Teacher assessment of student reading skills as a function of student reading achievement and grade

Stefan Johansson, University of Gothenburg, Department of Education stefan.johansson@ped.gu.se

Monica Rosén University of Gothenburg, Department of Education monica.rosen@ped.gu.se

Abstract

This paper investigates teacher ratings in relation to the student test score in PIRLS 2001. Teacher assessment relates to concerns of equity and equality, and of making correct judgments of the students' skills. It is thus crucial for a teacher to think about these issues. But do teachers assess the same skills equally? Previous research suggests that teacher assessments lack in equality. As an extension of the PIRLS 2001 study, Swedish teachers rated each and every student on 18 different statements about students' literacy skills on a scale ranging from 1-10. It is therefore possible to investigate the relationship between teacher ratings and student achievement in reading literacy. The analysis was conducted in three steps. First a two-level model was fitted to the data and the variance of the teacher ratings were studied by creating a latent model of the teacher ratings. Second students' achievement was used to explain variance between both students and teachers. In a later step effects of students' grade and teacher background were examined. Results show that the variance among the teacher ratings was high and that a substantial part of the variance was due to between class differences. The variance at the within class level was to a large degree accounted for by student achievement. No substantial effect arose when introducing explanatory indicators, such as teachers education or experience to account for the between class differences in teachers ratings. However, grade had a significant effect on the level of ratings between classes. Grade 3 teachers tended to rate their students higher than did the grade 4 teachers at the same performance level.

Keywords: *Teacher assessment, Student reading achievement, Equality, PIRLS 2001, Two-level modeling,*

Introduction

Teachers' ability to assess students' knowledge and skills in equal and fair ways is important for both formative and summative purposes. In Sweden, The National Agency for Education (2007) recently reported a study which indicates that there is variation in assessment of students' skills between different teachers and schools. Several international studies report similar results (Cizek, Fitzgerald, & Rachor, 1996; Gipps, Clarke, & McCallum, 1998). Equal assessment is, however, a difficult phenomenon to investigate and studies are often done in relation to national tests.

In Sweden national tests are used in the final grades of the compulsory school and at the upper secondary school as aids to establish equality of grading among teachers and schools, but they are not intended to serve examination purposes (Wikström, 2005). For grades 2 to 5, the National Agency for Education offers teachers an observation scheme that they could use for diagnostic purposes and to keep track of students' literacy development. In this study, we have used teacher ratings of some of the stated ability aspects in that observation scheme. The teacher rating instrument was included in the PIRLS 2001 test design as a national extension and here the variability of the ratings within and across classrooms is analyzed.

The Swedish school system

In the Swedish school system the students do not get any school marks until grade 8. The grade 3 and 4 students in the current study have thus not gotten any marks yet, and their teachers are not used to grade students for school marks. However, the students are nevertheless to receive feedback about their school performance in relation to the goals of the curriculum. To their support, teachers have access to the national diagnostics materials mentioned above, and the content of this observation scheme may be regarded as a clarification of the goals in the national curriculum for these grades.

The last few decades has implied several pedagogical reforms within Swedish education (Lindblad, Lundahl, Lindgren, & Zackari, 2002). The teacher education has been under continuous criticism and debate, and has been revised several times during the last decades. It is reasonable to expect that teachers' skills in assessment of students reading abilities differs quite a lot from one another, especially when they have gone through different programs of teacher education. Some teachers were educated for a school system where teachers assigned school marks in primary school, and many teachers who teach students in the lower grades have an education from before introduction of the goal oriented school system that was implemented in 1994. Furthermore, teachers at the primary level often teach the same students' from first to third grade in all subject domains, This structure reflects to a large degree the Swedish school system during the 70th-ies and is gradually undergoing changes. From grade 4 the teacher structure is more variable, some teach grades 4-5, others grades 4-6, while it is rare that the teachers only teach one grade. From grade four it becomes successively more common to have specific teachers in certain subjects, for example in athletics, crafts and music.

