Solving multiplication and division problems: latent variable modeling of students' solution strategies and performance

Fagginger Auer, M.F.

— 3 —

Using LASSO penalization for explanatory IRT: An application on instructional covariates for mathematical achievement in a large-scale assessment

Abstract

A new combination of statistical techniques is introduced: LASSO penalization for explanatory IRT models. This was made possible by recently released software for LASSO penalization of GLMMs, as IRT models can be conceptualized as GLMMs. LASSO penalized IRT shows special promise for the simultaneous consideration of high numbers of covariates for students' achievement in large-scale educational assessments. This is illustrated with an application of the technique on Dutch mathematical large-scale assessment data from 1619 students, with covariates from a questionnaire filled out by 107 teachers. The various steps in applying the technique are explicated, and educationally relevant results are discussed.

3.1 Introduction

Data with very high numbers of covariates can be analyzed using regularization methods that place a penalty on the regression parameters to improve prediction accuracy and interpretation, making this type of regression known as penalized regression. A popular form of penalized regression is LASSO (least absolute shrinkage and selection operator), where more and more regression parameters become zero as the penalty increases, thereby functioning as a covariate selection tool (Tibshirani,

This chapter is currently submitted for publication as: Fagginger Auer, M. F., Hickendorff, M., & Van Putten, C. M. (submitted). *Using LASSO penalization for explanatory IRT: An application on covariates for mathematical achievement in a large-scale assessment.*

The research was made possible by the Dutch National Institute for Educational Measurement Cito, who made the assessment data available to us.

1996). LASSO has so far been applied in many (generalized) linear models, but has only recently been extended to generalized linear mixed models (GLMMs), allowing for the modeling of correlated observations (Groll & Tutz, 2014; Schelldorfer, Meier, & Bühlmann, 2014).

In the present study, we utilize this GLMM extension of LASSO to introduce penalized regression for explanatory item response theory (IRT) models, making use of the possibility of conducting IRT analyses with general GLMM software demonstrated by De Boeck and Wilson (2004). This first use of LASSO penalized explanatory IRT is demonstrated with an application to a large-scale educational dataset, a type of data for which this technique promises to be especially useful as it allows for the simultaneous consideration of high numbers of potentially relevant covariates while optimally modeling achievement.

3.1.1 Explanatory IRT with LASSO penalization for large-scale assessment data

In large-scale educational assessments, achievement in an educational domain is assessed for a large representative sample of students to enable evaluation of the outcomes of an educational system (often that of a country), and to make comparisons to past outcomes or to outcomes of other educational systems. The analysis of achievement data from assessments usually requires the linking of different subsets of a total item set. These can be both subsets of the large complete item set within an assessment and item sets of successive assessments, and can be done using IRT (e.g, Mullis, Martin, Foy, & Akora, 2012; Mullis, Martin, Foy, & Drucker, 2012; OECD, 2013; Scheltens et al., 2013). IRT models achievement by placing persons and items on a common latent scale, and the probability of a correct response depends on the distance between the ability θ_p of a person p and the difficulty β_i of an item i in a logistic function: $P(y_{pi} = 1|\theta_p) = \frac{\exp(\theta_p - \beta_i)}{1+\exp(\theta_p - \beta_i)}$. This basic IRT model is the Rasch model (Rasch, 1960), which can be extended in multiple ways.

One extension is to make it an explanatory model rather than just a measurement model, by including explaining factors for items' difficulty and persons' ability (De Boeck & Wilson, 2004). These explanatory variables can be labeled in various ways (e.g., as predictors), but we will refer to them as covariates. Whereas in a Rasch model a separate difficulty parameter is estimated for each item, in an item explanatory model (e.g., the linear logistic test model (LLTM); G. H. Fischer, 1973) item covariates that differ across items but not persons (such as number of operations required in a math item) are used to model item difficulty. Similarly,

person covariates that vary across persons but not items (such as gender) can be used to explain ability level, and finally, person-by-item covariates that vary across both persons and items (such as solution strategy use) are also possible. IRT can therefore be used not only to optimally model achievement in large-scale assessments, but also to gain more insight into the factors that affect achievement (e.g., see Hickendorff et al., 2009).

Collection of data on such factors is a part of many assessments, as these assessments include questionnaires on topics such as children's background and attitudes, teachers' characteristics and instructional practices, and the conditions in schools (Mullis, Martin, Foy, & Akora, 2012; Mullis, Martin, Foy, & Drucker, 2012; OECD, 2013; Scheltens et al., 2013). These many different factors contribute to achievement jointly, and should be considered simultaneously so that effects are evaluated while controlling for other covariates, and so that the importance of different covariates relative to each other can be determined. However, analyses with very high numbers of covariates can be challenging, especially with models that are already complex models such as explanatory IRT models.

Penalized regression

A common way to deal with the challenge of high numbers of covariates is through so-called penalized regression. As described by Tibshirani (1996), normal regression with ordinary least squares (OLS) estimates can be improved in terms of prediction accuracy and interpretation by penalizing regression coefficients by shrinking them or setting some of them to zero. This can be done in various ways. One way is subset selection, in which a model with a subset of the covariates is selected (through forward or backward selection). Though the reduced number of covariates in this situation facilitates interpretation, small changes in the data can lead to the selection of very different models, creating the risk of chance capitalization and compromising prediction accuracy. A second way, ridge regression, is more stable as regression coefficients are shrunk in a continuous process, but is also more complex in terms of interpretation as none of the coefficients become zero. Tibshirani (1996) proposed a third way, LASSO regression, which seeks to combine stability and interpretability by shrinking some regression coefficients and setting others to zero.

