Measurement numeracy education for prospective elementary school teachers : effects of inductive and deductive teaching on classroom interaction and student performance

Houwelingen, M.J. van

CHAPTER 6

Effects on student performance

Measurement numeracy improvement: effects of contrasting didactic approaches and teacher effects

6.1 Introduction

Improving classroom interaction in mathematics classes will probably have a positive effect on numeracy improvement (see Chapter 5). In Chapter 4, the development of two different types of lesson series on measurement for prospective elementary school teachers are described: one with a pure deductive and one with a pure inductive didactic approach to classroom interaction. In this chapter, a quasi-experiment with pretest-posttest design, used to estimate the effect of teachers and the didactic approach on students' measurement numeracy, is described. Since an inductive didactic approach induced more classroom interaction time, and more stimulating questions, than a deductive didactic approach (see Chapter 5), the effects of these variables on students' measurement numeracy were also estimated. Furthermore, Freeman et al. (2014) argued that active learning enhances student performance in mathematics, so we are also interested in the effect of student behavior during class. We also recorded migration status and previous education, because earlier research suggests that an inductive approach has a lower effect on improvement of low performing students (Slavin & Lake, 2008) and non-native students (Civil, 2014). Teachers' preference for a didactic approach were also recorded, because a mismatch between the preference and the didactic approach used in the lessons might have a negative effect.

The research questions to be answered in this chapter are as follows:

1. To what extent does the didactic approach (inductive / deductive) affect students' measurement numeracy?
2. To what extent do different teachers affect students' measurement numeracy?
3. To what extent does student behavior during class affect students' measurement numeracy?
4. To what extent does the type of teacher questions affect students' measurement numeracy?

5. To what extent does the amount of classroom interaction affect students' measurement numeracy?
6. To what extent does previous education of the student and of his parents relate to students' measurement numeracy?
7. To what extent do gender, age, and the student's home language relate to students' measurement numeracy?
8. To what extent does a mismatch between the teacher's preference for the didactic approach and the actual didactic approach used in the lessons affect students' measurement numeracy?

6.2 Method

6.2.1 Sample

All ten freshmen groups of the Rotterdam School of Education participated in the quasi-experiment. The sample for this chapter consisted of 153 students who completed both the pretest and the posttest (see Chapter 2). Their mean age was 19.5 years (min=16, max=32), and 84% were female. The mean WISCAT score (a mandatory national math test for elementary school teacher training college freshmen, the norm is 103) was 100.9. 23% of the students also spoke another language besides Dutch at home, 55% only spoke Dutch at home, and for 22% it is unknown. 36% of the students did not attend mathematics classes in their previous education, or scored insufficiently; 42% scored sufficiently or higher at mathematics in their previous education, and for 22% it is unknown. 44% of the students had MBO or lower as highest previous education, 56% had HAVO or higher. 44% of the students' mothers had MBO or lower as highest previous education, 25% had HAVO or higher, and for 31% it is unknown. 42% of the students' fathers had MBO or lower as highest previous education, 24% had HAVO or higher, and for 34% it is unknown.

6.2.2 Materials

Performance tests. Students' measurement skills were measured twice, using the pretest and posttest that were developed earlier (see Chapter 2). Skills were measured separately for three different aspects: *understanding relationships within the metric system* (metric), *calculating with scale* (scale) and *calculating length, area, and volume* (area). For each aspect, both the pretest and the posttest contained ten items (five unique items in the pretest, five unique items

in the posttest, and five items in both the posttest and the pretest). Example items are shown in Table 6.1.

Table 6.1. *Item examples per aspect.*

Understanding relationships within the metric system: 0.034 km = …… dm. 450 are = …… m².
Calculating with scale: The distance from Rotterdam to Paris is 450 km. My map has a scale of 1: 3,000,000. How many cm is the distance from Rotterdam to Paris on my map? Explain. On my map the area of the living room is 5 dm². In real life the area of the living room is 45 m². Explain how you find the scale that was used for my map.
Calculating length, area, and volume: The area of a rhombus is 16 dm². One diagonal is twice as long as the other. Determine the size of the diagonals, and explain. The volume of a pack of lemonade is 1.5 liter. The pack has a length of 0.75 dm and a width of 1 dm. Calculate the height of the pack, and explain.

