On the Validity of Reading Assessments

GOTHENBURG STUDIES IN EDUCATIONAL SCIENCES 328

On the Validity of Reading Assessments

Relationships Between Teacher Judgements, External Tests and Pupil Self-assessments

Stefan Johansson

UNIVERSITY OF GOTHENBURG
ACTA UNIVERSITATIS GOTHOBURGENSIS

ISBN 978-91-7346-736-0
ISSN 0436-1121
ISSN 1653-0101

Thesis in Education at the Department of Education and Special Education

The thesis is also available in full text on
http://hdl.handle.net/2077/32012

Photographer cover: Rebecka Karlsson

Distribution: ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222
SE-405 30 Göteborg, Sweden

Print: Ale Tryckteam, Bohus 2013

Abstract

Title: On the Validity of Reading Assessments: Relationships Between Teacher Judgements, External Tests and Pupil Self-assessments
Language: English with a Swedish summary
Keywords: Validity; Validation; Assessment; Teacher judgements; External tests; PIRLS 2001; Self-assessment; Multilevel models; Structural Equation Modeling; Socioeconomic status; Gender
ISBN: 978-91-7346-736-0

The purpose of this thesis is to examine validity issues in different forms of assessments; teacher judgements, external tests, and pupil self-assessment in Swedish primary schools. The data used were selected from a large-scale study—PIRLS 2001—in which more than 11000 pupils and some 700 teachers from grades 3 and 4 participated. The primary method used in the secondary analyses to investigate validity issues of the assessment forms is multilevel Structural Equation Modeling (SEM) with latent variables. An argument-based approach to validity was adopted, where possible weaknesses in assessment forms were addressed.

A fairly high degree of correspondence between teacher judgements and test results was found within classrooms with a correlation of .65 being obtained for 3rd graders, a finding well in line with documented results in previous research. Grade 3 teachers' judgements correlated higher than those of grade 4 teachers. The longer period of time spent with the pupils, as well as their different education, were suggested as plausible explanations. Gender and socioeconomic status (SES) of the pupils showed a significant effect on the teacher judgements, in that girls and pupils with higher SES received higher judgements from teachers than test results accounted for.

Teachers with higher levels of formal competence were shown to have pupils with higher achievement levels. Pupil achievement was measured with both teacher judgements and PIRLS test results. Furthermore, higher correspondence between judgements and test-results was demonstrated for teachers with higher levels of competence.

Comparisons of classroom achievement were shown to be problematic with the use of teachers' judgements. The judgements reflected different achievement levels, despite the fact that test-results indicated similar performance levels across classrooms.

Pupil self-assessments correlated slightly lower to both teacher judgement and to test results, than did teacher judgements and test results. However, in spite of their young age, pupils assessed their knowledge and skills in the reading domain relatively well. No differences in self-assessments were found for pupils of different gender or SES.

In summary, a conclusion of the studies on the three forms of assessment was that all have certain limitations. Strengths and weaknesses of the different assessment forms were discussed.

Table of contents

Acknowledgements

I am very grateful to many people, who at various stages commented on my manuscripts and thereby improved this thesis.

First, my sincere thanks to my supervisors. Monica Rosén has been my main supervisor throughout my PhD studies. Thank you for all good advice, for being very loyal, patient and understanding during the long process of becoming a researcher. Eva Myrberg has been my co-supervisor and I am endlessly grateful for the support you have given to me, and for sharing your profound knowledge about the complex educational science. It is no exaggeration to say that without my supervisors' commitment, this piece of research would not have been what it is today. Thank you.

I would also like to express my deepest gratitude to Jan-Eric Gustafsson for extremely valuable advice at many stages of my studies. Kajsa Yang-Hansen has been a great support throughout my studies, kindly guided me through an array of methodological issues. As a member of the FUR group, I am indebted to all people there, because they all generously offered their help and shared their knowledge to me.

Further, I would like to thank the discussants at my planning, mid-stage and final seminars, Gudrun Erickson, Lisbeth Åberg-Bengtsson and Viveca Lindberg. Thanks also to Professor John Hattie, Professor Dylan Wiliam and Professor Patricia Murphy who gave me many valid comments on my manuscripts that I presented at the conferences of the National research school for graduates in educational assessment. Special thanks to the "assessment people" at Stockholm University whom have arranged annual conferences on educational assessment within the research school. My friends and colleagues Rolf Strietholt, Robert Sjöberg, Nicolai Bodemer, and Cecilia Thorsen have provided invaluable support and have generously shared thoughts and ideas on various issues. Alastair Henry has been a great help with the English language.

Finally, I am grateful to my friends and family. My love Rebecka has always been by my side, supporting me and reminding me about the most important things in life.

Göteborg, January, 2013

Chapter One: Introduction and points of departure

My doctoral research started with an interest in issues of equality in assessment, with the overarching question of how assessment equality can be achieved in school. In the data material of the Progress in International Reading Literacy Study 2001 (PIRLS), I found a feasible way to study questions of validity in educational assessments. This thesis investigates how different forms of assessment function in the context of the Swedish primary school. Relationships between three different assessment forms have been explored; teacher judgements, external test results and pupil self-assessments. Although there are numerous ways of assessing pupil knowledge and skills these forms of assessments are prominent aspects of teaching, crucial for the assessment of learning as well as for promoting learning. In Sweden, teachers' assessments are of vital importance since no external tests for high-stake examinations or grade retention purposes exist. Moreover, teachers have been considered as the single most powerful determinant for pupil learning (Hattie, 2009). Because of the vital role played by teachers in assessment, in the current thesis particular interest is directed to teacher assessment.

To understand the context of the present thesis it is worth rewinding to the educational context at the time of the data collection in 2001. At this point in time, the curriculum introduced in 1994[1] was fully implemented and the deregulation and decentralization of the school system had taken effect. In addition, a new generation of teachers had entered schools, graduates of a revised teacher-training program launched at the end on the 1980s. Furthermore, from being a school system regulated by sharp and distinctive criteria, since 1994 teachers have had to adapt to new assessment criteria, and a new grading system[2]. In the former system the formulations of the attainment goals were detailed, while in the Lpo 94, looser frames implied greater responsibility on the part of the teacher to interpret goals and assess pupil knowledge and skills (Tholin, 2006). It did not take long before serious validity concerns were raised

[1] Curriculum for the compulsory school, preschool class and the leisure-time centre, (Lpo 94)

[2] The criterion referenced grading system. This system did not focus selection as the former norm-referenced system. The new criterion referenced system was constructed with the purpose of giving information about pupil achievement measured against centrally formulated goals and locally defined criteria (Klapp-Lekholm, 2008).

regarding teachers' assessments. At least two circumstances contributed to an intensified discussion.

First, the interpretation of the goals and criteria was problematic from the perspective of equality. Tholin (2006) demonstrated that, when no grading criteria were explicit, the goals and criteria for grade eight varied considerably between schools. Grading criteria for the ninth grade had to be reformulated for use in grade eight, as the students there were also awarded grades. Selghed (2004) showed that teachers had not fully adapted to the new criterion-referenced grading system, but remained in former the norm-referenced strategies of grading. Different interpretations of criteria were probably also present in the school grades prior to grade eight. Issues of equality in grading have also been highlighted by the national authorities (see for example, The Swedish National Agency for Education, 2007, 2009; Swedish School Inspectorate, 2010, 2011). The Swedish National Agency for Education (2007, 2009) has concluded that teacher assessments differ from one teacher to another, even though test-results indicate that pupils have similar performance levels. When summative assessments differ between teachers, it is likely that teachers' formative feedback will be different too, since in practice these concepts often work together (Newton, 2007; Taras, 2005).

Second, parallel to the concerns about equality in teacher assessments, international comparative studies have been indicating an achievement trend in Sweden which is declining in both the science and the reading domains (Gustafsson, 2008; Gustafsson & Rosén, 2005; Gustafsson & Yang-Hansen, 2009; Rosén, 2012). While Sweden's overall achievement declined, the criterion-referenced assessments made by teachers did not however indicate an achievement drop. Indeed pupils were being awarded higher and higher grades; grade inflation was thereby present in most subjects in the Swedish schools (Gustafsson & Yang-Hansen, 2009).

The results of research on the criterion-referenced system and the results of the international studies have contributed to a deepened interest in validity issues of teachers' assessments. This, in turn, has consequences for teachers' assessment practice and teaching professionalism. For example, in order achieve a more uniform assessment practice among teachers, national tests have been implemented in a greater range of subjects than previously, and in earlier school-years. Furthermore, a new authority, the Schools Inspectorate, was established in 2008 and tasked with monitoring and controlling, amongst other things, teachers' assessments.

It can be concluded that the increased interest in valid assessments around the turn of the millennium has been intensified over the past decade, and the discussion about how to validate inferences drawn from teachers' judgements is vibrant (e.g., Gustafsson & Erickson, in press; The Swedish School inspectorate, 2010, 2011). With a background in these discussions, the present thesis aims to contribute further to knowledge about the crucial issue of validity in educational assessment.

Purpose

The overall purpose of the thesis is to contribute to the knowledge about how different forms of assessment function in Swedish primary school. Focus is directed to teacher judgements, pupil self-assessments and a standardized external test.

The thesis consists of an overarching discussion and four separate empirical studies. The relationships between the assessment forms are investigated in the four studies, where, even though the research questions do not concern validity explicitly, validity is nevertheless a common theme. The purpose of the overarching discussion is to provide a comprehensive picture of the validity of the three assessment forms. It has been written with the aim of elaborating and summarizing the results from the studies and could be read independently for those who do not want to immerse themselves in the studies.

The overarching discussion focuses on a number of issues explored in the four sub-studies:

1. How do teacher judgements of reading achievement work within classrooms and for classroom comparisons in grades 3 and 4?
2. How well do primary school pupils assess their own reading achievement?
3. How is pupil gender and socioeconomic status related to teacher judgements and pupil self-assessment?
4. How is teacher competence related to pupil achievement and to the teachers' judgement practice?

Guidance for readers

Swedish PhD theses that have focused on issues of validity in assessment have often concerned secondary and upper secondary school, or university education (e.g., Jönsson, 2008; Klapp-Lekholm, 2008; Selghed, 2004). However, there is a need to investigate these issues in primary school too, particularly in light of the trend towards earlier grade assignment. Moreover, very few studies have

investigated assessment practices within classrooms and between classrooms (teachers) simultaneously. One reason for this may be a lack of analytical techniques for decomposing the variance of the performances into individual and aggregated levels. The development of multilevel structural equation modeling (SEM) with latent variables makes it possible to simultaneously consider and estimate the effects of individuals (social characteristics, achievement) and effects at the class level (group achievement, teacher characteristics).

In this thesis, all measures of achievement concern knowledge and skills in the reading domain. Reading is considered as a fundamental knowledge which is the basis for performances in other subjects too. In the PISA study, performances in reading were shown to correlate highly with performances in mathematics and science (The Swedish National Agency for Education, 2001). This is a reason why measures of reading literacy are well fit to indicate school achievement. The IEA (International Association for the Evaluation of Educational Achievement) provides high quality reading achievement data from 9-10 year olds and it is this data that has been used in the thesis. Data from the Swedish PIRLS 2001 study have been particularly useful, since this assessment included some national additions among which a unique material was distributed to the teachers on which they assessed each and every pupil's reading achievement in their own classroom.

In the overarching discussion, the theoretical framework consists of three parts. The chapter 'Assessment of educational achievement' elaborates some of the definitions of the concept of assessment and provides a context for the types of teacher assessment in focus in the thesis. Thereafter, the chapter 'Validating measures of achievement' is devoted to validity theory and models for validation. An argument-based approach to validation is adopted. The starting point is that individual analyses with information from a variety of sources should be combined to provide strong arguments for sound interpretations of assessment results. The final part of the theoretical framework, 'Relations between different forms of assessment: an overview', discusses results of research on the relationship between different forms of assessment, particularly the relationship between teacher assessments and test scores/self-assessment. A methodology chapter follows the theoretical part, where the data and the methods used in the different studies are presented. Thereafter, the 'Result and Discussion' chapter summarizes and discusses the results of the thesis. In the chapter 'Conclusions' a number of methodological challenges are highlighted and directions for future

research are suggested. Then follows a Swedish summary and finally the four studies in full.

Chapter Two: Assessment of educational achievement

Although assessment in education is currently a hotly debated phenomenon, systematic assessments have been made for a long time. In fact, assessment is a central part of everyday life, and a number of things, such as speech, clothes and behaviour are things people continuously assess. However, education provides a setting where assessments have particular importance. Educational assessments can be made at many different levels (e.g., teachers assessing pupil knowledge, principals assessing teachers, school inspectorates assessing schools, and so forth) and for many different purposes (promoting learning, selection, certification, etc.). Educational assessments can be traced back to China 2000-3000 years ago, where performance-based examinations were conducted to assign different positions in the society (e.g., Lundahl, 2006; Madaus & O'Dwyer, 1999).

Even though assessment was present in ancient societies, it was in the first half of the 20th century, the major developments in the area of assessment were first made. The need for measuring aptitude and achievement increased and many assessments focused on selection and certification. In response to these new demands, the development of psychometrics took off (e.g., Binet & Simon, 1916; Spearman, 1904).

Further, the objectives of assessment have developed towards monitoring the outcomes of education and with the purpose of driving both curricula and teaching (Gipps, 2001). Ball (2003, 2010) has described a change in the governing of knowledge resulting in new demands for schools and teachers. New regulations entail an intensified use and gathering of performance data from large-scale assessments like the PISA studies and national evaluation systems, such as for example school inspection programs. In recent decades, an increasing focus on improving 'outputs' in education and on competition between schools has emerged in Sweden. Older policy technologies like bureaucracy and teacher professionalism have made way for newer policy technologies; market, managerialism and performativity (Englund, Forsberg, & Sundberg, 2012; Myrberg, 2006; Sjöberg, 2010). Government, schools and teachers are now held accountable for results of assessments of various kinds. In Sweden, the School Inspectorate holds schools accountable not only for

violation against rules and regulations, but also for unsatisfactory achievement results. Also, the trend internationally has been that the information about quality and efficiency affect ways in which educational systems are monitored and reformed at every level and in every sector (Ball, 2003).