Both in LASSO and ridge regression, the sum of a specific function of the regression parameters has to be smaller than or equal to a tuning parameter t. With ridge regression, this is the sum of the squared coefficients, $\sum_j \beta_j^2 \leq t$, and

with LASSO regression, the sum of the absolute coefficients, $\sum_j |\beta_j| \leq t$. With this restriction, the sum of the squared differences between the observed and predicted y's, $\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2$, is minimized. Incorporating the restriction explicitly in the latter equation, this can be alternatively formulated as $\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j \beta_j^2$ or $\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$. This whole equation is minimized, which in the case of a λ of 0 results in ordinary regression, but with increasing values of λ in a higher and higher penalization for the sum of the coefficients (until finally all penalized coefficients are zero). The different restrictions on the regression coefficients in ridge and LASSO result in shrunken coefficients in both cases, but generally, only with LASSO coefficients are set to zero (Tibshirani, 1996).

Recently, software has become available that allows for LASSO (but as far as we know, not ridge) penalization for GLMMs (Groll & Tutz, 2014; Schelldorfer et al., 2014). Schelldorfer et al. (2014) implemented GLMM LASSO in an `R` package entitled `glmmixedLASSO`, and demonstrated the efficiency and accuracy of their algorithm using various simulations with both relatively low (e.g., 10 and 50) and very high numbers of covariates (e.g., 500 and 1500) in logistic and Poisson models. They note that the mixed aspect of GLMMs causes a problem for LASSO, as the shrinkage of regression coefficients can severely bias the estimation of the variance components. They address this issue with a two-stage approach: first the LASSO is used as a variable selection tool, and then in a second step an unpenalized model with the selected variables is fitted using a maximum likelihood method, to ensure accurate estimation of the variance components.

The availability of LASSO for GLMMs makes LASSO penalization for explanatory IRT models possible. IRT models were not developed as a special case of GLMMs but in a separate line of research, with specialized IRT software such as BILOG, PARSCALE and TESTFACT (Embretson & Reise, 2000). However, more recently, De Boeck and Wilson (2004) have described how to formulate IRT models as GLMMs and how to estimate them using general purpose GLMM software, enabling a wider application of this class of models. Therefore, LASSO penalization for explanatory IRT models is now possible, and it can be used for the simultaneous consideration of high numbers of covariates for achievement in large-scale assessment data. In the present study, we apply this new combination of techniques for this purpose. We use it to investigate the effects of various factors on mathematical achievement in a large-scale assessment: children's and teachers' characteristics, as-

pects of teachers' instruction, and the solution strategies that children use to obtain item answers. The existing literature on the effects of these factors on achievement will now be succinctly described.

3.1.2 Covariates for mathematical achievement

Children's characteristics and achievement

Various characteristics of children have been found to be related to mathematical achievement. As for other achievement measures, children with a lower socioeconomic status (SES) generally perform less well in mathematics than children with a higher SES (e.g., Sirin, 2005). Children's general intelligence and mathematical achievement are also positively related (e.g., Primi, Eugénia Ferrão, & Almeida, 2010). While stereotypes still suggest that girls perform less well in mathematics than boys, no general gender differences in mathematical achievement for children are indicated (e.g., J. S. Hyde, Lindberg, Linn, Ellis, & Williams, 2008), though in some countries such differences do exist (e.g., the Netherlands; Scheltens et al., 2013).

Effects of teachers on student achievement

There is large amount of research on the effects that teachers and their instruction methods can have on achievement, in which many different aspects of the teaching process are considered. One obvious indicator of instruction is the formal curriculum provided by the mathematics textbook that is used. However, as noted by Remillard (2005), a distinction must be made between this formal curriculum and what actually takes place in the classroom (i.e., the intended versus the enacted curriculum). A review of the existing research on effective programs in mathematics by Slavin and Lake (2008) demonstrated very limited effects of textbooks, but much stronger effects of programs that targeted the instructional processes in which teachers and children interact in the classroom. Positive effects were found of interventions that concerned classroom management, keeping children engaged, promoting cooperation among children, and supplementary tutoring. In another review, the Royal Netherlands Academy of Arts and Sciences (2009) similarly concludes that there is little support for meaningful effects of the formal curriculum and more for effects of the actual teaching process.

However, these reviews for an important part concern studies with small samples, which could bias results as small studies with null or negative results may be

likely to remain unpublished and therefore not included in reviews (Slavin, 2008). Large-scale assessment data can, though correlational rather than experimental, supplement these findings with its very large and representative samples. This has been done for the investigation of the relation between teacher behaviors and children's achievement in what is called the process-product literature. Studies of this kind have indeed shown that certain teaching practices affect children's achievement, and have for example found a consistent positive effect of time spent on active academic instruction rather than other activities (Hill et al., 2005). The related notion of opportunity to learn (Carroll, 1963) posits that the assessed achievement in a domain depends on the time students have spent in learning about that domain relative to the time they need to learn it. The process-product literature can be contrasted with the educational production function literature, where not processes but the resources of children, teachers and schools are related to student outcomes. These can be resources such as children's SES and teachers' education or their years of teaching experience. Generally, the results on the effects of such factors have been mixed, indicating modest effects at best (Wenglinsky, 2002).