Teacher competence

The role of teacher competence can be shown for example as the teachers' education and experience. The research field about teacher competence is fairly broad and in the USA there have been a number of studies published concerning how teacher competences, such as education and experience affect student achievement in school. (Croninger, Rice, Rathbun, & Nishio, 2007; Darling-Hammond, 2000a; Darling-Hammond & Berry, 2006; Darling-Hammond, Berry, & Thoreson, 2001)

Some studies have shown that teaching experience influences how well teachers evaluate their students' skills. One previous study of the Swedish PIRLS 2001 data found no significant effects on reading achievement of the variable "teaching experience" as measured by number of years of teaching experience (Myrberg, 2007). However, others have, through meta-analyses, found that teaching experience has a positive influence on students' achievement (Klitgaard & Hall, 1975; Murnane & Phillips, 1981). Researchers, such as Darling-Hammond (2000b) have pointed out that the teachers' with less than three years of experience are less efficient than their colleagues with more experience and that the impact of experience seems to weaken after about five years. Others has found that the teaching experience has a positive effect over the first few years, but that the gain declined after the first five years (Rivkin, Hanushek, & Kain, 2005).

The reports from PIRLS 2001 show that Swedish teachers make relatively little use of tests or other more objective systematic observations for assessment, they rather rely on their own judgment and experience (Martin, Mullis, Gonzalez, & Kennedy, 2003; Rosen, Myrberg, & Gustafsson, 2005).

Equally assessed

Issues of equality are often related to concerns about unfair conditions for students to learn and achieve equally (Coleman & et al., 1966; Darling-Hammond, 2007; Darling-Hammond & College Entrance Examination Board, 1985). Darling-Hammond points out that many schools in exposed areas have less educated teachers and lower achievements scores. In this paper we focus on how teachers assess achievement at an equal level of performance on the PIRLS 2001 reading literacy test, which is also an issue of equality.

The main aim of this paper is to investigate the variability of teachers' assessments of reading achievement within and between classrooms, in relation students' reading achievement in the PIRLS test 2001 and in relation to students' grade (3 or 4). Furthermore, we explore mediating factors such as teacher education and experience which are assumed to affect the rating of students reading literacy skills.

Methodology

The Progress in International Reading Literacy Study (PIRLS) is the International Association for the Evaluation of Educational Achievement (IEA) regularly recurring assessment of reading achievement of grade 4 students. The international design of the PIRLS 2001 study is described in the PIRLS 2001 framework (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001) as well as in the technical report (Martin, Mullis, & Kennedy, 2003). In 2001 there were 35 countries participating in the survey. The data base holds information given by students, their parents, their teachers and their school masters.

The data for the current study come from Sweden's participation in PIRLS 2001. In contrast to the other countries, Sweden participated with two samples, one in grade 3 and one in grade 4 (Rosen et al., 2005). Furthermore, as a national extension in Sweden, teachers were asked to assess different aspects of literacy skills of their students. The scale comprised 18 items, each of which required a rating on a scale from 1-10. The items were based on the previously mentioned diagnostic observation scheme for grades 2-5 provided by the National Agency for Education (National Agency for Education, 2002), and concerned the level of the students' reading and writing abilities, and also their listening and comprehension abilities (Rosen et al., 2005). The diagnostic material identifies a number of skills at different difficulty levels, which the teachers are supposed to consider and describe in qualitative terms for each student. These observations were then to be communicated as feedback to the student or the students' parents, and/or used as a base for adjustments in the teaching. In the national extension of PIRLS these observation aspects were reformulated as statements, and instead of describing the students' ability level in qualitative terms a scale from 1 to 10 was offered for each aspect. The 18 items of different observation aspects are described in Table 1 below, along with the mean and standard deviation for each grade respectively.

[Insert table 1 about here]

The response rates of the questions in Table 1 range around 97-98 percent in both grades which is considered as very satisfying for the purpose of this study. Most of the means presented in the table are well above the midpoint of the scale. The variable Swe11 (Recognizes the letter/connects sound) appears to be the most easy item with a mean of 9.48 for third graders and 9.25 for fourth graders with a quite smallish standard deviation. The variable Swe17 (Can improve own text) is most difficult for the third grade students' according to the teachers, with a mean of 7.11. Of the variables included in the analyses, Swe11 is also most difficult for the fourth graders.

