission and conclusion 2 are:

> “The Committee’s mission was the following: To survey what is known about the relationship between mathematics education and mathematical proficiency based on existing insights and empirical facts. Indicate how to give teachers and parents leeway to make informed choices, based on our knowledge of the relationship between approaches to mathematics teaching and mathematical achievement.”
>
> “2. The public debate exaggerates the differences between the traditional [TME] and realistic [RME] approaches to mathematics teaching. It also focuses erroneously on a supposed difference in the effect of the two instructional approaches whereas in fact, no convincing difference has been shown to exist.”

This basically fits the Slavin & Lake (2008) methodology and conclusions on the USA.

3.3. The situation in Holland

Hickendorff’s review study, chapter 1 in the thesis, selects 25 studies (18 experimental and 7 curriculum) that would relate to the Dutch situation. Test-psychologist Ben Wilbrink would like to impose stricter criteria:

> Google Translate 2018: “The committee therefore seems to be a bit sloppy: there is no research available that makes it possible to say something sensible about the effectiveness of different didactics. This is certainly something different from the first two sentences in the report, cited above, suggesting that research would have been done, aimed at the existence of differences, where no differences were found.” [56]

Let me shortly indicate the three most relevant curriculum studies. The Harskamp 1988 thesis found a modest effect size of 0.09 at CITO in favour of RME, apparently not looking at later algebra. The Gravemeijer et al. 1993 MORE study found an effect size of 0.32 in favour of TME. (Wilbrink criticises this MORE study [57] but I find that he misrepresents the work Van Hiele. [58]) PPON 1997 still had some TME textbooks, and RME had an effect size over TME in

[55] https://openaccess.leidenuniv.nl/handle/1887/17979
[56] http://benwilbrink.nl/projecten/realistisch_kolomrekenen.htm, search on Lenstra
[57] http://benwilbrink.nl/projecten/more.htm
[58] https://boycottholland.wordpress.com/2015/09/05/pierre-van-hiele-and-ben-wilbrink/

the range of 0.22 to 0.53 depending upon textbook (p43). There again is no discussion of the relation to algebra later in highschool.

KNAW (2009) should have advised to abolish the Freudenthal Head in the Clouds Realistic Mathematics Institute in Utrecht, that pushed RME without proper testing on arithmetic and its relation to later algebra. Freudenthal and his institute clearly were motivated by ideology and not science. As said, logic also points to TME as the base line. KNAW (2009) however seems to have followed the reasoning by the psychometricians that you cannot say that the Moon has another side when statistical evidence is lacking.

3.4. Selective use of sources

The VSNU (Dutch joint universities) and Leiden code of research integrity has:

> Google Translate 2018: "4.5 A scientific practitioner is only a defender of a certain scientific point of view if that position has been sufficiently scientifically substantiated, and in addition, rival positions must be reported and explained." [59]

There was a conference in 2008 [60] that resulted in a special May 2009 issue of the peer review *Tijdschrift voor Orthopedagogiek* (TvO). The issue contains VPH (2009) and the paper by astrophysicist and teacher of mathematics Liesbeth van der Plas (2009). [61] The report KNAW (2009) was published on November 4. In the KNAW (2009) list of references on p91-94 we find VPH (2009) who neglect above window of opportunity for algebra. In the list of references of KNAW (2009), p91-94, *we do not find reference* to the Van der Plas (2009) paper and contribution to the 2008 conference, who warns about above window of opportunity for algebra and the effect measure. [62]

Thus the KNAW (2009) report, that was intended to deal with the math war, appears to be biased itself, and appears to be in violation of the VSNU and Leiden code of research integrity. Reference to Van der Plas (2009) is also missing in the Hickendorff (2011) thesis. The word “algebra” is entirely missing in the thesis too.

Van der Plas (2009) shows *in a didactically valid manner* that the scoring method also used by VPH (2009) is invalid. Van der Plas doesn’t refer to VPH but she discusses the scoring method of CITO that VPH use too. This CITO method only considers the outcomes of sums and not the algorithm, while the latter is relevant for learning algebra in subsequent education. Pupils might score better by the use of the calculator and trial and error as allowed by RME but this would still maim them mentally for highschool and the rest of their lives in the compentence w.r.t. algebra. Yet this criticism by Van der Plas was neglected by VPH and “thus” the KNAW committee.

It isn't only that the research record is tainted by (deliberate) neglect (exclusion). Let me add that there have also been costs to Van der Plas for not being referred to properly. What would have happened when VPH (and subsequently the KNAW report) had referred properly ? Then others would have seen the key relevance of this paper too. In a specialising world: when you are excluded from the key overview, then you likely aren't noticed anymore.

The conference paper VPH (2009) does not refer to Van der Plas (2009) either. It might be seen as fair that papers presented at a conference in 2008 don’t refer to each other. On the

---

[59] http://media.leidenuniv.nl/legacy/vsnu-code-wetenschapsbeoefening-2004-%282014%29-def.pdf "4.5. Een wetenschapsbeoefenaar is pas verdediger van een bepaald wetenschappelijk standpunt als dat standpunt voldoende wetenschappelijk is onderbouwd. Rivaliserende standpunten dienen daarnaast te worden gemeld en toegelicht."

[60] http://www.wiskundebrief.nl/471.htm

[61] http://www.liesbethvanderplas.nl/rekenvaardigheid-in-relatie-tot-wiskunde

[62] In the same issue of *Tijdschrift voor Orthopedagogiek* in May 2009 there is an article by Gerard Verhoef who has a similar point on the effect measure. There is also an article by Jan van de Craats who could have made the point but doesn't, perhaps because of specialisation (and there is no need to repeat what others have said). In 2015 Van der Plas repeats her comment as something that is rather obvious for didacticians: http://www.wiskundebrief.nl/720.htm#5

other hand, there were months between the conference and the actual publication. The idea of a conference with peers is that when there is criticism that invalidates your analysis, then you would at least adapt the paper with a discussion of the criticism.

Perhaps after the conference in 2008 VPH were so busy with the KNAW committee that they did not have time to listen to criticism ? Perhaps the only reason for VPH and thus KNAW (2009) to neglect the argument by Van der Plas (2009) may have been that she did not use statistics ? This is unclear, and as far as I know VPH publicly neither discussed it nor explained why they excluded it. For completeness: we can guess at other confusions [63] but none of these confusions would be valid either. The Hickendorff (2011) thesis refers to "empirical studies" but it may be that she confuses this with statistical studies only. Van der Plas (2009) clearly is an empirical study too, since it looks into the issue and its effect measure. A problem is that these psychometricians have no background in the didactics of mathematics and may not recognise the validity of the argument by Van der Plas (2009).

3.5. Invalid reasoning and A.D. de Groot's Forum Theory

Based upon their statistical analysis VPH infer that they cannot diagnose a difference in effectiveness in RME and TME at the end of elementary school. Their method is innovative in that they look at the pupils's exam papers rather than final answers to classify which approach each used. However, as clarified above, such classification differs from proper scoring. Teachers of mathematics and test researchers would have some points of doubt:

(1) The official exam rule is that "non-context questions" consist of arithmetic sums only while "context questions" have verbal formulations (narratives), and that only the latter may be done by calculator. This creates a bias towards RME that invented the very notion that "context warrants the calculator" and that relies on context and thus the calculator and trial and error.
(2) VPH don't grade steps and thus have another bias in favour of RME. Categorising students on the methods used in their exam papers cannot replace the basic didactic consideration that one anyhow grades steps to evaluate competence.
(3) A categorisation on techniques cannot discriminate between RME or TME didactics anyhow. By 2009 all textbooks were of the RME kind only, and all kids had been trained in RME fashion.

