Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

The relationship between students' self-assessed reading skills and other measures of achievement

Stefan Johansson

Correspondence:
Stefan.johansson@gu.se
Department of Education and Special Education, University of Gothenburg, Göteborg, Sweden

Abstract

Background: This study explored the credibility of Swedish third-grade students' self-assessments of their reading achievement by relating those assessments to two different criteria—teachers' judgments and students' reading test scores. Student gender and socioeconomic status (SES) were introduced to determine to what extent, if any, these variables were associated with the accuracy of the self-assessments once students' attitudes toward reading had been controlled for.

Methods: The data, drawn from the Swedish participation in the 2001 iteration of the Progress in International Reading Literacy Study (PIRLS), conducted by the International Association for the Evaluation of Educational Achievement (IEA), included information obtained from students ($N = 5{,}271$) and teachers ($N = 351$). The main method of analysis was two-level structural equation modeling (SEM) with latent variables.

Results: The magnitude of the correlation between student self-assessments and teacher judgments/test scores was similar and amounted to about 0.6. The relationship between teachers' judgments and students' test scores was slightly higher. Neither gender nor SES seemed to be significantly related to the self-assessments, indicating that the students assessed themselves in a fairly equal manner across groups.

Conclusions: The findings demonstrate that, despite their young age, third-graders' self-assessment of their reading literacy skills can be considered as fairly reliable indicators of those skills. In Sweden, the fact that Grade 3 students and their teachers have spent almost three years together in school may contribute to a shared understanding of what literacy knowledge and skills are important.

Keywords: Teacher judgment; Student self-assessment; Reading achievement; PIRLS 2001; Structural equation modeling

Background

A stated aim of the Swedish school curriculum is for students to develop "the ability to assess their own results and relate these and the assessments of others to their own achievements and circumstances" (The National Agency for Education, 2011, p. 19). Another aim is for students to develop increasing responsibility for their learning. These aims align with researchers' suggestions that students' self-assessment of their work is an important contributor to students' learning gains (see, for example, Black & Wiliam, 1998; Boekaerts, 1991; Gielen et al., 2010). This process also requires students

to interpret teachers' evaluations of their work and, from there, decide which strategies they will use in their further learning efforts.

Hattie (2009) found that when students and teachers have a shared understanding not only of when goals are reached but also of which goals to work toward next, learning improves. However, despite students' self-assessments being expressed as a goal in Swedish education and despite these assessments having been documented as beneficial for student learning, teachers still express doubts about both their accuracy and use (Ross, 2006). One of the main concerns focuses on how self-assessments correlate to other measures of achievement, such as teacher judgments and tests.

Black & Wiliam (1998) are two of the researchers who agree that students are likely to take more responsibility for their own learning when they have opportunity to self-assess their learning. If this is indeed the case, providing students with opportunities to assess their own work in an accurate and reliable manner seems an essential part of pedagogical practice. Ongoing feedback from teachers remains of considerable importance in this regard because it helps students make accurate self-assessments (Hattie & Timperley, 2007; Ross et al., 1999).

However, in the likely event of teacher presence becoming less marked as learning becomes more individualized, other factors such as students' home backgrounds tend to become increasingly important with respect to the effectiveness of assessing one's own knowledge and skills (Hansson, 2011). Over the years, students' socioeconomic status (SES) has been increasingly seen as a particularly powerful determinant of achievement differences among students (Myrberg & Rosén, 2006).

The main aim of this study was to investigate the accuracy of Grade 3 Swedish students' self-assessments of their reading skills by looking at those assessments firstly within the context of teachers' judgments of those skills, and secondly in relation to the students' results on the Progress in International Reading Literacy Study (PIRLS) test of reading literacy. In the present study, self-assessments thus relate to two different criterion measures—teacher judgments and standardized test results—and so allow the validity of the use of self-assessments to be addressed from two different angles. I also took into account student gender and SES when conducting my analyses, in order to investigate whether accuracy of self-assessment differed across these groups. Finally, I controlled for students' attitudes to reading.

Students' self-assessments and their alignment with other forms of assessment

Student self-assessment may be a successful route for further learning. Positive effects on achievement have been documented when students are trained in self-assessment (McDonald & Boud, 2003; Ross, 2006; Ross et al., 1999). When Ross et al. (1999) taught Canadian Grades 4 to 6 students the processes involved in self-assessment over periods encompassing 8 to 12 weeks, they found that these students outperformed control groups in narrative writing. McDonald and Boud (2003) also found positive effects of self-assessment on achievement across a range of subjects for older students. However, despite positive findings such as these, several researchers, among them Blatchford (1997) and Fredriksson et al. (2011), have found only modest correlations between student self-assessment and criterion variables. In the primary school years, predictive validity is perhaps not an issue, but it is still essential that self-assessments are a fairly

accurate indicator of skill level if they are to serve as a basis for students' further learning.

Researchers have generally addressed the relationship between students' self-assessments and other modes of assessment from a validity perspective. Within the context of self-assessment, validity typically means correspondence with teacher judgments. Ross (2006) describes this relationship as the "gold standard" for comparisons. Analysis of correspondence between self-assessments and standardized test results are, however, relatively few.

Blatchford (1997) studied the relationship between self-assessments and standardized tests and found that they did not correlate for young students (i.e., seven years of age). Among the reviews that elaborate the relationships between students' self-assessments, teachers' judgments, and test results is one by Shrauger and Osberg (1981), who reviewed 50 such studies. They found that when the task was to predict intellectual achievement or job performance, self-assessments were of similar validity to teacher judgments and test results.

Falchikov and Boud (1989) examined 57 studies that compared self-assessed marks with teacher marks and found substantial correlations between the two. Although the strength of the correlations varied across the studies, the overall conclusion was that self-assessments agree relatively well with teacher judgments. The two researchers also found that the strength of the agreements between students' self-assessments and their teachers' gradings increased as students moved up through the school system.

Sperling, Howard, Miller, and Murphy (2002), however, presented contradictory findings. They examined the self-assessments of American students in Grades 3 to 9 and found that agreement between the assessments and the students' achievement scores on tests were generally low. However, the researchers pointed out that the self-assessments, which reflected how aware students were about their learning, in general, might have measured something other than achievement. They concluded that the self-assessment statements were not domain specific but related to knowledge of cognition generally. Their conclusion accords well with the findings of Butler and Lee (2006), who found higher correlations between students' self-assessments of oral English language, teachers' judgments, and test scores when the self-assessments were more specifically articulated and task related.