Student questionnaires. For every lesson, students reported the following information about themselves: during the lesson a) the number of questions the student asked, b) the number of student interactions with a peer about the mathematical subject at hand, c) the attention (scale 1-5) the student had for the instruction, d) the attention (scale 1-5) the student had for the exercises, and e) the attention (scale 1-5) the student had for the lesson in general.

Classroom interaction measures. For every two-minute slot of every lesson, two different aspects of classroom interaction were recorded: 1) the number of seconds one or more students talked about the mathematical subject at hand, and 2) the type of teacher questions (Nelissen, 2002) (see Chapter 5).

6.2.3 Design and procedure

A quasi-experiment, with a pretest-posttest design, was used to estimate the effect of two didactic approaches on student performance. Five teachers all taught one group using the deductive approach, and one group using the inductive approach (see Chapter 5). Teachers were pre-assigned to the 10 pre-existing student groups by the Rotterdam School of Education

administration. Under these conditions, student groups were randomly assigned to the didactic approach, using the following procedure: for every teacher, one of the student groups was assigned at random (by flipping a coin) to either the inductive or the deductive approach. The other group would automatically be taught using the other didactic approach. We could not randomly assign students to a didactic approach because there were pre-existing groups. However, since there were ten groups in the sample, cluster randomization (the next best thing in terms of bias and power) is good enough (Van Breukelen, 2013). The procedure was as follows: first students took a pretest, then they took classes (with either a deductive or an inductive didactic approach, depending on their student group), completed a questionnaire after each lesson, and finally took a posttest (two weeks after the final lesson).

6.2.4 Statistical analyses

Multilevel analyses. To find out if multilevel analyses are necessary (usually, multilevel analyses are used with pre-existing groups), tests were performed with MLWIN (Rasbash, Steele, Browne, & Prosser, 2015) to see if the data (student scores) showed significant proportions of variance on the teacher level and/or on the student group level. Tests were performed separately for scores on the three different aspects (*metric system*, *scale calculation*, and *length, area, and volume calculation*). In the nested structure, students were given subscript i, student groups (10 groups) were given subscript j, and teachers (5 teachers) were given subscript k. The models for predicting student scores on the *metric system* aspect were as follows: Model 1 has only the student level, model 2 has a student level and a group level, model 3 has a student level and a teacher level, and model 4 has a student level, a group level and a teacher level.

Model 1. $Post\_metric_{ijk} = \beta_{0i} * constant + \beta_1 * (pre\_metric\text{-}gm)_{ijk}$

$(\beta_{0i} = \beta_{11} + e_{0ijk})$

Model 2. $Post\_metric_{ijk} = \beta_{0ij} * constant + \beta_2 * (pre\_metric\text{-}gm)_{ijk}$

$(\beta_{0ij} = \beta_{21} + u_{0jk} + e_{0ijk})$

Model 3. $Post\_metric_{ijk} = \beta_{0ik} * constant + \beta_3 * (pre\_metric\text{-}gm)_{ijk}$

$(\beta_{0ik} = \beta_{31} + v_{0k} + e_{0ijk})$

Model 4. $Post\_metric_{ijk} = \beta_{0ijk} * constant + \beta_4 * (pre\_metric\text{-}gm)_{ijk}$

$(\beta_{0ijk} = \beta_{41} + v_{0k} + u_{0jk} + e_{0ijk})$

Notes.

pre_metric-gm is the deviation from the grand mean of the pretest score for the aspect *metric system*.

u_{0jk} and v_{0k} are student group-specific and teacher-specific deviations from the mean. If they are not significantly different from zero, there is no reason to use multilevel analyses.

The -2*loglikelihood of the models 2, 3 and 4 were compared with the one from model 1. A difference (delta-2LL) of less than 2.706 (since variances cannot be negative, we tested with 10% instead of 5%) (Hox, 2010, p. 49) would mean that we do not need to use multilevel analyses (in that case adding the extra level in that model does not significantly improve the model). Table 6.2 shows that there is no reason to use multilevel analyses. Consequently, though it might still be possible that different (group or other) variables cancel each other out, chances are that no group variable has a significant effect on student performance.

