

Beyond measurement:

Applications of explanatory IRT models to primary school mathematics tests

Marian Hickendorff

Leiden University, The Netherlands

Supervisors

- **Willem J. Heiser** - Leiden University, The Netherlands
- **Cornelis M. van Putten** - Leiden University, The Netherlands
- **Norman D. Verhelst** - CITO, National Institute for Educational Measurement

first: measurement with IRT

individual person effects

ability scale



individual item effects

25 x 22

109 x 87

9.7 x 6.3

- commonly (in MML formulation)
 - person effects: random [e.g., $\vartheta \sim N(0, \sigma_{\vartheta}^2)$]
 - item effects: fixed [e.g., $(\beta_1, \dots, \beta_i, \dots, \beta_l)$]
- but in principle, all combinations possible

		items	
		fixed	random
persons	fixed	FP-FI (JML)	FP-RI
	random	RP-FI (MML)	RP-RI

De Boeck, 2008

beyond measurement: explanatory IRT models (De Boeck & Wilson, 2004)

explanation of

- person effects
by person properties

ability scale

- item effects
by item properties

M_{boys}



M_{girls}



whole
numbers

decimal
numbers

Commonly

- person properties: fixed effects (e.g., latent regression)
- item properties: fixed effects (e.g., LLTM)

descriptive vs. explanatory IRT models

- descriptive IRT models
 - measurement of individual differences
- explanatory IRT models: effects of predictors / covariates
 - person properties *gender*
 - item properties *problem format*
 - person-by-item properties *strategy used (observed)*

→ interesting for educational psychologists:
testing substantive hypotheses

- do boys have a higher level of mathematical ability than girls?
- does presenting a mathematics item in a context make the item more difficult?
- are algorithmic solution strategies more accurate than non-algorithmic ones? (Hickendorff et al., 2009, *Psychometrika*)

current presentation

- applications of explanatory IRT models
- to data sets on primary school mathematics
- different models: combinations of
 - different levels of external predictors:
person / item / person-by-item
 - with fixed or random effects
 - different modes of randomness: persons / items
 - with different dimensionality: 1 or 2
 - with different software: SAS-NLMIXED, R-lme4, SAS-GLIMMIX

data set

- national assessments of end-of-primary-school mathematics (PPON-EB)
 - 2 most recent assessments: 1997 and 2004
 - problems on complex multiplication and division

	1997	2004	total
<i>N</i> students	551	995	1546
<i>k</i> items	■ 11 $\xleftrightarrow{4 \text{ common}}$ ■ 10 ■ 10 $\xleftrightarrow{5 \text{ common}}$ ■ 13		■ 16 ■ 19
assessment design	complete	incomplete	

data set - variables

- responses: for each trial (= student-by-item combination)
 - accuracy: incorrect vs. correct
- predictor variables
 - student characteristic
 - year of assessment
 - item characteristic
 - operation: multiplication vs. division
 - student-by-item characteristic:
 - solution strategy (nominal – 4 categories):
algorithmic / non-algorithmic / no written working / other
NB. observed variable!

part of the data set:

data set

stu- dent	year	mult – it 1		mult – it 2		div – it 1		...
		strat	score	strat	score	strat	score	
1	1997	Alg	1	-	-	N-Alg	1	...
2	1997	Alg	0	-	-	Alg	1	...
3	2004	-	-	NWW	1	N-Alg	0	...
4	2004	NWW	1	-	-	-	-	...
...

multiplication problem

What is the total price for 5 CD's of € 19.95 each?

$$\begin{array}{r} 19.95 \\ \underline{\quad 5x} \\ 442 \\ 99.75 \end{array}$$

algorithmic
[traditional]