Common notions of educational assessment

There are many concepts related to the notion of assessment. 'Assessment', and 'evaluation' are commonly used and, sometimes, even used interchangeably. In the UK 'assessment' refers to judgements of pupil work, and 'evaluation' to the process of making such judgements (Taras, 2005). Broadfoot (1996) noted that some authors distinguish between 'assessment' as the actual process of measurement and 'evaluation' as the following interpretation of such measurements against particular performance norms. 'Evaluation' is often associated with aggregated levels, such as when school or countries are being evaluated. Scriven (1967) defined evaluation as:

> Evaluation is itself a logical activity which is essentially similar whether we are trying to evaluate coffee machines or teaching machines, plans for a house or plans for a curriculum. The activity consists simply in the gathering and combining of performance data with a weighted set of goal scales to yield either comparative or numerical ratings (Scriven, 1967, p. 2-3).

This definition could also apply to the concept of 'assessment', and may be a function of the time and place when it was written. In general, there is little consensus as to when to use 'assessment' and when to use 'evaluation'. Scriven (1967) emphasized the goals which performances should be compared to, which Sadler (1989) has subsequently expanded upon by describing the multiple criteria that often are used in relation to evaluations intended to support pupil learning. Multiple criteria have been characterized to be fuzzy rather than sharp, that each criterion should not be decomposed in parts, and that only a small subset are to be used at the time.

Furthermore, as Gipps (1994) pointed out, 'assessment' may also refer to a wide range of methods which are used to evaluate pupil knowledge and skills, for example, large-scale studies, portfolios, teachers' assessments in their own classrooms, and external test-results. Assessments of pupil achievement made by teachers are often called *teacher assessments*. However, in the US, 'teacher assessment' refers to the assessment of teachers' competencies (Gipps, 1994).

The varying uses of 'teacher assessment' is perhaps one reason why the term 'teacher judgement' is commonly used to label statements about pupil achievement in previous research (e.g., Feinberg & Shapiro, 2009; Hoge &

Coladarci, 1989; Martínez, Stecher, & Borko, 2009; Südkamp, Kaiser, & Möller, 2012). Teacher judgement is also in the present thesis used to denote the assessments teachers carry out. The term teacher ratings could also have been used, but ratings refer rather to single observations of different aspects of a construct. A judgement encapsulates any given information with bearing on the assessment carried out (Taras, 2005). When assessment outcomes i.e., test-result, observations, and portfolios, are being aggregated and interpreted by the teacher, the inferences (from many different ratings) lead to a judgement about pupil achievement.

Furthermore, the term assessment often embodies a summative and formative meaning, and a distinction between these two concepts has been made in literature. Summative and formative evaluation were coined by Scriven (1967), who underlined that these two concepts can be used in many various contexts, and at many different levels. Thus, summative and formative forms of assessment are not merely associated with assessments of pupil knowledge and skills, which has been the dominating area of use in the past few years.

While summative judgements do not always improve learning, they are nevertheless a necessary condition for learning. Judgements or test results which are summative and are used for selection and grades could also be used in a formative way (see for example, Harlen, 2011; Newton, 2007; Stobart, 2011). Scriven (1967) and Taras (2005) have emphasized that the assessment process basically leads to a summative judgement and that it is possible that the assessment is solely summative if the assessment stops with the judgement. For an assessment to be formative, a feedback component is required, however, assessment cannot be solely formative without a summative judgement preceding it. In a situation where the goal is to promote learning, feedback is information about the gap between actual knowledge level and a reference level, and is used in attempts to lessen the gap (Ramaprasad, 1983; Sadler, 1989). Newton (2007) has described assessments as either summative or descriptive—and not formative—arguing that the formative concept should be seen as a purpose of an assessment. Thereby, talk about summative and formative assessments can be misleading since method and purpose are not separated.

Assessing reading literacy in Swedish primary schools

Since the 1970s, several Swedish language diagnostic materials have been available as support for teachers' assessments in primary school (see for example, Pehrsson & Sahlström, 1999). One reason to use diagnostic materials was to help teachers to follow-up pupil language development in a systematic way, while

another had the aim that pupil performances should be assessed in an equal manner independently of which school the pupil attended, which books were used in teaching or which teaching methods had been applied. Moreover, the diagnostic materials should highlight individual pupils' strengths and weaknesses within a given subject and in this way contribute to the effective planning of further education (The Swedish National Agency for Education, 2002). In 2001, when data was collected for PIRLS, the Swedish National Agency for Education provided assessment support for the subjects Swedish and Swedish as a second language for grades 2 and 7; in addition to this, national subject tests were provided, but only in grade 5 and 9. In order to facilitate a systematic assessment practice in the primary school years, the Swedish National Agency for Education developed a diagnostic scheme which was to be used over a longer period of time. The diagnostic material launched in 2002[3], was more comprehensive, applying to all years of primary school prior to grade 6 (The Swedish National Agency for Education, 2002). Parts of this material was used in the present thesis.

In the context of the present study, the Swedish PIRLS 2001 report indicates that over 90% of the teachers (grades 3 and 4) placed great importance on their own professional judgement when assessing pupil achievement in reading (Rosén, Myrberg, & Gustafsson, 2005). Some 10% of the teachers ascribed great importance to written tests (teacher-made or textbook). One reason that teachers on average trusted their own professional judgement to such a great extent might have been due to their length of experience (m= 17.5) and long education (Rosén et al., 2005). Given the open frames for assessment in the beginning of the 21st century, many teachers most likely trusted their own observations and intuition. Gipps, Brown, McCallum and McCallister (1995) explored the teacher assessment models in the UK primary schools and identified three main models, the 'intuitives', the 'evidence gatherers' and the 'systematic planners'. The 'intuitives' tended to rely on their 'gut reaction', which basically implies that they memorized what children could, and could not do. The 'evidence gatherers' collected as much evidence as possible and from a variety of sources. They felt accountable to parents and principals and tended therefore to rely on written evidence. The 'systematic planners' devoted some part of the school week for assessment. These teachers used many and varied assessment techniques. For these teachers, assessment was a kind of diagnosis of how the pupils were doing

[3] At the time of the data collection 2001, the syllabuses did not include criteria for pupil minimum achievement levels in grades 1-4. Requirement levels were introduced in grade 3 with the latest curriculum 2011.

on the tasks, with the teacher taking notes and planning accordingly for the next activity.

Based on the primary school teachers' reports, and given that teachers in grade 3 and 4 in 2001 did not have explicit criteria or national tests to rely on, the results of the PIRLS report seem to be in accordance with the practice of the teacher-type Gipps et al. (1995) describe as 'intuitives'. However, in Gipps et al's study 'intuitives' did not adapt to the criterion-referenced system, while 'systematic planners' on the other hand, had adapted to the criterion-referenced system. These teachers believed in carrying out ongoing formative assessment and note-taking. Relying solely on memory was a strategy they found untrustworthy. The "PIRLS teachers" in general had a lengthy education and long experience and it seems reasonable that they could be flexible and rely on their intuition and expert judgements. Indeed, great flexibility is needed in teaching and assessment for learning to be efficient (Pettersson, 2011). The introduction of the diagnostic materials 2002 in the Swedish primary school was a step toward more systematic observations in teacher assessment, since the diagnostic material was meant to support teachers with criterion referenced assessment. In 2001, and in connection with the PIRLS 2001 study, an initiative to test the diagnostic material was undertaken by letting teachers rate pupil knowledge and skills on the different aspects in the diagnostic material. This dataset is exploited in the current thesis. The observational aspects in the diagnostic material are described in more detail in the Methodology chapter and can be viewed in the Swedish PIRLS report (Rosén et al., 2005).

Chapter Three: Validating measures of achievement

Cross-validation of different assessment forms can provide information about how well the results from one assessment can answer certain questions. Already in 1963, Cronbach stated that the greatest service evaluation can perform is to identify aspects of a program where revision is desirable. This statement is thus related to the formative aspects of assessment. However, validity must be determined before one can improve assessment forms of different kinds. Via mutual validation of teacher judgements, external test results and pupil self-assessments, it is possible to identify the strengths and weaknesses in the different assessment forms. For example, if the inferences of teacher judgements are found invalid for a particular use (e.g., classroom comparisons), the information about invalidity can be used to shape teachers' judgements. Different assessment forms can also be more or less useful at different levels of the educational system. Assessment of individual pupils may require other methods than the evaluation of classrooms or schools. In order to investigate the quality of assessments, validation is powerful, useful, but also necessary. The following section provides a background to validity theory and a framework for validation. First, focus is placed on a general understanding of the concept and thereafter Toulmin's model of arguments is used as a framework for validation.

Validity

Validity is no longer seen solely as a property of an assessment, but rather in terms of the interpretations and inferences drawn from assessment results. To evaluate the soundness of inferences based on different forms of assessment, validation is required.

Messick's (1989) framework has been proposed as a suitable theory for validating assessments in an educational context (see for example, Klapp-Lekholm, 2008; Nyström, 2004). One reason for this is that Messick takes the consequences of assessment into account, which, without doubt, are important in many educational settings. In formative assessment, for example, validity hinges on how effective learning/improvement takes place (Stobart, 2011). This therefore becomes an important aspect of consequential validity. However,

Messick's theory provides limited guidance on how, in practice, these consequences can be investigated (Bachman, 2005). It is also beyond the scope of this thesis. Validity theory and validity in practice have been shown limited overlaps and this gap has increased with the introduction of broader perspectives of validity (Wolming & Wikström, 2010). Taking the standpoint that validation requires evidence from multiple sources and because it is a never-ending enterprise, the argument-based approach (Kane, 1992, 2006; Toulmin, 1958/2003) for validating performances provides a logical set of procedures for articulating claims and for collecting evidence to support these claims. These are described in detail below. However, the first part of this chapter describes the concept of validity and its development from the early 20th century onwards.

In the present thesis, construct validity is treated as a unified form of validity. Initially, in order to describe how a unified view of validity has emerged, an account of how validity was previously broken down into three different subtypes is provided. In measurement science, a sharp distinction is sometimes drawn between validity and reliability. Most often reliability is taken as a direct evidence of validity, and the two are sometimes regarded as equivalent (Lissitz, 2009). Already in 1954, Cureton stated that validity has two aspects, which he labelled relevance and reliability. In the present thesis, reliability is regarded as a part of the validity concept and as a necessary, but not sufficient, condition for validity (Messick, 1989). The technical aspects of reliability are not covered in any detail here.

Early definitions of validity

The first definitions of validity were very straightforward. Guilford's (1946) definition of the concept was that a test was valid for anything which it correlates with. Guilford's definition was further developed by Cureton, (1951) who emphasized the relevance of the test purposes and uses:

> The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third. Hence, we cannot label the validity of a test as "high" "moderate" or "low" except for some particular purpose" (Cureton, 1951, p. 621).

These two definitions of validity point out that, for example, if a test designed to measure word knowledge is highly correlated with the construct of intelligence, the test would be a valid measure of intelligence. Cureton's definition points to the importance of the purposes with a test. It is therefore not possible to draw the conclusion that a particular test is invalid without knowing what the test was purported to measure. Up to the mid-20th century, validity was viewed as a

property of the test itself (Wolming, 1998). However, in the 1950s a more elaborated view of validity emerged.

The concept of validity has typically been broken down into three types, one of which comprises two subtypes (Messick, 1989). These are content validity, criterion related validity and construct validity. Between 1920 and 1950, criterion validity came to be the gold standard for validity (Angoff, 1988; Cronbach, 1971), although over time development drifted towards a unified view, where construct validity was equal to validity.

Criterion validity

The criterion model is often divided into concurrent and predictive validity. Concurrent validity indicates how well performances for the same or similar constructs correlate, e.g., correlations of standardized test scores and teacher judgements. It can be used to validate a new test which would then be compared to some kind of benchmark, i.e., criteria or earlier tests. Predictive validity refers to how well criteria are suited to predict future performance. The Swedish Scholastic Assessment Test for admission to higher education (SweSAT) is an example of a test which aims at predicting future study success. The main limitation of the criterion model is that it is difficult to obtain an adequate criterion, and ways of evaluating it. For example, it can be problematic to conceptualize and operationalize a satisfactory criterion for a latent trait, such as reading ability. The criterion model is useful in validating secondary measures, given that some primary measure can be used as a criterion. However, it cannot be used to validate the criterion, which has to be validated in another way (Kane, 2006).

Content validity

The content model interprets how well performances in a particular area of activity can be an estimate of overall ability in that activity. Content validity is dependent on how well the performance or tasks in a specific domain can be used to draw inferences about a larger domain. One of the main criticisms of the content model is that the evidence tends to be subjective. Content-based analyses tend to rely on expert judgements about the relevance of test tasks. Furthermore, test developers have a tendency to confirm their proposed interpretations (Kane, 2006).