Table 6.2. *Multilevel check: -2*loglikelihood for the four models, for three different aspects.*

	metric	scale	area
student level (model 1)	673.641	615.176	703.064
student+group (model 2)	671.177	615.174	702.393
delta-2LL	2.464	0.002	0.671
student+teacher (model 3)	671.598	615.176	702.316
delta-2LL	2.043	0	0.748
student+group+teacher (model 4)	670.975	615.174	702.235
delta-2LL	2.666	0.002	0.829

ANCOVA or Repeated Measures ANOVA. Since we can conclude that multilevel analyses are not necessary, analyses (for each aspect separately) can be performed with either ANCOVA (using the posttest scores as dependent variable and the pretest scores as covariate, which adds to the power), or repeated measures analyses (using the difference between the posttest score

and the pretest score as dependent variable). Since these two methods are the same if there are no group effects at the pretest (and that is the case in this study, see Table 6.3), and correction for measurement error in the pretest comes down to an ANCOVA if there are no true group differences at the pretest (Van Breukelen, 2013), analyses will be performed using ANCOVA.

Pre-existing differences between the two conditions. In total, 153 students completed both the pretest and the posttest (80 students were in the deductive group, and 73 students were in the inductive group). On average, students in the deductive group performed worse at the pretest on the metric aspect (M=4.9) than students in the inductive group (M=5.7). This difference was not significant: $t(151)=-1.89$, $p > .05$. On average, students in the deductive group performed worse at the pretest on the scale aspect (M=5.1) than students in the inductive group (M=5.3). This difference was not significant: $t(151)=-0.38$, $p > .05$. On average, students in the deductive group performed better at the pretest on the area aspect (M=3.2) than students in the inductive group (M=2.7). This difference was not significant: $t(151)=1.17$, $p > .05$ (see Table 6.3). On average, the WISCAT score of students in the deductive group (M=101.5) was higher than the WISCAT score in the inductive group (M=100.2). This difference was not significant: $t(141)=0.33$, $p>.05$ (see Table 6.4).

Table 6.3. *Independent t-test for pretest score differences in ten pre-existing groups ($N_{deductive}=80$, $N_{inductive}=73$).*

	Deductive		Inductive				
	M	*SE*	*M*	*SE*	*t*	*df*	*p*
metric	4.9	0.3	5.7	0.3	-1.89	151	.061
scale	5.1	0.4	5.3	0.4	-0.38	151	.705
area	3.2	0.3	2.7	0.3	1.17	151	.244

Table 6.4. *Independent t-test for WISCAT score differences in ten pre-existing groups ($N_{deductive}=76$, $N_{inductive}=67$).*

	Deductive		Inductive				
	M	*SE*	*M*	*SE*	*t*	*df*	*p*
WISCAT	101.5	2.6	100.2	3.4	0.33	141	.746

Pre-existing differences between the five teachers (results of one-way ANOVA). On average, students in the groups of teacher1 (M=89.7) performed worse in the WISCAT than students in other groups, particularly when compared to students in the groups of teacher3 (M=108.9)

(see Table 6.5). The difference between teachers was significant: $F(4,138)=3.20$, $p<.05$ (see Table 6.6).

Table 6.5. *Mean WISCAT scores per teacher.*

	N	*M*	*SE*
teacher1	33	89.7	4.0
teacher2	30	97.6	4.9
teacher3	31	108.9	4.4
teacher4	28	104.6	3.7
teacher5	21	106.4	5.8
Total	143	100.9	2.1

Table 6.6. *One-way ANOVA to check for significant differences in WISCAT score per teacher.*

	SS	*df*	*MS*	*F*	*p*
Between Groups	7494.62	4	1873.66	3.20	.015

Note: post-hoc test revealed a significant difference ($p=.018$) between the WISCAT scores of students in classes of teacher1 and teacher3.

On average, students in the groups of teacher1 ($M=3.5$) performed worse at the pretest for the aspect *scale* than students in other groups, particularly when compared to students in the groups of teacher3 ($M=6.2$) (see Table 6.7). The difference between teachers was significant for the aspect *scale*: $F(4,148)=3.85$, $p<.05$ (see Table 6.8). Student scores on the aspects *metric* and *area* did not differ much per teacher (see Table 6.7, and Table 6.8).