In October 2013, the then-chairperson of NVvW Marian Kollenveld gave this criticism on the final exam arithmetic test ("Rekentoets"):

> Google Translate 2018: "(...) they are multiple-choice questions and short-answer questions in which the dissolution process is not assessed (only a good answer counts, regardless of the complexity) - in case of a complex question, there are sometimes 4 steps that can all be right or wrong, and can stand for differences in the student's skill, which are not measured now, this also contributes to a minimal score)." [64]

[63] VPH might have potential confusions to neglect the Van der Plas (2009) article. None of these confusions are valid but we may list some. (1) It does not explicitly and concretely refer to their work. (2) It does not refer to papers in peer reviewed journals. (3) It does not provide statistics. (4) It might look like a personal opinion. (5) It discusses the link between primary and secondary education, instead of only primary education. (6) It looks at didactics and not student learning. (7) It does not *fully* develop the issue on the effect measure (because it also looks at other issues, like the relation of arithmetic to algebra). These seven possible confusions are invalid, because Van der Plas (2009) remains relevant for the issue of interest, and her analysis implies that the VPH (2009) paper has an invalid approach. If Hickendorff chooses to associate herself strongly with CITO, we may conclude that Van der Plas (2009) actually has *concrete criticism* w.r.t. VPH (2009).

[64] "(...) het zijn meerkeuzevragen en kort-antwoordvragen waarbij het oplosproces niet wordt beoordeeld (alleen een goed antwoord telt, ongeacht de complexiteit,- bij een complexe vraag zijn er soms wel 4 stappen die allemaal goed of fout kunnen, en ook kunnen staan voor verschillen in vaardigheid van de leerling, die nu niet worden gemeten, ook dit draagt bij aan een minimale score)." http://thomascool.eu/Papers/AardigeGetallen/2016-03-06-NVVW-bestuur-desinformeert-het-parlement-over-het-rekenen.pdf

The issues should have been resolved within the setting of A.D. de Groot's "forum theory". [65] [66] The Dutch journal Euclides since 1925 is online now, and there is the obituary of A.D. de Groot, in Euclides, 82 no 3. [67] He got a BSc in mathematics before he switched to psychology, and was a teacher of mathematics for a while. His is a key founder of CITO, where Hickendorff works parttime. I imagine that A.D. de Groot would be aghast to see how psychometricians maltreat the didactics of mathematics. De Groot would also be horrified by their lack of understanding of psychometrics itself. The first thing that a psychometrician should do is to explain that definitions determine the measurements. Thus when you claim to study education, then you define what is involved, and then you also specify what is a success and what is a failure, and you acknowledge the criticism that you also must score the intermediate steps when those are relevant for the strategy of answering a test question. And you should not confuse a technique used (cross-sectional) with didactics (longitudinal).

Forum theory hasn't been much implemented yet, and subsequently we also meet with researchers who refuse to answer to criticism. VPH might think that they are open to criticism, and can refer to discussions in TvO, Psychometrika or presentations at NVORWO Panama conferences – meetups commonly linked to "realistic" ideology. [68] Obviously I will not deny such communication, but it doesn't change the present criticism, which they neglect for some years now. I am not aware of other people complaining that VPH do not (adequately) respond to criticism. In fact, I am quite amazed that Liesbeth van der Plas, mathematician Jan van de Craats, mathematician Gerard Verhoef and test-psychologist Ben Wilbrink [69] [70] haven't really deconstructed the KNAW 2009 report, and subsequently Hickendorff's thesis [71] [72] and the discussion on these as well. I suppose that each might have his or her own reasons.

A key difference between these others and me is that I also wrote the books "Elegance with Substance" (2009, 2015) and "Een kind wil aardige en geen gemene getallen" (2012) and "A child wants nice and no mean numbers" (2015, 2018). [73] Thus my position in didactics of mathematics is much stronger compared to these other authors. Obviously also, as an econometrician, I am familiar with the basics of the IRT testing method that VPH employ. Holland is a small country. Also, I am an econometrician and look at these issues also from the viewpoint of (institutional) economics. [74] Thus perhaps it is unavoidable that I might be the only local researcher who can unravel the knotty problem created by the three groups involved: the ideologues of "realistic" mathematics education, the traditional mathematicians

[65] https://www.knaw.nl/nl/actueel/publicaties/het-forumwaarmerk-van-wetenschap
[66] https://boycottholland.wordpress.com/2015/11/24/a-general-theory-of-knowledge/
[67] https://archief.vakbladeuclides.nl/bestanden/82_2006-07_03.pdf
[68] http://www.nvorwo.nl/event/panamaconferentie/
[69] http://benwilbrink.nl/projecten/rekenproject.htm
[70] http://thomascool.eu/Papers/Math/2015-09-15-Breach-by-Jan-van-de-Craats-and-Ben-Wilbrink-wrt-scientific-integrity.html
[71] http://benwilbrink.nl/literature/hickendorff_2011.htm Google Translate 2018 of his comment of October 24: "Having read the entire dissertation at least once, it is striking that the research material - the calculations - is limited to what is typically tested in current mathematical education. Note: the current maths education is strongly marked by the ideas of realistic mathematical education. It is for me even the question whether the calculations in Hickendorff's research do belong in the [arithmetic] domain. In any case, they do not belong to the core of this: the 'realistic' mathematical education has been marginalized over the decades into convenient computing, calculating in school contexts that must also be called 'realistic' but that are not, and to mathematical education that is cut off from what later mathematics education presupposes [arithmetic] skills (knowledge of algorithms, basic [arithmetic] facts, can count with fractions). Hickendorff and others have good reasons for limiting the research to the typical calculations as can be found in realistic mathematical education, and unfortunately also in the Cito Eindoets Basisonderwijs and the PPON, but they still run the risk of giving the impression that this research is about calculating skills in the ordinary sense of the word, while in fact it concerns mathematical education as it is deformed under the influence of the not adequately empirically tested [arithmetic] ideas of Hans Freudenthal and his group."
[72] Wilbrink on Hickendorff's Chapter 1, Google Translate 2018: "The Lenstra committee also ignored the didactic theory, possibly because no opinion was possible within the committee. But what if a certain didactic theory is demonstrably based on failed psychology? I would like to keep this point at the center of discussion about mathematics education at all times. Marian Hickendorff undoubtedly, too, but she is understandably focused on empirical data about pupils." Wilbrink allows Hickendorff to get away with "testing without theory" ? What is the meaning of his "center of discussion" ?
[73] http://thomascool.eu/Papers/Math/Index.html
[74] https://boycottholland.wordpress.com/2015/10/31/the-power-void-in-mathematics-education/

and the psychometricans with their blinders. Yet I do not attend such Panama conferences. Before 2012 I had no real interest in primary education and its research. I never considered myself qualified for didactics in primary school, though a new law in 2016 declares that I am now. [75] The present analysis obviously is sound but still targetted at a very specific issue and point in research. My main point is that I pose questions and would like to hear some answers.

But I must also mention that much of what I say – in this case – really isn't new, see the reference to Van der Plas (2009) and the age-old discussion in statistics and testing about validity. The example on 100 / 4 above is so blatantly obvious, that I cannot see why VPH don't reply to this fundamental issue w.r.t. their research. Let me add that before 2016 I only superficially encountered Van Putten and Hickendorff once very briefly, namely at that KNAW 2014 conference, that looked back at the KNAW 2009 report. Thus, with more discussion in person, I could have gathered a better diagnosis of the situation. I can imagine that there might be communication issues between psychometricians and econometricians and didacticians, but I have done my share of looking into psychometrics as well (see the references in VTFD), and it would be no more than rational and scientically warranted if VPH did their share on didactics of mathematics.

