Multilevel latent class analysis for large-scale educational assessment data: Exploring the relation between the curriculum and students' mathematical strategies

Fagginger Auer, M.F.; Hickendorff, M.; Putten, C.M. van; Béguin, A.A.; Heiser, W.J.

Routledge
Taylor & Francis Group

Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: https://www.tandfonline.com/loi/hame20

Multilevel Latent Class Analysis for Large-Scale Educational Assessment Data: Exploring the Relation Between the Curriculum and Students' Mathematical Strategies

Marije F. Fagginger Auer, Marian Hickendorff, Cornelis M. Van Putten, Anton A. Béguin & Willem J. Heiser

To cite this article: Marije F. Fagginger Auer, Marian Hickendorff, Cornelis M. Van Putten, Anton A. Béguin & Willem J. Heiser (2016) Multilevel Latent Class Analysis for Large-Scale Educational Assessment Data: Exploring the Relation Between the Curriculum and Students' Mathematical Strategies, Applied Measurement in Education, 29:2, 144-159, DOI: 10.1080/08957347.2016.1138959

To link to this article: https://doi.org/10.1080/08957347.2016.1138959

Accepted author version posted online: 11 Jan 2016.
Published online: 03 Mar 2016.

APPLIED MEASUREMENT IN EDUCATION
2016, VOL. 29, NO. 2, 144–159
http://dx.doi.org/10.1080/08957347.2016.1138959

Routledge
Taylor & Francis Group

Multilevel Latent Class Analysis for Large-Scale Educational Assessment Data: Exploring the Relation Between the Curriculum and Students' Mathematical Strategies

Marije F. Fagginger Auer[a], Marian Hickendorff[b], Cornelis M. Van Putten[a], Anton A. Béguin[c], and Willem J. Heiser[a]

[a]Institute of Psychology, Leiden University; [b]Institute of Education and Child Studies, Leiden University; [c]CITO (Dutch National Institute for Educational Measurement)

ABSTRACT

A first application of multilevel latent class analysis (MLCA) to educational large-scale assessment data is demonstrated. This statistical technique addresses several of the challenges that assessment data offers. Importantly, MLCA allows modeling of the often ignored teacher effects and of the joint influence of teacher and student variables. Using data from the 2011 assessment of Dutch primary schools' mathematics, this study explores the relation between the curriculum as reported by 107 teachers and the strategy choices of their 1,619 students, while controlling for student characteristics. Considerable teacher effects are demonstrated, as well as significant relations between the intended as well as enacted curriculum and students' strategy use. Implications of these results for both more theoretical and practical educational research are discussed, as are several issues in applying MLCA and possibilities for applying MLCA to different types of educational data.

Latent class analysis (LCA) is a powerful tool for classifying individuals into groups based on their responses on a set of nominal variables (Hagenaars & McCutcheon, 2002; McCutcheon, 1987). LC models have a categorical latent (unobserved) variable, and every class or category of this latent variable has class-specific probabilities of responses in the categories of the different observed response variables. As such, each latent class has a specific typical response pattern where some responses have a higher and others have a lower probability, and different response profiles of individuals may be discerned based on this. For example, for a test covering language, mathematics and science, one latent class of students may have a high probability of correct responses for mathematics and science items but a lower probability for language items, while for an other latent class the probability of a correct response is high for language items and lower for mathematics and science items. These two classes then reflect different performance profiles.

Relatively recently, the technique of LCA has been extended to accommodate an additional hierarchical level (Vermunt, 2003): not only the nesting of variables within individuals is included in the model, but also the nesting of individuals in some higher-level group (e.g., students within school classes). This multilevel LCA (MLCA) is beginning to be applied more and more in various areas, such as psychiatry (Derks, Boks, & Vermunt, 2012), political science (Morselli & Passini, 2012), and education (Hsieh & Yang, 2012; Mutz & Daniel, 2011; Vermunt, 2003). In the current investigation, we describe a first application of MLCA to educational large-scale assessment data.

CONTACT Marije F. Fagginger Auer m.f.fagginger.auer@fsw.leidenuniv.nl Psychology, Methodology & Statistics, Leiden University, PO Box 9555, 2300 RB Leiden, Leiden, Netherlands.

MLCA for educational large-scale assessment data

MLCA can address several of the challenges of large-scale assessment data. A first challenge that many large-scale assessments offer is that they employ so-called incomplete designs: the complete item set is too large to be administered in full to students, and is therefore decomposed into smaller subsets. Relating these subsets to each other is difficult using traditional techniques, but is possible using a latent variable to which all items are related (Embretson & Reise, 2000; Hickendorff, Heiser, Van Putten, & Verhelst, 2009), such as the latent class variable in LCA. No imputation of missing responses on the items that were not administered is necessary, as the likelihood function of the analysis is only based on cases' observed responses (Vermunt & Magidson, 2005). A second challenge is the complexity of modeling cognitive phenomena that are not measured on an interval but on a nominal level (such as solution strategy use, item correctness or error types). Nominal response variables are naturally accommodated by (M)LCA.

The third challenge that MLCA addresses is the inherent multilevel structure of educational data (items nested within students, who are nested within teachers and schools). Previous applications of LCA (and also of other techniques) to students' responses on cognitive tests have generally ignored the teacher (or school) level in their modeling (e.g., Geiser, Lehman, & Eid, 2010; Hickendorff et al., 2009; Hickendorff, Van Putten, Verhelst, & Heiser, 2010; Lee Webb, Cohen, & Schwanenflugel, 2008; Yang, Shaftel, Glasnapp, & Poggio, 2005). Yet, the context of learning is vital to its outcomes. Zumbo et al. (2015) recently proposed an ecological model of item responding where responses are influenced by contextual variables at various levels: characteristics of the test, of the individual, of the teacher and school, of the family and ecology outside of school, and of the larger community. Based on this model, the authors demonstrate ecologically moderated differential item functioning (DIF) where different factors in this broader context play a role.

The consideration of a broader context fits in very well with MLCA, as its multilevel aspect makes it especially suited for the incorporation of contextual factors in models of students' item responses. Predictors at different hierarchical levels can be included in the model, a feature that is naturally called for in modeling the effects of both student and teacher characteristics on students' item solving.

In the current investigation, we therefore demonstrate the use of MLCA for educational large-scale assessment data, by applying it to data from the most recent large-scale assessment of Dutch sixth graders' mathematics. We investigate the relation between the curriculum on the one hand and students' use of solution strategies on the other (while controlling for student characteristics), and describe the technique of MLCA and some of the challenges in its application in more detail.