Table 6.7. *Differences per teacher in pretest score for the aspects metric, scale, and area.*

		metric		scale		area	
	N	*M*	*SE*	*M*	*SE*	*M*	*SE*
teacher1	34	4.7	0.4	3.5	0.6	2.2	0.4
teacher2	32	4.6	0.5	5.0	0.6	2.9	0.5
teacher3	34	5.8	0.4	6.2	0.5	3.6	0.5
teacher4	32	5.8	0.5	5.7	0.5	2.8	0.4
teacher5	21	5.5	0.5	5.8	0.6	3.5	0.7
Total	153	5.3	0.2	5.2	0.3	3.0	0.2

Table 6.8. *One-way ANOVA to check for significant differences per teacher in pretest score for the aspects metric, scale, and area.*

	SS	*df*	*MS*	*F*	*p*
metric	41.07	4	10.27	1.61	.174
scale	147.85	4	36.96	3.85	.005
area	42.13	4	10.53	1.49	.208

Note: post-hoc test revealed a significant difference between the pretest scores at the aspect scale in groups of teacher1 and teacher 3 (p=.005), and between the groups of teacher1 and teacher4 (p=.041).

As the pretest scores at the three aspects and the WISCAT score both reflect (parts of) mathematical skills, the correlations between the three pretest scores and the WISCAT score were estimated (see Table 6.9).

Table 6.9. *Spearman's rho correlation coefficients between pretest scores and WISCAT score (1-tailed).*

		metric	scale	area	WISCAT
metric	*r*				
	N				
scale	*r*	.43**			
	N	153			
area	*r*	.44**	.64**		
	N	153	153		
WISCAT	*r*	.40**	.71**	.61**	
	N	143	143	143	

**. Correlation is significant at the 0.01 level (1-tailed). All p's <.001.
Notes: variables are not normally distributed (see Table 6.10 and Figure 6.1). Correlation between WISCAT and pretest total: r=.72 (p<.001).

Table 6.10. *Normality (kurtosis and skewness) of the three pretest scores and the WISCAT score.*

		metric	scale	area	WISCAT
N	Valid	210	210	210	228
	Missing	49	49	49	31
Mean		5.06	5.05	2.91	99.03
SE		0.18	0.22	0.18	1.70
SD		2.61	3.23	2.62	25.68
Skewness		-0.27	-0.24	0.59	0.76
SE of Skewness		0.17	0.17	0.17	0.16
Z-skew		-1.59	-1.42	3.51*	4.72*
Kurtosis		-0.64	-1.16	-0.78	1.38
SE of Kurtosis		0.33	0.33	0.33	0.32
Z-kurtosis		-1.90	-3.47*	-2.34	4.31*

*. Significantly different (.001 level, i.e. Z>3.29) from normal. WISCAT looks normal, except for three very high scores. The pretest scores at scale and at area look skewed, because a large proportion of students did not answer any question correctly (floor effect).