Considering these various findings, the literature seems to suggest that effects of teachers on children's mathematics achievement are more in the actual process of how teachers interact with children, than in general characteristics of the teacher or of the curriculum. This is in line with findings about children's achievement in general, for which a large synthesis of thousands of studies by Hattie (2003) shows that teachers have the largest effects on children's achievement through the teaching behaviors of providing feedback and direct instruction, and through instructional quality.

Solution strategies and achievement

Children's solution strategies for mathematical items are also highly relevant to achievement. These strategies vary from formal algorithms with a fixed notation (such as long division), to informal approaches with a customized notation, to approaches that only comprise mental calculation in the head (see Table 3.1 for examples). Increased attention for children's own strategic explorations (rather than for a prescribed set of algorithmic strategies) is an important part of the reform in mathematics education that has taken place in various countries over the past decades (Gravemeijer, 1997; Kilpatrick et al., 2001; Verschaffel, Luwel, Torbeyns, & Van Dooren, 2007). As such, solution strategies are a crucial part of the instructional process, and they have received ample research attention (e.g., Barrouillet et

Table 3.1: Examples for the multiplication and division strategy categories.

	×	÷
digit-based algorithm	56	34/544\16
	23×	34
	168	204
	1120+	204
	1288	0
whole-number-based algorithm	56	544 : 34 =
	23×	340 - 10×
	18	204
	150	102 - 3×
	120	102
	1000+	102 - 3×+
	1288	0 16×
non-algorithmic strategies	1120 + 3 × 56	10 × 34 = 340
	1120 + 168	13 × 34 = 442
	1288	16 × 34 = 544
no written work	1288	544

al., 2008; Siegler & Lemaire, 1997; Torbeyns, Verschaffel, & Ghesquière, 2005).

In the present study, we therefore also devote attention to teachers' specific strategy instruction and to children's strategy use. We focus on strategies for answering multidigit multiplication and division items (items with larger or with decimal numbers, such as 23×56 or $31.2 \div 1.2$), as strategies in this domain have been shown to be highly relevant to achievement for the students in our sample (Dutch sixth graders). In particular, Hickendorff et al. (2009) and Hickendorff (2011) demonstrated a large accuracy advantage for multiplication and division strategies that involved writing down calculations compared to strategies that did not, and within these more accurate written strategies, a higher accuracy of the traditional digit-based multiplication algorithm than of other written approaches for multiplication. The accuracy advantage of written over non-written strategies was larger for children with a lower mathematical ability than for children with a higher ability, and girls wrote down calculations more often than boys.

3.1.3 Present study

In the present study, we consider these various types of covariates in our demonstration of the new combination of the techniques of LASSO penalization and explanatory IRT. We apply the LASSO penalized IRT to a large-scale educational dataset from the most recent (2011) national assessment of the mathematical ability of children at the end of primary school (sixth graders) in the Netherlands, for which no links between instruction and achievement have been investigated yet (Scheltens et al., 2013). Data on item responses, gender, ability and SES were collected for the children, and data on teacher characteristics and instructional practices were collected from the children's teachers.

Hypotheses

Based on our previous discussion of instructional effects on achievement, we expect that covariates that concern instructional practices during mathematics lessons are more strongly related to achievement than teacher characteristics or the mathematics textbook that is used. Several particular instructional practices covered in our covariates can be expected to have a positive relation to achievement. One is that of time spent on group instruction and not other activities, given the positive effect of active academic instruction from the process-product literature (Hill et al., 2005). Another is the frequency of practices that engage children in instruction (such as asking the class questions and letting children write out calculations on the blackboard), given the positive effects of keeping children engaged found in the review of effective programs in mathematics (Slavin & Lake, 2008). Another is practices that involve extra attention for weaker students, through extra support at or outside of school (and perhaps differentiation of instruction more broadly), given the positive effects of supplementary tutoring (Slavin & Lake, 2008).

For strategies, we expect to find written strategies to be associated with higher achievement than mental strategies, and possibly best achievement with the traditional digit-based algorithm (Hickendorff, 2011; Hickendorff et al., 2009). Accordingly, instructional practices focused on mental strategies may be negatively related to achievement, while practices that focus on the digit-based algorithm, or more generally, a single standardized approach rather than multiple approaches, may be positively related to achievement. Since previous research indicates interactions between strategy use and accuracy and children's characteristics (e.g., smaller accuracy difference between written and mental strategies for stronger students; Hickendorff et al., 2009), these interactions were also included in the analyses.

3.2 Method

3.2.1 Sample

Schools were selected for participation in the 2011 mathematics assessment according to a random sampling procedure stratified by socioeconomic status, resulting in a total number of 2548 participating sixth graders (11-12-year-olds) from 107 schools. The children were presented subsets of a large set of mathematical items on a variety of topics, and subsets containing multidigit multiplication and division items were presented to 1619 of the children. These children were in the classes of 107 teachers (one teacher per school), which means that an average of 15 children per teacher participated. Of the 1619 children, 49 percent were boys and 51 percent 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 percent were to go to prevocational secondary education. In terms of SES, most children (88 percent) had at least one parent who completed at least two years of medium or higher-level secondary school (SES not low), while 12 percent did not (SES low).