Curriculum effects on students' mathematical achievement and strategies

Recent reviews of research on the effects of mathematics teaching have concluded that the influence of the intended curriculum (as it is formally laid down in curriculum guides and textbooks; Remillard, 2005) on achievement is very small, while changes in the enacted curriculum of daily teaching practices have a much larger influence (Slavin & Lake, 2008). These findings are based mainly on small experiments, and can be supplemented using large-scale assessment data, which does not allow for causal inference but does offer much larger samples and representative descriptions of the natural variation in daily teaching practices (Slavin, 2008).

Previous research has indicated that this variation in instruction has substantial effects on students' achievement growth (Nye, Konstantopoulos, & Hedges, 2004; Rowan, Correnti, & Miller, 2002). In identifying the factors that determine teachers' influence on students' mathematical achievement, a line of research called "education production function research" has focused on the effects of available resources. Generally, routinely collected information on teachers' resources (such as their education level) has failed to show consistent, sizable effects (e.g., Jepsen, 2005; Nye et al., 2004; Wenglinsky, 2002), while more in-depth teacher resource measurements (such as knowledge for mathematical teaching) show more consistent positive effects (Hill & Rowan, 2005; Wayne & Youngs, 2003). The more process-focused line of "process-product research" has most

notably found positive effects of active teaching, which involves teachers' direct instruction of students in formats such as lecturing, leading discussions, and interaction during individual work (as described by Hill et al., 2005; and Rowan et al., 2002), as contrasted with frequent independent work of students and working on nonacademic subjects. Also, positive effects have been found of reform-oriented classroom practice, which involves activities such as exploring possible methods to solve a mathematical problem (Cohen & Hill, 2000).

These results all concern curriculum effects on students' mathematical *achievement*, but the mathematical *strategies* of students that are the focus of this investigation are also of great interest. The various reforms in mathematics education that have taken place in a number of countries in the past decades (Kilpatrick, Swafford, & Findell, 2001) share a view on strategy use that moves away from product-focused algorithmic approaches toward process-focused approaches with more space for students' own strategic explorations (Gravemeijer, 1997). Investigating which instructional practices elicit particular patterns of strategy choices may shed light on how reforms actually affect students' behavior. On a more theoretical level, the literature on children's choices between and performance with mathematical strategies has so far focused on the effects of children's individual characteristics and of the nature of the mathematical problems that are offered (e.g., Hickendorff et al., 2010; Imbo & Vandierendonck, 2008; Lemaire & Lecacheur, 2011; Lemaire & Siegler, 1995), and may therefore be extended by also exploring the effects of instruction.

Multidigit multiplication and division strategies in the Netherlands

An illustration of the connection between mathematics reforms and changes in strategy choices is provided by previous research on multidigit multiplication and division strategies in the Dutch situation (Hickendorff, 2011; Janssen, Van der Schoot, & Hemker, 2005). Multidigit multiplication and division go beyond simple multiplication table facts (such as 5×6 or $72 \div 8$) and require operations on larger numbers or decimal numbers (such as 56×23 or $544 \div 16$). The Dutch mathematics education reform introduced new algorithmic "whole-number-based" approaches for these multidigit operations, where every step toward obtaining the solution requires students to understand the magnitude of the numbers they are working with (Treffers, 1987). This approach deviates from the more traditional "digit-based" algorithms, where the numbers are broken up into digits that can be handled without an appreciation of their magnitude in the whole number (see Table 1 for examples of both algorithms). In general, Dutch children's learning trajectory consists of first learning the whole-number-based multiplication and division algorithms, and later switching to the digit-based algorithm for multiplication (and in some schools, also for division; Buijs, 2008).

Using data from large-scale assessments, it was demonstrated that with growing adoption of reform-based mathematics textbooks in Dutch elementary schools, many primary school students abandoned the digit-based algorithms for multidigit multiplication and division and switched to answering without writing down any calculations (mental calculation; Hickendorff et al., 2010) instead. These mental calculation strategies were found to be much less accurate than written strategies (digit-based or other) (Hickendorff, 2011; Hickendorff et al., 2009), and were used more by boys, students with low mathematical proficiency, and lower SES students.

The present study

In the present study, MLCA is used to investigate the relation between both the intended and enacted curriculum and the use of solution strategies for multidigit multiplication and division items by 1,619 Dutch sixth graders (11–12-year-olds). The intended curriculum is operationalized as the mathematics textbook and the enacted curriculum as the self-reports on mathematics teaching practices of the students' 107 teachers. The data are from the most recent (2011) large-scale national assessment of the mathematical abilities of Dutch students at the end of primary school (Scheltens, Hemker, & Vermeulen, 2013).

Table 1. Examples the digit-based algorithms, whole-number-based algorithms, and non-algorithmic strategies applied to the multiplication problem 23 × 56 and the division problem 544 ÷ 34.

Strategy	Multiplication	Division
digit-based algorithm	56 23× 168 1120+ 1288	34/544\16 34 204 204 0
whole-number-based algorithm	56 23× 18 150 120 1000+ 1288	544: 34 = 340–10× 204 102– 3× 102 102– 3×+ 0 16×
non-algorithmic written strategies	1120 + 3 × 56 1120 + 168 1288	10 × 34 = 340 13 × 34 = 442 16 × 34 = 544

Hypotheses

Based on previous research on Dutch students' multiplication and division strategy use by Hickendorff (2011), we expect to find a considerable group of students who mostly answer without written calculations (with relatively many boys, students with low mathematical proficiency, and lower socioeconomic status [SES] students), one group where students mostly use the digit-based algorithm, and one group where students mostly use the whole-number-based algorithm or non-algorithmic approaches. Hickendorff (2011) considered multiplication and division in isolation, but we consider them simultaneously and can therefore analyze the relation between individual differences in strategy use on multiplication and division items. For example, there may be a group of students who prefer the digit-based algorithm for multiplication and the whole-number-based algorithm for division, matching the most common end points of the respective learning trajectories.

The lack of research on the effects of the curriculum on strategy use makes it hard to make strong predictions in that area, but a tentative generalization of curriculum effects on achievement suggests that the effects of the enacted curriculum might be greater than those of the intended curriculum—although this could be countered by the fact that the mathematics textbooks that form the intended curriculum are an important direct source of strategy instruction. As for the particular effects of the enacted curriculum, the previously discussed achievement literature described positive effects of direct instruction rather than independent work, so these activities might affect choices for more accurate (written) or less accurate (mental) strategies. Differentiated instruction might also have such effects, especially because of the association between ability and strategy choices. Furthermore, we expect effects of teachers' strategy instruction in algorithms, mental calculation, and strategy flexibility, because of the apparent direct connection to students' strategy use.