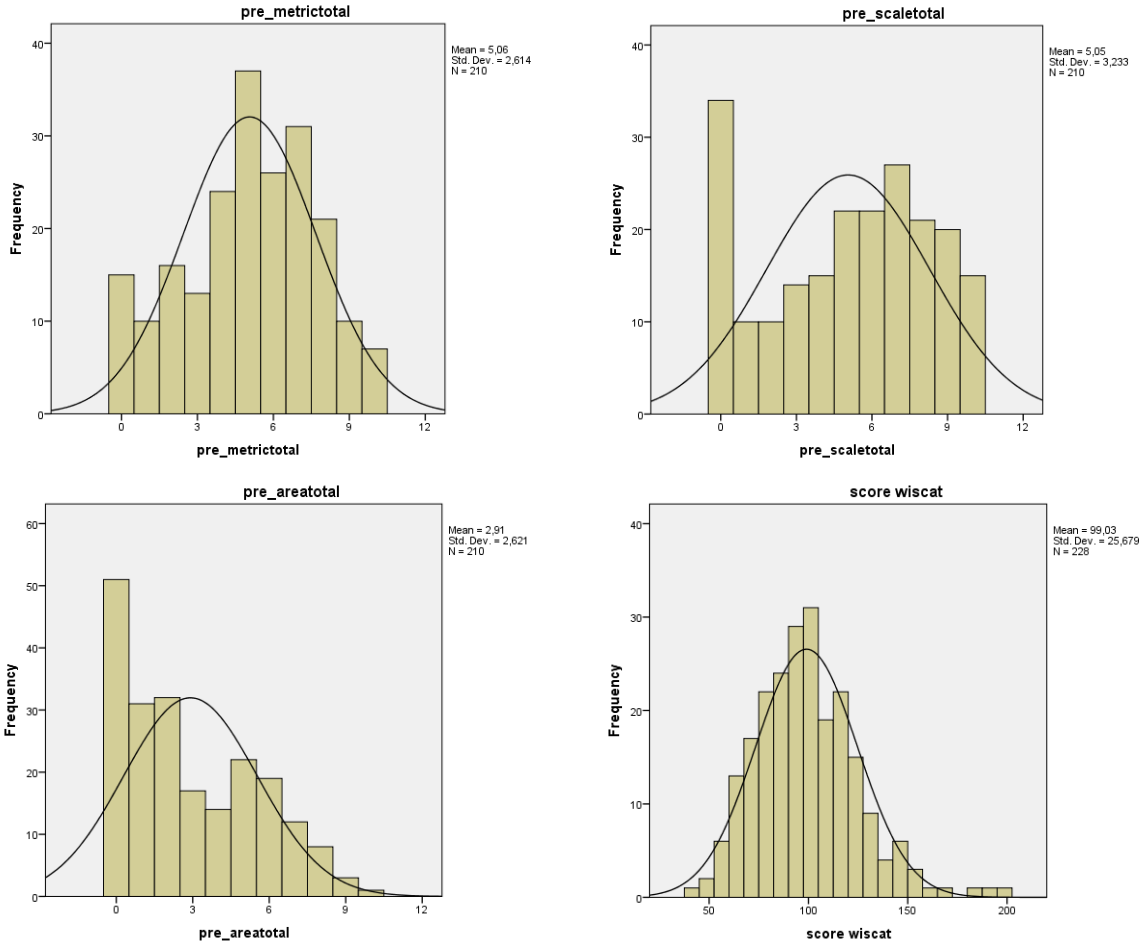

Figure 6.1. Distribution of the pretest scores of the three aspects and the WISCAT score

Using the pretest score as a covariate allows for estimating the effect of other variables on the posttest score, after controlling for the effect of the pretest score on the posttest score. The pretest score was therefore used as the first covariate to control for. Later, other continuous variables were also used as covariates. Categorical variables were used as fixed factors. In the first model, the effect of the condition and the teacher on the posttest score was estimated, after controlling for the effect of the pretest score on the posttest score. Next, other variables were added to see if they also had an effect on the posttest score. If a variable did not have a significant effect (at the .05 level) on the posttest score, it was removed from the model. The first four models (for the aspect *metric system*) were as follows:

1. $Post\_metric_i = b_0 + \beta_1 * (pre\_metric_i) + \beta_2 * (condition_i) + \beta_3 * (teacher_i) + error_i$.
2. $Post\_metric_i = b_0 + \beta_1 * (pre\_metric_i) + \beta_2 * (condition_i) + \beta_3 * (teacher_i) + \beta_4 * ((condition_i) * (teacher_i)) + error_i$.
3. $Post\_metric_i = b_0 + \beta_1 * (pre\_metric_i) + \beta_5 * (WISCAT_i) + \beta_2 * (condition_i) + \beta_3 * (teacher_i) + error_i$.
4. $Post\_metric_i = b_0 + \beta_1 * (pre\_metric_i) + \beta_5 * (WISCAT_i) + \beta_6 * (teacher\_question\_type_i) + \beta_2 * (condition_i) + \beta_3 * (teacher_i) + error_i$.

6.3 Results

Research question 1. The mean posttest score was higher than the mean pretest score for all three aspects (metric: ($M_{posttest}$=6.6, $M_{pretest}$=5.3; scale: $M_{posttest}$=6.5, $M_{pretest}$=5.2; area: $M_{posttest}$=5.4, $M_{pretest}$=3.0) (see Table 6.11). For all three aspects, the progress was substantial and significant, i.e. all effect sizes were large, and all *p*'s were <.001 (metric: $t(152)=5.97$, $r=.44$, $p<.001$; scale: $t(152)=6.47$, $r=.46$, $p<.001$; area: $t(152)=11.56$, $r=.68$, $p<.001$).

Table 6.11. *Mean pretest score and mean posttest score per aspect (N=153).*

	metric		scale		area	
	M	*SE*	*M*	*SE*	*M*	*SE*
pretest	5.3	0.2	5.2	0.3	3.0	0.2
posttest	6.6	0.2	6.5	0.2	5.4	0.2

Research question 2. The mean posttest score per aspect was roughly the same in both conditions (see Table 6.12). The mean posttest score for students of teacher 1 was roughly the same in both conditions. The mean posttest score for teacher 3 and teacher 4 (who both had a

preference for a deductive didactic approach, see Table 4.1) was higher in the deductive approach than in the inductive approach. For teacher 2 and teacher 5 (who both had a preference for an inductive didactic approach) the opposite applied (see Figure 6.2). However, the interaction effect between teacher and didactic approach was not significant, nor was a mismatch between the teacher's preference and the didactic approach (see Table 6.15).