The Dutch association of teachers of mathematics NVvW apparently did not debunk the KNAW (2009) report (a Google returns no results). They focused on the "Rekentoets" as applied in secondary education, which is in their own direct interest. The more serious implication however is for primary education: that if you don't properly teach and score the traditional algorithms, then you don't properly prepare students for algebra in secondary education and the rest of their lives. [76] The situation isn't helped much by that NVvW has turned out to be a seriously sick organisation. [77]

3.6. The Hickendorff email of 2014 and refusal to correct

A KNAW 2014 conference, looking five years back to 2009, caused me to contact Hickendorff, asking her about didactics and MER and the validity of her research and intended presentation. Hickendorff replied:

> Google Translate: "Dear Thomas Cool, Thank you for your mail. I fear that I can not find the time to view everything you send. In addition, I also wonder if you have come to the right place for me: I am not a didacticist but a psychological researcher, and I also try to stay out of the discussion about didactics as much as possible because I do not believe that that is my expertise. Kind regards, and until a.s Monday, Marian Hickendorff" [78] [79]

---

[75] https://www.delerarenagenda.nl/blog/weblog/weblog/2016/20160902-vo-docenten-bevoegd-voor-basisschool-%E2%80%98goede-methode-voor-talentvolle-kinderen%E2%80%99

[76] http://www.wiskundebrief.nl/721.htm#5

[77] http://thomascool.eu/Papers/Math/2016-06-28-Letter-to-NVVW-with-Red-Card.pdf For some Dutch readers, the curious clash in 2016 between the state secretary of education Sander Dekker and the board of NVvW, about a supposed "agreement" on the new highschool test on arithmetic, might be another eye-opener on the disfunctionality of NVvW.
http://www.volkskrant.nl/binnenland/wiskundeleraren-dekker-komt-afspraken-over-betere-rekentoets-niet-na~a4429909
https://boycotholland.wordpress.com/2017/04/08/update-van-bestuur-nvvw-verzint-een-afspraak-met-de-staatssecretaris/

[78] From: "Hickendorff, M." [...]
To: "Thomas Cool / Thomas Colignatus" [...]
Cc: J.A.Bergstra [at] uva.nl, "Craats, Jan van de" [at] uva.nl
Subject: RE: T.b.v. a.s. maandag (KNAW reken-onderwijs)
Date: Fri, 27 Jun 2014 [...]
Beste Thomas Cool, Bedankt voor uw mail. Ik vrees dat ik niet de tijd kan vinden om alles wat u stuurt te bekijken. Daarnaast vraag ik me ook af of u bij mij hiervoor aan het juiste adres bent: ik ben geen didacticus maar psychologisch onderzoeker, en probeer ook zo veel mogelijk buiten de discussie over didactiek te blijven omdat ik niet meen dat dat mijn expertise is. Vriendelijke groeten, en tot a.s. maandag, Marian Hickendorff

[79] I take this statement as it is. If she meant something else, then she should have said something else. Also, I have explained at various locations what her statement implied, and alerted her to this, so she

Originally, I praised Hickendorff for her modesty that she refrained from a discussion that wasn't her expertise. Hickendorff does not clarify her lack of expertise in the thesis itself, and apparently hasn't told this to the minister of education who might look differently at the KNAW report of 2009 now. I did question her because she involved herself nevertheless, and I observed that she couldn't avoid using an effect measure in her research, which effect measure can only be based upon didactic concerns. She did not reply to this, which is a clear breach of integrity of science.

> Google Translate 2018: "6.2 Academic practitioners are honest and loyal about the quality they deliver and they contribute to internal and external assessments of their research." [80]

My questions to Hickendorff amount to an external assessment and she rejected a reply basically by the argument that it was external to her. An analogy: A foot surgeon performing heart surgery rejects answering questions on this by saying, modestly, that he is only a foot surgeon.

Getting clarity on the effectiveness of the didactic approaches of TME versus RME was one of the main objectives of the KNAW report. Curiously, in her thesis Hickendorff claimed such expertise namely by claiming to do a review. What would be the basis of such claim ? Yet in her email to me she claimed keeping a distance for lack of expertise. This is inconsistent. Clearly my earlier praise for modesty must be withdrawn. She has wrongly informed me, and she should reply to the question on content. She should do so in public. Her disinformative email and the subsequent refusal by her and Van Putten to consider the criticism is in violation of the basic rule in science that researchers must be open to questions and criticism. Perhaps they only follow psychometric convention but keeping a field accountable runs via individual research ethics since one cannot address all at the same time.

Page xvi of Hickendorff's thesis clearly states that *she did report on the effect of didactics on results*.

> "The thesis starts with Chapter 1 reporting a research synthesis of empirical studies that were carried out in the Netherlands into the relation between mathematics education and mathematics proficiency. This chapter is based on work that was done for the KNAW (Royal Dutch Academy of Arts and Sciences) Committee on Primary School Mathematics Teaching [ftnt], whose report came out in 2009. Starting with an overview of results of Dutch national assessments and the position of Dutch students in international assessments, the main body of the chapter is devoted to a systematic review of studies in which the relationship between instructional approach and students' performance outcomes was investigated. The main conclusion that could be drawn was that much is unknown about the relation between mathematics programs and performance outcomes, and that methodologically sound empirical studies comparing different instructional approaches are rare, which may be because they are very difficult to implement. In the remainder of this thesis, the focus is shifted to other determinants of students' mathematics ability related to contemporary mathematics education, such as the strategies students used to solve the problems and characteristics of the mathematics problems. [ftnt: I worked as an associate researcher supporting the Committee. In particular, the Committee requested me to carry out the systematic literature review that formed the basis of chapter 4 in the report. Chapter 1 in the current thesis is based on this work.]"

It is didactics that deals with "the relationship between instructional approach and students' performance outcomes". See also the Dutch translation on p274-275. [81] Her study was a

---

could have corrected me in public since 2014 that she should have been more precise w.r.t. what she actually wanted to express.

[80] "6.2. Wetenschapsbeoefenaren laten zich eerlijk en loyaal de maat nemen over de door hen geleverde kwaliteit. Zij werken mee aan in- en externe beoordelingen van hun onderzoek." http://media.leidenuniv.nl/legacy/vsnu-code-wetenschapsbeoefening-2004-%282014%29-def.pdf

[81] Dutch p274-275: "Hoofdstuk 1 van dit proefschrift bevat een onderzoekssynthese van resultaten van Nederlandse empirische studies naar de relatie tussen rekendidactiek en rekenvaardigheid. Dit hoofdstuk is gebaseerd op literatuuronderzoek dat is uitgevoerd voor de adviescommissie

review, but for a review you still must have some qualifications and there are criteria for being critical. In her 2014 email to me she now suggests that she was unqualified to do such a review. Also observe a potential reduction of “empirical studies” to the use of statistics only.

VPH might hold that they only *reviewed* cause-effect research by others, and did not do this kind of research themselves, but this is not relevant here, because in their review they did not criticise the effect measure, as they should have. They might not criticise the effect measures by these other authors because of their own lack of knowledge about didactics of mathematics. When they exclude Van der Plas (2009) for their review study too, then clearly they exclude information about what a valid effect measure would be.

To some extent I can imagine that Hickendorff wants to keep some distance from didactics, since the math war between TME and RME has turned this field into a quagmire indeed. However, the proper response is not neglect but protest and re-engineering. [82] Obviously, this starts from an interest in didactics of mathematics indeed, and an interest in psychology itself might be less encouraging, but the point remains that she started studying arithmetic test scores.

3.7. When it becomes an issue of research integrity

My diagnosis is that VPH (i) use selective sources, (ii) use the wrong effect measure so that claimed outcomes are invalid, (iii) have inadequate knowledge about and respect for didactics of mathematics while their topic requires those, (iv) neglect criticism on (i) – (iii). I have documented the case in Dutch [83] and English. [84] Leiden University rejected mediation and thus I submitted the case to the Leiden committee on research integrity. [85]

It can be observed that procedures on scientific integrity are not well developed yet. [86] Society has shifted from an agricultural to an industrial to a service economy. The conduct of "information workers" becomes ever more important, but regulations on these are lagging. This is awkward especially for specialists, when only a few persons deal with an issue, and when issues of conduct (like also rules of proceedings like these) might have a

---

Rekenonderwijs op de basisschool [ftnt] ingesteld door de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW), wier rapport in 2009 is uitgekomen. Deze systematische kwantitatieve onderzoekssynthese laat geen eenduidige conclusies over het effect van verschillende rekeninstructiemethoden of rekencurricula toe. Enerzijds zijn er weinig methodologisch degelijk opgezette interventiestudies waarin de effecten van verschillende instructieaanpakken vergeleken worden. De wel beschikbare studies zijn bovendien beperkt in verschillende aspecten, zoals steekproefgrootte of inhoudsdomein. Ook zijn didactische kenmerken en instructiekenmerken vaak met elkaar verweven in de programma’s die vergeleken zijn, zodat het onmogelijk is de unieke effecten van verschillende kenmerken vast te stellen. Anderzijds zijn de curriculumstudies waarin de uitkomsten van leerlingen die verschillende rekencurricula (rekenmethodes) gevolgd hebben worden vergeleken, beperkt in de mate van controle over de implementatie van het curriculum en in de mogelijk tot het corrigeren voor verstorende variabelen. Hoewel er dus geen algemene hoofdconclusie getrokken kan worden, zijn er wel wat specifieke patronen die uit de bestudeerde onderzoeksresultaten naar voren komen. Ten eerste is het opvallend dat de prestatieverschillen binnen een bepaald type instructieaanpak groter zijn dan tussen verschillende aanpakken. Blijkbaar spelen didactische principes een kleinere rol dan de praktische implementatie door de leerkracht en de interactie tussen de leerkracht en de leerling, bevindingen die in overeenstemming zijn met die van bijvoorbeeld Slavin en Lake (2008) in hun grootschalige internationale onderzoekssynthese."