1997: 65%
2004: 45%

$$\begin{array}{l} 5 \times 20 = 100 \\ 5 \times 0.05 = 0.25 \\ 100 - 0.25 = 99.75 \end{array}$$

non-algorithmic

1997: 7%
2004: 18%



no written work

1997: 17%
2004: 25%

estimating the IRT models

- IRT (Rasch) model in Generalized Linear Mixed Model (GLMM) framework
 - linear component: $\eta_{pi} = \theta_p + \beta_i X_i$
 - linking component: $P(y_{pi} = 1) = \frac{\exp(\eta_{pi})}{1 + \exp(\eta_{pi})}$
 - random component: $y_{pi} \sim \text{Bernoulli}(P)$

- software for GLMMs
 - SAS – proc NLMIXED
 - SAS – proc GLIMMIX
 - R – lmer function from lme4 package

GLMM – packages for fitting IRT-models

analyses

	NLMIXED	lmer	GLIMMIX
approximation in estimation	integral: Gauss-Hermite quadrature	integrand: Laplace	data: PQL / MQL
accuracy of results	best	biased	biased (seriously)
speed of computation	slow	fastest	OK
crossed random effects possible?	no	yes	yes
multidimensionality	M < 4 (time constraints)	max ??	max ??
discrimination parameters?	yes	no	no
multicategory response?	yes (ORD / NOM)	no	yes (ORD / NOM)
multilevel possible?	no	yes	yes

application 1:

4 descriptive IRT models

- data set selection:
 - multiplication problems, 1997 assessment
 - 551 students; 11 items → 6061 observations
- 4 descriptive IRT models
 - fixed persons – fixed items (FP-FI) → JML-version Rasch
 - fixed persons – random items (FP-RI) → person measurement
 - random persons – fixed items (RP-FI) → MML-version Rasch
 - random persons – random items (RP-RI) → crossed random effects model
- 3 model fit packages: NLMIXED, lmer, GLIMMIX