Table 6.12. *Mean posttest scores for metric, scale, and area, by teacher by condition.*

		metric	*SD*	scale	*SD*	area	*SD*
teacher1	deductive (*N*=16)	5.6	2.0	5.9	2.7	4.0	3.3
	inductive (*N*=18)	5.7	2.5	5.8	2.1	3.9	3.1
teacher2	deductive (*N*=23)	6.5	2.4	5.7	2.5	4.8	2.6
	inductive (*N*=9)	8.1	1.8	7.4	2.4	6.4	2.2
teacher3	deductive (*N*=15)	7.7	1.7	7.5	1.4	7.0	2.2
	inductive (*N*=19)	6.9	2.6	6.4	2.6	5.6	3.3
teacher4	deductive (*N*=14)	7.9	2.3	7.3	2.4	6.9	2.6
	inductive (*N*=18)	6.3	2.2	5.9	2.2	4.9	2.3
teacher5	deductive (*N*=12)	5.3	2.4	7.1	1.4	5.5	3.2
	inductive (*N*=9)	6.9	1.9	7.8	1.1	6.7	2.9
Total	deductive (*N*=80)	6.6	2.3	6.6	2.3	5.5	3.0
	inductive (*N*=73)	6.6	2.4	6.4	2.3	5.2	2.9

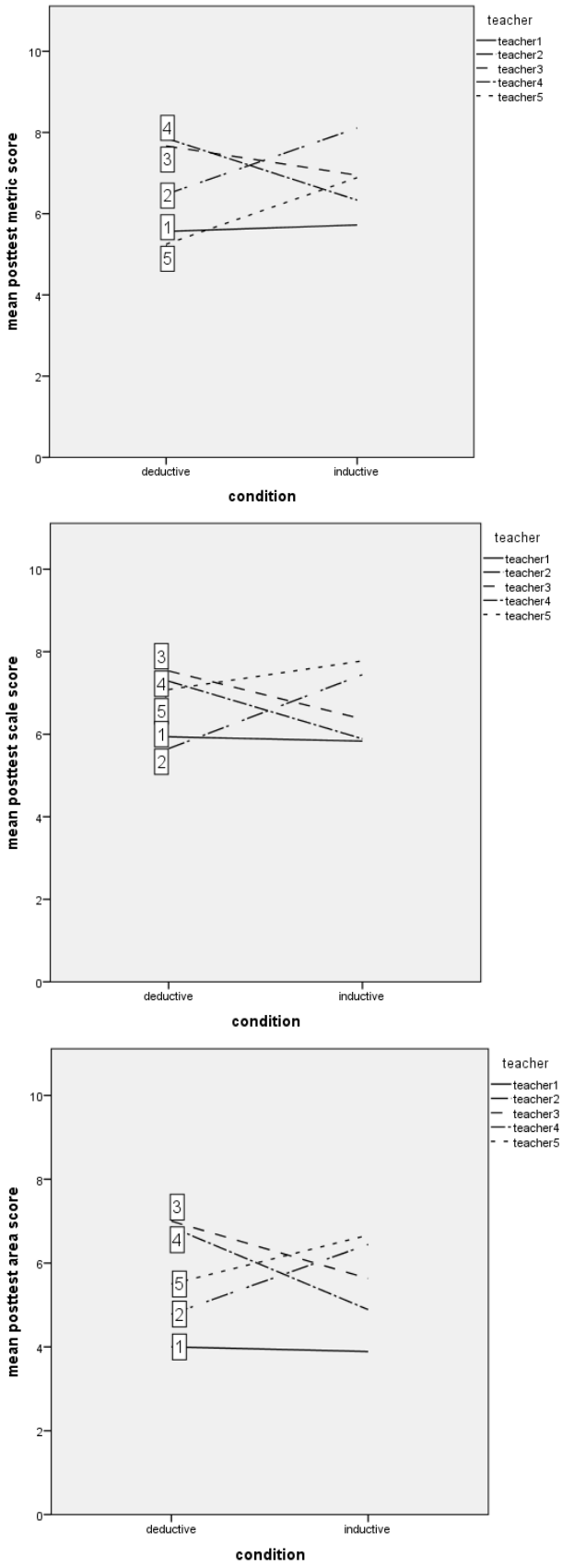

Figure 6.2. Mean scores on posttest by teacher and condition for metric (top), scale (middle), and area (bottom)