Issues in applying MLCA

The application of MLCA with predictors which is the focus of the present study comes with several practical issues that require attention. The first is the specification of the multilevel effect in the model. The common parametric approach specifies a normal distribution for group (in our case, teacher) deviations from the overall parameter value, but this distributional assumption is strong and the interpretation of such group effects is abstract. The nonparametric approach proposed by Vermunt (2003) instead creates a latent class variable for the groups (in addition to the latent class variable for the individuals), requiring less strong distributional assumptions, making computations less intensive, and allowing for easier substantive interpretation. Therefore, we will use the nonparametric approach.

The second issue is the inclusion of predictors in the model, as discussed by Bolck, Croon, and Hagenaars (2004). In the so-called one-step approach, the measurement part of the model (the part of the model without predictors) and the structural part (the predictor part) are estimated simultaneously. While this leads to unbiased effect estimates, the number of models that needs to be fitted and compared can quickly become unfeasible (all combinations of lower level and higher level latent class structures, combined with all predictor structures). In addition, the structural part of the model may influence the measurement part: individuals' class membership may be different with and without predictors. These problems do not occur in the three-step approach, where the measurement model without any predictors is fitted first, then individual class membership predictions are computed, and finally these class membership predictions are treated as observed variables in an analysis with the predictors. However, this approach treats class membership as deterministic and leads to systematic underestimation of the effects of the predictors. This can be corrected by taking into account the misclassification in the second step during the final third step (Asparouhov & Muthén, 2014). Therefore, we will use this corrected three-step approach.

The third issue is the selection of the best model. This is usually done based on information criteria that consider model fit and complexity simultaneously, such as the popular Aikaike and Bayesian Information Criterion (AIC and BIC). However, these criteria penalize model complexity differently and therefore often identify different models as optimal (Burnham & Anderson, 2004). The issue is further complicated with the introduction of a multilevel effect, because the BIC penalization depends on sample size, and it is then unclear whether to use the number of individuals or groups for that (Jones, 2011). Lukočienė and Vermunt (2010) investigated this issue and demonstrate optimal performance of the group-based BIC, and underestimation of complexity by the individual-based BIC and overestimation by the AIC. In our analyses, model selection with all three criteria is compared.

Method

Sample

For our data from the most recent large-scale assessment of the mathematical abilities of Dutch students, 107 schools from the entire country were selected according to a random sampling procedure stratified by socioeconomic status. From a total of 2,548 participating sixth graders (11–12-year-olds) in those schools, 1,619 students from the classes of 107 teachers (one teacher per school, between 5 and 25 students per school in most cases) solved multidigit multiplication and division problems (because of the incomplete assessment design, not all students solved this type of problems). Of the 1619 children, 49% were boys and 51% were girls. Fifty percent of the children had a relatively higher general scholastic ability level, as they were to go to secondary school types after summer that would prepare them for higher education, while the other 50% were to go to vocational types of secondary education. In terms of SES, most children (88%) had at least one parent who completed at least two years of secondary school, while 12% did not.

Different mathematics textbooks were used on which the children's mathematics instruction was based. These textbooks are part of a textbook series that is used for mathematics instruction throughout the various grades of primary school, and are therefore not (solely) determined by the sixth grade teacher. All textbooks in our sample could be considered reform-based, but they differ in instruction elements such as lesson structure, differentiation, and assessment. Textbooks from six different methods were used in our sample: Pluspunt (PP; used by 37% of the teachers in our sample); Wereld in Getallen (WiG; 30%); Rekenrijk (RR; 14%); Alles Telt (AT; 11%); Wis en Reken (6%); and Talrijk (2%).

Materials

Multiplication and division problems

The assessment contained 13 multidigit multiplication and eight division problems, of which students solved systematically varying subsets of three or six problems according to an incomplete design (see Hickendorff et al., 2009, for more details on such designs). The problems are given in Table 2, including whether the problem to be solved was provided in a realistic context (such as determining how many bundles of 40 tulips can be made from 2,500 tulips). Students were allowed to write down their calculations in the ample blank space in their test booklets, and these calculations were coded for strategy use. Six categories were discerned: the aforementioned digit-based and whole-number-based algorithms, written work without an algorithmic notation (such as only writing down intermediate steps), no written work, unanswered problems, and other (unclear) solutions (see Table 1 for examples). The coding was carried out by the first and third author and three undergraduate students, and interrater agreement was high (Cohen's κ's (J. Cohen, 1960) of .90 for the multiplication and .89 for the division coding on average, based on 112 multiplication and 112 division solutions categorized by all).

Teacher survey about classroom practice

The teachers of the participating students filled out a survey about their mathematics teaching practices. The 14 questions in the survey that concerned multiplication, division, and mental calculation strategy instruction were used to create four scores (by taking the mean of the standardized responses to the questions), as were the 10 questions that concerned instruction formats, and the 10 questions that concerned instruction differentiation. The Appendix gives the questions that were used to create each score.

Table 2. The content of the 13 multidigit multiplication problems and eight multidigit division problems in the assessment, and the strategy use frequency on each item.

			Strategy use (percent)						
	Problem	Context	Digit	Number	Non-alg.	No written	Unanswered	Other	*n*
M01	9 × 48 = 432	yes	39	4	24	30	2	2	368
M02	23 × 56 = 1288	yes	45	6	21	17	5	6	358
M03	209 × 76 = 15884	no	49	5	24	12	7	3	344
M04	*35 × 29 = 1015*	yes	40	4	28	23	3	2	353
M05	*35 × 29 = 1015*	no	43	4	23	24	3	3	352
M06	24 × 37.50 = 900	no	39	2	31	18	6	5	352
M07	*9.8 × 7.2 = 70.56*	no	40	3	17	27	10	3	352
M08	*8 × 194 = 1552*	yes	43	3	25	27	2	1	355
M09	6 × 192 = 1152	no	33	2	33	23	4	5	352
M10	1.5 × 1.80 = 2.70	yes	1	0	13	79	3	4	353
M11	0.18 × 750 = 135	no	41	2	16	27	12	2	356
M12	6 × 14.95 = 89.70	yes	32	1	29	34	2	2	359
M13	*3340 × 5.50 = 18370*	yes	41	3	23	18	10	5	359
D01	*544 ÷ 34 = 16*	yes	18	32	5	27	10	7	368
D02	*31.2 ÷ 1.2 = 26*	no	9	10	6	50	18	7	369
D03	11585 ÷ 14 = 827.5	yes	17	30	4	32	10	7	345
D04	1470 ÷ 12 = 122.50	yes	19	25	11	31	12	3	350
D05	*1575 ÷ 14 = 112.50*	no	17	30	16	22	12	3	355
D06	*47.25 ÷ 7 = 6.75*	yes	17	25	10	33	10	5	352
D07	*6496 ÷ 14 = 464*	yes	16	24	5	36	12	7	354
D08	2500 ÷ 40 = 62	yes	12	15	11	45	6	11	359
total multiplication			37	3	24	28	5	3	4613
total division			16	24	9	35	11	6	2852