[82] https://zenodo.org/communities/re-engineering-math-ed/about/

[83] http://www.wiskundebrief.nl/718.htm#7
http://thomascool.eu/Papers/AardigeGetallen/2008-2016-plus-Afgewezen-door-de-WiskundE-brief.html#2016-10-08
http://thomascool.eu/Papers/AardigeGetallen/2016-01-17-Meta-opmerkingen-over-psychologie-en-wiskunde.pdf
http://thomascool.eu/Papers/AardigeGetallen/2016-01-31-Enkele-emails-rekentoets-psychometrie.pdf
http://thomascool.eu/Papers/AardigeGetallen/2016-02-10-Basisprobleem-in-pedagogie-onderwijs-en-didactiek-van-wiskunde.pdf
http://thomascool.eu/Papers/Math/2016-05-25-Email-exchange-with-Kool-Noteboom-Tijdeman.pdf

[84] http://thomascool.eu/Papers/Math/2016-05-09-Letter-to-VOR-and-Trainers-of-teachers.pdf

[85] http://thomascool.eu/Papers/Math/CWI-Leiden/2016-09-30-Letter-to-CWI-anonimised.pdf

[86] https://boycottholland.wordpress.com/2015/11/26/allea-defines-research-integrity-too-narrow

disproportionate impact. Major concerns w.r.t. breaches of integrity have always been interference with politics or religion or personal advantage for income and status. Such breaches can be seen as coming from external sources. In the present case we have an ivory tower, with tunnel vision, own-language (empirics = science = statistics) and group think. This can be seen as deriving from internal sources in science. Science itself may invite to specialise, but over- and misspecialisation lead astray.

In my view, a professional with personal integrity can still breach the integrity of science. Therefor, I have *specified* what the breaches by VPH have been. The language for such issues is not well developed yet, and one tends to run into confusions because of ambiguous words. (Especially when others start generalising.) For example, a medical doctor might make an error that might even cause the death of a patient. But this doesn't need to be a case of malpractice. It might be a honest mistake. Professionals need freedom and might make mistakes. What can turn this into a breach of integrity (of medicine) is when the doctor neglects criticism and refuses to acknowledge the error. For example, a driver of a car might cause an accident, but still be insured for liabilities. What may turn this in problematic behaviour is when the driver was warned about risky weather conditions, and that he or she took risks that the insurer actually didn't take into account. It becomes a breach of truthful behaviour, for the overall learning process, when the driver doesn't ackowledge the true diagnosis of having taken too much risk.

VPH should have given a reaction to my analysis, in time and in public. This would have been normal scientific procedure: there is criticism on content, and reply on content. Now, there is this discussion on content but in the context of a procedure on integrity, and with a focus on restoring integrity of science.

Originally I had the vague idea that perhaps the Hickendorff (2011) thesis might still be maintained, since the main point of not responding to criticism is from 2014 onwards. However, a thesis should show that the candidate has learned what science is. Clearly Hickendorff hasn't. The thesis is a product of an ivory tower apparently created by Willem Heiser and Kees van Putten. Thus now I put more emphasis on the selective references, i.e. the not-including of Van der Plas (2009) and other didactic considerations. The scientific record better be set straight, so that one could not refer to the present "thesis" as if belonging to the scientific literature. Potentially Hickendorff is the victim of a selective thesis commission, but she also is an apt learner of such selective practices. Thus, my present view is that the thesis should be annulled too. It would be up to another promotor to determine what material can be rewritten in what manner for a revision. This really would be the best decision. Hickendorff is relatively young while Heiser, Van Putten and Tijdeman are retired. Hickendorff potentially has many more years as a potential scientist, and it is better that she learns what science is. Actually, after my original letter to CWI, this should have been the proper response by VPH as well.

4. Causal modeling for the basics of didactics

4.1. A basic model, mention of psychology, exclusion of didactics

Let us consider the causal modeling behind this. Let me denote *s* for student behaviour (learning, solution strategies), *d* for teacher behaviour (direction, instruction), and *o* for other factors. There will be some feedback when a teacher observes some ineffective learning strategy and adjusts the directions. For now function *f* suffices as a summary what is studied in didactics:

$$s = f[d, o]$$

Different directions *d1* and *d2* would give different outcomes $s1 = f[d1,$ o] and $s2 = f[d2, o]$. Each such relation can be called “a didactic”. Could you study *s* while neglecting the functional relationship $s = f[d, o]$? This would be like studying a phenomenon without its causal factors. The differences $s1 – s2$ would only be "noise" that cannot be explained. For science this might be a first step but it soon becomes absurd. For example, *A* says to *B:* "You

should look at a map because you are driving into the wrong direction." And *B* answers: "No, I am driving. Looking at maps is something else."

This clarification of the definition of didactics shows that the KNAW 2009 committee with its mission quoted above “To survey what is known about the relationship between mathematics education and mathematical proficiency based on existing insights and empirical facts” had a deficient composition, for they lacked didacticians. Arrogantly choosing to reinvent the wheel they came up with “garbage in, garbage out” (GIGO).

In her email of 2014, Hickendorff describes herself as a psychologist. Her thesis also expresses a wish of building a bridge between psychology and psychometrics. We might interprete this as a claim that, in her research frame, psychology was more important than didactics. Teachers get some training on pedagogy but will tend not have a degree in psychology. [87] Thus she would study function *g* that uses factors in psychology:

$$s = g[\psi, o']$$

Instead, the true model is rather that didactics already takes account of student psychology, with a distinction between true ψ for the student and its assumption ψ' by the teacher.

$$s = f[d[\psi'], \psi, o'']$$

If we can assume that there are no crucial errors in judging psychological reactions for most students, then we can assume $\psi = \psi'$, and the latter reduces again to:

$$s = f[d, o]$$

Above we observed that Hickendorff reviewed “the relationship between instructional approach and students’ performance outcomes”, didn’t spot adequate studies, and then looked at alternative explanations like student strategies themselves. My criticism was that Hickendorff incorrectly reduced $d = d[\psi']$ to noise o'. She assumes direct causality from ψ on s but the main channel is via *d.* While the KNAW study argued for a key role of the teacher, the distinction on TME and RME was rejected, but on invalid grounds.

With this notation in formula's I don't want to suggest exactness. I only think that these schemes help to emphasise the causal presumptions. This should also clarify that psychology is obviously relevant. For Hickendorff it perhaps is a key factor, but didactics might not put much emphasis on this since psychology is only one of the factors.

The following diagrams may clarify Hickendorff's conceptual error. **Figure 1** gives what is likely the "true model" for dominant causality. Potentially there are arrows between all factors but I now give the hypothesis for the main paths. **Figure 2** gives Hickendorff's position of cutting out the function *f* studied in didactics. Her suggestion is the inference from "students do so" to "students may be competent to do so". Observe that her "try to stay out of the discussion" might still mean that she would respect and include the conclusions from such discussions, like the Van der Plas (2009) paper. However, when her *conduct* is that she *neglects* such discussions, then “try to stay out of” is a misrepresentation of what she actually does.