Most of the rating items included appear to coincide quite well with the aspects of reading achievement that the PIRLS reading tests aim to cover. However, there are three items which appear not directly relevant for reading and writing skills (Swe01, Swe02 and Swe04) which were therefore excluded from the current analyses. In order to simplify the analysis the 15 remaining items were randomly divided into three groups of five items and summed into parcel scores. The maximal score a student could have on each parcel score thus is 50. The three sums of the teachers' ratings are then used as indicators of the latent variable *Tea_rate,* which stands for "Teachers ratings of students reading literacy skills". This variable appears at both within and between level and is referred to as *Tea_rateW* for within and *Tea_rateB* for the between level.

Other manifest variables included in the analysis were Grade (3 or 4) and the teacher variables years taught all together (Tea_exp) and adequate teacher education (Tea_edu). Among the manifest variables is also the Tot_read variable included, and this variable contains a mean of the five plausible IRT values on the overall reading score at the PIRLS reading test.

Descriptive statistics for the variables included in the analyses are presented in Table 2.

[Insert table 2 about here]

The response rates for the variables included in Table 2 are overall high. The variables that form the latent variable *Tea_rate* have a response rate ranging around 99 percent for the third graders and around 97 percent for the fourth graders. Concerning Tea_exp and Tea_edu the response rate is around 95 percent for Tea_exp and around 90% for Tea_edu. Generally, the response rate is a couple of percentage points lower in the fourth grade. Variables Tot_read and Grade have a 100 percent respond rate. When examining the descriptive statistics, we can see that teachers are fairly experienced. In third grade the average is almost 18 years and around 15 years in fourth grade. Moreover, the mean of the reading score is 563 points for the fourth graders and 523 for the third graders.

Method of analysis

The primary tool for analyzing the data in order to investigate differences in teachers' ratings of students within and between classrooms is two-level structural equation modeling. With this technique an additive decomposition can be made of the total variance of the teacher ratings into components due to variability between and within classrooms, and independent variables may be brought into the model to account for the variability in the components.

The analysis was conducted in three steps. In the first step a two-level model with one latent variable was fitted which just included the three teacher ratings variables. From this model the amount of variability within and between classrooms was determined. In the next step the student reading achievement score was added as an independent variable and the teacher rating was regressed upon the reading performance variable both at within-class level and at class level. In the third step the effects of the grade variable and the teacher variables at the class level were investigated. All models were fitted with the Mplus 5 program (Muthén & Muthén, 1998-2007).

Case weights and the missing data option in Mplus (Muthén &Muthén, 1998-2007) were used in order to account for stratification, cluster effects and cases with incomplete data. Mplus was used under the modeling environment STREAMS (Gustafsson & Stahl, 2005).

Findings and Discussion

In the following, the results from the three steps of the analyses are presented.

The two-level measurement model was hypothesized to include one student-level latent variable with relations to the three sums of the teacher ratings within classes (Tea_rateW), and one class-level latent variable with relations to what effectively is the class means of three teacher rating variables (Tea_rateB). The model is presented in Figure 1.

[Insert figure 1 about here]

Because there are only three indicators of the latent variables the model is a so called just-identified model which fits the data perfectly. However, factor loadings were high and even at both the student level (.92 - .99) and at the class level (.95 - .97).

Table 3 presents the decomposition of variance in the total sum of the teacher ratings into four components: error variance at student level, systematic variance at student level, error variance at class level, and systematic variance at class level. This decomposition is based on the model parameters in the manner shown by Reuterberg and Gustafsson (1992).

[Insert table 3 about here]

As seen in Table 3, the within class variance is lower in the third grade, about 62% of the total variance, while almost 68% of the total variance is due to within differences in fourth grade. Then examining the between class variance, it is a higher variance in the third grade, about 35 percentages of the total variance occurs at the between level. Concerning the fourth graders 29 percentages of the total variance occurs at the between level. Thus, for both grades it is about 65 percent of the total variance of the teacher ratings that is caused by differences between students within classes, while about 32 percent of the variance is due to differences between classes. It may be noted that this is a much higher amount of between-class variance than is typically observed for reading achievement in Sweden which typically is around 10 percent. This suggests that there is a considerably higher variability between classes in teacher ratings than in reading achievement. However, in this study classes from both grades three and four are included which increases the amount of between-class variance. Nevertheless, the variance between classes is substantial and a second step in the modeling procedure is to introduce students' total reading test score from PIRLS to explain variance at both levels.