Construct validity as the whole of validity

The construct model of validity was proposed as an alternative to the criterion and content models (Cronbach & Meehl, 1955). Construct validity came to be seen as representing validity theory as a whole (Loevinger, 1957). Cronbach and Meehl suggested that construct validity must be used whenever no criterion or universe of content is accepted as adequate to define the quality being measured. It has been proposed that construct validity can be expressed as the correspondence between the theory for the construct and the instrument measuring the construct (Wolming, 1998). Messick (1989) further elaborated the concept of construct validity. He stated that construct validity is based on an integration of any evidence that bears on the interpretation or meaning of test scores. Messick's view of validity extends the boundaries of validity beyond the meaning of tests score to include relevance and utility, values and social consequences. Although Messick's model of construct validity has witnessed mainstream use, it has also attracted a fair amount of criticism. For example, it has been argued that the aspect of social consequences should not be mixed up with validity (Mehrens, 1997).

The current general view of construct validity theory is that it refers to the interpretations and actions that are made on the basis of assessment results (Cronbach, 1972; Messick, 1989; Kane, 2006). However, Borsboom, Cramer, Kievit, Scholten and Franic (2009) argued that this view is a misconception. Instead, they proposed that validity is a property of the measurement instruments and whether these instruments are sensitive to variation in the targeted attribute. This view of the concept is similar to how the concept of validity was first defined; a test being valid if it measures what it should measure. Borsboom et al. thus argue that validity is a property of the assessment itself, not a property of interpretations of assessment results. One problem with the common definition of construct validity (Cronbach & Meehl, 1955; Kane, 2006; Messick, 1989) is that, by regarding validity as a function of evidence, the interpretations of data could be valid under certain conditions but invalid under others (Borsboom et al., 2009). Thus, test results may represent a more "true" ability for some groups of pupils than for others. Furthermore, Lissitz and Samuelsen (2007) argued that the unitary concept of validity is too broad for educational assessments, and consider its main focus to be on the test itself. They suggested that validation of a test should be labelled as content validity. Another critique is that the inferences drawn from test interpretations could be unrelated to the test-scores (i.e., valid interpretations made on the basis on an invalid test). As Borsboom and colleagues (2009) made clear that if a test does

not measure anything at all, it can never be valid in the first place, and therefore it makes no sense to examine the validity of the inferences based on the interpretations of such tests.

There are many researchers who agree upon that the common understanding of the term 'validity' should be what a test purports to measure. Many textbooks also present this rather straightforward definition. The answer to whether a test measures what it purports to measure requires a degree of evidence. Previously, a single correlation coefficient often was accepted as sufficient (Shepard, 1993). However, viewing validity as a property of a test may lead to unreflected conclusions about validity as a whole. Kane (2006) described the unified concept of construct validity, pointing to three major positive effects with construct validation. First, the construct model focuses its attention on a broad array of issues which are essential to the interpretations and uses of test scores. Thus, the model is not simply based on the correlation of test scores with specific criteria in particular settings and populations. Second, construct validity emphasizes the general role of assumptions in score interpretations and the need to check these assumptions. Finally, it allows for the possibility of alternative interpretations and uses of test scores and other forms of assessment.

Threats to construct validity

The two major threats to construct validity are labelled construct under-representation and construct-irrelevant variance. Construct underrepresentation occurs, according to Messick (1995), when an assessment is too narrow and fails to include important dimensions or facets of the construct. An example of this would be a test that aims to capture reading literacy but focusing too much on word knowledge.

If an assessment suffers from construct irrelevant variance, it is too broad, containing systematic variance associated with other distinct constructs. It could also be related to method variance, in the sense that response sets or guessing propensities affect responses in a manner irrelevant to the interpreted construct. For example, non-cognitive factors such as behaviour and effort might be taken into consideration when teachers assess pupil reading achievement. Construct irrelevant variance could also concern bias in written test answers. Answers written in neat handwriting may bias teachers' judgements, and therefore, lead to conclusions about cognitive skills, based on misinterpretation of motorical skills. It is thus important to be aware of construct irrelevant variance in all educational measurements. As Messick (1995) pointed out, in particular it concerns the contextualized assessments and the authentic simulations of real-world tasks.

Validation

Critical validation is required when examining the proposed interpretations and uses of test scores. Validation is the process by which one validates the interpretations of data arising from a specific procedure. This implies that the test in itself is not subject to validation; rather it is the actions and inferences drawn from the test scores that form the focus of validation. For example, a reading test, could be used for grading purposes, or as a diagnosis for adjustments in teaching. Each application is based on different interpretations and evidence that justifies one application may not have relevance for another. Cronbach (1971) stressed that even if every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test "is valid".

Validation examines the soundness of all interpretations of a test – descriptive and explanatory interpretations as well as situation-bound predictions (Cronbach, 1971, p. 443). It is an ongoing process of investigation, and as Cronbach (1988) concluded, it is a never-ending enterprise. In practical terms it is merely possible to make a final statement about the validity of anything. Therefore, even though one may strive for strong evidence and arguments for reasonable judgements, interpretations of assessments may change over time as new knowledge is generated. However, accuracy in the validation process depends on the interpretations and the claims being made. If the results of the assessment have a direct and straightforward interpretation, little or no evidence would be needed for validation; that is to say if the interpretation does not go much beyond a summary of the observed performance. For example, if a teacher reports that a pupil managed to successfully identify 30 out of 40 words in a word knowledge test, this would probably be accepted at face value. A stronger claim about the performance, however, would require more evidence. If the performance was taken as evidence that the pupil had good reading comprehension, we might have to ask for a definition of reading comprehension and why this kind of performance is appropriate as a measure of reading comprehension in general for pupils of this age and gender. In validation, the proposed interpretations are of great importance and the arguments for the interpretations must be cohesive. To accept a conclusion without critical examination is known as the fallacy "begging the question of validity" (Kane, 2006).

Using an argument structure for validation

The argument-based approach to validity reflects the general principles of construct validity. Validation, according to Kane (2006), requires two kinds of argument. On the one hand, it requires an interpretive argument, which specifies the proposed interpretations and uses of assessment results by setting out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances. On the other hand, there is the validity argument, which provides an evaluation of the interpretive argument. To claim that a proposed interpretation or use is valid is to claim that the inferences are reasonable and the assumptions are plausible. In other words, the validity argument provides an evaluation of the interpretive argument and begins with a review of the argument as a whole as a means of determining whether it makes sense.

Theoretical models can be used to describe how assessment results can be interpreted and used. To illustrate the validation of the assessment process, Kane, Crooks and Cohen (1999) introduced the bridge analogy, which describes how interpretations must be reliable in three steps in order to make a conclusion valid. One rationale for this analogy was the fact that while a general validity problem can be very difficult to comprehend, if broken down into components it becomes less complex. The model is highly useful not only in relation to the validation of performance assessments, but also in other assessments where scoring, generalization and extrapolation need to be elaborated. In the present thesis, scoring of the different assessments has already been made, and other models for validations can be adequate. The questions in this thesis regard the validity of the inferences made on the basis of different forms of assessments. The research agenda is to either support or to problematize the different claims that are made on the basis of the different assessment forms. The Toulmin model (1958) provides a logical structure of arguments to support or reject claims about a performance. This model thus seems to be appropriate for the objectives of the current thesis.

Toulmin's structure of arguments

Toulmin (1958/2003) proposed a general framework and terminology for analyzing arguments which has been used in a variety of contexts. In the field of language testing, Bachman (2005) has expanded upon argument-based approaches by proposing an 'assessment use argument' (AUA) framework that links judgements to interpretations about language ability. AUA consists of two parts: a validity argument, which provides logical links from performance to

interpretation and a utilization argument, which links the interpretation to test use. In particular, the validity argument of AUA seems to be appropriate for use as a framework for investigation of the validity of the interpretations made on the basis, for example, of teacher judgements of pupil reading skills. This framework is grounded in Toulmin's (1958/2003) argument structure. For Toulmin, an argument consists of making claims on the basis of data and warrants. The assertion of a claim carries with it the duty to support the claim and, if challenged, to defend it or, as Toulmin (1958, p.97) puts it, "to make it good and show that it was justifiable". A diagram of the structure of arguments is provided in Figure 1 below.

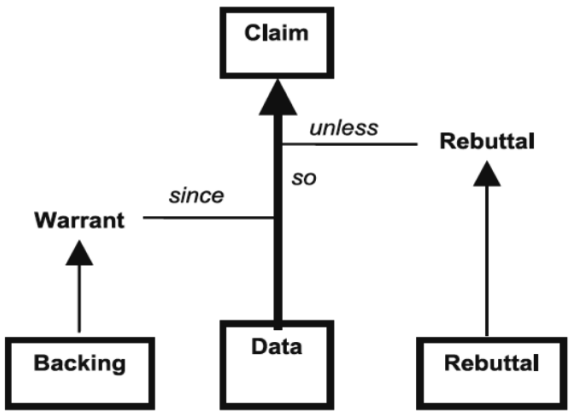


Figure 1. Toulmin diagram. Bachman (2005, p. 9)

A **Claim** is an interpretation of an assessment result; it concerns what the pupil knows and is able to do. **Data** are the pupil performances on the assessment and the characteristics of the assessment procedure, or, as Toulmin (1958, p.90) explains, the "information on which the claim is based". **Warrants** are propositions used to justify the inferences from the data that lead to the claim. **Rebuttals** are alternative explanations or counterclaims to the claim. Finally, **Backing** is the evidence used to support the warrant and weaken the rebuttal. Backing can be obtained from the test design and development process, as well as from evidence collected as part of research studies and the validation process. This model will be used as a method of analysis in this overarching discussion about the validity of the three different forms of assessments.

In the next chapter, a more concrete approach to validity is taken where previous research regarding the relation between teacher judgements, external tests and pupil self-assessments is presented.

Chapter Four: Relations between different forms of assessment: An overview

In this chapter, an overview of research on validity issues in different forms of assessment is provided. Previous research with a primary focus on assessments of reading achievement, and particularly in the primary school-years, is presented.

The research area of validation of assessments is very broad and includes studies using many different methods and samples. In the US particularly, there has long been interest in evaluating the quality of different assessment forms. In Sweden, studies with a focus on validity aspects of different assessment forms are fewer (Forsberg & Lindberg, 2010). Rather than covering a wide range of studies, the aim of this chapter is to focus on studies more closely related to the research objectives of the current thesis.

The principles underpinning searches of the assessment literature included, had as a starting point, the most relevant keywords with regard to the research questions in the current thesis. Systematic searches of the literature were conducted where keywords such as 'teacher judgement', 'teacher rating', and 'pupil self-assessment' were used. Primarily, Swedish studies, reviews of the literature, and meta-analyses have been selected. Although not all of these relate to primary school years and reading, they can however provide an overview of results, to which the current results can be compared. The references of the review studies have also been explored in some detail, many being found to be of particular importance for the current purposes. Typically these studies used similar assessment methods and related to the same subject domain as the current research.

The intention is to shed light on the complexity of assessments and what the different assessments can and cannot measure, in terms of scholastic performance at the individual as well as aggregated levels. The first part of the chapter elaborates the relationship between teacher judgements and standardized tests, and how different aspects—such as pupil and teacher characteristics—can influence the assessments. The next part of the chapter concerns pupil self-

assessments and their agreement with other forms of assessment. Here too, different factors that could affect the validity of self-assessments are discussed.

Teachers assessing pupil achievement

Teacher judgements are one of the most important activities for pupil learning outcomes (Hattie, 2009; Lundahl, 2011). Teacher judgements play an important role in making daily instructional decisions, conducting classroom assessments, determining grades, and identifying pupils with disabilities or those in need of special assistance. Because of their vital role in education, the quality of teacher judgements has been closely examined in various areas of research (e.g., Brookhart, 2012; Hoge & Coladarci, 1989; Harlen, 2005).

Much of the research that has examined the quality of teacher judgements has been in the context of the early identification of learning and reading difficulties. One reason for this may be the importance of the early identification of pupils with difficulties. The acquisition of early reading skills has proved to be crucial for future academic performance. Those who are able to read early are also likely to read more, which may trigger an upward spiral into motion (e.g., Cunningham & Stanovich, 2006).

Teachers have a particularly important responsibility for identifying pupils' skills in reading and many studies have examined the quality of teacher judgements in relation to external measures of achievement, such as standardized test results (e.g., Black & Wiliam, 2006; Harlen, 2005; Feinberg & Shapiro, 2009). In Sweden, such research is quite rare, especially for the primary school years. One reason for this may be that assessments of younger pupil abilities, in accordance with curricula, have been expressed in a qualitative manner, in, for example, individual education plans. Studies of the relation between teacher judgements and test results have, however, been conducted for the secondary and upper secondary school, where grades and national tests have been used.

The Swedish National Agency for Education (2007, 2009) has studied the correspondence between final grades and national tests in the final year of compulsory school and in upper secondary school. The results showed that most pupils got same national test grade as the final grade. The correlation amounted to about .80. However, the results indicated that the correspondence differed substantially from one teacher to another. This has raised questions concerning equality in assessment since different teachers seem to interpret criteria differently. As regards the correspondence within a classroom, Näsström (2005) has found that teachers in Swedish upper-secondary school are adept at estimating their pupils' national test grades in math. In her study, the four

grading steps (IG-MVG) were reformulated to a 12-point scale to allow for more nuanced estimations. The correlation between teachers' estimations of pupil national test results and pupil actual test scores amounted to .80. In contrast to the studies conducted by the Swedish National Agency for Education, teachers in Näsström's study were explicitly asked to estimate their students' national test performance. One might suspect that the overall mathematics subject grade include more non-cognitive aspects than do the test-score predictions, but given the consistent findings this seems not to be the case.

In a meta-analysis, Südkamp, et al. (2012) investigated 75 studies on the issue of the accuracy of teacher judgements. Although most of the studies included in their analysis were conducted in the US, studies from all continents except South America were represented. The authors concluded that the relationship between teachers' judgements of students' academic achievement and students' actual test performance was "fairly high", with a correlation of .63. However, because they found teacher judgements far from perfect and considering the unexplained proportion of variance, the authors advise that this result should be treated with caution. Further, Südkamp et al. found large variability in the correlation across different studies, a finding consistent with, for example, the results of Hoge & Coladarci's (1989) earlier review of the literature on teacher judgements. Moreover, Südkamp et al. (2012) suggested that judgement and test characteristics were two moderators of the relationship between teacher judgements and pupil achievement.