fit statistics and estimates of random effects

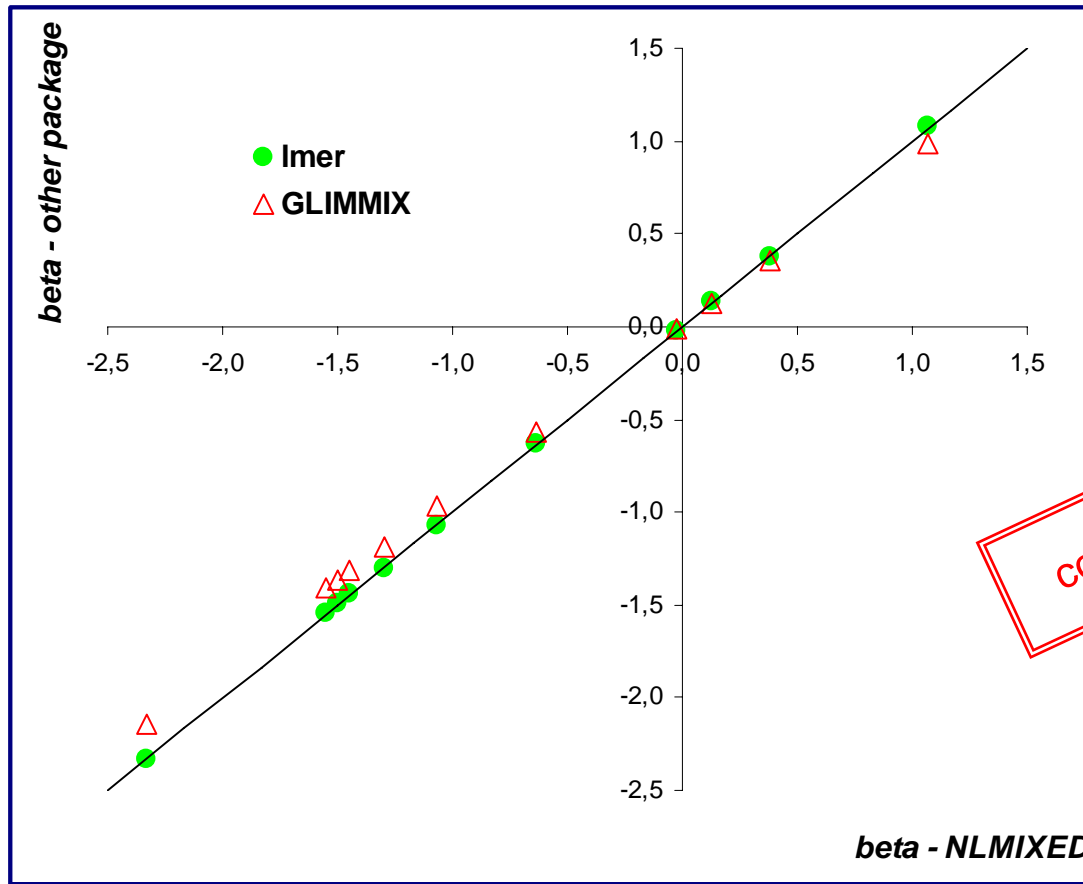
lowest BIC

Rasch –
MML formulation

		FP-FI	FP-RI	RP-FI	RP-RI
nr. of parameters		$2k = 22$	$k+2 = 13$	$k+1 = 12$	3
NLMIXED 20 Q-points	logLik	-2642	-2672	-3302	not possible (crossed random effects)
	σ_{θ}^2 (SE)	-	-	1.60 (.15)	
	σ_{β}^2 (SE)	-	1.18 (.15)	-	
Imer	logLik	-2642	-2672	-3307	-3339
	σ_{θ}^2 (SE)	-	-	1.55 (xx)	1.53 (xx)
	σ_{β}^2 (SE)	-	1.18 (xx)	-	0.93 (xx)
GLIMMIX	σ_{θ}^2 (SE)	convergence problems	convergence problems	1.26 (.12)	1.25 (.12)
	σ_{β}^2 (SE)			-	0.87 (.40)

comparison fixed difficulty parameters RP-FI (Rasch-MML) models

application 1



correlations > .9999

descriptive IRT models: conclusions

- psychometrically
 - fixed person models
 - lmer and NLMIXED identical results
 - random person models
 - comparable results among the 3 software packages
 - shrinkage effects R-lmer and GLIMMIX (approximation of the integrand)
- substantively
 - enough person-variation for measurement
 - according to AIC / BIC: better fit uncommon *fixed person* models than common *fixed item* models
 - next: beyond 'pure' measurement → explanation
 - what is the ability trend between 1997 and 2004?
 - how accurate are the different solution strategies?

application 2:

explanatory RP-FI models

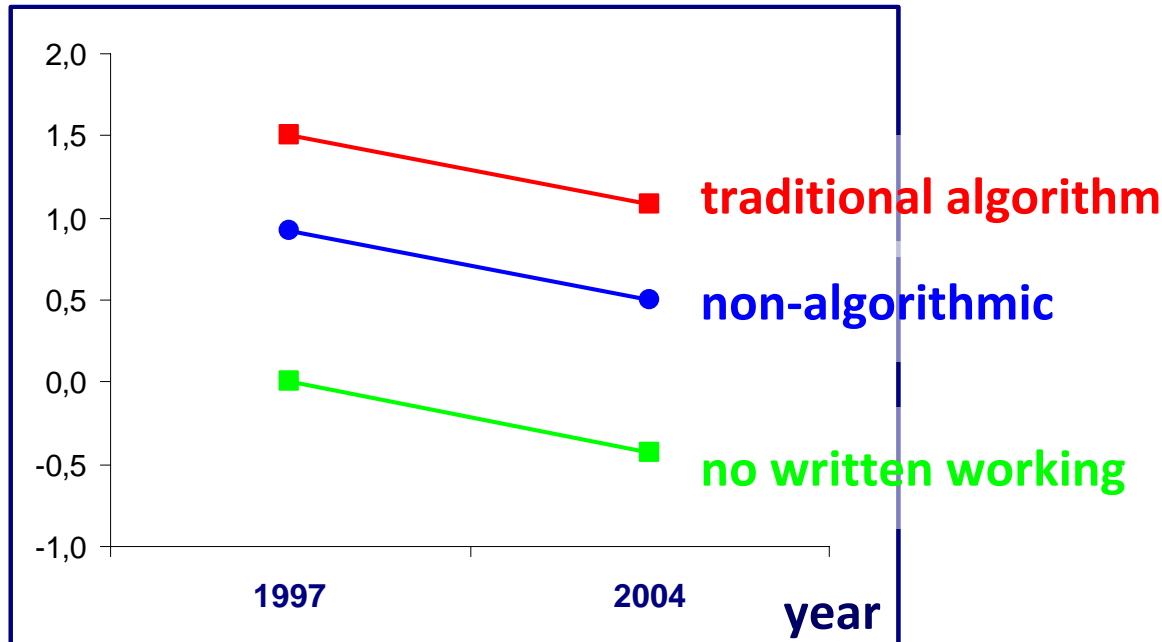
- data set selection:
 - multiplication problems, 1997 + 2004 assessment
 - trials with 'Other' strategies excluded
 - 1495 students; 16 items → 8706 observations

- explanatory IRT model:
 - fixed effects
 - item indicator [Rasch difficulties] *item level*
 - year of assessment [trend over time] *person level*
 - strategy used [strategy accuracy] *person-by-item level*
 - random effect
 - person error-variance *random over persons*

- 3 model fit packages: NLMIXED, lmer, GLIMMIX

graphical display fixed effect parameter estimates (NLMIXED)

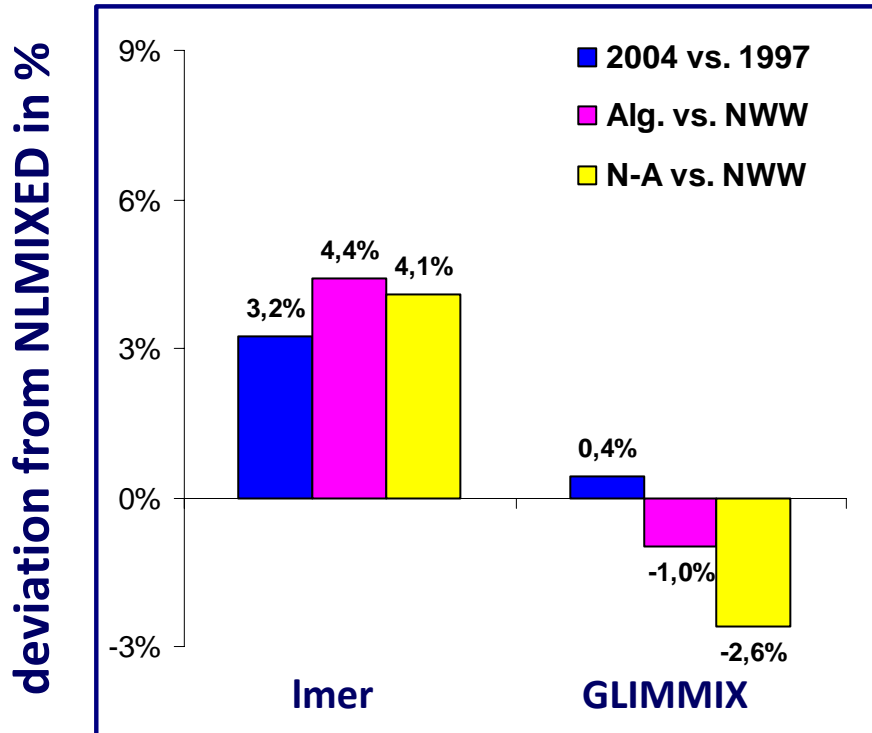
effect (logit scale)