3.2.2 Materials

Multiplication and division items

The assessment contained thirteen multidigit multiplication items and eight multidigit division items in total. These items were administered to children according to an incomplete design (see Hickendorff et al., 2009, for more details on such designs): children were presented systematically varying subsets of either three or six of these items. Table 3.2 provides information on the content of the items: the numbers with which the multiplication or division operation had to be performed and whether these numbers were presented in a realistic context describing a problem situation (such as determining how many bundles of 40 tulips can be made from 2500 tulips) or not. The items were presented in booklets in which children could also write down their calculations and solutions. The children were not given any other paper to write on and were explicitly instructed that if they wanted to write down calculations, they could use the (ample) blank space next to the items in the booklet.

Following the assessment, these calculations were coded for strategy use. For this, five different categories were distinguished. The first two categories are for

algorithmic solutions: the more traditional digit-based algorithm and the newer whole-number-based algorithm. The third category consists of written work without an algorithmic notation, such as writing down only intermediate steps. Table 3.1 gives examples for multiplication and division strategies in these three categories. The two remaining categories are solutions with no written calculations, and a small other category, consisting mostly of unanswered items.

The strategy coding was carried out by three undergraduate students and the first and third author. Parts of the material (112 multiplication and 112 division solutions) were coded by all coders to determine the interrater reliability. Cohen's κ (J. Cohen, 1960) was found to be .90 for the multiplication and .89 for the division coding on average, which indicates high levels of interrater agreement.

Teacher questionnaire about classroom practice

The teachers of the participating children filled out a questionnaire about their mathematics teaching. A total of 39 questions were selected from this questionnaire (see the Appendix) that were either relevant to the mathematics lessons in general (teacher characteristics, mathematics textbook used, and general instructional practices during the mathematics lessons), or that specifically concerned multiplication, division, or mental strategies (the latter because of the aforementioned large achievement difference between strategies with and without written down calculations). Questions that were excluded specifically concerned mathematical domains other than multiplication or division (e.g., addition or percentages) or concerned attitudes or opinions rather than concrete characteristics or instructional practices (e.g., opinion on class size rather than actual class size). Dummy variables were made for questions with nominal response categories and scores were standardized for the other questions (missing values were imputed with the variable mode, because multiple imputation was not feasible given the complex LASSO IRT analyses).

3.2.3 Statistical analysis

The `R` package `glmmixedLASSO` (Schelldorfer et al., 2014) was used to conduct the LASSO penalized explanatory IRT analysis. As described by De Boeck and Wilson (2004), the explanatory IRT model was specified by using a binomial model with the solution accuracy (incorrect or correct) as the dependent variable, and a random person intercept for the latent ability variable and fixed item effects for the

Table 3.2: The content of the thirteen multidigit multiplication items and eight multidigit division items in the assessment and the percentage of correct solutions.

item	context	*N*	%
$9 \times 48 = 432$	yes	368	76
8 × 194 = 1552	yes	355	72
$6 \times 192 = 1152$	no	352	70
35 × 29 = 1015	yes	353	69
$6 \times 14.95 = 89.70$	yes	359	66
$1.5 \times 1.80 = 2.70$	yes	353	65
35 × 29 = 1015	no	352	64
$23 \times 56 = 1288$	yes	358	58
$209 \times 76 = 15884$	no	344	54
$24 \times 37.50 = 900$	no	352	47
$0.18 \times 750 = 135$	no	356	36
9.8 × 7.2 = 70.56	no	352	26
3340 × 5.50 = 18370	yes	359	21
total multiplication		4613	56
544 ÷ 34 = 16	yes	368	56
47.25 ÷ 7 = 6.75	yes	352	47
1575 ÷ 14 = 112.50	no	355	41
$1470 \div 12 = 122.50$	yes	350	40
$2500 \div 40 = 62$	yes	359	32
31.2 ÷ 1.2 = 26	no	369	30
6496 ÷ 14 = 464	yes	354	29
$11585 \div 14 = 827.5$	yes	345	26
total division		2852	38

Note: The items in italics are slightly modified parallel versions of items that have not yet been released for publication by Cito because they may be used in subsequent assessments.

item easiness parameters. The person covariates that were added were children's gender (boy or girl), general scholastic ability level (lower or higher) and SES (not low or low), and the questions from the teacher questionnaire. The person-by-item covariate that was added was that for strategy use on the item (with dummy variables for the aforementioned multiplication and division strategy categories). In addition, interactions between strategy use and the three student characteristics (gender, ability and SES) were added. The penalization was not imposed on all covariates: the fixed item effects were specified as unpenalized, so the IRT part of the model remained intact regardless of the degree of penalization. The children's characteristics (gender, general scholastic ability level and SES) were also unpenalized, so that these were always controlled for in evaluating the effects of the instruction and strategies.

The final element of the model to be specified is the degree of penalization, which is determined by λ (as discussed in the introduction). We did this using the approach taken by Schelldorfer et al. (2014), the authors of the `glmmixedLASSO` package: we used the Bayesian Information Criterion (BIC) to select the model that provided the best balance between model parsimony and fit to the data. The BIC is calculated by taking the log-likelihood (LL) of the observed data under the model and imposing a penalty for the number of parameters (k) in the model, weighed by the logarithm of the number of cases (N) (individuals, in our case): $-2LL + log(N) \times k$ (and asymptotically, the BIC is equivalent to k-fold cross-validation with some optimal k; Shao, 1997). The lower the BIC, the better the trade-off between model fit and complexity, so the model with the lowest BIC was selected.