Research question 3. After each lesson, students were asked to answer five questions about classroom interaction and about their attention to the lesson (almost no attention, a little, average, much, full). Table 6.13 and Table 6.14 show the results of the student questionnaires. Students said the attention they had for the lesson, in both approaches, was between average and much. In the inductive approach, students said they asked more questions (1.7 compared to 1.3 questions), and talked more to their peers (4.0 times compared to 3.7 times).

Table 6.13. *Results of student questionnaires in the deductive approach (drawn after each lesson).*

		Classroom interaction		Attention (scale 1-5) to:		
		# questions	# peer talk	instruction	exercises	total
N	Valid	79	78	79	79	79
	Missing	1	2	1	1	1
M		1.3	3.7	4.0	3.6	3.5

Note: Scale (1-5): 1=no, 2=little, 3=average, 4=much, 5=full.

Table 6.14. *Results of student questionnaires in the inductive approach (drawn after each lesson).*

		Classroom interaction		Attention (scale 1-5) to:		
		# questions	# peer talk	instruction	exercises	total
N	Valid	72	72	73	73	73
	Missing	1	1	0	0	0
M		1.7	4.0	3.7	3.6	3.4

Note: Scale (1-5): 1=no, 2=little, 3=average, 4=much, 5=full.

Research question 1 to 8. The ANCOVA analyses showed that – after controlling for the pretest score and the WISCAT score, which both had a significant effect on the posttest score in all models –only one variable had a significant effect on the posttest score, and this was only the case for one aspect (see Table 6.15 and Table 6.16). This variable was mother's highest education, and it only had a significant effect ($F(2,133)=4.437$, $p=.014$) on the posttest score for the area aspect. This was a small effect ($\eta=.063$). On average, students whose mother's education was HAVO or higher scored higher on the area aspect ($M=6.2$, $SE=.46$, $N=38$) than students whose mother's education was MBO or lower ($M=5.0$, $SE=.35$, $N=68$). None of the other measured independent variables had a significant effect on the posttest score, with one exception: in the base model without controlling for the effect of WISCAT, the teacher had a significant effect on the posttest score of the metric aspect. However, after a

Bonferroni[5] correction (since there were three aspects, p=.031*3=.093), this effect was not significant.

Table 6.15. *Effect on posttest score for metric, scale, and area.*

				metric		scale		area	
	Effect	*N*	*df*	*F*	*p*	*F*	*p*	*F*	*p*
Base model	pretest	153	1	21.87	.000	82.67	.000	64.12	.000
	condition	153	1	0.18	.671	0.68	.413	0.00	.984
	teacher	153	4	2.73	.031	0.96	.430	1.98	.100
Base model +	condition*teacher	153	4	2.09	.085	1.27	.287	1.55	.191
Base model +	WISCAT (=Model3)	143	1	12.14	.001	12.54	.001	27.30	.000
Model3 +	question type	143	1	0.57	.453	0.34	.563	0.13	.716
Model3 +	interaction time	143	1	0.11	.736	0.03	.864	1.27	.262
Model3 +	# student questions	142	1	1.22	.271	2.01	.158	0.58	.448
Model3 +	# peer interaction	140	1	0.08	.780	0.03	.868	0.34	.559
Model3 +	attention instruction	142	1	0.34	.560	0.44	.507	0.02	.883
Model3 +	attention span total	142	1	0.80	.372	0.05	.827	0.67	.415
Model3 +	attention exercises	142	1	1.05	.308	0.51	.476	0.35	.557
Model3 +	gender	143	1	0.00	.974	0.01	.906	0.03	.865
Model3 +	home language	143	2	0.38	.684	0.03	.970	0.86	.426
Model3 +	math prev. education	143	2	0.32	.727	1.19	.307	1.55	.217
Model3 +	prev. education	143	1	0.78	.378	0.17	.680	0.02	.881
Model3 +	prev. education mom	143	2	0.74	.478	0.09	.915	4.44	.014
Model3 +	prev. education dad	143	2	0.68	.506	0.34	.712	0.94	.393
Model3 +	age	143	1	0.16	.691	1.68	.197	0.52	.472
Model3 +	presence	143	1	0.20	.657	1.79	.183	1.00	.319
Model3 +	mismatch teacher's preference/condition	143	1	3.46	.065	0.01	.912	1.14	.288