Parallel versions of problems not yet released for publication are in italics.

Multilevel latent class analysis

We estimated latent classes of students reflecting particular strategy choice profiles using MLCA, which classifies respondents in latent classes that are each characterized by a particular pattern of response probabilities for a set of problems (Goodman, 1974; Hagenaars & McCutcheon, 2002). For our case, let Y_{ijk} denote the strategy choice of student i of teacher j for item k. A particular strategy choice on item k is denoted by s_k. The latent class variable is denoted by X_{ij}, a particular latent class by t, and the number of latent classes by T. The full vector of strategy choices of a student is denoted by $\mathbf{Y_{ij}}$ and a possible strategy choice pattern by $\mathbf{s}$. This makes the model:

$$P\left(\mathbf{Y_{ij}} = \mathbf{S}\right) = \sum_{t=1}^{T} P\left(X_{ij} = t\right) \prod_{k=1}^{K} P\left(Y_{ijk} = s_k \middle| X_{ij} = t\right). \tag{1}$$

In this model, the general probability of a particular pattern of strategy choices, P ($\mathbf{Y_{ij}}$ = $\mathbf{s}$), is decomposed into T class-dependent probabilities, $\prod_{k=1}^{K} P$ (Y_{ijk} = $s_k|X_{ij}$ = t). These class-dependent probabilities are each weighted by the probability of being in that latent class, P (X_{ij} = t). The interpretation of the nature of the latent classes is based on the class-dependent probabilities of strategy choices on each of the problems, P ($Y_{ijk} = s_k|X_{ij} = t$). The model is extended with a multilevel component by adding a latent teacher class variable, on which students' probability of being in each latent student class (P (X_{ij} = t)) is dependent. Predictors at the teacher and student level that influence class probabilities can also be added, as described by Vermunt (2003, 2005). For such a multilevel model with one teacher-level predictor Z_{1j} and one student-level predictor Z_{2ij}, let W_j denote the latent teacher class that that teacher j is in, with m denoting a particular teacher class. The model then becomes:

$$P(X_{ij} = t|W_j = m) = \frac{\exp(\gamma_{tm} + \gamma_{1t}Z_{1j} + \gamma_{2t}Z_{2ij})}{\sum_{r=1}^{T} \exp(\gamma_{rm} + \gamma_{1r}Z_{1j} + \gamma_{2r}Z_{2ij})}. \tag{2}$$

See Henry and Muthén (2010) for graphical representations of this type of models.

The MLCA was conducted with version 5.0 of the Latent GOLD program (Vermunt & Magidson, 2013). All 13 multiplication and eight division strategy choice variables were entered as observed response variables and a teacher identifier variable as the grouping variable for the multilevel effect. Models with latent structures with up to eight latent student classes and 11 latent teacher classes were fitted, and the model with the optimal structure was selected using the AIC and BICs. Using the three-step approach (Bakk, Tekle, & Vermunt, 2013), this measurement model was then fixed and curriculum and student predictors were added to the model in groups, because of the high number of predictors. The successive models were compared using information criteria and the best model was investigated in more detail by evaluating the statistical significance of each of the predictors with a Wald test. The practical significance of the predictors was evaluated based on the magnitude of the changes in the probability of class memberships associated with different levels of the predictors. Effect coding was used for all predictors.

Results

The latent class measurement model

For the LC measurement models fitted on the strategy data, both the AIC and BICs (see Table 3) show that adding a multilevel structure greatly improves model fit, signifying a considerable within-teacher dependency of observations. While the AIC identifies a very complex model as optimal (10 latent teacher classes and six latent student classes), the BICs are in near agreement on a more simple model (four latent teacher classes and three or four latent student classes). Of these simpler models,

Table 3. Fit statistics for the non-parametric and parametric multilevel latent class models.

Latent classes					BIC	
Teachers	Students	Log-likelihood	Parameters	AIC	Individual-based	Group-based
1 (no multi-level effect)	2	−9801	209	20020	21146	20587
	3	−9388	314	19403	21096	20242
	4	−9165	419	19169	21427	20289
	5	−8964	524	18976	21800	20376
2	2	−9717	211	19856	20993	20419
	3	−9253	317	19141	20849	19988
	4	−8912	423	18670	20950	19800
	5	−8713	529	18484	21335	19898
3	2	−9707	213	19839	20987	20408
	3	−9207	320	19054	20779	19910
	4	−8819	427	18491	20792	19632
	5	−8614	534	18295	21173	19723
4	2	−9705	215	19840	20999	20415
	3	−9178	323	19002	**20743**	19865
	4	−8790	431	18441	20764	**19593**
	5	−8585	539	18248	21153	19688
5	2	−9705	217	19844	21013	21965
	3	−9220	326	19092	20849	19963
	4	−8866	435	18257	21189	19711
	5	−8584	544	18234	21167	19689
parametric	2	−9708	210	19836	20968	20397
	3	−9205	316	19042	20745	19887
	4	−8861	422	18566	20841	19694
	5	−8661	528	18377	21223	19789

The lowest BICs are bold. The lowest AIC was for 10 teacher and 6 student classes.

the model with four student classes has a much clearer interpretation and is also favored by the group-based BIC that is optimal according to Lukočienė and Vermunt (2010). This model has an entropy R^2 of .87 for the latent student classes and .82 for the teacher classes, which both indicate a high level of classification certainty (Dias & Vermunt, 2006).