Kuncel et al. (2005) meta-analysis of ability to make accurate self-assessments across different student groups was inconclusive as to whether gender and/or other demographic variables influence the validity of self-assessments. In a study from Sweden, Fredriksson et al. (2011) found the association between general self-assessment and scores on a reading test was almost the same for both boys and girls in both Grades 3 and 8. However, there were indications that Grade 3 girls were slightly better than Grade 3 boys at making general self-assessments of their reading skills. Another finding in this study was that a majority of the students overestimated their skills. Kuncel et al. (2005) also discussed socioeconomic status (SES) differences in students' self-assessments after having found that the self-reported grades and actual grades of minority groups of students were lower than those of nonminority students. In the study by Kuncel and his colleagues, the minority groups typically came from the lower SES backgrounds. Although the research community has not widely studied the influence of SES on students' self-assessments, it seems possible that SES may relate to the ability to make accurate such assessments.

Another factor potentially influencing the relationship between self-assessment and academic achievement is students' attitudes toward the school-subject. For example, various studies, such as those by Gustafsson and Rosén (2004) and Swalander and Taube (2007), show strong relationships between students' attitudes and students' achievement in reading, and the relationship is probably reciprocal. A number of studies, including the one by Swalander and Taube (2007), also show that among diverse groupings of students, girls and children of well-educated parents tend to have the most positive attitudes toward reading.

Because of the inconclusive results of previous research, ongoing contributions to the current discussion about the relationship between self-assessment and other measures of achievement seem desirable. The increasing demand for self-assessment in primary education also makes the current study relevant. From my reading, I decided to further explore the relationship between self-assessments, scores on standardized tests, and teacher judgments of ability and skills in order to illustrate the credibility of *young* students' self-assessments. I also wanted to look more closely at how explanatory variables such as SES and gender relate to the relationship between self-assessment and other measures of achievement. My specific research questions were:

1. How does student self-assessment correspond to other measures of the same construct?
2. How do group differences (in my case, gender and SES) relate to young children's ability to assess their own skills?

Method

The battery of tests in the Progress in International Reading Literacy Study (PIRLS), a regularly recurring assessment of reading achievement of Grade 4 students conducted by the International Association for the Evaluation of Educational Achievement (IEA), had a number of design characteristics that served the current study well. I describe these below.

Data sources

The data in the current study emanate from Sweden's participation in the 2001 iteration of PIRLS. The international design of PIRLS 2001 is described in the PIRLS 2001 framework (Campbell et al., 2001), as well as in the study's technical report (Martin et al., 2003).

In 2001, 35 countries participated in the survey. The PIRLS database holds information provided by students, parents, teachers, and school principals. In contrast to the one-only student sample (i.e., Grade 4) drawn in most of the other participating countries, Sweden selected two samples of Swedish students, one for Grade 3 and one for Grade 4 (Rosén et al., 2005). The current study drew on the Grade 3 data, provided by 5,271 students and 351 teachers. Data were also obtained from the students' parents or guardians.

One reason why I decided to use data from Grade 3 is that the students in this grade had been taught by the same class teacher for almost three years while those in Grade 4 had been taught by their respective teachers for only a semester because students change

teacher after the end of Grade 3. I also thought that the benefits of self-assessments might be more likely to accrue when teachers and students have a shared understanding of criteria. Because the Grade 3 students and teachers would have spent a lot of time together, they would have been more likely than their Grade 4 counterparts to have achieved a shared understanding of assessment criteria. For example, over the three-year period, students would have had ample time to interpret teachers' judgments of their (the students) abilities.

Selection of variables

Teacher judgment was an important variable in my study, especially as Sweden included it as a "national extension" in PIRLS 2001[a]. Teacher judgment was used as a criterion against which to compare students' self-assessments. In Sweden at the time of the data collection, no national tests were being administered to students, which meant that teachers were entirely re for assessing students' achievement. However, since 2008, Grade 3 students have been required to take national tests in a number of core subjects. (Note, however, that the tests are used mainly for formative purposes within a classroom.) I also considered the judgment variable important because it ties in with the Swedish curriculum and diagnostic material created to support teachers' assessment practice (The National Agency for Education, 2002).

In the Swedish national extension of PIRLS, Sweden's PIRLS research team reformulated observation aspects from this diagnostic material as statements. Instead of describing students' ability level in qualitative terms, teachers were required to refer to these statements and then use a rating scale, ranging from 1 to 10, to set down their judgment of the achievement of each of their students per statement. The statements focused on reading, writing, and listening abilities (Rosén et al., 2005). The scale was defined by given endpoints and a midpoint, and its appearance was similar to that of a Likert-scale, which is often used in surveys designed to capture attitudes.

During my study, I used 12 of the PIRLS test items relating to students' reading and writing skills. The items I chose aligned with the competencies measured by the PIRLS reading test. Johansson et al. (2012) used the same items to model a teacher-judgment construct and suggested parceling the items because a one-factor model of all 12 items did not fit the data well. Item parceling is a procedure that involves combining single items and then using these combined items as the observed variables. The main reason why researchers use this procedure is to improve measurement properties (Little et al., 2002).

I parceled the 12 items into four summed scores of three items each, and then used the four parcels as indicators of the latent variable *teacher judgment* (i.e., teacher judgments of students' reading literacy skills). Student self-assessment is also a latent variable (*self-assessment*), which I created from four indicators ("Self_assess1" to "Self_assess4"). Here, students estimated their reading abilities on a four point Likert scale ranging from "agree a lot" to "disagree a lot." Table 1 presents, along with descriptive statistics, a more detailed description of the items.