As recommended by Schelldorfer et al. (2014), we then reran the model with the selected covariates with the `R` package `lme4` (Bates & Maechler, 2010), for an unbiased estimation of the random effects. In this model, a random intercept was also added for the teachers to account for the nesting of children within teachers (see Doran, Bates, Bliese, & Dowling, 2007), which is not yet possible in `glmmixedLASSO`. This model was used for final interpretation of the covariate effects.

Expressed mathematically, the explanatory model for the probability of a correct response with J person covariates j (which can be both at the child and teacher level) for child p with teacher t (denoted Z_{ptj} with regression parameter ζ_j), H person-by-item covariates h for child p of teacher t and item i (denoted W_{ptih} with regression parameter δ_{ih}), and I item dummy variables X_i with easiness parameter β_i, is then:

Table 3.3: Use and (observed and estimated) accuracy of the multiplication and division strategies.

	observed				estimated P(correct)		
	use (%)		correct (%)		×	÷	
	×	÷	×	÷		boys	girls
digit-based algorithm	37	16	68	61	.76	.61	.63
number-based algorithm	3	24	68	63	.75	.65	.64
non-algorithmic strategies	24	10	59	50	.62	.51	.45
no written work	28	35	51	22	.50	.22	.16
other	8	15	2	2	.05	.05	.07

$$P(y_{pti} = 1|Z_{pt1}...Z_{ptJ}, W_{pti1}...W_{ptiH}, X_1...X_I) = \frac{\exp(\eta)}{1 + \exp(\eta)} \tag{3.1}$$

with

$$\eta = \Sigma_{j=1}^{J}\zeta_j Z_{ptj} + \Sigma_{h=1}^{H}\delta_{ih}W_{ptih} + \Sigma_{i=1}^{I}\beta_i X_i + \epsilon_p + \epsilon_t \tag{3.2}$$

3.3 Results

3.3.1 Descriptives

Overall, 56 percent of the multiplication items was solved correctly (varying between 21 percent correct for the item 3340×5.50 and 76 percent for 9×48), and 39 percent of the division items (varying between 26 percent correct for $11585 \div 14$ and 56 percent for $544 \div 34$) (see Table 3.2). Multiplication items were most often solved using the digit-based algorithm, which was also (together with the whole-number-based algorithm) the most accurate strategy with 68 percent of correct solutions (see Table 3.3 for strategy descriptives). Solutions without written work were also frequent (and relatively inaccurate, with 51 percent correct solutions), as were non-algorithmic written strategies (59 percent correct). Division items were most often solved without written work, an approach that was very inaccurate (22 percent correct). Application of the whole-number-based algorithm was also frequent, followed by application of the digit-based algorithm, and both these strategies were relatively accurate (63 and 61 percent correct respectively).

3.3.2 Covariate selection using penalized regression

The LASSO IRT model with penalization on the teacher and strategy covariates was estimated with different settings of λ. All penalized coefficients were shrunk to zero for $\lambda \geq 240$, so models with all (integer) λs from 0 (no penalization) to 240 (all penalized covariates dropped from the model) were estimated. Figure 3.1 shows the shrinking of penalized regression coefficients over this range, with each line representing one coefficient. The optimal amount of penalization indicated by the BICs (also see Figure 3.1) was found to be at $\lambda = 35$. The 18 penalized covariates with non-zero regression coefficients at this λ are the questions from the teacher questionnaire marked with asterisks in the Appendix and the multiplication and division solution strategy use, and the interaction between division strategy use and student gender.

3.3.3 Effects in the final model

The results of running an explanatory IRT model with the unpenalized and the selected covariates are given in Table 3.4 (the selected questions from the teacher questionnaires are numbered as in the Appendix). Of the unpenalized covariates, performance was found to be significantly related to children's general scholastic ability: higher ability children had a significantly higher probability of a correct response ($P = .58$) than lower ability children ($P = .33$), $z = 13.1$, $p < .001$. Gender did not have a significant effect, $z = 1.1$, $p = .29$, nor did SES, $z = -1.6$, $p = .10$.

Of the selected teacher covariates, the strongest positive effect was of the amount of time spent on group instruction in mathematics lessons ($P = .40$ for 1 SD below the mean and $P = .50$ for 1 SD above the mean). The strongest negative effect was of the amount of support at home ($P = .49$ for 1 SD below the mean and $P = .41$ for 1 SD above the mean).

There were strong effects of the employed solution strategy on the probability of a correct response, both for multiplication and division (see Table 3.3 for the estimated probability per strategy). The accuracy of the whole-number-based algorithms was comparable to that of the digit-based algorithms, while non-algorithmic strategies, strategies without any written work and other strategies (mostly leaving items unanswered) were less accurate (with the smallest accuracy difference for non-algorithmic strategies and the largest for other strategies). There was also an interaction between division strategy use and student gender: most notably, the