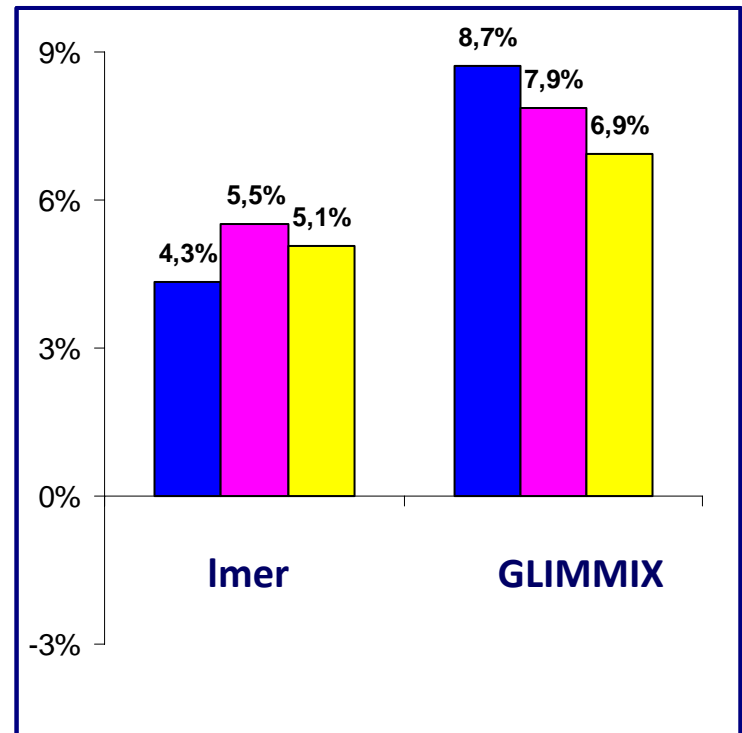
- interaction Year x Strategy n.s.
- results in 2PL-formulation very similar
- estimates σ_{θ}^2
 - NLMIXED: .97
 - lmer: .87
 - GLIMMIX: .69

fixed effect parameter estimates: deviations from NLMIXED

deviation z-values
(estimate / SE)





deviation effect sizes
(estimate / σ_{ϑ})