The table shows very high respondent rates for all items. The mean values of the 12 teacher judgment items are also quite high, which indicates that students are, according to their teachers, fairly able readers. The self-assessment variables also have quite high mean values, indicating that students assess themselves as good readers. Self_assess2

Table 1 Descriptive statistics for the 12 teacher judgment items and the four self-assessment items

Variable	Question/statement			
	Pupil can …	***N.***	**Mean**	**SD**
Tch01	Construct sentences correctly	5,208	7.67	2.16
Tch02	Recognize frequently used words in an unknown text	5,213	8.35	1.93
Tch03	Connect a told story with an experience	5,162	8.26	1.85
Tch04	Use the context to understand a written text	5,207	8.05	2.05
Tch05	Write a text continuously fluently	5,209	7.84	2.18
Tch06	Understand the meaning of a text when reading	5,124	8.30	2.00
Tch07	Recognize the letter/connect sound	5,136	9.48	1.27
Tch08	Read unknown words	5,133	8.11	2.03
Tch09	Reflect on a written story	5,083	8.09	1.90
Tch10	Read fluently	5,135	8.32	2.10
Tch11	Improve own written text	5,072	7.11	2.24
Tch12	Use a reasonably large vocabulary	5,132	8.30	1.89
Variable	**Question/statement**	***N.***	**Mean**	**SD**
Self_assess1	Reading is very easy for me	5,138	3.45	0.64
Self_assess2	I do not read as well as other pupils in my class	5,121	3.02	1.01
Self_assess3	I understand almost everything I read, when I read on my own	5,128	3.49	0.69
Self_assess4	To read aloud is very hard for me	5,138	3.06	1.00

and Self_assess4 are reversed items. I therefore recoded these into the same direction as the other items.

All but one of the statements required students to consider their abilities in the absence of reference points. The variant item allowed students to compare their reading skills against the reading skills of the other students in their class. (Note that the label "TotAch" on the figures in this paper is used to denote students' reading achievement on the PIRLS reading test).

I also obtained data on student gender, SES, and attitudes toward reading from the PIRLS questionnaires. Table 2 provides information pertaining to these variables. Several of the attitude items in the table are negatively phrased, but I recoded them. Table 3 provides descriptive statistics—the number of respondents, mean values, and standard deviations—for the student background and attitude variables.

The response rates for the variables included in Table 3 are generally high. Although the variables from the home questionnaires have lower rates of response, frequencies are nevertheless high at about 90%. The number of books in the home and family's financial position are ordinal variables with alternatives from 1 to 5, where 5 is the highest. It appears that most of the PIRLS parents reported having quite a few books in their homes, and estimated themselves to be fairly well-off financially relative to other families.

Most responses regarding annual income and educational level in Table 3 are well above the midpoint of the scale. Students' SES was operationalized with a latent variable using five indicators of SES. In the present study, I measured SES through indicators similar to those suggested by Yang (2003), who studied the SES concept using data from IEA's Reading Literacy Study of 1991. The chosen indicators also seem to agree well with indicators traditionally used when measuring SES (Sirin, 2005).

Table 2 Description of the items included in the analyses

Variable	Information/question/statement	Source
Reading achievement	Pupils' test result on the PIRLS 2001 test.	Pupil
Gender	Pupil gender (Girl = 1, Boy = 0)	Pupil
Number of books at home	About how many books are there in your home? Ordinal variable—1–5: 0–10, 11–25, 26–50, 51–100, more than 100	Parent
Well-off financially	How well off do you think your family is compared to other families? Ordinal variable—1–5: Not at all well-off, Not well-off, Average, Somewhat well-off, Very well-off	Parent
Annual income	Household annual income. Ordinal variable—1–6: Less than $20,000, $20,000–$29,999, $30,000–$39,999, $40,000–$49 999, $50,000–$59,999, $60,000or more	Parent
Highest education	Highest educational level in the home. Ordinal variable—1–8: Some compulsory school, Completed compulsory school, Two years of upper-secondary education, Three years of upper-secondary education, Post-secondary education, Two years of university studies, University studies—candidate level, University studies—Master's level	Parent
Highest occupational level	Highest occupational level in the home. Ordinal variable—1–3: Blue collar, White collar, Academic	Parent
Attitude1	I read only if I have to. Ordinal variable. Four alternatives: Agree a lot (1), Agree a little (2), Disagree a little (3), Disagree a lot (4)	Pupil
Attitude2	I think reading is boring. Ordinal variable. Four alternatives: Agree a lot (1), Agree a little (2), Disagree a little (3), Disagree a lot (4)	Pupil
Attitude3	I would be happy if someone gave me a book as a present. Ordinal variable. Four alternatives: Agree a lot (1), Agree a little (2), Disagree a little (3), Disagree a lot (4)	Pupil
Attitude4	I enjoy reading. Ordinal variable. Four alternatives: Agree a lot (1), Agree a little (2), Disagree a little (3), Disagree a lot (4)	Pupil

The mean values and the standard deviations for students' attitudes toward reading presented in Table 3 indicate that most students estimated themselves as quite strong readers. Nevertheless, the variation in all variables made them useful for my purpose, that is, as indicators of latent variables.

I used the variables indicating teacher judgments and students' self-assessments, SES, and attitudes toward reading in latent variable models. These models appear in the results section of my paper.

Table 3 Means, standard deviations, and number of cases for the explanatory variables

	Mean	SD	*N*
Pupil gender	0.50	0.50	5,271 (boys = 2,640)
Reading achievement	523.60	72.60	5,271
Attitude1	2.77	1.13	5,106
Attitude2	3.51	0.85	5,072
Attitude3	3.33	0.87	5,085
Attitude4	3.49	0.81	5,114
Number books at home	4.00	1.10	4,701
Well-off financially	3.47	0.88	4,649
Annual income	3.92	1.58	4,557
Highest Education	5.71	1.92	4,676
Highest occupational level	2.16	0.78	4,607

Analysis

My main method of analysis was two-level structural equation modeling (SEM), conducted using the Mplus 6 program (Muthén & Muthén, 2007–2012). I used case weights and the missing data option in Mplus in order to account for stratification, cluster effects, and cases with incomplete data. In order to take advantage of all available information, I used the FIML (full information maximum likelihood). The use of SEM has many benefits compared to an approach with manifest variables. A particular strength is that latent variables are said to be free from measurement error, because the unique part of the variance is separated from the unexplained part (Gustafsson, 2009).