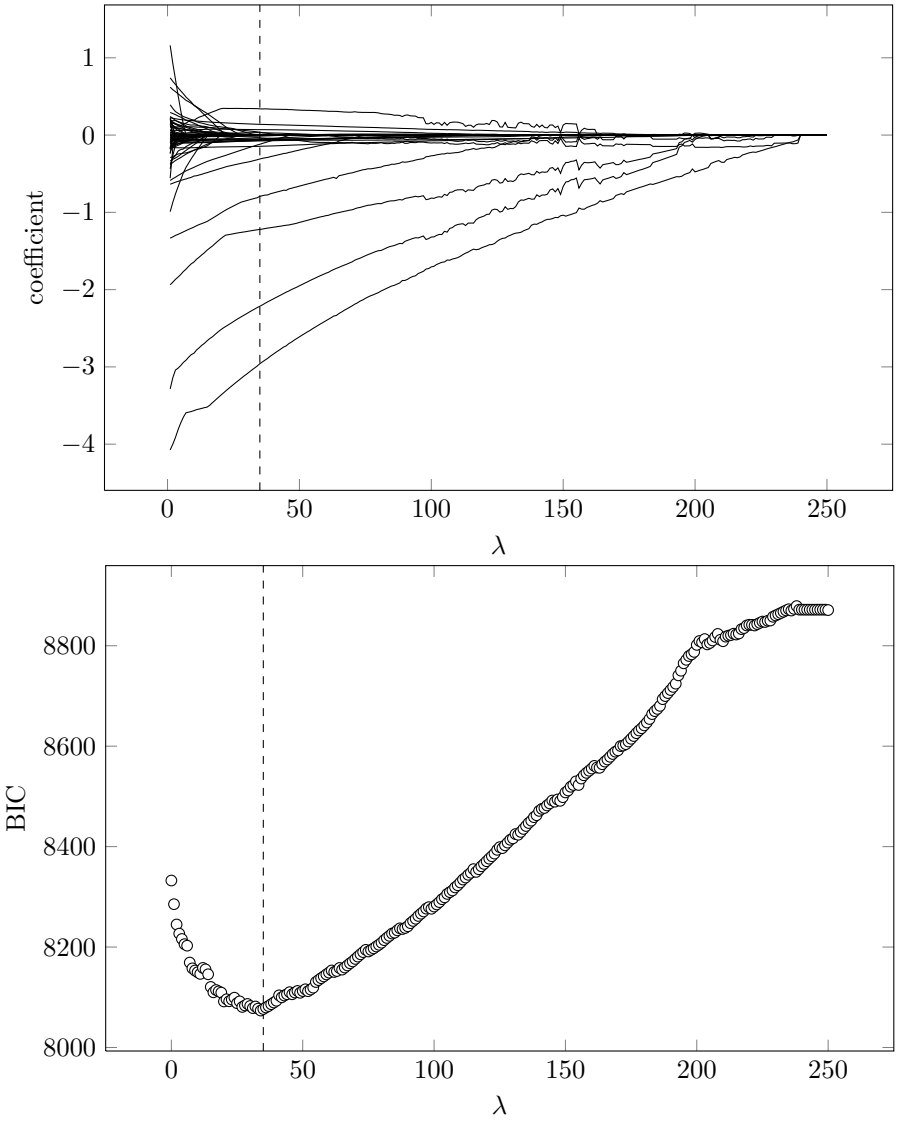

Figure 3.1: Penalized regression coefficients and BICs for the different settings of λ in the LASSO penalized IRT model (dashed vertical line at optimal $\lambda = 35$).

Table 3.4: Effects of the student characteristics and selected teacher covariates.

covariate		levels		estimate (S.E.)
		reference	target	
student char.	gender	boy	girl	0.10 (0.09)
	ability	lower	higher	1.05 (0.08)
	SES	not low	low	-0.20 (0.12)
teacher char.	1. age			0.04 (0.06)
	2. gender	male	female	-0.16 (0.10)
	5. years grade 6			-0.10 (0.05)
general instr.	12. time group instr.			0.18 (0.06)
	13. time indiv. instr.			-0.08 (0.05)
	15. ask class questions			0.01 (0.05)
	16. blackboard solutions			-0.05 (0.05)
	18. discuss errors			-0.01 (0.05)
instr. differ.	19. lesson diff.			-0.09 (0.05)
	22. support at home			-0.17 (0.06)
	23. external support			-0.09 (0.05)
strategy instr.	25. division alg.			0.05 (0.05)
	30. strat. multidigit ÷	one	multiple	-0.11 (0.10)
	32. ment. mul. div.			-0.07 (0.07)
	35. ment. smart strat.			-0.12 (0.08)
strategy use	multiplication	digit.	number.	-0.03 (0.24)
			non-alg.	-0.63 (0.11)
			no writ.	-1.14 (0.11)
			other	-4.07 (0.27)
	division	digit.	number.	0.19 (0.21)
			non-alg.	-0.40 (0.26)
			no writ.	-1.69 (0.18)
			other	-3.43 (0.33)
	gender×division	digit.	number.	-0.14 (0.21)
			non-alg.	-0.34 (0.31)
			no writ.	-0.51 (0.21)
			other	0.33 (0.41)

difference in accuracy between the digit-based algorithm and strategies without any written work was larger for girls ($P = .63$ vs. $P = .16$) than for boys ($P = .61$ vs. $P = .22$).

3.4 Discussion

In the present study, we introduced LASSO penalization for explanatory IRT models. This was made possible by recently released software that allows for LASSO penalization of GLMMs (Groll & Tutz, 2014; Schelldorfer et al., 2014), as IRT models can be conceptualized as GLMMs (De Boeck & Wilson, 2004). We argued that this new combination of techniques is especially useful for simultaneous consideration of the effects of the high numbers of covariates for students' achievement that are collected in large-scale educational assessments. This was illustrated with an application of LASSO penalized explanatory IRT to data from the most recent national large-scale assessment of mathematics at the end of primary school in the Netherlands. The various steps involved in applying the technique were explicated and educationally relevant results were discussed.