[87] IGPME looks into psychology and mathematics education. http://www.igpme.org/index.php/home

Figure 1. Likely a good model of dominant causality

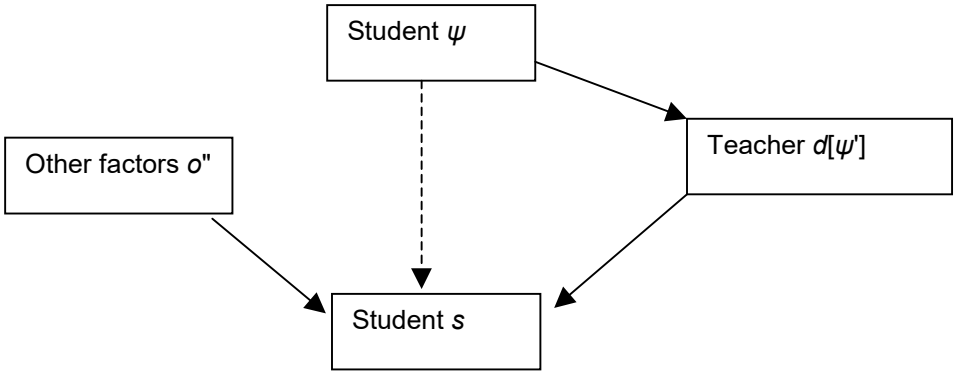
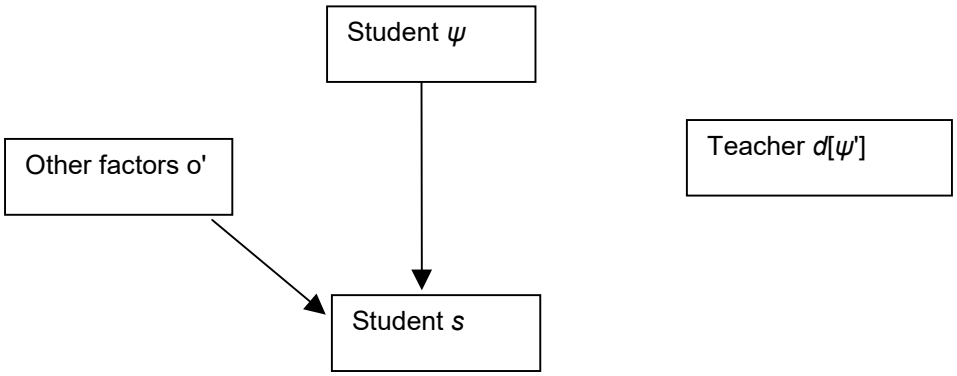


Figure 2. "No didactician" and "try to stay out of the discussions": students invent ("objectively given") algorithms themselves without didactics



4.2. Evaluation

There is a simple model in Item Response Theory (IRT) that has questions as items and student answers as responses. This looks at *s* only. I discussed this kind of modeling in my book *Voting Theory for Democracy* (VTFD), Colignatus (2001, 2014). IRT has the nice property that the test tells about both the competence of students and the adequacy of the test itself. However, IRT is only a limited model, and the proper analysis looks wider. Psychometricians focusing on only *s* are at risk of misrepresenting their field of study and the conditionality of their findings.

- Didactics obviously is focused on *affecting* learning behaviour by students.
- It is quite silly to argue (a) that a teacher only does his or her thing, and (b) that what students do is entirely independent.

Secondly, we can only *describe* the very *s* by using information from *d.*

- It are the *d* that define *what a strategy actually is*. The only theory that provides a rationale for what it means "solving correctly" derives from didactics. There is both the algorithm that students use, and possibly an independent *golden standard* provided by a computer programme, but both are designed by didactics.
- If you don't know about *d* then anything that students do is basically random behaviour (with some mean and dispersion).
- Proper didactics also assigns points for the intermediate steps in the algorithm. This valid effect measure would show that students *using the algorithm* would score much better.

(The use of the calculator would only give a few points for a right answer.) VPH would argue that this would be irrelevant for their research on *s* ?
- Why would a psychometrician select only *s* and the invalid effect measure of "answering a sum correctly" ? A psychometrician claiming to look at only *s* is at danger of creating his or her own universe of *s,* while neglecting that *s* only is meaningful because of *d.*
- Thus VPH used a statistical exercise to argue against the relevance of *d,* but their exercise was based upon the assumption that *d* was not relevant (for it neglected the discussion on the effect measure).

If a student solves 100 / 4 by means of traditional long division or "realistic" trial and error, then the use of these "strategies" would be random for psychometricians looking at only *s*, because these researchers would not have the didactics *d* that define what the proper algorithms are. Without the use of the algorithm, and only looking at the outcomes, they might determine that 100 / 4 is an “easy” question (with a higher rate of success) and that 57 / 3 is (perhaps) a “hard” question (with a lower rate of success). Without the algorithm such distinction would remain unexplained. Potentially psychometricians might think that "everyone knows what long division is", so that they don't need to check with didacticians of mathematics. In such an ivory tower they might reduce didactics into "ways to teach students about obviously clear techniques, given from heaven". This would be improper research, because it would neglect outcomes from an adjacent field of research (didactics).

This discussion might be contaminated by the context of the Dutch regulations about what is expected from children at the end of primary education. The standard is the CITO test. Hickendorff is associated with CITO. VPH might say that their definition of arithmetic is what CITO has chosen. This might boil down to "testing without theory". Then psychometrics reduces to behaviourism again. However, whatever these test-for-the-test philosophers claim, there is still a distinction between the CITO tests and the didactic objectives that have been selected, as what pupils should be able to do. In this case the objectives w.r.t. algebra in secondary school are clearly important. In that case Hickendorff *as a scientist* might have to criticise CITO instead of embracing it. It is not impossible that CITO has incompetent didacticians of mathematics too.

Clearly, when properly evaluated, the data in the KNAW (2009) report or the Hickendorff (2011) thesis chapter 1, or the evaluations by VPH (2009) in their own (non-review) research on such solution strategies, would generate other conclusions about the mathematical competence of the students (and by implication on the $s = f[d, o]$ relation).

Obviously the other factors *o* can be dominant (Van de Grift), but, in the case of comparing traditional didactics and "realistic" didactics in arithmetic (the present issue of concern), there is a clear dependence:

- Students don't simply invent the traditional algorithms of say long division or solving problems like 1/3 + 1/5. They must be taught via some *d,* and mastery comes from adequate training.
- If you apply the *proper* measure of success (scoring steps in the algorithm) then the difference between *s1* and *s2* will be highly correlated with the difference between *d1* and *d2.* This argument is based upon logic and not in need of a statistical study, and thus cannot be excluded as supposedly being “non-empirical”.
- If you apply an invalid measure of success then you might not see that correlation. In that case you might erroneously conclude that the statistical evidence doesn't support a distinction in effectiveness of either didactic method.

4.3. Possible confusions by psychometricians

In itself, when there is a math war between "realistists" and traditionalists, who actually both neglect both empirical research and statistics, and who don't care to design a distinguishing experiment, then I can imagine that psychometricians decide to focus on *s.* It is the kind of research that psychometricians have been creating a tradition in themselves based upon the Item Response Theory (with the risk of tunnel vision). Potentially it might generate results. That said, they still should be open to criticism, that one cannot just focus on *s* while

neglecting *f* and *d.* If the math war is a problem, then the math war should be resolved (and not neglected). Thus, when the psychometricians observe such a math war, then they should protest (too) instead of (only) neglect it. (My advice is an enquiry by parliament. [88])

This neglect of the role of didactics (with the example of long division) links up with the notion that various fields of research are looking into arithmetic: from neuroscientists to psychologists to didacticians. The suggested implication that other fields step in but that didactics might be neglected is a gross generalisation, and quite invalid.

- For example, I have warned neuroscience to beware of conclusions on number sense, when there are some crooked features in current arithmetic. For example, *two and a half* is 2 + ½ but it is written as *two times a half* or 2½ (compare *2a* or 2 km). A conclusion should not be that children have difficulty learning arithmetic, if the cause of learning problems lies in so-called arithmetic itself. See also the issue of pronunciation of the numbers. (New would be a discussion on the errors by Van Putten & Hickendorff and also CITO on 2 + ½, but I have deliberately chosen to first deal with the present conventional points.)
- It requires didactics to grow aware of such issues. Thus multidisciplinary research is welcome and ivory tower research might soon run astray.

Psychometricians should not be so singular as to claim that they can do this research themselves, with only other scientists who they select themselves, while using an invalid generalisation as "others neglect didactics and thus we can do so too". When other scientists join the party on their own initiative and utter criticism, then there is scientific reason to pay attention to the arguments.