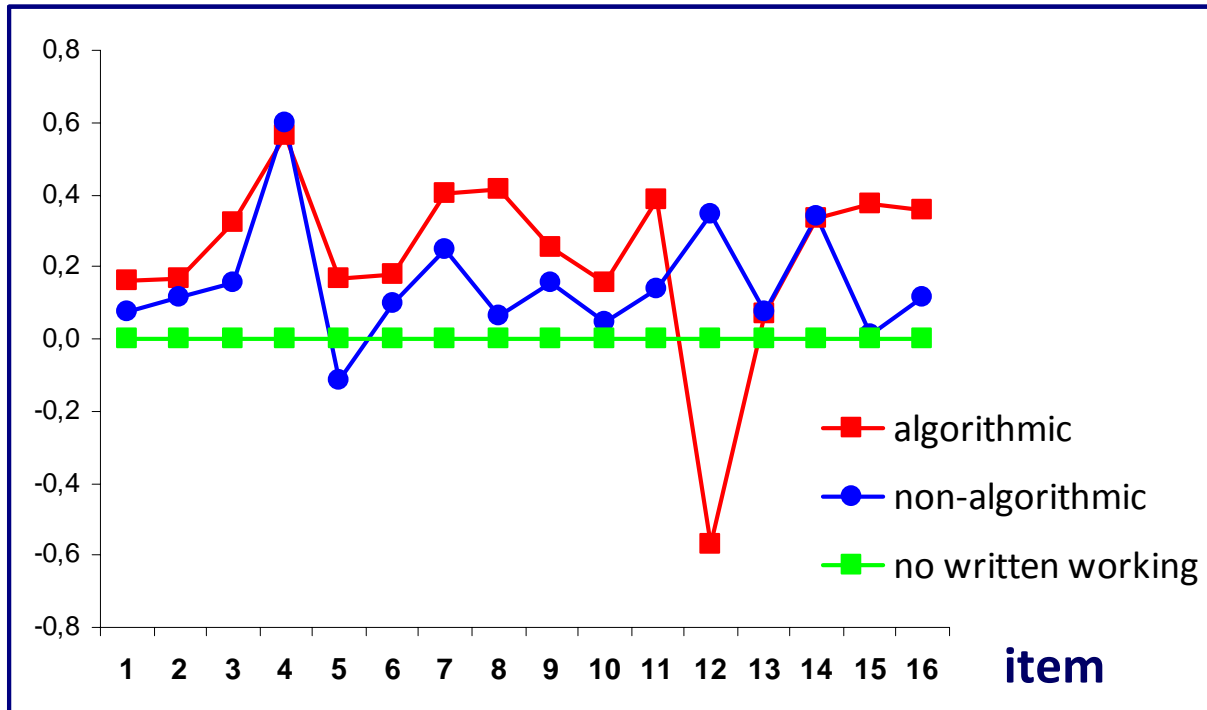
explanatory RP-FI models: conclusions

- psychometrically
 - comparable results among the 3 software packages
 - but shrinkage effects lmer and GLIMMIX
 - lmer and GLIMMIX seem to result in upwardly biased z-values and effect sizes, compared to NLMIXED

- substantively
 - corrected for shifts in solution strategy use between 1997-2004,
 - a significant ($z = 4.44$) and medium-sized ($ES = -.43$) decrease of ability between remains
 - significant differences between the accuracy of the strategies
 - traditional > non-algorithmic > no written working
 - but: assumed that they were equal for all items 
 - next: interaction strategy x item, modeled with random effect 

variation strategy accuracies over items

deviation P (correct) from No Written Working



application 3: explanatory RP-RI models

- data set selection: as in application 2
- explanatory IRT model:
 - fixed effects
 - year of assessment [trend over time] *person level*
 - strategy used [strategy accuracy] *person-by-item level*
 - random effects
 - strategy used [item difficulty *per strategy*] *random over items*
 - person error-variance *random over persons*
- 2 model fit packages: lmer + GLIMMIX

variance – correlation matrix

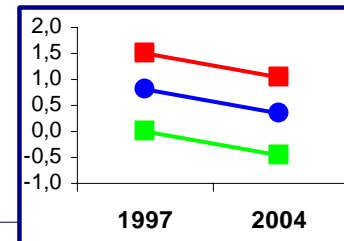
strategy effects random over items (Imer)

	alg.	non-alg.	NWW
algorithmic	.66		
non-algorithmic	.77	.61	
no written working	.75	.67	.81

- model with random item difficulties *per strategy* better fit (AIC / BIC) than model with random item difficulties (cf. $\sigma_{\theta}^2 = .58$)
- results GLIMMIX: comparable to Imer (but smaller variances)

explanatory RP-RI models: conclusions

- psychometrically
 - making a subject-by-item predictor (strategy) random over items works
 - accounts for inter-item differences in 'strategy' effects → better fit
 - but results in larger SEs of fixed effects 'strategy'
- substantively
 - there is significant variation between items in strategy accuracy differences
 - items differ from each other in difficulty level per strategy; the largest variation is in the difficulty level given the 'No Written Work' strategy
 - if an item is more difficult with one strategy, it is also more difficult with the other two strategies
 - fixed main effects Year and Strategy remain significant...



application 4: between-item multidimensional models

- data set selection
 - multiplication + division problems, 2004 assessment
 - 995 students; 23 items → 8904 observations
- multidimensional IRT model:
 - fixed effects
 - item indicator [item difficulty] *item level*
 - random effects
 - multiplication ability
 - division ability