Tot_read was in the next step introduced at both the within and the between level as an explanatory variable of the latent teacher rating variable. This model fitted excellently (Chi-square = 131.104, Df = 4, P =0.00, RMSEA =0.053)

The standardized regression coefficient for Tot_readW on *Tea_rateW* was .66 which was highly significant. This estimate implies that 44 % of the variance in the latent teacher rating variable within classes was accounted for by the variability in student reading test scores within classes.

At the between level the standardized regression coefficient for Tot_readB on *Tea_rateB* was .24, which was significant. This estimate implies that 6 % of the variance in the latent teacher rating variable between classes was accounted for by variation in the student reading test scores between classes.

These results indicate that teachers are very good at differentiating between the levels of reading achievement of the students within their classes, whilst they are not good at differentiating the level of performance of their own class in relation to other classes in Sweden. The model indicates that there is variance left in the teacher rating variable at both levels to be explained by other variables after reading achievement has been taken into account.

In the third step of the modeling sequence further explanatory variables were added at the class level, in order to investigate if they could account for some of the variance left in the teacher ratings between classes.

First the variable Grade was added to the model as a dummy variable. This variable had a significant (t=-2.87) negative relation (-.20) to Tea_rateB. This result indicates that at equal levels of performance the grade 4 classes obtained a lower teacher rating than did the grade 3 classes.

Next the Tea_edu variable was added to the between-class part of the model. However, the regression coefficient for Tea_rateB was non-significant (t = 1.10). The regression coefficient for Tea_exp also was non-significant (t = 0.16). Thus, neither teacher education, nor teacher experience as represented by the indicators here affected the level of teacher rating of the classes.

Conclusion and Implications

The purpose of the current study was to investigate differences in teachers' ratings of students reading literacy abilities using a two-level multivariate approach, where both manifest and latent variables were used as independent explanatory variables. The first step in the analysis was to create a latent variable with the sums of teacher ratings and investigate how they varied across classes and within classes. The results revealed that the teacher ratings varied more within a class than between classes. About 65 percent of the variance was due to variations between students within classes, while 32 percent was due to between class differences.

In the next step of our analysis, we entered the students' results of the PIRLS test as an independent variable and explained a part of the variance in the teachers' ratings with this. At the within level, the test score explained about 44 percent of the total variance, but at the between level, only a small amount of the variation could be explained.

Since the variance in the ratings between classes is substantial even when differences in student performance between classes are taken into account, there seem to be difficulties to rate equally among the teachers. For helping teachers to rate students' equally, the National Agency for Education (2002) published a diagnostic material, containing goals that every student shall achieve. This material was also what the rating questionnaire was based on.

The variability between classes can not to a high extent be explained by the variables used in this study. The total reading literacy score in PIRLS accounts for roughly half of the variation in teachers' rating of the students which could be considered as a fairly substantial part. The remaining variance could be an indication that the school subject "Swedish" in school is more complex and involves more knowledge, than the PIRLS test comprises. An additional or alternative explanation is indicated by research that suggests that not only the students performance is assessed when teachers are rating students abilities (Brookhart, 1993; Klapp Lekholm & Cliffordson, 2008; McMillan, Myran, & Workman, 2002).

In this study it was found that teachers rated the fourth graders lower than third graders at an equal level of performance. Reasonable explanations for this can be that teachers in the different grades have different education and they have probably also known their students different periods of time; many grade three teachers have educated the students for almost three years, while the four grade teacher only known the students' for about one semester. It can also be that grade four teachers tend to ask for more when they rate their students. Grade three teachers have followed the students' development during three years, and have, in contrast to the grade four teachers, seen a progression. The students' progression is also something that the National Agency for Education (2002) focuses on in their diagnostic material. As a teacher you have to be concerned about the students' progression and take this into account in the rating of the student. This can be a possible explanation of some of the discrepancy between the different grades. Grade four teachers often teach up to the 6th or 7th grade and this can be an explanation why they tend to judge their students' in a harder manner.