In samples of educational data, students are often nested in classrooms, and classrooms are nested in schools, and so forth. In a hierarchical structure of this kind, individual observations are not independent. Because of different selection mechanisms such as SES, students in the same classroom tend to be more similar than dissimilar to one another. Students also share a common history in terms of going to the same school and being taught in the same classrooms by the same teachers. However, statistical tests are normally premised on the assumption of independence. If this assumption is violated, the estimates of the standard errors will be too small, a situation that can result in many spurious findings (Hox, 2002). Multilevel modeling is a useful means of accounting for this problem because it allows for dealing with dependencies at different levels.

Given the many possible goodness-of-fit indices that exist, the usual advice is to assess model fit by inspecting several fit indices that derive from different principles (Hox, 2002). In this study, the χ^2 goodness-of-fit test was used. However, as the χ^2 is sensitive to sample size, and with large samples will almost certainly be significant, it was combined with three other fit indices.

RMSEA (root mean square error of approximation) takes both the number of observations and free parameters into account and is considered as a robust measure to use when assessing model fit (Jöreskog, 1993). An acceptable value on the RMSEA is generally below 0.08, while a close fit is about 0.05 or below (Loehlin, 2004).

The CFI (comparative fit index) is a fit index that depends on the average size of the correlations in the data. It should be as close to 1.0 as possible; values below 0.95 are not considered acceptable (Bentler, 1990).

SRMR (the standardized root mean square residual), which is a measure of residuals compared separately for within and between levels, was also used. SRMR should be below 0.08 for the model to be accepted (Hu & Bentler, 1999).

In the current study, self-assessments were set as a dependent variable. Students' reading achievement scores on the PIRLS test and teachers' judgments were used as criteria in separate analyses. The first step was to formulate measurement models and, thereafter, study the relationship. The later steps involved introducing explanatory variables in order to examine whether differences in self-assessments were due to group differences.

Results

I begin this presentation of the results of my analyses by fitting measurement models of students' self-assessment and teachers' judgments. I then compare the self-assessments to the test scores of the Swedish Grade 3 students who participated in

PIRLS 2001 and to their teachers' judgments of their reading skills. These analyses allowed me to address questions about the accuracy of the assessment measures, particularly how well the self-assessments correlated with the other measures of achievement.

Modeling of the students' self-assessments

The latent variable *self-assessment* was based on four statements (Self_assess1 to Self_assess4) concerning students' perceptions of their own reading abilities, as previously described in Table 1.

The intra-class correlations for the four self-assessment variables showed modest estimates of 0.02 to 0.03, indicating that only small between-classroom effects could be found in the data. Two-level modeling may therefore not be warranted (Muthén, 1994). In order to account for cluster effects, which were present for the other variables, I used two-level SEM, but fitted the models only at the within level and allowed the between variables to co-vary. When Yang-Hansen et al. (2006) used this self-assessment measurement model, they obtained an excellent fit to the data.

The model fit in the present study was $\chi^2 = 1.88$, P = 0.17, df =1, RMSEA = 0.01, CFI = 1.00, SRMR = 0.004. All of the factor loadings were significant ($p < 0.001$) and moderate to high. The item "I do not read as well as other students in my class" had the highest factor loading, indicating that students relate their self-assessments to other students' knowledge and skills in the own classroom.

During the next step, I formulated the latent variable teacher judgments in a similar measurement model. I then built on the four manifest parcel variables (to which I had previously assigned the 12 judgment items) by fitting one latent judgment variable. The factor loadings turned out to be high and even, and an acceptable model fit was obtained (model fit: $\chi^2 = 103.08$, P = 0.00, $df = 4$, CFI = 0.99, RMSEA = 0.07, SRMRw = 0.01, SRMRb = 0.02).

The credibility of students' self-assessments

To address the question of the validity of students' self-assessments, I related the self-assessment variable to teachers' judgments and students' test results in PIRLS. The relationship between *self-assessment* and *TotAch* amounted to 0.58, which was similar to the relationship between *teacher judgments* and *self-assessments* (0.59). The model, presented in Figure 1, showed that teacher judgments also related to TotAch. Note that this correlation, at 0.65, was slightly higher than the relationship between the teachers' judgments and their students' self-assessments. All estimates in the model depicted in Figure 1 were significant at $p < 0.001$ (model fit: $\chi^2 = 280.20$, P = 0.00, $df = 24$, RMSEA = 0.05, CFI = 0.99, SRMR = 0.02).

Although slightly weaker, the correlations between self-assessment and the other measures of achievement nevertheless provide support for the contention that students can estimate their reading achievement, despite their young age. The nature of the questions measuring students' estimations of their reading skills seemed suitable for nine-year-olds. It appears that most of the students were aware not only of whether they were good or weak readers, but also of the standing of their performance in relation to the performance of other students in their class.

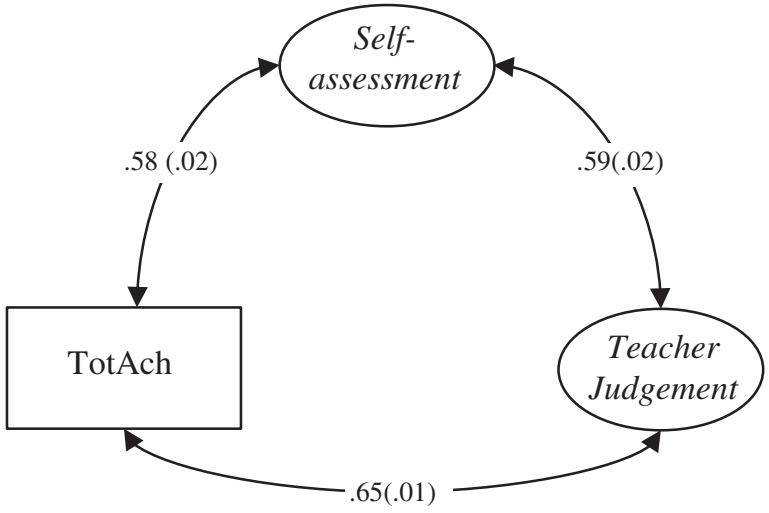

Figure 1 Relationship between three measures of achievement. Note: Standard errors are within parentheses.

The next question addressed whether there were any differences in the ways that boys and girls and students from different SES backgrounds estimated their reading skills. This step focused on whether some groups tended to overestimate their skills compared to the criterion (i.e., teachers' judgments or test scores). I carried out these analyses in two different models. In the model displayed in Figure 2, *TotAch* is the criterion; in Figure 3, *teacher judgments* is the criterion.