NB. Actually, the situation is that the original party had been organised by didactics, and it are the psychometricians who created their own subparty, trying to take over. Let me refer to above quote from chapter 1 of Hickendorff's thesis:

> "(...) a research synthesis of empirical studies that were carried out in the Netherlands into the relation between mathematics education and mathematics proficiency."

Thus the issue is within the realm of didactics of mathematics, and the psychometricians are hired guns to illuminate aspects by their expertise. (They might use the same techniques as for language or other issues.) It can happen that the agent takes over from the principal, or that the lieutenant ("stadhouder") takes over from the king (William of Orange vs Philip II), but in this case, didactics has a sound position that they aren't fulfilling the contract and doing the job properly.

4.4. The causal models and the situation in Holland

Our discussion of these causal models might not be understood without the reference to the developments in Holland.

(1) These insights might be seen as differences in opinion in approaches to research. It might be seen as if VPH (2009) only have a different opinion other than Van der Plas (2009) or me. However, the true problem with VPH are the breaches w.r.t. research integrity w.r.t. the points mentioned above.
(2) VPH might argue that their research would only concern tests on learning. Van der Plas and I provide criticism from didactics, which thus might not apply to their research on learning. The present discussion should clarify that didactics also looks at learning. Thus, if VPH would suggest that criticism from didactics would not apply to their research on learning, then they again would show that they lack in understanding of didactics. Also, such suggestion would be disingenuous since VPH and KNAW (2009:10) point 2 clearly draw conclusions w.r.t. the effectiveness of TME and RME, and thus encroach upon didactics, even while the KNAW committee did not have members with a background in didactics of mathematics.

[88] https://www.ipetitions.com/petition/tk-onderzoek-wiskundeonderwijs/

(3) The causal models are useful for this analysis in institutional economics on the math war. VPH still presented an analysis on *s* and the cause *d,* as if there would be no evidence for a relevant difference of effect size between TME and RME, while didactics clearly shows that TME has logic on its side. We also see the problem of the many hands and shared responsibility, when a committee takes over. It were mathematician and chairman Jan Karel Lenstra (without a background in didactics of mathematics) and his full KNAW committee (including Van Putten with assistance by Hickendorff), who supported the invalid analysis. Committee members should respond to criticism *also afterwards,* and not hide behind the committee itself.

5. Development in 2017-2018

5.1. A 2017 study for NRO and IvhO

Hickendorff et al. (2017:24) [89] is a repeat review study commissioned by the Inspectorate of Education (IvhO) with administrative intermediary NRO. The authors qualify their review as "narrative" as opposed to a quantitative meta-analysis. Remarkably, this 2017 "narrative review" still excludes Van der Plas (2009) or my criticism (which one migh qualify as "narrative" too since those don't rely on statistics but on logic). Hickendorff et al. (2017) finally acknowledge, still confusing "empirical" with "statistical" (p24):

> Google Translate 2018: "Finally, the focus on empirical research limits the scope of the research by not addressing important theories about learning in general and [didactics of arithmetic] in particular." [90]

Thus, while KNAW (2009) deliberately restricted its attention to statistical findings, Hickendorff at al. (2017) finally agree that such an approach has limited meaning. Yet, not for their own study in 2017 but as recommendation for future research.

However, their comment tends to imply a claim that they are competent to judge upon the importance of didactic theories. Hickendorff already stated her lack of expertise. Co-author T.M.M. Mostert has a MSc degree in "Education and Child Studies", that indeed looks into "factors that effect reading and arithmetic", but this might not be didactics of arithmetic. [91] Co-authors C.J. van Dijk and L.L. van der Zee apparently have no Leiden page. Co-author L.L.M. Jansen [92] has a MSc degree in "Education and Child Studies", and some of her keywords are "mathematics" and "mathematics education" while these do not seem to be covered by her training. Co-author M.F. Fagginger Auer [93] has a background in developmental psychology and a Ph.D. in "methodology and statistics", [94] and its topic appears to be related to the thesis by Hickendorff. My inference is that these authors likely don't have the expertise to really judge that didactics of arithmetic would be relevant. It must be a cheap remark. A symptom is that they did not include such a researcher in their review team.

The subsequent critical question for Hickendorff et al. (2017) would be: who would be the judges on didactics of arithmetic ? If you hire TME then they will reject RME and if you hire RME then they will reject TME. Since the KNAW (2009) word of power there tends to be a new attitude "to take the best of each", without clear criteria what would be "the best", thus with a soup that neglects the discussion before that KNAW (2009) misdirection. Hickendorff et al. do not discuss this moot question who would judge about didactics. Potentially these authors might still think that statistical outcomes would determine which didactics would be

[89] https://www.nro.nl/nro-projecten-vinden/?projectid=405-17-920-rekenen%20op%20de%20basisschool
[90] Dutch original: "Ten slotte beperkt de focus op empirische onderzoeken de reikwijdte van het onderzoek doordat niet wordt ingegaan op belangrijke theorieën over leren in het algemeen en rekenwiskundedidactiek in het bijzonder."(p24)
[91] https://www.universiteitleiden.nl/en/staffmembers/terry-mostert#tab-1 and https://www.linkedin.com/in/terry-mostert-617830ba/
[92] https://www.universiteitleiden.nl/en/staffmembers/lisa-jansen#tab-1 and https://www.linkedin.com/in/lisa-jansen-2146ab65/
[93] https://www.linkedin.com/in/marijefaggingerauer/
[94] https://www.narcis.nl/research/RecordID/OND1344773/Language/en

“the best” (with some thin air to drop whoever frames the test questions and determines what the proper answers would be).

Overall, Holland has heavily invested in educational degrees such as “Education and Child Studies” and “education management” and Holland suffers a math war, but Holland never got around to set up a decent research line in the empirical science of didactics of mathematics. KNAW (2009) should have advised to abolish the Freudenthal Head in the Clouds Realistic Mathematics Institute at the University of Utrecht, that pushed RME without proper testing, but the misery continued thanks to the incompetence and arrogance of these psychometricians and child educationalists. [95]

5.2. Their claimed result

The Hickendorffet al. (2017) main conclusion is:

> Google Translate 2018: “This means that in the current situation no more than 10 percent of the differences in [arithmetic] performance can be explained by (influenceable and non-influenceable) factors in the educational process.” (p95) [96]

They used TIMSS 2015 (Grade 4) en PPON 2011 (Grade 6). [97] We already observed that by 2009 all Dutch textbooks used RME, and thus it should not surprise that these data show less variation in 2011. The TME textbook “Reken Zeker” was started in 2010, but their students reached Grade 6 only in 2016. Why did Hickendorff et al. (2017) not use my suggestion on using the results of 2016 ? Perhaps though, such would be “original research” and not a “review” study, and if the principal asks for a review then you as an agent might not offer the idea that something better is possible.

Their effect measure is still the outcomes of sums, and they do not explicitly refer to the intended algebra in highschool. Hickendorff et al. (2017) still accept the current tests as valid, though we have seen that they are biased towards RME. These researchers claim to study “education in arithmetic” while in fact they study what RME has created under this false label.

After the KNAW (2009) criticism that adequate studies lacked, the education researchers in Holland in particular the Freudenthal Head in the Clouds Realistic Mathematics Institute (FHCRMI) in Utrecht did not succeed in setting up an adequate study in 2010-2016, and Hickendorff et al. (2017) still only find Slavin & Lake (208) as the only relevant one. They refer uncritically to the math war in the USA, see our discussion below:

> Google Translate (2018): “The teaching method used is often part of a debate about mathematical education (Slavin & Lake, 2008). [Only one single] review was found of the effects of teaching methods on the [arithmetic] performance of primary school students. Slavin and Lake (2008) concluded on the basis of the median of the effect sizes found that [arithmetic] methods have a negligible to small effect on mathematical performance. Such small positive effects were found for various types of [arithmetic] methods. In general, this review therefore provides little evidence for the proposition that different [arithmetic] methods have different effects on [arithmetic] performance. A comparison with other

[95] Dutch readers may benefit from criticism by Imelman, Wagenaar and Meijer 2017, http://webwinkel.vangorcum.nl/NL_toonBoek.asp?PubIID=5095-0 http://www.beteronderwijsnederland.nl/vakwerk/2018/02/imelman-politiek-pedagogiek/ https://www.beteronderwijsnederland.nl/nieuws/2016/09/in-gesprek-met-prof-dr-imelman/ and also these sources: https://historiek.net/vier-pioniers-van-de-pedagogiek/49844/ https://www.dub.uu.nl/nl/artikel/langeveld-de-tragiek-van-een-befaamd-hoogleraar