In the US, Meisels, Bickel, Nicholson, Yange, and Atkins-Burnett (2001), examined the relationship between teacher judgements on a curriculum-embedded assessment of language and literacy and a standardized measure from kindergarten and through 3rd grade. They concluded that teacher judgements of pupils' performance could be trusted, since they correlated well with external measures. Teacher judgements were strong predictors of achievement scores, and accurately discriminated between pupils who were at risk and those who were not. In another study from the US, Llosa (2007) investigated the relationship between standards-based classroom assessments and standardized tests of English reading proficiency in grades 2-4. The teacher-assessed scores and standardized test scores were aligned to the same standards, and via a multivariate analytic approach, Llosa concluded that the correspondence between the two measures was high. Beswick, Willms, and Sloat (2005) used correlational analysis to examine the correspondence between the information derived from teacher ratings and from a standardized test with prior evidence of construct validity. Beswick et al. were positive about finding a correlation

between the two achievement measures of .67, but raised concerns regarding findings showing that teacher judgements were systematically affected by extraneous variables, such as pupil and family characteristics. Teachers rated boys and pupils from lower SES lower than the standardized test results indicated. Consequently the researchers advise caution in the use of teacher ratings in grade retention decisions.

Most studies that have examined validity issues of teacher judgements have used an approach that has focused either on the extent to which judgements correlate with standardized test measures (Beswick et al., 2005; Brookhart, 2012; Coladarci, 1986; Hoge & Coladarci, 1989; Meisels et al., 2001; Taylor, Anselmo, Foreman, Schatschneider, & Angelopoulos, 2000) and/or the extent to which judgements accurately predict future performance (Gijsel, Bosman, & Verhoeven, 2006; Hecht & Greenfield, 2002; Taylor et al., 2000). The principal focus of these studies has been general teacher judgements of pupil achievement (Hoge & Coladarci, 1989; Perry & Meisels, 1996), emerging reading and literacy skills (Bates & Nettleback, 2001; Beswick et al., 2005; Meisels et al., 2001), and reading and learning disabilities (Reeves, Boyle & Christie, 2001; Taylor et al., 2000).

Standardized test results have often been used as a criterion to measure teacher judgements, rather than the other way around. In this sense, standardized test results are often viewed as more objective and a more valid measure of achievement. However, low correspondence between test results and teacher judgements may also be caused by low reliability of tests (e.g., Harlen, 2005). Furthermore, to achieve high construct validity of external test-results, tests need to be aligned to the constructs stated in the curricula and syllabi. If they are not, a mismatch between teacher judgements and external test results may appear. In the context of exploring the construct validity of assessment interpretations, an important question is whether the content in standardized tests accords with the content of the subject assessed by the teachers.

For example, when results from PIRLS are to be interpreted and used in a national context, like Sweden, it is important to compare the PIRLS framework not only with the Swedish curriculum (Lpo 94) but also the syllabus for Swedish. If the correspondence is high there are good grounds to use the results from PIRLS to articulate claims about pupil reading achievement, as well to use the results as a basis for discussion about and development of reading comprehension in Swedish schools. If the correspondence is low, there is a risk that the test fails to capture constructs that may be specific to the particular national setting (The Swedish National Agency for Education, 2007).

Another way to express this is to ask whether the framework in the international studies reflects the content and form of Swedish school education. Such analyses have been carried out by the Swedish National Agency for Education (2006) who explored the alignment between the content in PIRLS 2001 and the Swedish syllabus. More specifically, they investigated the agreement between the framework for reading in PIRLS and that in the Swedish curriculum and syllabus, specifically the goals to be attained at the end of the fifth[4] year of compulsory school. The Swedish National Agency for Education found the purpose of PIRLS to be well in line with the criteria in Swedish primary schools. This conclusion is also mentioned in a report from the same agency in 2007, although in this report it is emphasized that the PIRLS test cannot comprise the whole Swedish language subject domain, which may also not be the goal of PIRLS. A more in-depth study of the type of knowledge and skills that PIRLS comprises has been conducted by Liberg (2010) in which she examined the reading tasks in the questionnaires used in PIRLS 2006. Her findings suggested that most tasks in PIRLS involved knowledge regarding identification of information in the text and the ability to link different routes to find a context within the text. On the other hand, few items tested the ability to read between the lines, to use one's own experiences and to creatively interpret the text. However, Liberg (2010) also pointed out that if such tasks were allowed it would be difficult to correct the tests in an equal manner across different cultures.

Factors influencing teacher judgements

The Swedish Education Act (2010) states that there shall be educational equality between schools irrespective of school type and where in the country education is provided. Equality in education means that, for example, pupils with a disability or handicap should not be denied appropriate schooling. Furthermore, irrelevant aspects, such as for example gender, socioeconomic status or other non-cognitive factors should not be allowed to influence assessment and grading. If teachers have different frames of reference, given the same achievement levels, their assessments will nevertheless differ from one classroom to another. This could in turn mean that a pupil in one classroom might be provided with adequate assistance while a different pupil in another classroom might not. Consequently, it is crucial that teachers' judgements are in agreement, otherwise equality of education will be jeopardized. This concerns an aspect of

[4] The attainment goals in grade 5 were used, since these goals were not provided to the school-years prior to grade 5 in 2001.

the inter-rater reliability, an indication of how well different judgements of similar knowledge and skills are in agreement. However, even though teachers might consistently assess the same knowledge and skills, it does not follow that validity will be high since the construct validity of the assessed knowledge and skills might be low. Enhanced inter-rater reliability has been claimed when teachers have access to adequate scoring rubrics. Jönsson & Svingby (2007) reviewed the literature regarding scoring rubrics and arrived at the conclusion that the reliable scoring of performance assessments could be enhanced by the use of rubrics. However, their review concluded that rubrics did not facilitate valid judgements per se.

As previously mentioned, teachers' interpretations of goals and criteria have been shown to be problematic in Sweden (Selghed, 2004; Tholin, 2006). Interpretation of criteria is likely to be influenced by the length of teachers' education and amount of experience. In 2001, teacher characteristics varied largely in Swedish primary schools (Frank, 2009; Rosén et al., 2005). Teachers' characteristics are one cause of variation in assessment of pupil knowledge and skills (e.g., Llosa, 2008). However, the characteristics of individual pupils can also affect teachers' judgements. If teachers take account of non-achievement factors it may threaten the validity of the inferences drawn from teacher judgements. These problems are further elaborated below.

Teacher characteristics

Teachers with higher competence levels are likely to have pupils with higher achievement levels (Hattie, 2009; Darling-Hammond & Bransford, 2005). One hypothesis is that these teachers can more accurately identify their pupil knowledge and skills and thereby are better at adjusting their teaching to pupil's different knowledge levels. Relatively few studies have investigated the role of formal teacher competence for teachers' judgements of pupil achievement, perhaps because it has been hard to define and establish what a competent teacher is. Further, relevant data indicating teacher competence may be difficult to access.

Hanushek (1989, 2003) as well as Hattie (2009) have demonstrated that teachers have a powerful influence on pupil achievement. However, previous research has sometimes arrived at different conclusions about the impact of teacher competence. A reason for this may be the lack of consistency of the indicators and approaches of measuring teacher competence. For example, competence can be measured in terms of pupil outcomes; the higher pupil performances are, the higher the teacher competence. One of the advocates of this view is Hanushek (2003) who claimed that it is teachers' persona, rather than

university degrees or other educational qualifications, that is the key factor. Teacher competence has been measured by, for example, teacher certificate, academic degrees and experience but also by performance (e.g. principals, parents and pupils). In the present thesis a construct of teachers' formal competence was adopted. A similar construct was used by Frank (2009), who found strong effects of teacher competence on pupil reading achievement.

Results from studies using approaches with single indicators, such as education and experience, are sometimes unpredictable because the length and content of teacher education varies across studies. Different impact of teacher competence can also be due to the school subject under investigation. In general, stronger effects of teachers' formal competence have been found in the maths domain than in the reading domain (Wayne & Youngs, 2003). A review by Wayne and Youngs (2003) demonstrated that teachers with a master degree were likely to have pupils with higher achievement. Darling-Hammond (2000) has also presented support for the contention that a major in the subject field has an impact on achievement. In addition, it was shown that certification status is also of importance for pupil achievement gains. Contradictory results have however been found by Goldhaber & Brewer (2000), although their conclusions were drawn on the basis of a limited data material.

In Sweden, Myrberg (2007) has shown that an appropriate teacher education degree has a significant effect on pupil reading achievement. The effect size of holding a subject-appropriate degree amounted to no less than 0.33. Frank (2009) has operationalized teacher competence in a latent variable model that included several indicators of teachers' formal training and experience. She estimated significant effects of teacher competence on pupil reading achievement.

The analytical challenges when studying the effects of competence on teacher judgement are many. Teacher judgements need to be compared to a criterion (standardized test results) and the relationship between the two measures of achievement must be studied with respect to how the relationship varies in relation to different teacher characteristics.

In a study from the US, Martínez, Stecher, and Borko (2009) included several teacher characteristics in their analyses of the relationship between teacher judgements of pupil achievement and standardized test scores in grades 3 and 5. Using multilevel modeling analyses, they concluded that teacher judgements varied significantly across classrooms and that pupil achievement on tests could not explain much of the variance. Analyses using so-called random slopes revealed, however, that some teachers assessed pupil achievement in ways that

corresponded more closely to standardized test scores than others. The judgements of teachers that were based on more test-like classroom work had a higher degree of correspondence with pupil standardized test-scores than those that did not. Teachers' educational level (Bachelor degree, Master degree) was not found to have any significant effect on the agreement between teacher judgements and pupil test scores. D'Agostino, Welsh, and Corson (2007) suggested that, with pupil background differences controlled for, teachers who adhered to recommended assessment practices and whose teaching mirrored the test had students with higher achievement. To achieve more credible teacher judgements of pupil achievement, Harlen (2005) suggested training and moderation of teachers' assessments. For example, it was hypothesized that by discussing assessment related questions in teacher teams, common assessment practices and frames of references could be developed.

Student characteristics

Evidence from a vast amount of research using different methods indicates that assessments not only reflect pupil subject knowledge, but also pupil characteristics (Brookhart, 2012; Cross & Frary, 1999; Klapp-Lekholm & Cliffordson, 2008; Llosa, 2008; Thorsen & Cliffordson, 2012). Brookhart (2012), who reviewed literature on the use of teacher judgements for summative assessments in the US, found that non-achievement factors such as behaviour and effort were considered in teachers' judgements.

Klapp-Lekholm and Cliffordson (2008; 2009) studied grading in the final year of compulsory school in Sweden. They used two-level confirmatory factor analysis and identified a common grade dimension in teachers' grading, which they suggested was due to non-achievement factors (gender, family background and motivation), that was included in teachers' grading. Already in the early seventies, Svensson (1971) demonstrated that girls received slightly better grades than was justified by the national test results. This trend seems to be relatively stable in Sweden. In a study by Emanuelsson and Fischbein (1986) similar patterns were found. In Reuterberg and Svensson's (2000) study of gender differences in mathematics, results from previous research was confirmed. Regarding assessment of different SES-groups achievement, Svensson (1971) and Reuterberg and Svensson (2000) showed that, for different SES-groups, national test results corresponded fairly well with the actual grade level.

Different interpretations of the results from these studies have been proposed (Wernersson, 2010). One interpretation, for example, is that girls are awarded grades higher than justified by achievement. It might be some trait (behaviour, motivation) that teachers take into account during assessment (e.g.,

Fischbein & Emanuelsson, 1986). Another interpretation is that girls complete assigned course work more successfully, thus generating higher judgements (Wernersson, 1989).

In summary, much research show that teacher judgements of pupil achievement are often affected by elements which are not reflected in test-results. These factors can be due to both pupil and teacher characteristics. While factors such as pupil effort and attitude are certainly important for pupil achievement in school, they should not be the subject of teachers' assessment. It is crucial for the teaching profession that teachers validly and reliably assess pupil achievement. In the US at least, confidence in teacher judgements is low. Brookhart (2012) suggested that one implication of results showing low credibility for teacher judgements may a more extensive use of standardized tests. Other implications could be that the teaching profession will be less autonomous in relation to the assessment of pupils' knowledge and skills.

Pupils assessing their own achievement

While teacher assessments and tests are used both for summative status reports and feedback, pupil self-assessments are mainly used for formative purposes. However, pupil self-assessments could also be considered as summative assessments of pupil achievements (Taras, 2009). Asking pupils to self-assess their own knowledge and skills is a relatively easy way to obtain information about pupil performance in school. Klenowski (1995) and Ross (2006) suggest that the benefits of self-assessments are more likely to increase if three conditions are met: 1) that teachers and pupils have common understanding of goals and criteria; 2) that teacher – pupil dialogues focus on evidence for judgements; and 3) that self-assessments (in collaboration with teacher assessment) contribute to a grade.