The results with regard to TotAch as the criterion showed no significant effect on the self-assessments for either gender or SES, indicating that boys and girls and students from different SES backgrounds gave fairly similar estimates of their skills. Figure 2 presents the model along with standardized regression coefficients. All estimates in the model were significant at $p < 0.001$, unless otherwise stated. The model fit was $\chi^2 = 212.76$, P = 0.00, $df = 39$, RMSEA = 0.03, CFI = 0.98, SRMR = 0.02.

A similar pattern emerged for the analysis in which *teacher judgments* was the criterion. However, the relationship between self-assessments and SES was slightly positive, but not significant. All estimates in this model (presented in Figure 3) were significant at $p < 0.001$, unless otherwise stated. The model fit was $\chi^2 = 791.58$, P = 0.00, $df = 71$, RMSEA = 0.04, CFI = 0.98, SRMR = 0.03.

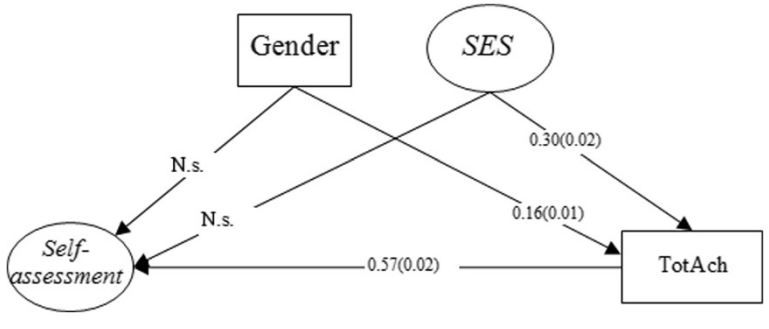

Figure 2 Gender and SES differences in self-assessments, with students' reading test scores (TotAch) as the criterion. Note: Standard errors are within parentheses.

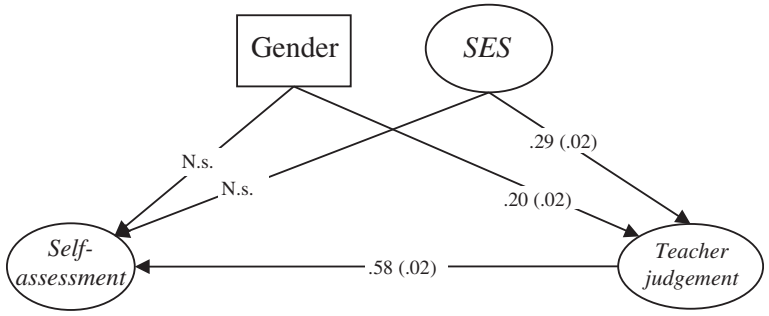

Figure 3 Gender and SES differences in self-assessments, with teacher judgments as the criterion.
Note: Standard errors are within parentheses.

When conducting these analyses, I controlled for gender-related and SES-related achievement differences. In the absence of controlling for achievement differences, girls and high-SES students were more likely than boys and low-SES students to rate their reading skills highly. Moreover, the relationships between the criteria teacher judgments/TotAch and gender and SES were highly positive, indicating higher achievement levels for girls and for students from higher SES backgrounds.

In order to further determine the presence of differences on the basis of gender and SES, I introduced *students' attitudes toward reading* into the model. Strong associations have been found between attitudes and gender differences in reading performance (Gustafsson & Rosén, 2004). Figure 4 presents the measurement model of students' attitudes toward reading. All factor loadings in the model were significant at $p < 0.001$. The model fit was $\chi^2 = 42.08$, P = 0.00, df = 2, RMSEA = 0.06, CFI = 0.97, SRMR = 0.02.

The depiction of the latent two-level model of students' attitudes in the Figure 4 model shows that the items *I enjoy reading* and *I think reading is boring* were those that obtained the highest factor loadings, an outcome that seems to have good agreement with the attitude construct. Because the relationship between (1) *attitude* and students' self-assessments and (2) *attitude* and the reading achievement variables may have been reciprocal, I considered correlating the residuals of these variables to be an adequate approach. Moreover, in order to control for plausible differences in attitude between boys and girls and students with different SES, I related gender and SES to *attitude* as independent variables.

For my final model, I ran two analyses. The first had *TotAch* as the criterion (Figure 5), and the second had *teacher judgments* (Figure 6). All estimates depicted in Figure 5 were significant at $p < 0.001$, unless otherwise stated. The model fit was $\chi^2 = 1174.01$ P = 0.00 df = 81 RMSEA = 0.05 CFI = 0.93, SRMR = 0.04. All estimates in the Figure 6 model were also significant at $p < 0.001$, unless otherwise stated. The model fit was $\chi^2 = 1668.28$, P = 0.00, $df = 125$, RMSEA = 0.04, CFI = 0.96, SRMR = 0.04.

As I hypothesized would be the case, strong correlations emerged between attitudes and self-assessments and between attitudes and test results. When I introduced *attitude* into the model, the effect of SES and gender on self-assessments remained nonsignificant ($p < 0.01$). When I used *teacher judgments* rather than TotAch as the criterion, the estimates were similar. I did find a slight negative effect for gender on