Notes:
Base model and Model3 (for the aspect *metric system*):
Base model: $Post\_metric_i = b_0 + \beta_1 * (pre\_metric_i) + \beta_2 * (condition_i) + \beta_3 * (teacher_i) + error_i$.
Model3 = Base model + WISCAT.
The pretest and WISCAT effect were significant in all models.
After controlling for pretest and WISCAT, none of the models showed significant condition or teacher effects (all p's > .05).
Levene's tests were not significant in all analyses.
There were no violations of homogeneity of regression slopes in any of the analyses, i.e. no significant factor-covariate interaction effects (three aspects were tested, so we used a Bonferroni correction, and tested with a .003 significance level).
Due to lack of degrees of freedom, no full factorial model was used.

[5] A Bonferroni correction was used to compensate for multiple hypothesis testing. Since we tested for three aspects on the same sample, we multiplied *p* by 3.

Table 6.16. *Model3 effect sizes (partial eta squared) and regression weights, of pretest and WISCAT, for metric, scale, and area.*

	metric		scale		area	
	η^2	*B*	η^2	*B*	η^2	*B*
pretest	.059	.221	.150	.308	.073	.292
WISCAT	.082	.028	.085	.028	.168	.052

6.4 Conclusion and discussion

Although student characteristics like previous education, mathematical history, home language, and gender have a significant and substantial effect on their pretest score (see Appendix), these effects on the posttest score disappear after controlling for the pretest score and the WISCAT score. Although these characteristics have significant effects on students' initial score, they do not have an effect on performance gains. After controlling for the pretest score and the WISCAT score, the didactic approach (inductive / deductive) (Klahr, 2009; Sweller, Kirschner & Clark, 2007), the teacher, student behavior during class (Freeman et al., 2014), the type of questions the teacher asks (Nelissen, 2002), and the classroom interaction time all had no significant effect on students' measurement numeracy (research questions 1, 2, 3, 4, and 5). Nor did students' and parents' previous education, mathematical background, gender, age, or the student's home language have a significant effect on students' measurement numeracy (although there was a small significant effect of the students' mother's education for the area aspect) (research questions 6 and 7). Since we did not find a significant difference between the conditions, after controlling for the pretest score (and the WISCAT score), we cannot confirm that a deductive didactic approach might be a better choice for a group of (mathematically) low performing students (Slavin & Lake, 2008) or for non-native students (Civil, 2014). Finally, a mismatch between the teacher's preference for the didactic approach and the actual didactic approach used in the lessons did not have a significant effect on students' measurement numeracy either (research question 8).

Even though an inductive didactic approach induced more classroom interaction time, and more stimulating questions, than a deductive didactic approach (see Chapter 5), we found no significant measurement numeracy improvement difference between inductive and deductive classroom interaction. However, the mean classroom interaction time in our experiment was rather high in both conditions: 48 per 120 seconds in the deductive approach, and 62 per 120 seconds in the inductive approach (see Chapter 5). Therefore, this study could not estimate the effect of very low classroom interaction intensity. We did not find a teacher

effect, but this might be due to the instructions (didactic approach, teacher manual, PowerPoint sheets) they were given (all teachers followed the instructions for teaching quite well; although teacher 4 made some other choices, he complied reasonably with the instructions, see Chapter 5). If the aim were to estimate the teacher effect on measurement numeracy, it would have been necessary for teachers to have more freedom of choice in their classes.

A limitation of this study is that the participants in our sample were all Rotterdam School of Education students. We had no students from other colleges or elementary school pupils in our sample, which might slightly compromise the external validity. Furthermore, lesson attendance was rather low (lessons were not compulsory for students). The reasons for not attending lessons were not recorded (perhaps students were not motivated, perhaps they did not need the lessons). However, lesson attendance was controlled for in the analyses, and the effect was not significant. Finally, we could not randomly assign students to groups, because students were in pre-existing groups, and the design was unbalanced. However, the internal validity of the results of this study is acceptable, as student characteristics were reasonably equal across conditions (see Chapter 2). We conclude that further research is needed on the effect of classroom interaction on numeracy improvement, in order to empirically substantiate claims of positive effects on learning gains.