[96] Dutch:“Dat betekent dat in de huidige situatie hoogstens 10 procent van de verschillen in rekenprestaties verklaard kan worden door (beïnvloedbare en nietbeïnvloedbare) factoren uit het onderwijsleerproces.” (p95)

[97] Scheltens et al. (2013) (with contribution by Hickendorff) “Balans van het rekenwiskundeonderwijs aan het einde van de basisschool 5. Uitkomsten van de vijfde peiling in 2011", PPON-reeks nummer 51, CITO. https://zoek.officielebekendmakingen.nl/blg-219337.pdf

studies in the review showed that the associated instructional guidance is a more important factor." (p59) [98]

In their study, TME is called "direct instruction" and RME is called "constructivist instruction". The didactics are also referred to as "calculation methods", likely without intending to be denigrating but nevertheless still condescending w.r.t. didactics of mathematics. In Holland, the term "method" is also used for a particular textbook (-series). In the USA the term "curriculum" may be used for a textbook as well. The PPON 2011 study introduces a confusion by using the word "calculation methods" for textbooks too. Its table 9.2 on page 300 compares "calculation methods" but this is erroneous, because this compares textbooks that all use RME. There is no comparison between RME and TME on arithmetic. The conclusion of PPON 2011, that there is hardly difference between the "methods", should not be seen as a conclusion pertaining to the difference between TME and RME, but only pertains to different RME textbooks. When Hickendort et al. (2017) page 13 & 19 also claim that "calculation methods" hardly differ in results, they might adopt this confusion of PPON 2011 too.

By again excluding Van der Plas (2009) and my criticism, Hickendorff et al. (2017) again manage to conclude that "robust" results would be lacking, while TME has logic on its side:

> Google Translate 2018: "It is striking that there are no robust research results with regard to subject matter or calculation method: neither in the international literature nor in the further analyzes of PPON-2011 and TIMSS-2015. Although the importance of these factors is obvious (see also Hiebert & Grouws, 2007; Van Zanten & van den Heuvel-Panhuizen, 2014), it seems difficult to investigate this in a targeted manner. This may be due to the fact that the terms are very broad, the curriculum is strongly related to the legal reference levels and therefore there is little variation in supply because of the used calculation method is related to other school and teacher factors that affect the effects of calculation method can not be determined accurately, or because teachers vary the calculation method use." (p96) [99]

While the authors squeeze in a reference to some didactics, as RME Van Zanten & van den Heuvel-Panhuizen 2014, they still refuse to mention Van der Plas (2009).

In their recommendations for future research they include "calculation methods" – which might mean "didactics of arithmetic" (Dutch "didactiek van de rekenkunde") – but again fail to mention my suggestion to look at the 2016 results on "Reken Zeker".

> Google Translate 2018: "We recommend further research on the following themes: the pedagogical subject knowledge of the teacher, the role of the [arithmetic]

[98] Dutch: "De gebruikte lesmethode is vaak onderdeel van debat over het rekenonderwijs (Slavin & Lake, 2008). Naar de effecten van lesmethoden op de rekenprestaties van basisschoolleerlingen is één review gevonden. Slavin en Lake (2008) concludeerden op basis van de mediaan van de gevonden effectgrootten dat rekenmethoden een verwaarloosbaar tot klein effect hebben op rekenprestaties. Dergelijke kleine positieve effecten werden voor diverse soorten rekenmethoden gevonden. Over het algemeen komt uit deze review dus weinig bewijs naar voren voor de stelling dat verschillende rekenmethoden verschillende effecten hebben op rekenprestaties. Uit een vergelijking met andere studies in de review bleek dat de bijbehorende instructiebegeleiding een belangrijkere factor is." (p59)

[99] Dutch:"Opvallend is dat er geen robuuste onderzoeksresultaten zijn met betrekking tot leerstofaanbod of rekenmethode: noch in de internationale literatuur, noch in de nadere analyses van PPON-2011 en TIMSS-2015. Hoewel het belang van deze factoren voor de hand ligt (zie ook Hiebert & Grouws, 2007; Van Zanten & van den Heuvel-Panhuizen, 2014) lijkt het moeilijk deze gericht te onderzoeken. Mogelijk komt dit doordat de begrippen heel breed zijn, het leerstofaanbod sterk samenhangt met de wettelijke referentieniveaus en er daarom weinig variatie in aanbod bestaat, doordat de gebruikte rekenmethode samenhangt met andere school- en leerkrachtfactoren waardoor de effecten van rekenmethode niet zuiver te bepalen zijn, of doordat leerkrachten de rekenmethode verschillend gebruiken." (p96)

coordinator and [arithmetic] policy / vision of the school, and the calculation methods (content and use by teachers)." (p25) [100] (See their p101.)

5.3. The math war in the USA

Slavin & Lake (2008) is a meta-study that included 87 studies. It is quite possible that these studies do not deal properly with the distinction between TME and RME (and properly re-engineered mathematics education). Slavin currently is director of the Center for Research and Reform at John Hopkins. By training, Slavin is a psychologist and Lake a sociologist. It is not clear to me what their research in didactics of mathematics has been.

S&L p431: "The purpose of this review is to examine the quantitative evidence on elementary mathematics programs to discover how much of a scientific basis there is for competing claims about the effects of various programs. (...) A broad literature search was carried out in an attempt to locate every study that could possibly meet the inclusion requirements." It is important to realise that the USA has still much variety of TME and RME, compared to the dominance of RME in Holland. Thus the USA is better placed to show a difference. My problem is not the use of quantitative methods but the validity of what is measured. For example, KNAW (2009) excluded Van der Plas (2009) perhaps because of lack of statistics but the study was of key importance for validity. We might run into the same problem with the S&L study.

On the other hand, S&L p436 is informative on the math war in the USA. It relates how the NSF funded "reform mathematics" programs but without requiring proper testing: "Yet, experimental control evaluations of these and other curricula that meet the most minimal standards of methodological quality are very few. Only five studies of the NSF programs met the inclusion standards, and all but one of these was a post hoc matched comparison." The post hoc approach suffers the risk of selection bias or censoring, with schools dropping a textbook that doesn't work for them. [101]

S&L show that they are not quite aware of didactics of mathematics and the relation of arithmetic to algebra, when they state (p482): "This is not to say that curriculum is unimportant. There is no point in teaching the wrong mathematics. The research on the NSF supported curricula is at least comforting in showing that reform-oriented curricula are no less effective than traditional curricula on traditional measures, and they may be somewhat more effective, so their contribution to nontraditional outcomes does not detract from traditional ones." Do their studies grade algorithms by steps or do they only look at the outcomes ?

While S&L indicate that RME would give a slightly better median effect size of 0.1, the following indicates that TME could do better with a particular effect size of 0.22.

Namely, my problem now is that Hickendorff et al. (2017) refer to Slavin & Lake (2008) of 9 years earlier. If they had studied the S&L paper more thoroughly, they would have seen that S&L refer to a What Works Clearinghouse 2006 study that wasn't published yet at the time when S&L were writing. In 2017, Hickendorff et al. could have looked. For example, I find this 2013 NCEE Evaluation Brief "After two years, three elementary math curricula outperform a fourth". [102] The outperformed textbook / curriculum is called "*Investigations*" supported by TERC [103] and it is of the RME kind, while the other three are of the TME kind. The Brief p7: "This 0.22 difference (also known as an "effect size") means that a study student at the 50th percentile in math would score 9 percentile points higher as a result of being taught in 1st and 2nd grade with Math Expressions, Saxon, or SFAW/enVision instead of with Investigations."