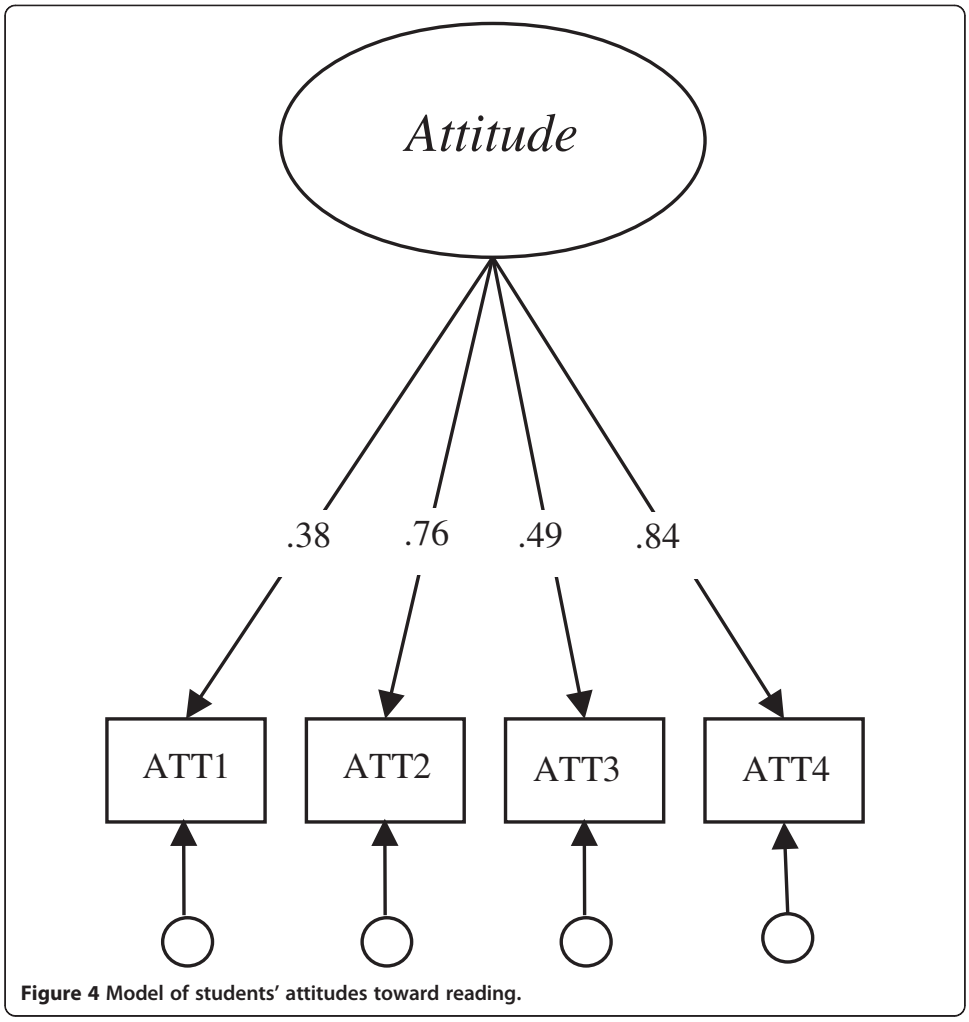

Figure 4 Model of students' attitudes toward reading.

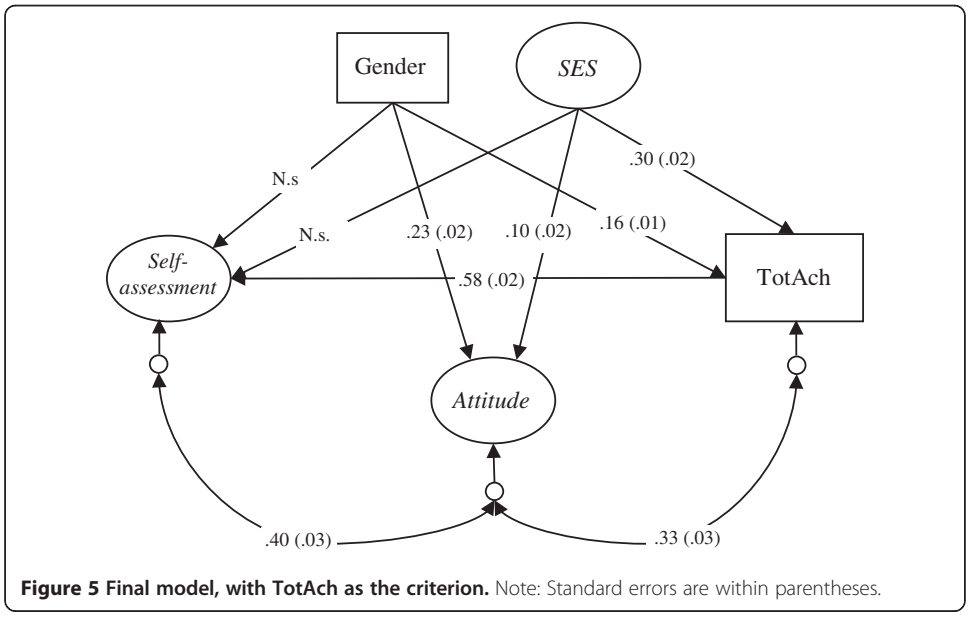

Figure 5 Final model, with TotAch as the criterion. Note: Standard errors are within parentheses.

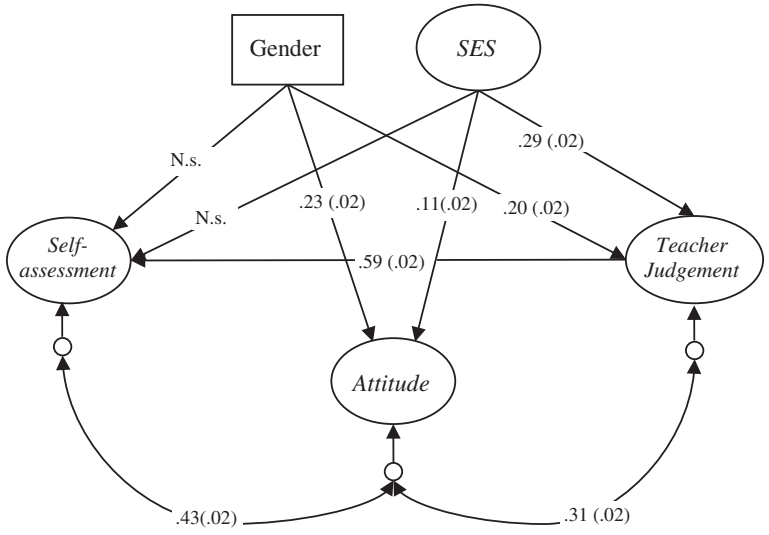

Figure 6 Final model with teacher judgments as the criterion. Note: Standard errors are within parentheses.

students' self-assessments—$p < 0.10$ (t =1.87)—a finding which indicates that when achievement and attitudes toward reading were controlled for, the boys seemed slightly more likely than the girls to give higher estimates of their skills. The relationship between gender and teachers' judgments showed a slightly higher estimate (0.20) than the estimate for TotAch and gender (0.16), as shown in Figure 5.