For pupils to be able to correctly assess their own skills, they have to become aware of what they need to learn and where to go next, which is the basis of effective feedback (Hattie & Timperley, 2007). It has been suggested that pupil self-assessment can form an important and integrated part of learning. Black and Wiliam (1998) have argued that if formative assessment is to be effective, pupils need to be trained in self-assessment. They can then understand the main purposes of their learning and how goals can be achieved. Klenowski (1995) defines self-assessment as "the evaluation or judgement of 'the worth' of one's performance and the identification of one's strengths and weaknesses with a view to improving one's learning outcomes" (Klenowski, 1995, p. 146). This definition focuses on the improvement aspect of self-assessments and thus on

the consequential aspects of validity (Messick, 1989). The self-assessment concept is closely related to self-concept and self-efficacy, two concepts widely studied in psychological research. Self-concept is multidimensional and formed through experiences of the environment (James, 1890/1998). It is influenced especially by environmental underpinnings and the evaluations of significant others (Shavelson, Hubner, & Stanton, 1976). Self-concept judgements involve an evaluation of, among others personal characteristics, skills and abilities. Self-efficacy is a more specific construct that primarily concerns the cognitively perceived capability of the self and could be considered as the cognitive dimension of self-concept. Many self-concept researchers have considered academic self-concept to be an explanatory variable for pupil educational outcomes, whereas others assert that self-concept is mainly a consequence, not a cause, of pupil academic achievement (see for example, Bong & Clark, 1999).

Although pupil self-assessment has been an explicit goal contained in recent Swedish curriculums and syllabuses, it is also, in the current curriculum[5], stated in the knowledge requirements that in the end of year 3, pupils should be able to assess their own and others' competencies.

> …pupils in response to questions can give simple assessments of their own and others' texts, and also on the basis of responses work on and clarify their texts in a simple way" (The Swedish National Agency for Education, 2011, p. 216).

Whether self-assessment should be used at all in primary education is a contested issue. Teachers have sometimes argued that self-assessments are not sufficiently accurate to be used for feedback purposes (Ross, 2006). The use of self-assessment may be warranted if it is a high correspondence to other achievement measures, such as teacher judgements.

In the 1991 IEA Reading Literacy Study (RLS, 1991) pupils were asked to assess their own reading skills. In most countries, the correlation between performance and self-assessment was between 0.25 and 0.55 for narrative and expository scores on the reading test, and slightly less for document scores (Elley, 1992). There are also overviews on the accuracy of self-assessment. In Shrauger's and Osberg's (1981) review of 50 studies, it was found with regard to predictions of academic achievement, vocational choice and job performance that the validity of self-assessments was comparable to other forms of assessment such as teacher assessments and tests. In samples of older students, Falchikov and Boud (1989) reviewed 57 studies that compared self-assessed marks with teacher-marks, finding substantial correlations between the two. In

[5] Curriculum for the compulsory school, preschool class and the leisure-time centre (Lgr 11).

consequence, it was concluded that self-assessments provide trustworthy evidence of pupil achievement. Falchikov and Boud also found better agreement between pupil self-assessments and teachers' assessments at more advanced educational levels. However, in a Swedish study Fredriksson, Villalba, & Taube (2011) found a weak association between grade 3 pupil self-assessments of reading skills and reading test results. The correlation amounted to about .3 between self-assessments and test results. Swalander (2006) estimated effects of academic self-concept on grade 8 pupil reading achievement in IEA's reading literacy study in 1991. The beta values were estimated to .42 for the main sample and .56 for the cross validation sample.

From this review, support for the validity of self-assessments is certainly not overwhelming. Nevertheless, in terms of predictive validity, pupils have been shown to make reasonably trustworthy predictions of their achievement. The modest correlations (.20-.30) presented, may cast doubt on the validity of self-assessments and their use in school and there are reasons to believe that self-assessment varies between pupils with some more accurate than others. In the next section, this is examined more closely.

Factors influencing pupil self-assessments

It is a high complexity involved in the self-assessment of knowledge and skills and it can vary due to a number of factors. Factors at the individual level, such as gender and SES have previously been shown to influence self-assessments (e.g., Reuterberg & Svensson, 2000). Moreover, age and ability may also influence self-assessments. Influences at the system level may come from the teacher and the school, as well as from the school-type, making self-assessment a multilevel problem.

Previous research demonstrates that pupil self-assessments vary between gender (Swalander, 2006), between pupils with different home background characteristics (Kuncel, Crede, & Thomas, 2005; Swalander & Taube, 2007), and between school-subjects (Marsh, 1986). The differences between subject domains may be a cause of different frames of references to the own achievement among pupils. Marsh (1986) found that pupils who achieved better in one school-subject (i.e., better in maths than in Swedish) tended to overestimate their achievement in their "best" subject and underestimate their achievement in their "weaker" subject; the pupils estimations of two different self-concepts being uncorrelated. The results of this model have been repeated with similar results (e.g., Brunner, Lüdtke, & Trautwein, 2008).

Studies of gender differences in general academic self-concepts have, however, often yielded inconclusive results (Skaalvik, 1997). Fredriksson, et al. (2011) did not find any significant gender differences for pupils in 3rd grade. Swalander (2006) reported higher general academic self-concept for boys, whereas girls had higher verbal self-concept. Reuterberg & Svensson (2006) showed that boys tended to overestimate their mathematics skills, as did low achieving pupils with disadvantaged backgrounds. While pupils with high abilities estimated their achievement reasonably well, Kuncel et al. (2005) found less accuracy of the self-assessments of low ability pupils. Older pupils have shown better accuracy in their estimations than younger pupils (Butler & Lee, 2006; Fredriksson et al., 2011). However, in this matter results are inconclusive. For example, Sperling, Howard, Miller, and Murphy (2001) investigated the agreement between self-assessment and teacher judgements in grades 3-9, finding the relationship to be lower in the older populations. One cause for the divergent results is that the criteria, as well as the methods for self-assessments, are different for pupils of different ages and for different subjects.

A limitation of many studies on teacher judgements and pupil self-assessments is that they often rely on only one type of analysis for validation—mainly correlation between teacher judgements and an external criterion (Hoge & Coladarci, 1989; Südkamp, et al., 2012; The Swedish National Agency for Education, 2007, 2009), or content investigations on the alignment between the test content and the national standards. These kind of studies have for example been conducted by the Swedish National Agency for Education (2006, 2007, 2010). Even though the information gathered by both approaches is important, when considered individually the information they provide is limited. Validation requires research that relies on multiple sources of evidence. As Kane (2006) pointed out:

> Individual studies in a validity argument may focus on statistical analyses, content analyses, or relationships to criteria, but the validity argument as a whole requires the integration of different kinds of evidence from different sources" (p. 23).

This is particularly important when assessments are used for more than one purpose.

Chapter Five: Methodology

One of the most crucial aspects of construct validity is how different concepts are measured. Construct validation is concerned with validity of inferences about unobserved variables, (the constructs) on the basis of observed variables (their presumed indicators) (Pedhazur & Pedhazur, 1991). The capacity of the indicators is thus of particular importance for the quality of the measurement of a construct. The operationalization of the concepts used in this thesis are thus of great importance and will be described in detail, especially the three measures of achievement.

Data

The empirical work in this thesis is based on data from the PIRLS (Progress in International Reading Literacy Study) 2001 study, performed by the IEA (International Association for the Evaluation of Educational Achievement). Sweden participated with two samples, one from grade 3 and one from grade 4. A total of 35 countries participated in PIRLS 2001. The studies in the current thesis draw exclusively on the Swedish data. The number of participating schools, teachers and pupils is presented in Table 1 below.

Table 1. Valid N – Schools, Teachers and Pupils[6]

	Grade 3	Grade 4	Total
Schools	144	146	290
Teachers	351	344	695
Pupils	5271	6044	11315
Girls	2631	2965	5596
Boys	2640	3079	5719

Beside pupils' reading literacy skills, PIRLS also provides extensive information about the school context and the home environment of the pupils.

[6] Source: Rosén, et al. 2005, p 32.

Variables

Pupils, their parents and teachers have provided information on a large number of questions. Information about the questionnaires and variables are available in the Swedish PIRLS report (Rosén, et al., 2005), the international user-guide (Gonzalez & Kennedy, 2003), as well as in the technical report (Martin, Mullis, & Kennedy, 2003). The teacher judgement items, the reading test, the pupils' self-assessment items and home background variables, and the teacher background variables selected for the studies in this thesis will be described below.

Teacher judgements

The 2001 Swedish database included a national extension, a questionnaire, in which teachers were asked to assess pupil language skills on a number of aspects. This questionnaire was developed from the national diagnostic material (*Språket lyfter*). The diagnostic material contained observation aspects used by teachers for assessing and monitoring pupil knowledge and skills in the Swedish language domain. Some adjustments of the observation form were however required to be feasible for large scale comparative purposes. Instead of teachers' written comments for each pupil, they were asked to rate pupil language achievement on a 10-point scale. The original diagnostic instrument included 18 aspects of Swedish language skills, which in this adjusted version were rephrased into 18 different statements teachers had to consider. The original instrument is available in the Swedish national report of PIRLS 2001 (Rosén et al., 2005). Selected as indicators of teacher judgements were those items relating to either aspects of reading or writing, eight about reading and four about writing. The latter was warranted by the fact that some of the PIRLS test items also required a certain amount of written responses. These teacher judgement items were used in all four studies of this thesis. The rating items and descriptive statistics are presented below.

Table 2. Descriptive statistics for the 12 items of the teacher judgement scale

Variable	Question/Statement	Grade 3			Grade 4		
	Pupil can…	N	Mean	SD	N	Mean	SD
01	Construct sentences correctly	5208	7.67	2.16	5856	7.47	2.25
02	Recognize frequently used words in an unknown text	5213	8.35	1.93	5855	8.05	1.99
03	Connect a told story with an experience	5162	8.26	1.85	5840	8.01	1.93
04	Use the context to understand a written text	5207	8.05	2.05	5812	7.78	2.15
05	Write a text continuously fluently	5209	7.84	2.18	5860	7.66	2.22
06	Understand the meaning of a text when reading	5124	8.30	2.00	5767	8.08	2.08
07	Recognize the letter/connect sound	5136	9.48	1.27	5779	9.25	1.46
08	Read unknown words	5133	8.11	2.03	5778	7.85	2.11
09	Reflect on a written story	5083	8.09	1.90	5768	7.88	1.98
10	Read fluently	5135	8.32	2.10	5777	8.36	2.11
11	Improve own written text	5072	7.11	2.24	5766	6.96	2.31
12	Use a reasonably large vocabulary	5132	8.30	1.89	5774	8.06	1.98

It may be noted that most items have high mean values, which indicates that teachers consider pupils to be on average good readers. The statement: "pupils can recognize the letter/connect sound" was rated highest by the teachers with means well above nine. To "improve own written text" was regarded as more challenging for the pupils, which seems reasonable. The assessment instrument described has been used to indicate an overarching latent construct; *teacher judgements of Swedish language skills*. The properties of the single items are therefore of secondary significance.

The reason for labelling the variable derived from these items 'teacher judgements' and not 'teacher rating' was because 'rating' refers to single estimations rather than a global measure, which the items will ultimately form. Thus, when clustered and modelled into a latent variable, the term 'judgement' was preferred over 'rating'.

Reading test

The concept of reading literacy was coined by the IEA when launching the 1991 reading literacy study. The definition of reading literacy includes a description what it means to be an able reader. A central part in the definition of reading literacy is to understand and use different written forms, both in order to learn, to be a functioning member of the society and to be able to read texts for one's own enjoyment. The definition of reading literacy in the PIRLS 2001 framework was formulated as follows:

> …the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers, and for enjoyment (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001, p.3).

This approach to reading includes a number of theories on reading literacy, which is conceived as a constructive and interactive process in which readers actively construct meaning and assumed knowledge. Readers have a positive attitude to reading and they read both for recreational purposes and to retrieve information. Further, reading experience can be seen as constructing a sense of interaction between the reader and the text. The reader has a number of skills, cognitive and metacognitive strategies, and knowledge of the background. The context in reading situations promotes commitment and motivation to read, and there are often specific requirements on the reader.

PIRLS examines three aspects of reading literacy, namely 1) reading comprehension processes, 2) purposes of reading and behaviour, and 3) reading habits and attitudes. While the first two aspects together form the basis for the written test, the third aspect is addressed in the background questionnaires.
The reading test results in PIRLS 2001 were used in the studies included in this thesis. The item pool comprised 98 tasks, where pupils had to read various texts and answer questions, both multiple choice and open-ended.

The PIRLS design uses a matrix sampling technique, which implies that not all the questions were answered by every pupil. Pupils' scores on each booklet must be combined on a common scale for an overall picture of the results in each country. The item response theory (IRT) scaling procedure yielded five imputed scores or plausible values for each pupil's reading literacy skills. Because each pupil responded only to a subset of the assessment item pool, the generated scores were not sufficiently reliable for reporting the results. The plausible value proficiency estimates were used to obtain reliable scores (Martin, et al., 2003).

Furthermore, the derived international scale has a mean value of 500 with a standard deviation of 100. The Swedish pupil mean score in grade 4 was about 563 points and in grade 3 about 521 points with SD of 63 and 71 respectively (Mullis, Martin, Gonzalez, & Kennedy).

Pupil variables

From the pupil questionnaire, information about pupil reading ability was available from their self-assessments. The items are presented below, along with descriptive statistics for grade 3 pupils.

Table 3. Variables defining pupil self-assessments

Variable	Question/Statement	N	Mea	SD
Self_assess1	Reading is very easy for me	5138	3.45	0.64
Self_assess2	I do not read as well as other pupils in my class	5121	3.02	1.01
Self_assess3	I understand almost everything I read, when I read on my own	5128	3.49	0.69
Self_assess4	To read aloud is very hard for me	5138	3.06	1.00

The self-assessment items consider how well pupils estimate their own skills, both with and without reference to other pupils. The four statements concern reading skills in general, rather than in relation to specific aspects in the subject Swedish. The rating scale goes from agree a lot (1) – disagree a lot (4). Items Self_assess1 and Self_assess3 were recoded to get same direction of the scale as the two other variables. The mean values are quite high for all variables, indicating that the 3rd grade pupils assess themselves to be able readers. Compared to the teachers' assessments, pupil self-assessments are broader, which is also reasonable given that younger pupils are not expected to evaluate their knowledge on the basis of the more complex statements to which the teachers responded. The self-assessment items were used in Study IV.