[100] Dutch: "Wij bevelen nader onderzoek aan op de volgende thema's: de pedagogisch vakinhoudelijke kennis van de leerkracht, de rol van de rekencoördinator en rekenbeleid/-visie van de school, en de rekenmethoden (inhoud en gebruik door leerkrachten)." (p25)
[101] S&L p434: "Despite all of these concerns, post hoc studies were reluctantly included in this review for one reason: Without them, there would be no evidence at all concerning most of the commercial textbook series used by the vast majority of elementary schools."
[102] https://ies.ed.gov/ncee/pubs/20134019/pdf/20134019.pdf
[103] https://www.terc.edu

Using a conversion table: [104] with a class of 25 this means 2 more students switching from Fail to Pass. “Even Cohen’s ‘small’ effect of 0.2 would produce an increase from 50% to 58% – a difference that most schools would probably categorise as quite substantial.”

NYC Hold is of the TME conviction, and their 2008 review [105] [106] of *Investigations* indicates that the statistical exercise by NCEE / IES was rather superfluous, and needlessly unkind to the pupil guinea pigs, like Ralph Nader testing whether car safety belts really are useful. This only concerned Grade 1 and 2. In itself the 0.22 standard deviation is less than I would expect, but this would also require a look at the Rock & Pollack 2002 ECLS-K test [107] used, getting us further from our present focus on the math war in Holland and getting distracted by the math war in the USA. For due process, let me refer to a remarkably positive EdReport’s review [108] of *Investigations* and a reply by the authors on remaining criticism. [109]

6. Conclusions

For this analysis in institutional economics, the causal modeling on didactics and testing on competence in arithmetic and algebra, with a focus on long term memory of pupils, appeared illuminating for understanding the role of formal and informal institutions. Agents in formal institutions on education and its research are most likely influenced by informal institutions that are given by durable ideas and conceptions that do not change easily, in this case on traditional and “realistic” approaches to mathematics education and its research, and on notions what exactly would constitute scientific research and ideas how logic and statistics relate to empirics. The causal modeling provided a framework to understand empirical developments in Holland on mathematics education and its research, also as factors in the overall economy – again see *Elegance with Substance* (2009, 2015).

Psychologists Van Putten & Hickendorff (VPH) and Hickendorff (2011) incorrectly excluded Van der Plas (2009) from their (review) study by confusing empirical science and statistics, while the empirical science of didactics of mathematics would warrant its inclusion. The KNAW (2009) committee had a biased composition without didacticians of arithmetic and algebra and did not correct the error. VPH and KNAW neglect criticism on their conceptual error which is a breach of research integrity. The scientific record must be corrected by removing these "publications" VPH (2009) and KNAW (2009) and Hickendorff (2011) that have been produced with these breaches.

Given that I have no reason to question personal or professional integrity of these psychometricians, my most likely explanation is the ivory tower, in which VPH really adopt these distorted concepts from conventional psychometrics, to insulate themselves from criticism. [110] But this ivory tower or tunnel vision is not science. Science is open minded. It actually doesn't quite matter what confusions VPH have chosen to neglect criticism. Fact is that they breach scientific integrity by selecting their sources and neglecting criticism. That Dutch procedures on research integrity are deficient has not been discussed here.

The KNAW (2009) conclusion that the empirical data in 2009 did not show a difference in effectiveness of TME and RME is false and based upon invalid research and deliberate neglect of information to the contrary. Their position in 2009 can be compared to a position in 1950 that “there is no statistical study that shows that the Moon has another side”. With proper tests, that score points for steps in the traditional algorithms in arithmetic, TME should obviously score better than RME that has insufficient training on those algorithms. KNAW (2009) confuses an issue of logic with statistics. Statistics are relevant for effect sizes on particular cases but have limited value for decisions upon principles for curriculum design.

---

[104] www.leeds.ac.uk/educol/documents/00002182.doc
[105] http://www.nychold.com/terc.html
[106] http://wgquirk.com/TERC.html
[107] https://eric.ed.gov/?id=ED470320
[108] https://www.edreports.org/math/investigations-in-number-data-and-space-3rd-edition-2017/index.html
[109] https://investigations.terc.edu/wp-content/uploads/2017/05/AuthorResponse.pdf?x71805
[110] Maltreat $s = f[d, o]$. Science = statistics. Take the effect measure as outcome only and neglect steps. Neglect future algebra. Expertise is a flexible concept.

Measurements are relevant for student diagnostics which didactics would work for them for particular stages in a curriculum, and such measurements might also be used for statistical reporting, but one should not confuse the purpose of this exercise for something else. Diagnosing students is something else than the KNAW (2009) exercise of trying to stop the social nuisance of a math war between ideologues who misrepresent propaganda as scientific research.

The Freudenthal Head in the Clouds Realistic Mathematics Institute (FHCRMI) at Utrecht University should be abolished as unscientific and comparable to astrology, alchemy or homeopathy. The RME section there has teamed up since 2009 with the STEM researchers so that there is more body to empirical research in education, but this remains a cover up of the unscientific RME core. After being warned by KNAW (2009) they still did not manage in 2010-2016 to set up a distinguishing experiment, as Hickendorff et al. (2017) observes. Holland better sets up a Simon Stevin Institute for mathematics education and its research.

There remains the statistical question of the unknown effect size of TME over RME in a PPON registration. This likely can be found by looking at the Dutch PPON 2016, and going back to the school archives to recover the data on SES and other variables for the 20 schools that adopted the textbook “Reken Zeker” in 2010, and a control group of normal (RME) students. It must be regretted that this suggestion by Colignatus (2015c) for PPON 2016 was not adopted in time. The VPH neglect of criticism was a factor in the neglect of that suggestion.

References

PM 1. Colignatus is the name in science of Thomas Cool, econometrician (Groningen 1982) and teacher of mathematics (Leiden 2008).
PM 2. References in footnotes need not be mentioned here.
PM 3. Commentary by anonymous researchers is acknowledged and has contributed to the quality of the present analysis.

Colignatus (2001, 2014), “Voting Theory for Democracy”, https://zenodo.org/record/291985
Colignatus (2009, 2015), “Elegance with Substance”, mijnbestseller.nl, https://mpra.ub.uni-muenchen.de/66012/ and https://zenodo.org/record/291974
Colignatus (2014, 2015), "Pierre van Hiele, David Tall and Hans Freudenthal: Getting the facts right", http://arxiv.org/abs/1408.1930
Colignatus (2015, 2018), “A child wants nice and no mean numbers”, https://doi.org/10.5281/zenodo.774272, 2nd edition 2018 forthcoming
Colignatus (2015c), "Het rekenexperiment op kinderen moet en kan worden beëindigd" http://www.wiskundebrief.nl/721.htm#5
Grift, W.J.C.M. van de (2010), “Ontwikkeling in de beroepsvaardigheden van leraren”, https://www.rug.nl/education/lerarenopleiding/onderwijs/oratie-van-de-grift.pdf
Hickendorff, M. (2011). "Explanatory latent variable modeling of mathematical ability in primary school : crossing the border between psychometrics and psychology", https://openaccess.leidenuniv.nl/handle/1887/17979
Hickendorff, M., T.M.M. Mostert, C.J.van Dijk, L.L.M. Jansen, L.L.van der Zee, & M.F. Fagginger Auer (2017), “Rekenen op de basisschool. Review van de samenhang tussen beïnvloedbare factoren in het onderwijsleerproces en de rekenwiskundeprestaties van basisschoolleerlingen”, https://www.nro.nl/wp-content/uploads/2017/12/405-17-920-010-Rapport-NRO-Review-Rekenenen.pdf
KNAW (2009), "Rekenonderwijs op de basisschool", https://www.knaw.nl/nl/actueel/publicaties/rekenonderwijs-op-de-basisschool
Plas, L. van der (2009), “Rekenvaardigheid in relatie tot wiskunde”, Orthopedagogiek (TvO), may 2009 (no 5), 48, 205-211, http://www.liesbethvanderplas.nl/rekenvaardigheid-in-relatie-tot-wiskunde/
Putten, C.M. van (2008), "De onmiskenbare daling van het prestatiepeil bij de bewerkingen sinds 1987", http://media.leidenuniv.nl/legacy/putten_reactie%20op%20Treffers%20in%20PPost.pdf
Putten, C.M. van, M. Hickendorff (2009), “Peilstokken voor Plasterk: Evaluatie van de rekenvaardigheid in groep 8”, Tijdschrift voor Orthopedagogiek, 48, 184-194
Slavin, R.E., and C. Lake (2008). Effective programs in elementary mathematics: A best evidence synthesis. Review of Educational Research, 78(3), 427–455