3.4.1 Substantive conclusions

A first result that was found is that the LASSO did not select formal curriculum covariates as important covariates for students' achievement: at the optimal degree of penalization, the coefficients for the textbook covariate were shrunk to zero. This is in accordance with findings of Slavin and Lake (2008) and the Royal Netherlands Academy of Arts and Sciences (2009) of very limited effects of the formal curriculum. A positive effect of the amount of time the teacher spends on group instruction was found, concordant with the positive effect of time spent on active academic instruction rather than other activities in the process-product literature (Hill et al., 2005). Though we expected practices that involve extra attention for weaker students to be beneficial because of the positive effects of supplementary tutoring (Royal Netherlands Academy of Arts and Sciences, 2009; Slavin & Lake, 2008), the amount of support that students received at home according to their teachers was negatively related to achievement. This could suggest that home support affects achievement negatively, but could also indicate that weaker students receive more home support. However, the teacher reported on the amount of home support only at the class level, and a proper investigation of this effect should be conducted with support measures at the student level.

Children's use of mathematical strategies was also found to play an important role. Strategies with written work were found to be much more accurate than those without written work, as was also found by Hickendorff et al. (2009) and Hickendorff (2011). Within written strategies, these authors found an accuracy advantage of the

digit-based algorithm over other written approaches, and we refined this finding by dividing the other written approaches into the whole-number-based-algorithm and non-algorithmic written strategies. This showed the accuracy of the whole-number-based algorithms to be comparable to that of the digit-based-algorithms, while the non-algorithmic approaches were less accurate. An interaction between gender and division strategy use was also found: strategies without written work were found to be relatively more inaccurate for girls than for boys. Fortunately, girls appear to use strategies without written work less frequently than boys (Fagginger Auer, Hickendorff, Van Putten, Béguin, & Heiser, in press; Hickendorff et al., 2009). It should be noted, however, that the accuracy estimations of the strategies could be biased by the ability of the students using the strategies and the difficulty of the items the strategies are applied to (bias by selection effects; Siegler & Lemaire, 1997), though a statistical correction for such bias is carried out with the inclusion of student ability and item easiness parameters in the model.

3.4.2 Limitations and future directions

The present study also has several limitations, some of which provide directions for future investigation and development.

Mediation student and teacher effects by strategies

A first limitation is substantive in nature. We investigated the effects of student and teacher covariates on student achievement, but some effects may have been obscured because they occurred through strategy use. For example, we found no significant effect of gender per se, but boys do make more use of the inaccurate strategy of answering without any written work (Fagginger Auer et al., in press). As for teacher effects, the sociocultural context is an important determinant of strategy use (Verschaffel et al., 2009), and teacher covariates are significantly related to students' strategy use (Fagginger Auer et al., in press). Given the large differences in students' achievement with different strategies, this means that teacher covariates can exert effects on achievement through strategy use, and these effects may go undetected when strategy use is also in the model. Though hard to incorporate in our current LASSO penalized explanatory IRT analysis, a more thorough investigation of this chain of effects could be done with a mediation analysis. However, it should also be noted that the impact of this issue may be limited, as teachers appear to exert relatively little influence over the strategy that has the largest negative con-

sequences for achievement - answering without any written work (Fagginger Auer et al., in press).

LASSO for correlated covariates

A second limitation is that when LASSO is used for covariates that are (highly) correlated, the selection of covariates can be to some extent random: when there is a near perfect correlation between two covariates, selection of either covariate results in nearly equal prediction of the dependent variable. This limitation is true for LASSO in general and not particular to our LASSO penalized explanatory IRT. However, in their successful simulation tests of the `glmmixedLASSO` procedure, Schelldorfer et al. (2014) included correlations among the covariates of up to .20, and the vast majority (90 percent) of correlations among our teacher covariates fell within that range. Less than one percent of the correlations was large ($\geq .50$), none of which concerned covariates that were found to be significant. Therefore, our results should not be affected too much by correlations among the covariates.

More random effects

A third limitation is that only one random effect could be specified for the LASSO penalization. While this is enough for a basic IRT model, in an educational context (with students nested in classes in schools) a random effect for the teacher or school level is also called for. In addition, in some contexts it makes more sense to model the item effects as random than as fixed - for example when items can be considered random draws from a domain, such as the items in this study from the domain of multidigit multiplication and division (De Boeck, 2008). A larger number of possible random effects (e.g., as in the package `lme4`; Bates & Maechler, 2010) would therefore be an important improvement for LASSO penalized explanatory IRT.

Cross-validation

A fourth limitation is the way in which the optimal degree of LASSO penalization was determined. We did this using the BIC (as in Schelldorfer et al., 2014), but a more common approach is to use cross-validation (e.g., it is a standard option the `R` package `penalized`; Goeman, 2010). With cross-validation, overfitting is prevented through fitting the model on one part of the data, and evaluating the prediction error of the model on another part of the data (Colby & Bair, 2013). Implement-

ing cross-validation in LASSO GLMM packages would provide an important tool for selecting the amount of penalization in LASSO penalized IRT. One problem with implementing this, however, is that the LASSO penalized IRT is already very computationally intensive with the estimation of just one model for each value of λ, but this should be resolved with ongoing improvements in computational power. Another problem is that cross-validation for GLMMs is not straightforward, but several approaches have been proposed to deal with this issue (Colby & Bair, 2013).