Selected from the pupil and parent questionnaire were also items indicating gender (Girl=1, Boy=0), socioeconomic status, and pupil attitudes towards reading. In the present thesis, 5 indicators of socioeconomic status were used. The SES indicators were available in the parent questionnaire. The indicators were chosen based on the suggestions of previous research (Sirin, 2005). The variables are presented in more detail in Table 4 below.

Table 4. Description of the SES indicators included in the analyses

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
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.
Annual income	Within which span are your household's annual income. Ordinal variable - 1-6: Less than 180 000sek, 180 000 – 269 999sek, 270 000-359 999sek, 360 000-449 999sek, 450 000-539 999sek, 540 000sek or more
Highest Education	Highest educational level in the home. Ordinal variable - 1-8: Some compulsory school, completed compulsory school, 2 years of upper secondary education, three years of upper secondary education, post-secondary education, 2 years of university studies, University studies – candidate level, University studies – Master level.
Highest Occupational level	Highest occupational level in the home. Ordinal variable - 1-3: Blue collar, white collar, academic.

The SES indicators form a latent variable, which was used in Study I, II and IV. The attitude variables were used in Study IV and form a latent variable consisting of five indicators.

Table 5. Pupil attitudes toward reading

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)
Attitude2	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)
Attitude3	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)
Attitude4	I need to read well for my future. Ordinal variable. Four alternatives: Agree a lot (1), Agree a little (2), Disagree a little (3), Disagree a lot (4)
Attitude5	I enjoy reading. Ordinal variable. Four alternatives: Agree a lot (1), Agree a little (2), Disagree a little (3), Disagree a lot (4)

The attitude items indicated relatively high attitudes with the mean-values all well above the midpoint of the scale. Descriptive statistics are available in Study IV.

Teacher background variables

Four indicators of the concept of teacher competence were selected. These items were available in the teacher questionnaire. Selected items included information about teachers' education and experience, whether or not teachers were certified, and the extent to which reading pedagogy was a focus of their formal training. These variables were used as indicators of the latent variable *teacher competence*. The role of teacher competence for teachers' judgements was studied in Study II and III.

The effects of teacher competence are more likely to be revealed when they have taught their pupils for a period of time. Therefore, the effect of teacher competence was studied for the 3rd grade sample only. About 70% of the classrooms had one and the same teacher during the first three years of school education. Some 20% of the grade 4 classrooms had had their teacher for more than one year.

More detailed information and descriptive statistics on the items presented in the current sections can be found in the respective studies.

Methods of analysis

In the educational sciences, concepts are often rather abstract and cannot be assigned a quantity in the way that a direct observation or measurement might.

For example, instead of observing one aspect of reading achievement, such as reading fluency, several indicators of the construct must be used. Via appropriate indicators it is possible to operationalize theoretical constructs; however, theoretical constructs are always simplifications of the 'real world', something that is important to bear in mind when complex phenomena are studied.

Latent variable modeling

Most concepts in the present thesis have been operationalized using latent variables and the relations have been analysed using Confirmatory Factor Analysis (CFA) and Structural Equation Modeling (SEM). CFA concerns *measurement models* that regard the relationships between the observed indicators (e.g., test-items, observational ratings) and the latent variable (e.g., motivation, attitudes). When measurement models are related to each other, the model becomes a structural model, which specifies relations between latent variables. The analytical tool used for this is SEM.

Given that a model fits the data, an advantage with latent variables is that in general they better represent the researchers' theoretical frame of reference. The most powerful advantage with latent variables, however, is that measurement error is accounted for and it is for this reason that latent variables are said to be free from measurement error (Gustafsson, 2009). Directly observable, or manifest variables generally contain measurement error. All observed variables are error-laden and the error indicates that there is unexplained variance in the manifest variable, which cannot be explained by the latent variable. The error in the indicator can be either random or systematic. The important point however is that the latent variable will not be affected with the errors from the indicators, since the 'latent' part has been separated from the 'error' part in the latent variable model.

Results of Confirmatory Factor Analysis (CFA) and Structural Equation Modeling (SEM) can provide evidence of the convergent and discriminant validity of theoretical constructs (Brown, 2006). Convergent validity is indicated by evidence that different indicators of theoretically similar constructs are interrelated. For example, indicators of attitudes towards reading load on the same factor. Discriminant validity is indicated by results showing that indicators of theoretically distinct construct are not highly inter-correlated. This implies that different attributes load on separate factors. These factors are not highly correlated as to indicate that a broader construct incorrectly has been separated into two or more factors. One way to strengthen reliability can thus be to use more indicators of the same construct. What can be problematic is that the

indicators must reflect the same construct and the more indicators used, the greater the risk of measuring something other than the desired construct.

CFA and SEM models are often presented graphically. Usually, latent variables are depicted by circles or ellipses and manifest variables by squares or rectangles. The relations are drawn by arrows, pointing in the direction of the dependent variable. Double-headed curved arrows indicate a covariance. The measurement errors are drawn by a short arrow, sometimes with a small circle attached to it. The relations between the latent variable and the manifest indicators are expressed by so-called factor loadings, which indicate the strength of the relationship between the indicator and the latent variable. A high factor loading implies that most of the variance in the indicator is captured by the latent variable. If a factor loading is low, the manifest variable does not contribute very much to the latent variable, either because it is a bad indicator of the construct, or because it is heavily error-laden. However, small factor loadings may also depend on measurement level of the indicator. A dichotomous variable or an ordinal scale with few steps holds much less information than, for example, a total score on the test. The construct of interest also influences the loading. The important point is that the latent variable absorbs the information in the indicator. Figure 2 displays two measurement models; 1) Pupil reading achievement rated by the teachers, and 2) pupil attitudes towards reading, which together form a structural model.

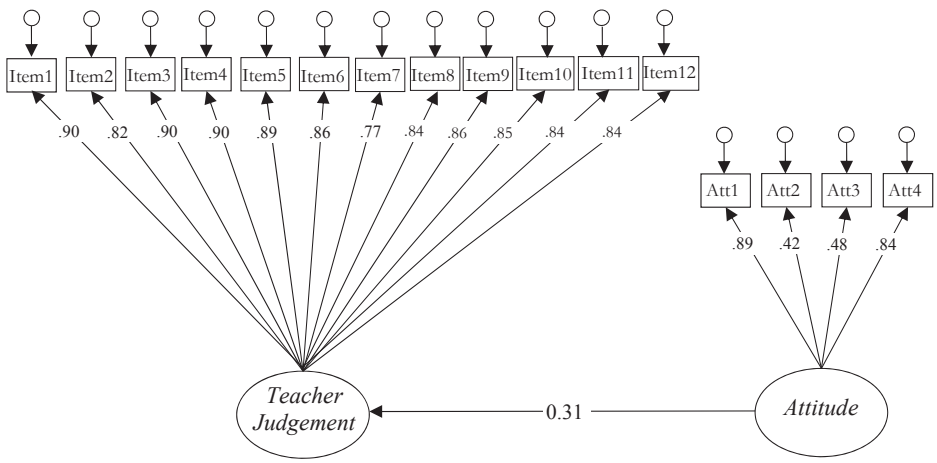


Figure 2. Model of two latent factors and the relation between them.

The measurement model specifies the relations between a latent variable and its indicators. In Figure 2, most factor loadings are high, especially for the *Teacher Judgement* variable, indicating that the latent teacher judgement variable accounts for a large proportion of the variance in the observed items.

The relation between the latent variables *Teacher Judgement* and *Attitude* depicts a relationship between the dependent variable, teacher judgements, and the independent variable, student reading attitudes. The standardized regression coefficient indicates the effect of attitudes on teacher judgements. The result shows that the higher attitude, the higher judgements from the teacher, which is reasonable since attitude is often highly correlated with student achievement.

Multilevel modeling

In the social and behavioural sciences, many phenomena that are of interest are nested within a hierarchical structure. Examples of hierarchies may be pupils nested within classrooms, classrooms nested within schools, patients nested within hospitals or workers nested within companies. The lowest-level of measurement is at the micro level and all higher-level measurements are at the macro level. Traditional statistical analyses assume that the observations are independent of each other. The assumption of independence implies that subjects' responses are not correlated with each other, and this may be the case when the data is drawn from a simple random sample from a large population. However, when people are clustered within naturally occurring units, such as schools and classrooms, the responses of people from the same cluster are likely to have more in common. With clustered data, traditional statistical analyses that assume independence will produce standard errors that are too small, which is incorrect. With standard errors that are too small, incorrect rejections of the null-hypothesis (Type I error) can be made (McCoach, 2010). In multilevel analysis, the degree of relatedness of observations is explicitly estimated within the same cluster, thereby correctly estimating the standard errors and eliminating the problem of inflated type I error rates.

The advantages with multilevel modeling are not only statistical. Multilevel analyses also allow advantage to be taken of the information in cluster samples to explain both the between- and within-cluster variability of an outcome variable of interest. The more advanced models allow for the use of predictors at both the individual level and the organizational level (classroom) to explain the variance in the dependent variable. It is also possible to test whether the relation between an independent variable and a dependent variable varies significantly across clusters. If the impact of an independent variable on the dependent variable varies across clusters, it is possible to attempt to explain the variability in this relation by using cluster level variables.

One way to deal with multilevel data is to aggregate or disaggregate variables to a common level. This, however, causes problems with loss of statistical power

and the aggregation and disaggregation may overlook the interdependency and interaction effects across levels. Multilevel structural equation modeling can help to realize conceptual and statistical demands. In this thesis, multilevel modeling has been applied in order to take account for the dependencies among pupils and between classrooms.

To investigate the variability in, for example, achievement across classrooms, variability in teachers' assessments, or other differences between macro units, two-level modeling can be applied. The basic principle of two-level modeling is to decompose the total variance into one between-group component and one within group component. The proportion of between class variance is a measure of the amount of similarity within groups, and this measure corresponds to the intraclass correlation. The intraclass correlation (ICC) is a measure of the degree of dependence between individuals and can be used to find out whether or not a multilevel framework is needed. The intraclass correlation can also be regarded as a measure of group homogeneity. The more individuals that share common experiences due to proximity and/or time, the more they tend to have in common. A high degree of dependence can, for example, be found for children born and raised in the same family. If the ICC is low, groups are only slightly different from each other; if it is zero, no group differences exist at all for the variables of interest (Kreft & De Leeuw, 1998).

In the example in Figure 3 below, a two-level measurement model of teacher judgements is presented. While one step to enhance the validity is to formulate the model at two levels, another involves parcelling the items included in the teacher judgement construct. One of the empirical advantages of parcels relatively to items can be the psychometric merits (see for example: Bagozzi & Heatherton, 1994; Hau & Marsh, 2004; Kishton & Widaman, 1994; Little, Cunningham, & Shahar, 2002). Advocates of parcels also argue that parcels are to be preferred in that fewer parameters are needed to define a construct. Due to the enhanced measurement properties the overall model fit usually becomes more acceptable when parcels are modelled, compared to when single items are used. By parcelling items and formulating the models at two-levels, an improved model fit, and thus a theoretically sounder model, have been obtained.

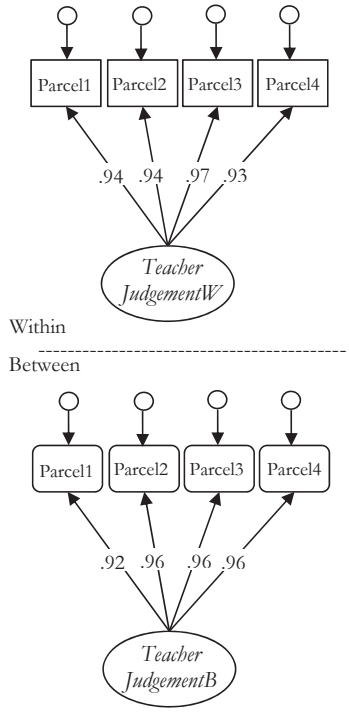


Figure 3. Two-level model of teachers' judgements of pupil reading achievement.

Figure 3 is divided into two parts; a within part and a between part. The parcel items at the between level are not directly observable as they are average values of the teacher judgements in a classroom. Thus the 'parcel-boxes' do not have sharp edges. There are several ways of displaying two-level models and there is no common standard. In this example, the appearance of the figure is guided by the models drawn in the user's guide of Mplus, the software used for estimating the latent variable models (Muthén and Muthén, 1997-2012).

Random slope modeling

In order to examine whether a within class relation between two variables varies across classrooms, multilevel models with random slopes (also known as varying slopes) can be used (Brown, 2006; Hox, 2002). The assumption to be tested with random slope modeling is whether a within-level relationship between a dependent variable Y and an independent variable X varies significantly across clusters. For example, the relation between teachers' judgements (Y) and pupil reading achievement (X) is assumed to vary across classrooms. If the correspondence between judgements and test results varies between different classrooms a random slope variable can be formulated at the between part of the

model. It is possible to relate variables at teacher-level to the random slope variable and thereby investigate effects of the teacher on relationships within classrooms (see also Hox, 2002).

If there is too little variation in the relationship between judgements and achievement (in most classrooms the relation is about .60) a random slope variable cannot be formulated.

The question addressed in Study III was whether teacher competence can moderate the relationship between judgements and achievement, that is to say whether highly competent teachers have higher agreement between their judgements and test-scores than their less competent counterparts. Figure 4 shows a random slope model.