The variables teacher education and teaching experience did not seem to give any substantial effects on the results. One possible explanation for this could be the fact that the variability in teacher data was rather low. Many teachers had a fairly long experience of teaching and this could be a plausible cause to the weak effect, given that previous research suggests that the impact of experience tends to be weaker after a few years (Darling-Hammond, 2000b). Another suggestion could be that the mediating variables are more relevant to a fair rating within classes. As the teacher becomes more experienced, (s)he might get used to differentiate between the students' skills in the class. However, this is nothing that affects the variation between classes, where teachers obviously have a different frame of reference when rating their students. Moreover, this can be a question for future research, where different teacher factors, such as education and experience at the class level are related to the within class variation with varying slopes for the relation between student achievement and the teacher rating.

The results thus indicate that the teachers interpreted the rating scales differently, and one reason for this may be that the different points on the scale did not have any signification, more than that 10 is the highest and 1 the lowest. The teachers had to interpret the scale themselves, and some teachers might not give a 10 to anything but than an extraordinarily gifted student, while some teachers may give a 10 to several students in a class.

This situation is quite similar to the one faced by teachers in ordinary school work. Teachers have to assess several goals in every subject, goals that the National Agency for Education has decided. To get a certain grade you have to fulfill several goals and criteria. Every teacher has to interpret these and then grade their students. The interpretation seems to be a problematic thing, and the goals and criteria seem to be to somewhat unclear and diffuse. Here is an example of some goals in the subject Swedish that a student has to fulfill to receive the "Pass" grade.

> Goals that pupils should have attained by the end of the fifth year in school
>
> Pupils should
> – be able to read with fluency, both aloud and to themselves, and understand events and

> meaning in books and non-fiction written for children and young persons, and be able to discuss their experiences from reading, as well as reflect over texts,
>
> – be able to produce texts for different purposes as a tool for learning and communication,
>
> – be able to orally relate and present something so that the contents are understandable and brought to life,
>
> – be able to apply the most common rules of the written language and the most common rules of spelling, as well as be able to use dictionaries.
>
> (National Agency for Education, 2008)

When reflecting over the goals it is likely that this is a source of variation in the grading of the students. Teachers are different as persons, some tend to be harder in their judgement, and some will be more "kind", therefore a variation in grading occurs.

References

Brookhart, S. M. (1993). Teachers' Grading Practices: Meaning and Values. *Journal of Educational Measurement, 30*(2), 123.

Campbell, J. R., Kelly, D. L., Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2001). *Framework and Specifications for PIRLS Assessment 2001-2nd Edition*. Chestnut Hill, MA: Boston College.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' Assessment Practices: Preparation, Isolation, and the Kitchen Sink. *Educational Assessment, 3*(2), 159.

Coleman, J. S., & et al. (1966). *EQUALITY OF EDUCATIONAL OPPORTUNITY*. Washington D. C.: National Center for Educational Statistics.

Croninger, R. G., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher Qualifications and Early Learning: Effects of Certification, Degree, and Experience on First-Grade Student Achievement. *Economics of Education Review, 26*(3), 312.

Darling-Hammond, L. (2000a). How Teacher Education Matters. *Journal of Teacher Education, 51*(3), 166.

Darling-Hammond, L. (2000b). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives, 8*(1).

Darling-Hammond, L. (2007). Race, Inequality and Educational Accountability: The Irony of "No Child Left Behind". *Race, Ethnicity and Education, 10*(3), 245.

Darling-Hammond, L., & Berry, B. (2006). Highly Qualified Teachers for All. *Educational Leadership, 64*(3), 14.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does Teacher Certification Matter? Evaluating the Evidence. *Educational Evaluation and Policy Analysis, 23*(1), 57.

Darling-Hammond, L., & College Entrance Examination Board, N. Y. N. Y. (1985). *Equality and Excellence: The Educational Status of Black Americans*.

Gipps, C., Clarke, S., & McCallum, B. (1998). *The Role of Teachers in National Assessment in England*.

Gustafsson, J. E., & Stahl, P. A. (2005). *STREAMS User' s Guide, Version 3.0 for Windows 95/98/NT*. Mölndal, Sweden: MultivariateWare.