The performance differences in reading between girls and boys became more pronounced when *teacher judgments* was the performance measure, as depicted by the high correlations between attitudes toward reading and reading performance in both Figures 5 and 6. The strength of the relationship between test scores and attitudes and between attitudes and teacher judgments was similar.

Discussion

The overall purpose of this study was to investigate the trustworthiness of Swedish third-grade students' self-assessments of their reading literacy skills. The comparison criteria were students' reading test scores in PIRLS 2001 and teachers' judgments of the students' skills. Overall, I found moderate correlations between students' self-assessments and their test scores on the PIRLS reading test, and between these self-assessments and the teachers' judgments of their students' general reading literacy abilities.

In previous research, answers to the question of whether self-assessments can be considered valid measures of student performance have been mixed. However in meta-analyses, the predictive validity of self-assessments is relatively high. Note, however, that most of these studies were conducted with samples of older students (see, for example, Falchikov & Boud, 1989; Shrauger & Osberg, 1981).

Although, in the present study, I initially found no differences between girls' and boys' self-assessments, a slight tendency for boys to overestimate their reading skills became apparent when I controlled for attitudes toward reading. In general, girls had a

clearly more positive attitude than boys toward reading. A previous study from Sweden, also conducted with a sample of primary school children, found the association between general reading performance and self-assessment to be fairly equal for boys and girls (Fredriksson et al., 2011).

I could find no evidence in my study of SES influencing the students' self-assessments relative to their overall test scores in PIRLS. However, this question needs to be followed-up, particularly as the achievement differences with respect to SES increased in Sweden (Gustafsson & Yang-Hansen, 2009; Myrberg & Rosén, 2006). Students' achievement level may be important with respect to the ability to self-assess: previous research conducted with students in Grades 6 through high school (e.g., Falchikov & Boud, 1989; Reuterberg & Svensson, 2000) suggests that high-achieving students tend to be more realistic and thus perhaps to underestimate their performances, while low-achieving students tend to do the opposite. This suggestion seems important to bear in mind given the clear differences in achievement between SES-groups found in previous research (Johansson et al., 2012; Sirin, 2005).

Because there was no variation in PIRLS across classrooms with respect to self-assessments, I did not have opportunity to determine the existence of possible differences in self-assessments across classrooms with children from different SES backgrounds. However, use of a sharper self-assessment instrument might have led to other conclusions about differences in the ability to self-assess skills. These considerations provide additional interesting questions for future research.

The correlation coefficient that emerged from my analyses for the relationship between teachers' judgments and students' test results in PIRLS was 0.65. Researchers who have found correlations of similar magnitude concluded that the teacher judgments under consideration were trustworthy measures of students' achievement (see, for example, Feinberg & Shapiro, 2009; Hoge & Coladarci, 1989).

The significant correlations that I found between the attitude variables and self-assessment align with the findings of research conducted by Swalander and Taube (2007). They found an effect of the same magnitude for attitude on eighth-grade students' academic self-concept. It may be that the attitude items share common characteristics with those of self-assessment, or it may simply be that students like learning content they are good at.

The results of the present study also showed a similar relationship between students' test scores on PIRLS and their attitudes, and between their attitudes and their teachers' judgments. Despite previous research indicating the need to undertake compensatory grading for low-achieving students with high attitudes (e.g., Klapp-Lekholm & Cliffordson, 2009), the current study did not provide such evidence. The reason for this pattern of findings may be that student attitudes are confounded in the gender and SES variables.

As indicated by the low between-classroom effects, the average value of students' self-assessments was similar in most of the classes that featured in my study. However, the classrooms' average achievement varied, as indicated by the PIRLS test results and the teachers' judgments. One could assume that high-achieving classes tend, on average, to give higher self-estimates of their reading skills than do low-achieving classes. But assessing one's own skills is a complex process, and overestimation and underestimation may be plausible occurrences across different groups of students. Researchers (e.g., Falchikov and Boud, 1989; Reuterberg & Svensson, 2000) who have conducted

research with samples of students older than the Grade 3 PIRLS participants arrived at just this conclusion.

Another explanation for the similar levels of self-assessment across classrooms may be that the PIRLS questionnaire items concerned comparisons of the students' own reading skills with those of other students in the same class. As such, the students would not have made absolute estimations of their skill or, to express this point another way, they would not have considered whether their reading skills were better or worse than those of the whole student population. It is also possible that more scale points and/or more items would have rendered larger variation in the average self-assessments, and thereby large between-classroom effects.

Conclusion

My findings demonstrate that, despite their young age, third-graders' self-assessment of their reading literacy skills can be considered as fairly reliable indicators of those skills. In IEA's 1991 Reading Literacy Study, the correlation between students' self-assessments and their reading scores was 0.25 to 0.55 for most of the countries that participated in the study (Elley, 1992). The correlation in the current study was thus somewhat higher, possibly because Swedish teachers and students work closely together across the years, possibly because teachers continuously provide students with information about their achievement, and possibly because teachers give students ample opportunity to identify their achievement levels. In Sweden, the fact that Grade 3 students and their teachers have spent almost three years together in school may also contribute to a shared understanding of what literacy knowledge and skills are important.

Also, teachers in Sweden have sole responsibility for interpreting and applying the assessment criteria, whereas assessment in other countries tends to be carried out by both teachers and external examiners. Given the different circumstances for student self-assessments across countries, it would be interesting to examine the relationship between these assessments and teacher judgments in a comparative study. In IEA's goldmines of educational data, there may be other adequate items and countries that researchers and other interested parties can use to shed more light on these and other salient issues regarding students' self-assessment.

Endnote

[a]Countries participating in large-scale assessment studies sometimes elect to introduce an element of particular interest to their country into their data gathering and analysis. These elements are typically called national extensions.

Competing interests
The author declares that he has no competing interests.

Received: 28 August 2013 Accepted: 3 September 2013
Published: 16 September 2013

References

Bentler, PM. (1990). Comparative fit indexes in structural models. *Psychological Bullentin, 107*(2), 238–246.

Black, P, & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–74.