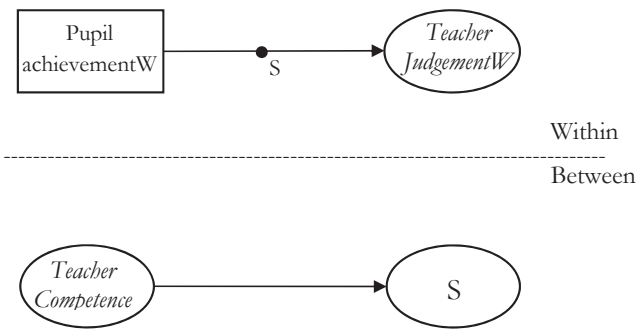


Figure 4. The relationships in a random slope model.

In this model 'S' depicts the relation between pupil test results and teacher judgements. The within relationship is assumed to vary between classrooms, which is depicted as the 'S' ellipsis at the between part of the model. For example, covariates such as teacher competence (Z) can be introduced at the between level part to investigate whether or not they have an influence on 'S'. The "varying slope" defines the relationship between teacher judgements and pupil achievement. If teaching competence is positively related to the slope the effect of pupil achievements on teachers' judgements is stronger for the more experienced teachers. This implies that for teachers with higher competence levels there is a higher correspondence between their judgements and pupil achievement. Conversely, if the effect of competence is small, for more competent teachers the pupil achievement effect will be smaller.

Assessing model fit

An important part of the validation procedure involves determining whether the hypothesized model fits the data. A model will fit the data if the observed covariance matrices can be reproduced from the model (Brown, 2006). Several goodness of fit indices indicate whether or not the model fits the data.

It has to be noted that, for model fit, there are no golden rules or definitive cut-off values, meaning that care should be taken not to reject a model without careful examination (Bentler; 2007; Goffin, 2007; Markland, 2007;). Theoretical considerations are helpful in directing the modification of an initial model, as are statistical tests. However, there are a number of guidelines usually followed by a majority of researchers.

Given the many possible goodness of fit indices available, the usual advice is to assess model fit by inspecting a number of fit indices that derive from different principles (Hox, 2002). In the current thesis the most commonly used indices were adopted. In addition to the use of the χ^2 goodness-of-fit test, because χ^2 is sensitive to sample-size and with large samples nearly always delivers a significant value, it was combined with three other fit indices. The RMSEA (root mean square error of approximation) takes account of both the number of observations and the number of free parameters. The cut-off value for the RMSEA has been suggested to be 0.08, while a value of 0.05 or below indicates a close fit (Loehlin, 2004; Marsh, Hau, & Wen, 2004). The CFI (Comparative Fit Index) is a fit index that depends on the average size of the correlations in the data. The CFI value should be as close as possible to 1.0, with values below 0.95 usually not considered as being satisfactory (Hu & Bentler, 1999). The SRMR (Standardized Root Mean Square Residual), a measure of residuals compared separately for within and between levels, was also used. For the SRMR values of 0.08 or lower are needed for the model to be accepted.

Missing data

Missing information on different variables is a common problem in educational research. A general problem with missing data is that it may affect a study's external validity. This implies that results that are built on a subset of the observations might not correspond to results that would have been obtained had all the observations been included in the analyses. Another problem is loss of statistical power, which implies that effects in the population cannot be registered because the sample is too small to give significant results. Missing data can either be at random or systematic. If data is missing completely independent of all the observed variables it is denoted MCAR (missing completely at

random), while if there is a probability distribution (i.e. the distribution of missingness depends on the observed part), it is called MAR (missing at random). If missing is systematic it is denoted MNAR (missing not at random) (Schafer & Graham, 2002).

There are several approaches to handling missing data. For example, listwise deletion removes cases with missing data on any variable. Although easy to use it often results in the loss of a considerable proportion of the original sample and, in turn, statistical power. The most widely preferred methods to handle missing data in SEM applications is maximum likelihood (ML) estimation and multiple imputation (Brown, 2006). These approaches make use of all the available data. ML has been regarded as the best method for handling missing data in SEM applications. Instead of excluding observations, ML uses an estimation method which uses all available information from all observations to handle the missing data.

Analytical stages

The purpose of this overarching discussion is to provide an overall picture of the validity of teacher judgements, pupil-self assessments and external test results. Furthermore, factors that could in particular influence the teacher judgements will be investigated. Sound validation requires a logical and systematic procedure. The Toulmin model (1958/2003) previously presented in the theory section is one way to structure the analytical steps. In order to justify a claim, i.e., that teacher judgements correspond to pupil actual level of performance, data must be provided. Such data might, for example, be statements about pupil achievement (e.g., grades). It needs to be justified, or 'warranted', that the 'relevant' knowledge and skills have been assessed. Rebuttals address threats to a valid claim. Rebuttals may be many and can be explored by relevant research questions or by hypotheses. Support can be provided by evidence suggesting that the threat, expressed in terms of a rebuttal, is false.

The structure of arguments

For the sake of simplicity, only the analytical steps involved in validating teacher judgements are presented. A diagram of the structure of arguments is set out in Figure 5.

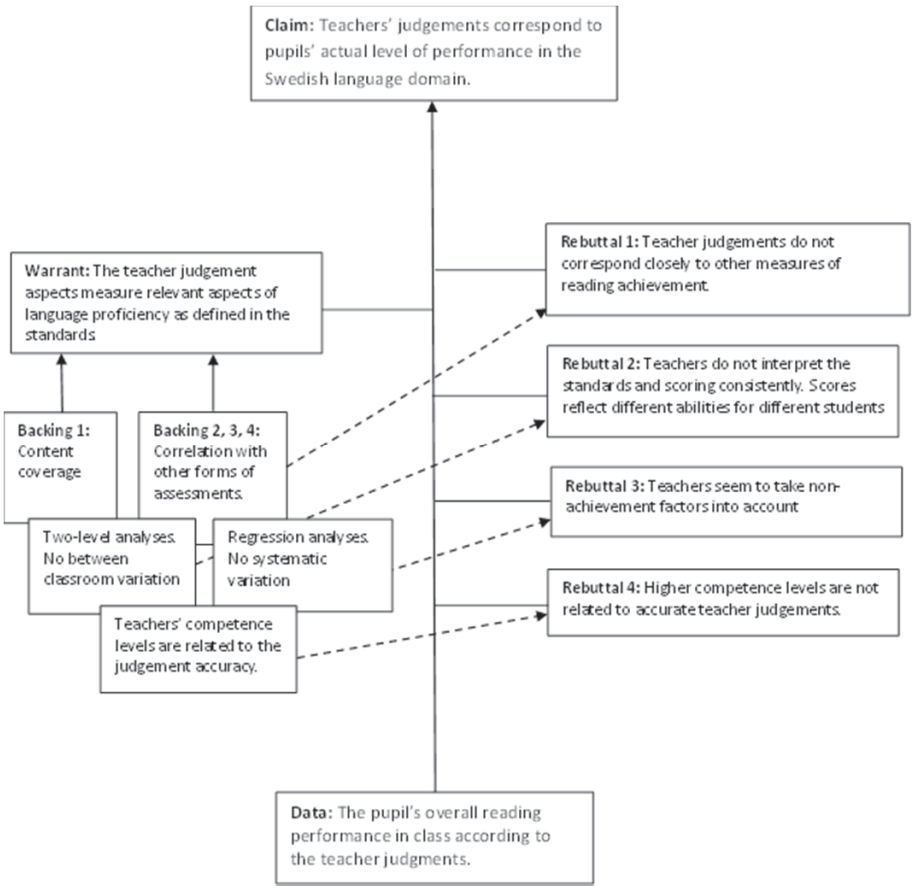


Figure 5. Analytical steps when validating inferences of teacher judgements.

The arrow from the data to the claim represents an inference that is justified on the basis of the warrant. The **claim** here is that teachers' judgements correspond to the pupil actual knowledge in the Swedish language domain. The data is "the pupil's overall language performance in the classroom according to teachers' judgements. In this case, the warrant that justifies the inference from the data to the claim is the following:

Warrant: The observational aspects in the assessment material measure relevant literacy aspects, stated in the syllabus. The teacher judgements should thus include the appropriate knowledge and skills. If the data consisted of a pupil's performance on the PIRLS test, the warrant could be that the test is well aligned to the Swedish syllabus and that it measures reading and writing aspects which are critical for pupils in the primary school years. If the data consisted of pupil self-assessments of reading achievement, the warrant might be that the

assessment statements strongly relate to reading achievement and that it is possible for primary school children to answer them reliably.

In addition to the warrant that justifies the inference from the data to the claim, a validity argument includes rebuttals or alternative explanations that might account for the observed performance and scores on an assessment. Rebuttals weaken the validity of the intended inferences. Specific rebuttals must be articulated and investigated to gather evidence about how teacher judgements might have affected the score-based interpretations. In this thesis, in the case of the teacher judgements, four rebuttals are investigated:

Rebuttal 1: Teacher judgements do not correspond closely to other measures of reading achievement.

Rebuttal 2: Teachers do not rate pupil performances consistently, and as a result, the judgements reflect different abilities for different pupils. Thus, the teacher judgements reflect differences among teachers' interpretations instead of individual pupil performances.

Rebuttal 3: Teachers seem to take non-achievement factors into account when assessing their pupils. Thus, the judgements do not reflect achievement only, but also, for example, take into account gender and SES.

Rebuttal 4: Teacher judgements do not seem to be related to competence. Teachers with higher levels of competence do not assess pupil achievement more accurately than others.

To weaken or eliminate the rebuttals and to support the warrants, adequate analyses are required. In this thesis support is provided by means of the following analytical steps:

Backing 1 – Content coverage: The teacher judgement aspects in the diagnostic material should be aligned with the syllabus.

Backing 2 – Correspondence: The teacher judgements should correlate significantly with measures of the same ability. Evidence of correspondence could be collected by examining the relationship between the teacher judgements and the PIRLS test.

Backing 3 – Two-level modeling: Small between class effects indicate that teacher judgements work similarly for different teachers, i.e., teacher judgements can be used for comparison purposes across classes and schools.

Backing 4 – Non-random external influences: Little or no systematic variation should occur when gender and SES are included in the analyses as independent variables.

Backing 5 – Random slope analyses: The correspondence between judgement and achievement should be higher for teachers with higher competence levels.

Chapter Six: Results and Discussion

In this section the findings of the research are summarized and discussed. The results of the four studies are presented together. The results are presented in accordance with the research questions and the analytical steps outlined in the argumentation model presented in the previous chapter.

In order to examine validity issues in teacher judgements, the reading test results in the PIRLS 2001 study were used as a criterion. If the validity of the test-results was in focus, the teacher judgements could be used as a criterion for the test. As indicated by previous research, standardised test results have been used to explore issues of validity in teacher judgements. However, the teacher judgements should also be possible to use as a criterion for the PIRLS test.

Validating teacher judgements for use within classrooms and for classroom comparisons

Different sources warrant that the teacher judgement and the PIRLS test are valid measures of achievement. For example, the test has been suggested to measure relevant aspects of the curriculum (the Swedish National Agency for Education, 2006). The time spent with pupils, observing and documenting their knowledge and skills gives credibility to the teachers' judgements (Gipps, 1994). However, mutual investigation of the validity can also be made statistically by investigating the correspondence between the two measures.

Assessment within classroom

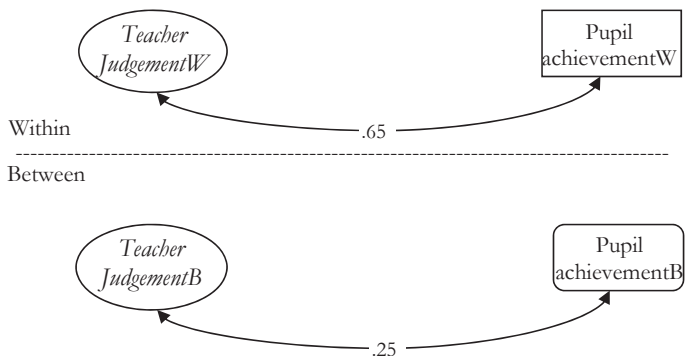
The first step of the analysis was to address the rebuttal stating that teacher judgements did not correspond to other measures of the same construct. The same rebuttal can be specified for the PIRLS test results and the self-assessments. By correlating the teacher judgements and the PIRLS test results, the first step involves the mutual investigation of the validity of inference of teacher judgements and external tests.

The relationship between teacher judgements and pupil test scores on reading achievement was the main focus for the analysis in Study I, which also investigated this relationship in grade 3 and grade 4.

Initially, a latent model for teacher judgements was defined. This was carried out in order to get rid of the low reliability of a single indicator approach.

Further, it extrapolated the single ratings to an overall judgement about pupil language ability. As the intraclass correlation indicated high variability in the judgements between classrooms, the model was formulated in a two-level model where the total variance was decomposed into within and between group variances. Initially, a two factor model separating reading and writing was fitted to the data. As the two factors were very highly correlated the model was reformulated with the 12 manifest variables. Thus, one single latent factor defines teacher judgement of overall reading and writing ability. Even though this step improved model fit, a close fit was not obtained. Therefore, the model was modified. In order to improve the psychometric properties of the model, items were parcelled. The 12 items were randomly assigned into four parcels of three items in each. Since the focal interest was not at item-level, it seemed theoretically sound to parcel the items in order to obtain a global construct of the teacher judgements. This model was accepted and used in all four studies.

The next step was to relate pupil achievement on PIRLS 2001 to the teacher judgement factor at both the within and between classroom level. Figure 6 shows a model of the analysis. The relationship at the within classroom level investigated the correspondence between the teacher judgements and the pupil achievement. The correlation was about .65 in grade 3 and about .60 in grade 4. These results accord reasonably well with other correlational studies of the relationship between teacher judgments and pupil achievement (Hoge & Coladarci, 1989; Meisels et al., 2001; Südkamp, et al., 2012).