Klitgaard, R. E., & Hall, G. R. (1975). Are There Unusually Effective Schools? (Vol. 10, pp. 90-106): University of Wisconsin Press.

Klapp Lekholm, A, & Cliffordson, C. (2008). Discrepancies between School Grades and Test Scores at Individual and School Level: Effects of Gender and Family Background. *Educational Research and Evaluation, 14*(2), 181.

Lindblad, S., Lundahl, L., Lindgren, J., & Zackari, G. (2002). Educating for the New Sweden? *Scandinavian Journal of Educational Research, 46*(3), 283.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2003). *PIRLS 2001 Technical Report*. Chestnut Hill, MA: Boston College.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary Teachers' Classroom Assessment and Grading Practices. *Journal of Educational Research, 95*(4), 203.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools*. Chestnut Hill: Boston College.

Murnane, R. J., & Phillips, B. R. (1981). Learning by doing, vintage, and selection: Three pieces of the puzzle relating teaching experience and teaching performance. *Economics of Education Review, 1*(4), 453-465.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

Myrberg, E., & Rosen, M. (2006). Reading Achievement and Social Selection in Independent Schools in Sweden: Results from IEA PIRLS 2001. *Scandinavian Journal of Educational Research, 50*(2), 185.

Myrberg, E. (2007). The Effect of Formal Teacher Education on Reading Achievement of 3rd-Grade Students in Public and Independent Schools in Sweden. *Educational Studies, 33*(2), 145-162.

National Agency for Education. (2002). *Språket lyfter!* [Diagnostic material in Swedish and Swedish as a second language for the years before the 6th year]. Stockholm: Author.

National Agency for Education. (2007). *Provbetyg-Slutbetyg-Likvärdig bedömning.* [Test grades – Final grades -Equal assessment?] Stockholm: Author.

National Agency for Education. (2008). *Syllabuses for compulsory school.* Retrieved July 24, 2008, from: http://www3.skolverket.se/ki03/front.aspx?sprak=EN&ar=0708&infotyp=23&skolform=11&id=3890&extraId=2087

Reuterberg, S.-E., & Gustafsson, J.-E. (1992). Confirmatory Factor Analysis and Reliability: Testing Measurement Model Assumptions. *Educational and Psychological Measurement, 52*(4), 795.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417-458.

Rosen, M., Myrberg, E., & Gustafsson, J.-E. (2005). *Läskompetens i skolår 3 och 4. Nationell rapport från PIRLS 2001 i Sverige*. Göteborg: Göteborgs universitet.

Wikström, C. (2005). *Criterion Referenced Measument for Educational Evaluation and Selection.* Umeå University, Umeå

Table 1. Descriptives for the 18 items of the teacher rating scale

Variable	Question/Statement	Grade 3			Grade 4		
		N	Mean	SD	N	Mean	SD
Swe01	LISTEN CONCENTRATED	5213	8.68	1.822	5857	8,34	2,005
Swe02	UNDERSTANDS INSTRUCTION	5212	8.30	2.073	5854	7,99	2,197
Swe03	IS WILLING TO WRITE	5211	7.58	2.358	5853	7,35	2,415
Swe04	TALKS SPONTANEOUSLY IN GROUPS	5209	7.36	2.462	5856	6,89	2,665
Swe05	CONSTRUCTS SENTENCES CORRECTLY	5208	7.67	2.164	5856	7,47	2,246
Swe06	RECOGNIZES FREQUENTLY USED WORDS	5213	8.35	1.933	5855	8,05	1,994
Swe07	CONNECTS A STORY WITH EXPERIENCE	5162	8.26	1.850	5840	8,01	1,926
Swe08	USES THE CONTEXT	5207	8.05	2.054	5812	7,78	2,146
Swe09	WRITES CONTINUOUSLY	5209	7.84	2.178	5860	7,66	2,216
Swe10	UNDERSTANDS THE MEANING OF A TEXT	5124	8.30	2.002	5767	8,08	2,078
Swe11	RECOGNIZES THE LETTER/CONNECTS SOUND	5136	9.48	1.273	5779	9,25	1,458
Swe12	UNDERSTAND THE MEANING OF THE TEXT IN A STORY	5130	8.98	1.523	5768	8,74	1,638
Swe13	CAN READ UNKNOWN WORDS	5133	8.11	2.034	5778	7,85	2,110
Swe14	CAN MAKE REFLECTIONS OF THE STORY	5083	8.09	1.899	5768	7,88	1,978
Swe15	READS FLUENTLY	5135	8.32	2.099	5777	8,36	2,111
Swe16	LIKES TO READ/TAKES OWN INITIATIVES	5135	7.84	2.380	5768	7,51	2,514
Swe17	CAN IMPROVE OWN TEXT	5072	7.11	2.240	5766	6,96	2,305
Swe18	HAS A REASONABLY LARGE VOCABULARY	5132	8.30	1.890	5774	8,06	1,984