We also estimated measurement models with a parametric rather than a non-parametric teacher effect (see the bottom part of Table 3). The parametric model with the lowest group-based BIC also had four student classes, and the class-specific probabilities of these classes were very similar to those of the classes in the non-parametric model (indicating very similar nature of the classes), but the classes differed considerably in size in the two approaches (by 13, 4, 25, and 15 percentage points, respectively). Latent teacher classes cannot be compared as there are none in the parametric approach, which also prevents later easy substantive interpretation of the multilevel effect. The fit of the best parametric model was not better than that of the best non-parametric model according to the information criteria, and the entropy R^2 for the student classes of the parametric model was lower (.80).

Latent student classes

Overall, students solved multiplication problems most often with the digit-based algorithm, while solutions without written work were most frequent for division (see Table 2 for frequencies for each strategy). The class-dependent probabilities of choosing each strategy in each of the four latent student classes are given in Table 4, which shows that every latent student class is dominated by high probabilities of choosing one or two strategies.

The largest student class (with a class probability of .31, i.e., containing 31% of students) is characterized by a high probability of answering without written work for every item, and also a considerable probability of leaving problems unanswered (especially division problems). Because of this, we label this class the "no written work class." The second largest student class (probability of .29) is characterized by a high probability of solving multiplication problems with the digit-based

Table 4. The mean probabilities of choosing each of the six strategies for the multiplication and division problems for each latent class.

	Mean probability of strategy choice (proportion students in class)							
	No written work class (.31)		Mixed algorithm class (.29)		Non-algorithmic class (.21)		Digit algorithm class (.20)	
Strategy	×	÷	×	÷	×	÷	×	÷
digit-based algorithm	.06	.01	**.71**	.01	.04	.03	**.68**	**.70**
whole-number-based alg.	.01	.02	.02	**.54**	.14	**.37**	.02	.01
non-algorithmic written	.25	.03	.15	.10	**.68**	.21	.16	.03
no written work	**.52**	**.65**	.10	.24	.08	.22	.10	.17
unanswered	.13	.23	.02	.06	.03	.08	.03	.03
other	.04	.05	.02	.05	.04	.10	.02	.06

The highest strategy probability per operation within a class is in boldface.

algorithm and a high probability of solving division problems with the number-based algorithm (the "mixed algorithm class"). The third largest student class (probability of .21) is characterized by a high probability of solving multiplication problems with non-algorithmic written strategies and a mixture of the number algorithm, non-algorithmic written strategies and no written work for the division problems (the "non-algorithmic written class"). The smallest student class (probability of .20) is characterized by a high probability of solving both multiplication and division problems with digit-based algorithms (the "digit-based algorithm class").

Latent teacher classes

The latent student class probabilities (or sizes) from Table 4 are the mean for all the teachers. Within the four latent teacher classes, the student class probabilities differ greatly. As can be seen in Table 5, the probability of the digit algorithm class varies most over teacher classes (between .00 and .74), followed by that of the mixed algorithm class (between .00 and .61), and that of the non-algorithmic written class (between .03 and .51). The probability of the no written work class varies relatively little over teacher classes (between .23 and .38). The largest teacher class (size of .39) is characterized by a high probability of the mixed algorithm class, the second largest teacher class (.30) by a high probability of the non-algorithmic written strategy class, the third largest teacher class (.19) by a high probability of the digit-based algorithm class, and the smallest teacher class (.12) by substantial probabilities for all classes except the non-algorithmic written class.

These insightful results on the magnitude and nature of teachers' effects illustrate one of the advantages of the nonparametric specification of the multilevel effect.

Adding predictors to the latent class model

Next, the structural part was added to the model: predictors for students' probability of being in a particular latent strategy class. First the relation between the intended and enacted curriculum (textbook and instruction) was investigated, using a MANOVA with textbook as the between-

Table 5. The latent student class probabilities in each of the four latent teacher classes.

	Latent student class probability			
Latent teacher class	No written work class	Mixed algorithm class	Non-algorithmic class	Digit algorithm class
1 (P = .39)	.27	**.61**	.11	.00
2 (P = .30)	.38	.08	**.51**	.02
3 (P = .19)	.23	.00	.03	**.74**
4 (P = .12)	.34	.22	.09	**.36**
Total	.31	.29	.21	.20

The highest latent student class probability within a latent teacher class is in boldface.

Table 6. Fit statistics for the latent class models with successively added predictors.

Predictors added to the model		Log-likelihood	Parameters	AIC	BIC	
					Individual	Group
none		−1651	15	3333	3414	3373
student characteristics	gender, ability, SES	−1569	24	3186	**3315**	**3250**
intended curriculum	textbook	−1550	36	3172	3366	3268
enacted curriculum	strategy instruction	−1517	48	3129	3388	**3257**
	instruction formats	−1500	60	3120	3443	3280
	instruction differentiation	−1479	72	**3103**	3491	3295

The lowest information criteria are in boldface.

group independent variable and the twelve teachers' instruction scores as the dependent variables. No significant relation was found, *Wilks'* λ = .57, F (48, 322) = 1.05, p = .39. Next, student characteristics and intended and enacted curriculum predictors were added to the model in a stepwise fashion. As can be seen in Table 6, according to both BICs model fit is best with only the student characteristics as predictors, whereas the AIC identifies the more complex model with all predictors as optimal. The group-based BIC is nearly as low for the model with the textbook and strategy instruction predictors added as for the model with only student predictors (3257 vs. 3250). Since curriculum effects were our primary interest, we chose to proceed with this more extensive model.

The statistical significance of the covariates in this model was evaluated with Wald tests, and the magnitude of the effects is illustrated by comparisons of the probabilities of membership of the latent student classes for individuals at the different levels of the predictors (see Table 7). These probabilities were calculated with all of the other selected predictors in the model set at their mean. For the interval-level instruction variables, probabilities are compared for students of teachers who score one standard deviation above the mean of that variable and students of teachers who score one standard deviation below the mean. Probabilities for the different levels of a predictor that differ by .10 or more are discussed.

Student characteristics

Student gender had a significant effect on class probabilities, W^2 = 107.1, p < .001, with the probability of being in the no written work class being .33 higher for boys than for girls. The probability of being in the mixed algorithm class was .17 higher for girls than for boys. Students' general scholastic ability also had a significant effect, W^2 = 53.0, p < .001, with the probability of being in the no written work class being .25 higher for students with a lower compared to a higher ability, and the probability of being in the non-algorithmic class .12 lower. SES also had a significant effect, W^2 = 8.4, p = .04, but class probability differences between children with a different SES were all smaller than .10.