Other IRT models

A final limitation is that not all IRT models can be specified as GLMMs (De Boeck & Wilson, 2004), and therefore that our currently outlined procedure for LASSO penalized explanatory IRT does not apply to all types of IRT models. For example, models that cannot be specified as univariate GLMMs are the popular two-parameter (2PL) model (with item discrimination parameters) and models for polytomous response data. However, there is still ample flexibility within the current Rasch (1PL) framework, as any combination of person, item, and person-by-item covariates that is of interest can be made (e.g., we did not include item covariates, but LLTM-like models that include many potential sources of item difficulty are possible). Therefore, with our current demonstration of LASSO penalized explanatory IRT, we aimed to introduce a new combination of techniques that is versatile and that can lead to insightful results regarding the factors that influence achievement.

3.A Teacher survey questions

(when the same response options apply to multiple questions, those options are given under the last question they apply to for brevity; and the questions selected with the LASSO are marked with asterisks)

3.A.1 Teacher characteristics

1. *What is your age? *(. . . years)*
2. *What is your gender? *(male / female)*
3. From which teacher education did you graduate? *(PABO (after 1985) / PA, kweekschool or kindergarten training (before 1985) / other)*
4. In which grade do you have most teaching experience? *(sixth grade / other grade)*

5. *At the end of this school year, how many successive years have you been teaching in the sixth grade? *(... years)*
6. Have you received extra training in the past five years? *(yes / no)*
7. If so, in what areas have you received extra training? *(optimizing the learning opportunities of students with different backgrounds / evaluating the level of progress of a class / school self evaluation / subject-specific / other)*

3.A.2 Textbook

8. Which textbook do you use (predominantly) for mathematics instruction? *(Pluspunt / Wereld in Getallen / Rekenrijk / Alles Telt / other)*

3.A.3 General instruction

9. How many students are in your class? *(... students)*
10. How much time do you spend on mathematics lessons in an average week? *(... hours)*
11. How many minutes do you spend on multiplication and division in your mathematics lessons in a week? *(<30 minutes / 30-60 minutes / 60-90 minutes / 90-120 minutes / >120 minutes)*
12. *How many minutes do you on average spend on group instruction in a mathematics lesson?
13. *How many minutes do you on average spend on individual instruction in a mathematics lesson?
14. How many minutes do your students on average spend on individual work in a mathematics lesson? *(<10 minutes / 10-20 minutes / 20-30 minutes / 30-40 minutes / >40 minutes)*
15. *How often do you ask the class questions during instruction?
16. *How often do you let students write out calculations on the blackboard?
17. How often do you ask students how they found an answer they gave?
18. *How often do you discuss frequent errors with the class? *(less than once a month / once a month / twice a month / once every two weeks / at least once a week)*

3.A.4 Instruction differentiation

19. *To what extent do you differentiate in your mathematics teaching by level or pace? *(generally no differentiation / differentiation in practice materials but not instruction / differentiation in instruction and materials for different groups / individual instruction and selection of materials)*

20. How much extra learning time do weak students get compared to average students? *(. . . minutes per week)*
21. Are there possibilities for extra individual support in mathematics for students in your school from a remedial teacher or a mathematics specialist? *(no / yes, a remedial teacher / yes, by a care coordinator or mathematics specialist / yes, a remedial teacher and a care coordinator or mathematics specialist)*
22. ⋆How intensive is the support of students at home, by parents or caretakers? *(no support / little support / medium support / frequent support / permanent support)*
23. ⋆How many students receive external support, for example in homework classes? *(. . . students)*

3.A.5 Strategy instruction

24. Which multiplication algorithm reflects the practice in your class most closely?
25. ⋆Which division algorithm reflects the practice in your class most closely? *(whole-number-based / both / digit-based)*
26. How often do you devote attention to mental calculation and estimation in your mathematics lessons? *(. . . times a week)*
27. Do your students use a single or multiple strategies for mental multiplication?
28. Do your students use a single or multiple strategies for mental division?
29. Do your students use a single or multiple strategies for multidigit multiplication?
30. ⋆Do your students use a single or multiple strategies for multidigit division? *(one strategy / multiple strategies)*
31. How much time do you devote to mental calculation and estimation per week? *(. . . minutes)*
32. ⋆How often do you devote attention to basic skills in multiplication and division in mental calculation and estimation?
33. How often do you devote attention to roughly estimating the solution of a problem?
34. How often do you devote attention to applying approximations, estimations and rounding off? *(never / less than once a month / once a month / twice a month / at least once a week)*
35. ⋆How often do you devote attention to finding and using smart number-dependent strategies in mental calculation and estimation?
36. How often do you devote attention to letting students use multiple solution strategies for a single problem type in mental calculation and estimation? *(never / less than once a month / once a month / twice a month / at least once a week)*

37. Are calculators or computer software used during mathematics lessons? *(only calculators / both calculators and computer software / only computer software / neither)*

38. Do you instruct your students in the multiplication function of the calculator? *(yes / no)*

39. Do you instruct your students in the division function of the calculator? *(yes / no)*