} *multivariate random over persons*

NB. Making an item property random over persons → MIRT model
- 3 model fit packages: NLMIXED, lmer, GLIMMIX

fit statistics and parameter estimates multidimensional IRT models

	logLik	$\sigma_{\vartheta_M}^2$ (SE)	$\sigma_{\vartheta_D}^2$ (SE)	r_{MD} (SE)
NLMIXED	-5129	1.56 (.19)	2.56 (.25)	.87 (.04)
Imer	-5145	1.37 (xx)	2.37 (xx)	.93 (xx)
GLIMMIX	-	1.08 (xx)	1.68 (xx)	.96 (xx)

variances

latent correlation

multidimensional IRT models: conclusions

- psychometrically
 - item difficulties NLMIXED / lmer / GLIMMIX similar ($r > .9999$)
 - but latent correlation estimates deviate
 - lmer / GLIMMIX upwardly biased, compared to NLMIXED
 - but substantially faster...
- substantively
 - the correlation between multiplication ability and division ability is very high
 - but not perfect: there is significant non-shared variation
 - not the same 1-dimensional construct

general discussion - psychometrically

- GLMM software usable to fit IRT models
 - NLMIXED most accurate, but
 - slow
 - no crossed random effects possible
 - lmer seems the best alternative
 - (approximate) log-likelihood close to NLMIXED
 - some shrinkage of fixed effects / random effects
 - disadvantage: no SEs for random effect variances/correlation (yet)
 - maybe for multidimensional models too much deviation?
 - GLIMMIX: not advisable
 - the largest deviations from NLMIXED
 - no (approximate) log-likelihood: comparison of models?
 - can also be quite slow / convergence problems
- model building with lmer → model check with NLMIXED

general discussion - psychometrically

- random item approach fruitful
 - large reduction in number of parameters
 - small cost in log-likelihood (better AIC / BIC)
 - approximately equal person variance
 - also substantive reasons for treating items random (see De Boeck, 2008)

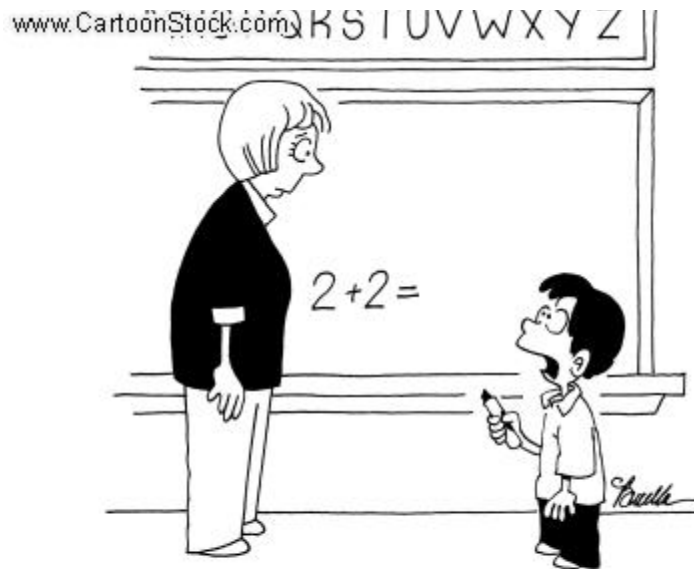
general discussion - substantively

- proficiency on multiplication problems by Dutch 6th graders
 - has decreased between 1997 and 2004
 - different solution strategies have different success rates (accuracies)
 - algorithmic > non-algorithmic > no written working
 - the differences in success rates vary over items
 - multiplication ability is highly related to, but still different from, proficiency on division problems

- two shifts contributed to the decline in achievement
 - increase of use of less accurate strategies
 - algorithmic ↓
 - non-algorithmic / no written working ↑
 - accuracy decrease within each strategy (!)

the end

thank you for your attention!



"I plan on becoming an automobile mechanic when I grow up. Would you settle for an estimate?"