Intended curriculum

Mathematics textbook had a significant effect, W^2 = 123.6, p < .001. Students being instructed from the PP textbook had a probability for the non-algorithmic class that is .14 higher than than that of the total, and a .13 lower probability for the digit-based algorithm class. Students with the RR textbook had a .16 lower probability for the digit algorithm class. Students with the AT textbook had a .16 lower probability of being in the mixed algorithm class and a .13 higher probability of being in the non-algorithmic written class. Students with other textbooks had .14 lower probability of being in the mixed algorithm class and a .14 higher probability of being in the digit algorithm class.

Enacted curriculum

All strategy instruction scores had significant effects. When comparing students whose teacher scored one standard deviation above the mean in their focus on the digit-based algorithm for multiplication to

Table 7. Students' probabilities of membership of the four latent student classes for different levels of the student characteristics and the intended and enacted curriculum predictors.

			Difference in probability of class membership [95% confidence interval]			
Predictor	Compared to		No written work	Mixed algorithm	Non-algorithmic	Digit algorithm
gender	boys	girls	+.**33** [+.31,+.34]	−.**17** [−.17,−.16]	−.09 [−.09,−.08] −.**12** [−.13,−.11]	−.07 [−.08,−.07] −.04 [−.05,−.04]
ability	lower	higher	+.**25** [+.23,+.26]	−.09 [−.09,−.09]	−.12 [−.13,−.11]	−.04 [−.05,−.04]
SES	low	not low	+.06 [+.03,+.09]	−.04 [−.05,−.03]	+.03 [+.02,+.05]	−.05 [−.07,−.04]
textbook	PP	total	+.04 [+.02,+.06]	−.05 [−.06,−.05]	+.**14** [+.13,+.14]	−.**13** [−.14,−.12]
	WiG	total	+.06 [+.04,+.07]	+.09 [+.09,+.10]	−.08 [−.07,−.09]	−.08 [−.08,−.07]
	RR	total	+.06 [+.03,+.09]	+.09 [+.07,+.11]	+.01 [+.00,+.02]	−.**16** [−.17,−.16]
	AT	total	+.03 [+.01,+.05]	−.**16** [−.16,−.16]	+.**13** [+.12,+.14]	−.01 [−.02,+.00]
	other	total	−.05 [−.08,−.02]	−.**14** [−.15,−.13]	+.04 [+.02,+.06]	+.**14** [+.11,+.16]
digit ×	+1*SD*	−1*SD*	−.08 [−.12,−.05]	+.**25** [+.18,+.27]	−.**14** [−.14,−.12]	−.02 [−.03,−.01]
digit ÷	+1*SD*	−1*SD*	+.03 [+.00,+.07]	−.**18** [−.18,−.17]	−.**12** [−.14,−.11]	+.**26** [+.24,+.29]
mental	+1*SD*	−1*SD*	−.05 [−.09,−.02]	+.**18** [+.18,+.18]	+.02 [+.00,+.04]	−.**15** [−.17,−.13]
more	+1*SD*	−1*SD*	+.**18** [.+.13,+.22]	−.**35** [−.36,−.33]	+.09 [+.08,+.10]	+.08 [+.05,+.10]

Probabilities for different levels of a predictor that differ by .10 or more are in boldface.

students whose teacher scored one standard deviation below the mean (and who were thus more focused on the whole-number-based algorithm for multiplication), their probability of being in the mixed algorithm class was .25 higher, while their probability of being in the non-algorithmic written class was .14 lower, $W^2 = 36.6$, $p < .001$. Students whose teacher scored above rather than below the mean for digit-based division had a .26 higher probability of being in the digit algorithm class, and a .18 and .12 lower probability of being in the mixed algorithm and non-algorithmic written class respectively, $W^2 = 100.9$, $p < .001$. Students whose teacher scored above rather than below the mean in their attention to various aspects of mental calculation had a .18 higher probability of being in the mixed algorithm class and a .15 lower probability of being in the digit algorithm class, $W^2 = 49.0$, $p < .001$. Students whose teachers scored above rather than below the mean for the use of multiple strategies per operation type, had a .35 lower probability of being in the mixed algorithm class and a .18 higher probability of being in the no written work class, $W^2 = 54.0$, $p < .001$.

Discussion

The present study demonstrated a first application of MLCA to educational large-scale assessment data. We argued that this technique is especially suitable for the challenges of this type of data and for evaluating contextual effects on problem solving (Zumbo et al., 2015). We demonstrated the added value of adequately modeling the multilevel structure inherent to educational data: though teacher effects are often ignored by researchers, we found them to be considerable. Model fit was much better with than without a multilevel structure for the teacher level, and latent teacher groups were found with large differences in students' probability of having a certain strategy choice profile. Ignoring teacher effects therefore seems to result in the omission of a crucial part of the model, and thereby in an incomplete representation of reality. The present study also demonstrated the relevance of the possibility of including predictors at different hierarchical levels in the model by simultaneously controlling for student characteristics and investigating curriculum effects, which led to interesting results relevant to both educational practice and theory.

Substantive conclusions

The results with regard to strategy choice profiles (or latent classes) that were found were largely in line with our hypotheses: there were profiles dominated by answering without written work, by the digit-based algorithm, by non-algorithmic approaches and the whole-number-based algorithm, and by both algorithms depending on the operation (multiplication or division). Students' probability of being in each of these classes was found to depend strongly on the teacher, because it varied considerably between latent teacher groups. The range was largest for the algorithmic classes and smallest for the no written work class. Therefore, teachers appear to have large effects on students' strategy use, but these effects unfortunately seem smallest for the inaccurate mental strategies without written work.

Intended and enacted curriculum predictors were added, controlling for student characteristics. Consistent with previous research findings, boys and students who were going to a lower secondary school level were more likely to answer without written work. The intended curriculum and enacted curriculum were not significantly related to each other, and were both found to be related to strategy choices, despite the suggestion from the literature of limited effects of the intended curriculum. As for the intended curriculum, the textbooks mostly appeared to be related to students' probability of using the different algorithmic and non-algorithmic written strategies.