Blatchford, P. (1997). Pupils' self-assessments of academic attainment at 7, 11 and 16 years: Effects of sex and ethnic group. *British Journal of Educational Psychology, 67*(2), 169–184. doi:10.1111/j.2044-8279.1997.tb01235.x.

Boekaerts, M. (1991). Subjective competence, appraisals and self-assessment. *Learning and Instruction, 1*(1), 1–17.

Butler, YG, & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *Modern Language Journal, 90*(4), 506–518.

Campbell, JR., Kelly, DL., Mullis, IVS, Martin, MO, & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001* (2nd ed.). Chestnut Hill, MA: Boston College.

Elley, WB (1992). *How in the world do students read?* The Hague, the Netherlands: International Association for the Evaluation of Educational Achievement (IEA).

Falchikov, N, & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430.

Feinberg, AB., & Shapiro, ES. (2009). Teacher accuracy: An examination of teacher-based judgments of students' reading with differing achievement levels. *Journal of Educational Research, 102*, 453–462.

Fredriksson, U, Villalba, E, & Taube, K. (2011). Do students correctly estimate their reading ability? A study of Stockholm students in Grades 3 and 8. *Reading Psychology, 32*, 301–321.

Gielen, S, Peeters, E, Dochy, F, Onghena, P, & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

Gustafsson, J-E (2009). Strukturell Ekvationsmodellering [Structural equation modeling]. In G. Djurfeldt & M. Barmark (Eds.), *Statistisk verktygslåda 2: Multivariat analys [Statistical toolbox 2: Multivariate analysis]* (pp. 269–321). Lund, Sweden: Studentlitteratur.

Gustafsson, J-E, & Rosén, M. (2004). *Förändingar i läskompetens 1991–2001: En jämförelse över tid och länder* (Changes in reading literacy 1991–2001: A comparison across time and countries). Gothenburg, Sweden: University of Gothenburg.

Gustafsson, J-E, & Yang-Hansen, K. (2009). Resultatförändringar i svensk grundskola. [Changes in achievement in Swedish compulsory school]. In L. M. Olsson (Ed.), *Vad påverkar resultaten i grundskolan?* (What influences achievement in compulsory school? pp. 40–84). Stockholm, Sweden: Skolverket.

Hansson, Å (2011). *Ansvar för matematiklärande: Effekter av undervisningsansvar i det flerspråkiga klassrummet [Responsibility for learning in mathematics: Effects of teaching responsibility in the multilinguial classroom]* (Unpublished doctoral thesis). Sweden: University of Gothenburg. Available online at http://hdl.handle.net/2077/26669.

Hattie, J (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge.

Hattie, J, & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hoge, RD, & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297–313.

Hox, J (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hu, L, & Bentler, PM. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Johansson, S, Myrberg, E, & Rosén, M. (2012). Teachers and tests: assessing pupils' reading achievement in primary schools. *Educational Research and Evaluation, 18*(8), 693–711. doi:10.1080/13803611.2012.718491.

Jöreskog, KG (1993). Testing structural equation models. In K. A. Bollen & J. Scott Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Klapp-Lekholm, A, & Cliffordson, C. (2009). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation, 15*(1), 1–23.

Kuncel, NR, Credé, M, & Thomas, LL. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis. *Review of Educational Research, 75*(1), 63–82.

Little, TD, Cunningham, WA, Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel? Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151–173.

Loehlin, JC. (2004). *Latent variable models* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Martin, MO, Mullis, IVS, & Kennedy, AM. (2003). *PIRLS 2001 technical report*. Chestnut Hill, MA: Boston College.

McDonald, B, & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education, 10*(2), 209–220.

Muthén, BO. (1994). Multilevel covariance structure-analysis. *Sociological Methods & Research, 22*(3), 376–398. doi:10.1177/0049124194022003006.

Muthén, LK., & Muthén, BO (2007–2012). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Myrberg, E, & Rosén, M. (2006). Reading achievement and social selection in independent schools in Sweden: Results from IEA PIRLS 2001. *Scandinavian Journal of Educational Research, 50*(2), 185–205.

Reuterberg, S-E, & Svensson, A (2000). *Köns- och socialgruppsskillnader i matematik - orsaker och konsekvenser* (Gender differences and SES differences in mathematics: Causes and consequenses). Mölndal, Sweden: University of Gothenburg.

Rosén, M, Myrberg, E, & Gustafsson, J-E (2005). *Läskompetens i skolår 3 och 4. Nationell rapport från PIRLS 2001 i Sverige [Reading literacy in Grades 3 and 4: National report from PIRLS 2001 in Sweden]*. Gothenburg, Sweden: University of Gothenburg.

Ross, JA (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research & Evaluation, 11*(10), 1–13. Available online at http://pareonline.net/getvn.asp?v=11&n=10.

Ross, JA, Rolheiser, C, & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6*(1), 107–132.

Shrauger, JS, & Osberg, TM (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin, 90*, 322–351.

Sirin, SR. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453.

Sperling, R. A., Howard, B. C., Miller, L. A., & Murphy, C. (2002). Measures of children's knowledge and regulation of cognition. *Contemporary Educational Psychology, 27*(1), 51–79.

Swalander, L, & Taube, K. (2007). Influences of family based prerequisites, reading attitude, and self-regulation on reading ability. *Contemporary Educational Psychology, 32*(2), 206–230.

The National Agency for Education. (2002). *Språket lyfter! Diagnosmaterial i svenska och svenska som andra språk för åren före skolår 6* (Progression in language! Diagnostic material for Swedish and Swedish as a second language for Grades 2–5). Stockholm, Sweden: Skolverket.

The National Agency for Education. (2011). *Curriculum for the compulsory school, preschool class and the leisure-time centre 2011*. Stockholm, Sweden: Skolverket.

Yang, Y (2003). *Measuring socioeconomic status and its effects at individual and collective levels: A cross-country comparision*. Gothenburg, Sweden: University of Gothenburg.

Yang-Hansen, K, Rosén, M, & Gustafsson, J-E. (2006). Measures of self-reported reading resources, attitudes and activities based on latent variable modeling. *International Journal of Research & Method in Education, 29*(2), 221–237.

doi:10.1186/2196-0739-1-3
Cite this article as: Johansson: **The relationship between students' self-assessed reading skills and other measures of achievement.** *Large-scale Assessments in Education* 2013 **1**:3.