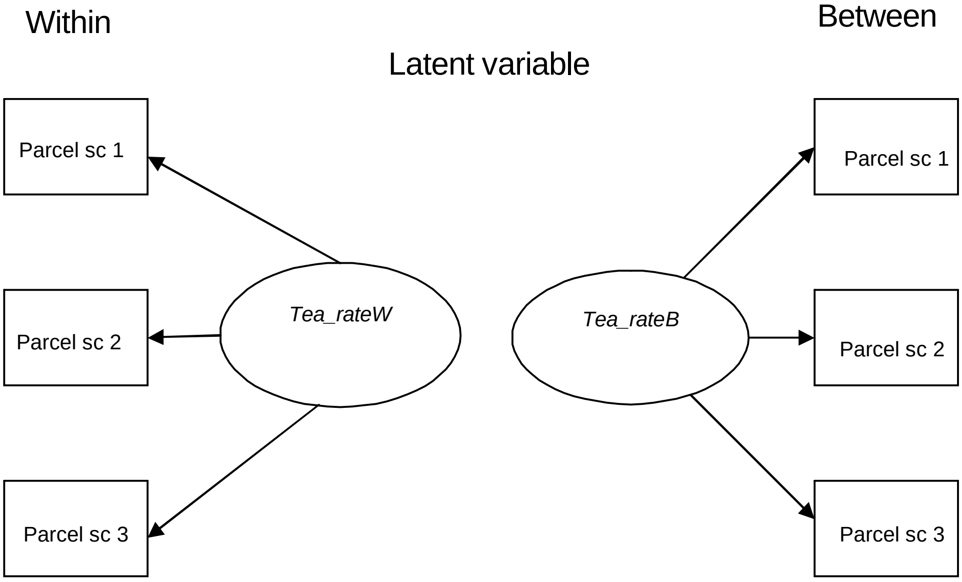
Table 2. Indicators used in the analysis

Variable	Label	Question/Statement	Valid N and respond rate in percentages					
			Grade 3			Grade 4		
Teachers rating of students' reading and writing skills			N	Mean	SD	N	Mean	SD
Parcel score 1	Sum of 5 random items	"Teachers ratings of reading literacy skills" (***Tea_rate***)	5217	39.46	9.947	5866	38.55	10.192
Parcel score 2	Sum of 5 random items		5217	39.82	8.663	5864	38.89	9.136
Parcel score 3	Sum of 5 random items		5217	41.48	8.634	5865	40.28	9.140
Variables from the teacher questionnaire								
Tea_exp	Years taught all together		5036	17.75	12.316	5730	15.29	11.03
Tea_edu	Adequate teacher education		4814	3.02	1.156	5373	1.88	1.042
Variables regarding the student								
Tot_read	Total score on intl scale		5271	523.57	72.610	6044	565.51	60.981
Grade	Student grade		5271	0	0	6044	1	0

Table 3. Variance decomposition of teachers' ratings of students reading literacy skills

	Grade 3		Grade 4	
	Estimate	%	Estimate	%
Total variance	772.52	100	822.87	100
Within class variance	475.98	61.6	556.99	67.7
Within class error	21.65	2.8	20.57	2.5
Between class variance	269.40	34.9	238.72	29.0
Between class error	5.50	0.7	6.59	0.8

Figure 1


Within
Between
Latent variable
Parcel sc 1
Parcel sc 2
Parcel sc 3
Tea_rateW
Tea_rateB
Parcel sc 1
Parcel sc 2
Parcel sc 3