As for the enacted curriculum, its relation to strategy use appeared somewhat stronger than that of the intended curriculum. Teaching digit-based algorithms was associated with an accordingly higher use of these strategies, while teaching whole-number-based algorithms appeared to have the unexpected side-effect of a higher use of non-algorithmic written strategies. Devoting more attention to mental strategies was associated with higher probability of the mixed algorithm class and lower

probability of the digit-based algorithm class. Teaching more than one strategy per operation was associated with lower probability of the mixed algorithm class and higher probability of the no written work class. Instruction formats did not have significant effects on strategy use, thereby not confirming our expectations regarding the effects of direct instruction versus independent work. Instruction differentiation also did not have a significant effect.

Limitations

A limitation of the present study could be the sample size, which is both relevant for the estimation of the complex MLCA models and the generalizability of the results. As for the sample size required for the estimation of MLCA models (or LCA models more generally), there are no general rules of thumb. Our sample of 1,619 students with 107 teachers seems to be of a similar order of magnitude as those in the examples used by Vermunt (2003) in his introduction of MLCA, where applications were featured with 886 employees in 41 teams, 2156 students in 97 schools, and 3584 respondents in 32 countries. A more precise estimate for a specific situation can be made using Monte Carlo simulations, where factors such as the number and type of problems, the separation of the classes and their relative sizes (approximately equal or not) and the amount of missing data play a role (Muthén & Muthén, 2002; Nylund, Asparouhov, & Muthén, 2007). Nylund et al. (2007) found particular problems with information criteria when a small sample (N = 200) was combined with unequal class sizes, as small classes then contain very few subjects. This is not the case in our sample.

Another limitation is the correlational nature of the large-scale assessment data. We of course had no influence on the intended or enacted curriculum, and therefore the causal nature of the found relations between curriculum and strategy use is uncertain and requires further (experimental) investigation. The present study does provide a starting point for such follow-up research. It should also be noted that the intended and enacted curriculum do not reflect (direct) effects of the teachers in our sample to the same extent, as the enacted curriculum is in the hands of the teacher, whereas the intended curriculum (the textbook) is determined on a school level.

Implications

The results suggest several implications (though the limited sample size should be noted). They suggest that models for strategy choices such as the Adaptive Strategy Choice Model (ASCM; Lemaire & Siegler, 1995) may need to be extended to include factors beyond the student and the problem (in line with suggestions by Verschaffel, Luwel, Torbeyns, & Van Dooren, 2009), and the same goes for other investigations of mathematical strategy use that have overlooked instructional factors so far (e.g., Hickendorff et al., 2010; Imbo & Vandierendonck, 2008; Lemaire & Lecacheur, 2011). The results also suggest that the investigations of curriculum effects on achievement may so far have omitted an important mediator: curriculum affects strategy use, and there are strong performance differences between strategies (Hickendorff, 2011; Hickendorff et al., 2009), so the curriculum may (in part) affect achievement through its effect on strategy use.

For educational reforms, our results suggest that although positive effects on achievement have been found of instructional practices congruent with reform ideas (Cohen, & Hill, 2000), reform-oriented instruction may also have unexpected side-effects: teaching that is more oriented toward the whole-number-based algorithms introduced by the Dutch mathematics education reform, is not only associated with more use of those algorithms, but also with more use of non-algorithmic strategies that have previously been shown to be less accurate than algorithms (Hickendorff et al., 2009). Finally, our finding that the effects of teachers and the curriculum on the proportion of students who mainly use mental strategies were small suggests that it might be challenging to reduce students' use of mental strategies through means of regular instruction, and that perhaps special interventions are necessary to promote their use of more accurate written strategies.

Conclusion

We would like to conclude by noting that our application of MLCA is relevant to applications of this technique to educational data more generally, and that several generalizations can be thought of: applications to other domains (e.g., strategies in spelling or reading), other types of nominal response data (e.g., error types), and also educational data from other sources than large-scale assessments (e.g., educational intervention studies with a large enough sample). With this article, we hope to have increased the attractiveness and accessibility of MLCA for educational researchers.

References

Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: 3-step approaches using Mplus. *Mplus Web Notes*, *15*, 1–24.

Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, *43*, 272–311. doi:10.1177/0081175012470644

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*, 3–27. doi:10.1093/pan/mph001

Buijs, C. (2008). *Leren vermenigvuldigen met meercijferige getallen* [Learning to multiply with multidigit numbers]. Utrecht, The Netherlands: Freudenthal Institute for Science and Mathematics Education.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*, 261–304. doi:10.1177/0049124104268644

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. doi:10.1177/001316446002000104

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record*, *102*, 292–343.

Derks, E. M., Boks, M. P. M., & Vermunt, J. K. (2012). The identification of family subtype based on the assessment of subclinical levels of psychosis in relatives. *BMC Psychiatry*, *12*, 71. doi:10.1186/1471-244X-12-71

Dias, J. G., & Vermunt, J. K. (2006). *Bootstrap methods for measuring classification uncertainty in latent class analysis*. COMPSTAT 2006—Proceedings in Computational Statistics, part I, 31–41.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Geiser, C., Lehman, W., & Eid, M. (2010). Separating "rotators" from "nonrotators" in the mental rotations test: A multigroup latent class analysis. *Multivariate Behavioral Research*, *41*, 261–293. doi:10.1207/s15327906mbr4103_2

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215–231. doi:10.1093/biomet/61.2.215

Gravemeijer, K. P. E. (1997). Instructional design for reform in mathematics education. In M. Beishuizen, K. P. E. Gravemeijer, & E. C. D. M. Van Lieshout (Eds.), *The role of contexts and models in the development of mathematical strategies and procedures* (pp. 13–34). Utrecht, The Netherlands: Freudenthal Institute.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge, England: Cambridge University Press.

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 193–215. doi:10.1080/10705511003659342

Hickendorff, M. (2011). *Explanatory latent variable modeling of mathematical ability in primary school: Crossing the border between psychometrics and psychology* (Unpublished doctoral dissertation), Leiden University.

Hickendorff, M., Heiser, W. J., Van Putten, C. M., & Verhelst, N. D. (2009). Solution strategies and achievement in Dutch complex arithmetic: Latent variable modeling of change. *Psychometrika*, *74*, 331–350. doi:10.1007/s11336-008-9074-z

Hickendorff, M., Van Putten, C. M., Verhelst, N. D., & Heiser, W. J. (2010). Individual differences in strategy use on division problems: Mental versus written computation. *Journal of Educational Psychology*, *102*, 438–452. doi:10.1037/a0018177

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371–406. doi:10.3102/00028312042002371

Hsieh, T.-C., & Yang, C. (2012). Do online learning patterns exhibit regional and demographic differences? *The Turkish Online Journal of Educational Technology*, *11*, 60–70.

Imbo, I., & Vandierendonck, A. (2008). Effects of problem size, operation, and working-memory span on simple-arithmetic strategies: Differences between children and adults? *Psychological Research*, *72*, 331–346. doi:10.1007/s00426-007-0112-8

Janssen, J., Van der Schoot, F., & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 4 [Fourth assessment of mathematics education at the end of primary school]*. Arnhem, The Netherlands: CITO.

Jepsen, C. (2005). Teacher characteristics and student achievement: Evidence from teacher surveys. *Journal of Urban Economics*, *57*, 302–319. doi:10.1016/j.jue.2004.11.001

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in Medicine*, *30*, 3050–3056. doi:10.1002/sim.v30.25

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up. Helping children learn mathematics*. Washington, DC: National Academy Press.

Lee Webb, M.-Y., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test III. *Educational and Psychological Measurement*, *68*, 335–351. doi:10.1177/0013164407308474

Lemaire, P., & Lecacheur, M. (2011). Age-related changes in children's executive functions and strategy selection: A study in computational estimation. *Cognitive Development*, *26*, 282–294.

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, *124*, 83–97. doi:10.1037/0096-3445.124.1.83

Lukočienė, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, L. Berthold, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241–249). Berlin, Heidelberg, Germany: Springer.

McCutcheon, A. L. (1987). *Latent class analysis*. Beverly Hills, CA: Sage Publications.

Morselli, D., & Passini, S. (2012). Disobedience and support for democracy: Evidences from the world values survey. *The Social Science Journal*, *49*, 284–294. doi:10.1016/j.soscij.2012.03.005

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620. doi:10.1207/S15328007SEM0904_8

Mutz, R., & Daniel, H.-D. (2011). University and student segmentation: Multilevel latent-class analysis of students' attitudes towards research methods and statistics. *The British Psychological Society*, *83*, 280–304.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237–257. doi:10.3102/01623737026003237

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 535–569. doi:10.1080/10705510701575396

Remillard, J. T. (2005). Examining key concepts in research on teachers' use of mathematics curricula. *Review of Educational Research*, *75*, 211–246. doi:10.3102/00346543075002211

Rowan, B., Correnti, R., & Miller, R. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record*, *104*, 1525–1567. doi:10.1111/tcre.2002.104.issue-8

Scheltens, F., Hemker, B., & Vermeulen, J. (2013). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 5* [Fifth assessment of mathematics education at the end of primary school]. Arnhem, The Netherlands: CITO.

Slavin, R. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*, 5–14. doi:10.3102/0013189X08314117

Slavin, R., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*, 427–515. doi:10.3102/0034654308317473

Treffers, A. (1987). Integrated column arithmetic according to progressive schematisation. *Educational Studies in Mathematics*, *18*, 125–145. doi:10.1007/BF00314723

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*, 213–239. doi:10.1111/some.2003.33.issue-1

Vermunt, J. K. (2005). Mixed-effects logistic regression models for indirectly observed discrete outcome variables. *Multivariate Behavioral Research*, *40*, 281–301. doi:10.1207/s15327906mbr4003_1

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 User's Guide*. Belmont, MA: Statistical Innovations.

Vermunt, J. K., & Magidson, J. (2013). *Latent GOLD 5.0 upgrade manual*. Belmont, MA: Statistical Innovations.

Verschaffel, L., Luwel, K., Torbeyns, J., & Van Dooren, W. (2009). Conceptualizing, investigating, and enhancing adaptive expertise in elementary mathematics education. *European Journal of Psychology of Education*, *24*, 335–359. doi:10.1007/BF03174765

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*, 89–122. doi:10.3102/00346543073001089

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, *10*, 12. doi:10.14507/epaa.v10n12.2002

Yang, X. D., Shaftel, J., Glasnapp, D., & Poggio, J. (2005). Qualitative or quantitative differences?: Latent class analysis of mathematical ability for special education students. *The Journal of Special Education*, *38*, 194–207. doi:10.1177/00224669050380040101

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly*, *12*, 136–151. doi:10.1080/15434303.2014.972559

Appendix: Teacher survey questions used to create the scores

For brevity, similar questions were combined in one sentence with the differences between the questions denoted by forward slashes (/); the number of questions used to create each score is given between brackets.

Strategy instruction

Focus on digit-based versus whole-number-based algorithm for multiplication (2). Which multiplication algorithm (whole-number-based, both, digit-based) reflects the practice in your class most closely? To what extent do you as a teacher prefer the whole-number-based or digit-based multiplication algorithm?

Focus on digit-based versus whole-number-based algorithm for division (2). Which division algorithm (whole-number-based, both, digit-based) reflects the practice in your class most closely? To what extent do you as a teacher prefer the whole-number-based or digit-based division algorithm?

Attention to various aspects of mental calculation (6). How many times a week do you pay attention to mental calculation and numerical estimation in your mathematics lessons? How often do you pay attention to these aspects of mental calculation: basic multiplication and division skills / smart number-dependent strategies / multiple strategies for one problem type / numerical estimation / applying approximations, estimations and rounding off?

Use of multiple strategies per operation type (4). Do your students use one or more strategies for mental multiplication / division? Do your students use one or more strategies for multidigit multiplication / division?

Instruction formats

Focus on group instruction (2). How important is giving group instruction in mathematics lessons to you? How much time do you spend on average on giving group instruction?

Focus on individual instruction (2). How important is giving (extra) individual instruction in mathematics lessons to you? How much time do you spend on average on giving (extra) individual instruction?

Focus on individual work (2). How important is letting students work individually in mathematics lessons to you? How much time do you spend on average on letting students work individually?

Actively involving students in instruction (4). How often do the following situations occur during mathematics lessons: you ask the class questions during instruction / you let students write out calculations on the blackboard / you ask students how they found an answer they gave / you discuss frequent errors with the class?

Instruction differentiation

Understanding students' thinking (3). How well do you understand the thinking of your students with low / average / high performance?

Differentiation within mathematics lessons (2). Do low-performing students get more learning time than average students? To what extent do you differentiate in your mathematics teaching by level or pace?

Extra support within the school (2). Are there possibilities for extra individual support in mathematics for students in your school from a remedial teacher or a mathematics specialist? How satisfied are you with the results of the school support that students receive?

Extra support outside the school (3). How intensive is the support of students at home, by parents or caretakers? Are there students who receive external support, for example in homework classes? How satisfied are you with the results of the support at home / external support that students receive?