Note. P<.05 unless otherwise stated. Model fit: χ^2=169.45 p=000 *df*=10 CFI=.99 RMSEA=.06 SRMRw=.01 SRMRb=.04

Figure 6. Model of the relationship between teacher judgement and pupil test results for PIRLS 2001.

The results showed that, for teachers in the 3rd grade, there was a higher association between their judgements and pupil achievement. In order to investigate whether the difference between grades was statistically significant, a multiple group model was formulated. The difference between grades was found to be statistically significant according to the chi-square test for nested models. The differences between grade 3 and 4 may be due to the different amount of time spent with the pupils, different types of teacher education vis-à-vis the different grades, but also different expectations of pupil knowledge and skills.

The results also indicate that the inferences based on the external test should be considered as valid. Speaking in Toulmin's (2003) terms, the correlational analysis provided support for both teacher judgements and PIRLS test results. The rebuttals stating that the teacher judgements/external test results were not valid measures of achievement were thereby weakened.

Despite the substantial correlation, the part of the variance in the teacher judgements unaccounted for was nevertheless larger than the part explained by the test results. However, high correlations are, due to several reasons, hard to obtain. First, neither of the assessments are perfect measures of achievement. The PIRLS results have been obtained from one single measurement, and give a snapshot of what pupil can do. Differences in administration of the test and pupil motivation taking the test can have affected the reliability of the test results. Second, the teacher judgements could contain non-achievement factors, as has been suggested by previous research (e.g., Brookhart, 2012). This relates to other rebuttals, which may be that factors other than achievement influence teacher judgements. Possible influences of such factors were investigated in a second set of analyses. First, however, attention is paid to the between level part of the model displayed in Figure 6.

Classroom comparisons

While the previous section concerned analyses at the individual level, the current section focuses on system-level analyses. The unit of analysis here is the classroom level, which means that the variables—judgements, test-results and so forth—are averaged measures of the individual observations. The variation existing within and between classrooms is separated so that it becomes possible to simultaneously conduct analyses at the different levels. Two-level modeling is a powerful tool when differences, for example in achievement, are attributed to different levels. If the school-system is largely homogenous, there is little variation in achievement levels between schools or classrooms. If the level of performance differs substantially between classrooms, it is useful to adopt two-

level modeling. The results, at a macro level, tend also to be more stable since individual observations are more attenuated with measurement error than is an aggregated observation.

At the classroom level a rebuttal is directed to how the different measures work for classroom comparisons. To provide support for the contention that teacher judgement would be similar for similar achievement levels across classrooms, attention was paid to the between-level part of the model. In the model pupils' average achievement were correlated with teacher average judgements (presented in Figure 6). The analyses revealed modest correlations; about .25 in grade 3 and about .18 in grade 4. Thus, the teachers' average judgements did not correspond well with the classrooms' average test scores. In other words, these results indicate that low-achieving classrooms may have higher average judgement than high-achieving classrooms and vice versa. Similar signs of mismatch between teachers' grading and national test results have been noted in the final year of compulsory school and in upper secondary school by the Swedish National Agency of Education (2007, 2009).

Strong validity claims about the usefulness of teacher judgements for comparisons across classrooms was not warranted by this analysis. Thus, no support was found that could weaken this rebuttal. However, there were some circumstances in the present study which may have made it difficult for equality of teacher judgements to be established.

Although each statement in the teacher rating instrument was anchored in the national syllabus no explicit criteria were attached to each scale-point. This probably made it more difficult to obtain consistent teacher judgements between classrooms. The conditions were quite similar to those which teachers face in their ordinary work. Since the reforms in the 1990s, Sweden has a highly decentralized system, meaning that each municipality is responsible for organizing and operating school services. A challenge for a decentralized school system is how to ensure equality and consistency in teacher assessment, especially since interpretations of criteria have been shown to vary between teachers and schools (Selghed, 2004; Tholin, 2006). At the time of the data collection, no explicit criteria were given to teachers in grades 3 and 4. In the most recent curriculum (Lgr 11) the levels of knowledge demands pupils are expected to attain are more explicitly stated. Although these changes in the curriculum warrants more equal assessment practices, it remains to be investigated if these specifications result in more consistent teacher judgements.

Additional ways to achieve greater uniformity in understanding and assessment between schools and teachers may be via moderation (Harlen, 2005;

Klenowski & Whyatt-Smith, 2010; Nusche, Halász, Looney, Santiago, & Shewbridge, 2011). The purpose of moderation is, via collegial discussion, to develop successful strategies for assessment as a means of obtaining a shared understanding of expectations of the pupil knowledge and skills and interpretations of standards. The results of Klenowski and Whyatt-Smith's study suggested different ways in which teachers discussed and interacted with one another to reach agreement about the quality of pupil performance. To interpret standards similarly between teachers is the key for achieving consistency of teacher judgement. Not only are clear criteria important for achieving consistency, time must be allowed for moderation. Nusche et al. conclude that in order to ensure that assessment of pupil knowledge and skills is reliable and fair, teachers' moderation practices need to be supported.

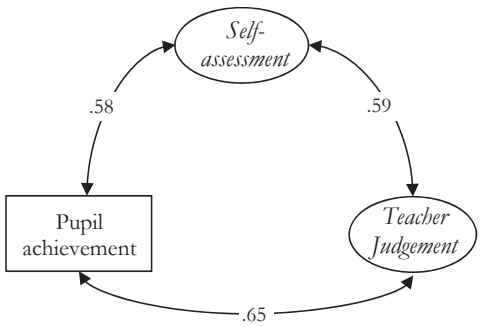
Pupil self-assessments in relation to other forms of assessment

The next analysis examined the correspondence between pupil self-assessments, teacher judgements and the PIRLS test results. This was also the primary purpose of Study IV. These analyses address an issue that relates to the accuracy of self-assessment. For self-assessments to be useful, they should correlate with teachers' judgements. Pupils interpret teachers' judgements and use the information to make decisions about their own performances (Black & Wiliam, 1998; Fredriksson et al., 2011). According to previous research (e.g., Ross, 2006), teacher judgements have most often been used as a criterion for pupil self-assessments, although comparisons to test scores also have been presented. Unlike many other studies, the present study compares pupil self-assessments to both teacher judgements and pupil test-scores.

The findings of the previous analysis (presented in Figure 6) showed that for teachers who taught their pupils for a longer period of time, there was a higher correspondence between their judgements and the test-results. Consequently, pupils who have had the same teacher for a longer period of time may be in a better position to assess their own knowledge and skills. Thus, for this reason, a grade 3 sample was selected.

The rebuttal to the credibility of pupil self-assessment is, primarily, that they do not correspond to their actual knowledge and skills. In order to explore whether pupil self-assessments were accurate predictions of reading achievement, they were correlated to teacher judgements and pupil test results. The model was set up at two levels. However the intraclass correlation (ICC) for the latent self-assessment variable indicated no between classroom variability.

The low ICC implies that the averages for self-assessments at the classroom level were relatively similar across classrooms. Consequently, analyses at a second level could not be conducted with these variables. In Figure 7, the model with the three forms of assessment is presented.



Note. P<.05 unless otherwise stated. Model fit: χ^2=280.20 P= .00 *df*=24 CFI= 0.99 RMSEA=0.05 SRMR=0.02

Figure 7. Model of the relationships between the three assessment forms.

The relationship between self-assessment and the other measures of achievement generated correlations close to .60. The relation to teacher judgements was .59 and to PIRLS result .58. As previously noted, the relationship between grade 3 teachers' judgements and pupil test results was .65 and, even though the relationship between pupil self-assessments and the two other measures was lower, it was nevertheless quite substantial. In relation to the IEA Reading Literacy study 1991, were similar analyses conducted. These analyses showed that, for a range of countries, the correlations amounted to .25-.55 (Elley, 1992). The results of the current research thus show that pupils in grade 3 were fairly capable of assessing their own knowledge and skills in the subject of Swedish, although only at the within classroom level.

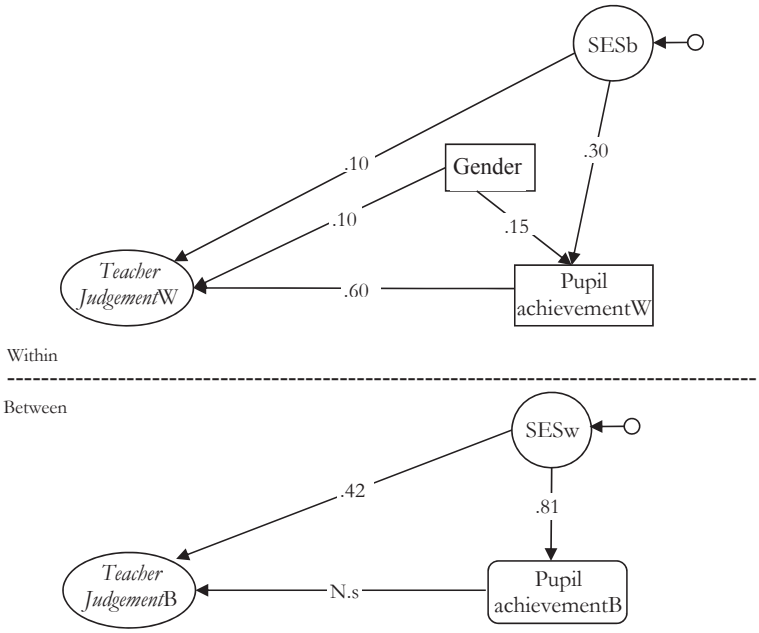
One use of self-assessment is for example to increase pupils' awareness of their own achievement level and thereby make them more responsible for their own learning. Hansson (2011) argues that the increased responsibility for own learning in school could be one cause of the trend of declining knowledge in mathematics. Her findings show that pupils from lower SES groups have most difficulties in adapting to the demand to take greater responsibility for their studies. Consequently, there may be reason to be cautious about increasing the use of self-assessment. In the current study, differences in the ways in which boys and girls and pupils from different socioeconomic status groups assess their

Swedish language abilities were therefore also investigated. In the following section, factors influencing both teacher judgements and pupil self-assessment are examined.

Factors influencing teacher judgements and pupil self-assessment

In order to further investigate rebuttals to the validity of teachers' judgements, pupil gender and socio-economic status (SES) were related to the judgement variable as independent explanatory variables. In this analysis, the correlation between teacher judgements and pupil test results was rephrased into a dependent-independent relationship, where teacher judgements were set as dependent of the test achievement. The reason to use pupil achievement as a criterion was that the primary focus of the analyses was on the validity of teacher judgements. If the PIRLS test results had been in focus, the effect of gender and SES on test-results could have been investigated when teachers' judgements had been controlled for. This step would investigate whether the test-results were associated with any gender and SES bias. However, validity issues in the test are typically investigated with other methods. The tests have been constructed so that the level of difficulty of the items should not depend on gender. Analyses of gender bias in the PIRLS test instrument have been conducted by Geske & Ozola (2009) who noted a large gender gap among Latvian pupils' achievement in PIRLS 2006. However, differential item functioning (DIF) analysis showed that the test instrument was not a cause for the gender gap in achievement.

A research question stated in the first of the four sub-studies of this thesis was whether teacher judgements and pupil achievement reflected factors other than actual achievement. The rebuttal here is thus that the judgements include other characteristics. SES could be addressed at both the within and between levels. Gender could only be included in the within part of the model, because there was no between class effect for the gender variable as the relative number of girls and boys was about the same in most classes. The model is presented in Figure 8 below.



Note. P<.01 unless otherwise stated. Model fit: χ^2=763.86 p=000 *df*=72 CFI=.97 RMSEA=.04 SRMRw=.03 SRMRb=.04

Figure 8. A two-level model of the influence of pupils' SES and gender on teacher judgements. Standardized regression coefficients for Grade 3.

The model is divided into two parts, the upper part highlighting the within classroom relationship and the lower part focusing on the relationship between classrooms.

Within classrooms, the relation between teacher judgements and test results decreased slightly when the other variables were introduced. Both gender and SES have a direct effect on achievement, implying that girls and pupils from homes with higher socioeconomic status have higher levels of achievement. This is in line with previous research on the matter (e.g., Rosén, 1998; Yang, 2003). The research question in focus at the within level was whether effects of student gender and SES can be estimated on teacher judgements when test-results are held at an equal level for all pupils.

Girls and pupils with higher SES had higher test-results, but they received higher teacher judgements when test-results were held equal for all pupils. The effects were statistically significant and not negligible; they indicate that girls and pupils with higher SES received higher ratings by the teacher which the test-results could not account for. However, some explanations for the results may be found in previous research.

That girls have been awarded higher subject grades than their test results can account for, has previously been shown in Swedish research (Klapp-Lekholm, 2008; Reuterberg & Svensson, 2000; Svensson, 1971). One interpretation has been that girls are awarded grades that are too high because they tend to conform to the teachers' wishes (e.g., Emanuelsson & Fischbein, 1986). However, another interpretation put forward has been that girls have performed ordinary course work in ways better than boys (e.g., Wernersson, 1989).

Following the steps of analysis that addresses the rebuttals, it must be concluded that there is no ringing endorsement for the contention that teachers do not take factors other than achievement into account. A variable measuring vocabulary or oral skills may be the mediating variable "behind" the effect of gender and SES. It can be hypothesized that girls and pupils with higher SES have more highly developed skills in these areas. Motivation, effort and behaviour may also be possible mediators. These factors are all crucial for continued study success. If teachers do take such aspects into account, it does not imply that "favoured" pupils would manage any less well in school. However, the principles of equality in education would be challenged. An interesting avenue for further research would therefore be to identify possible mediators that can explain teacher-judged achievement differences between boys and girls and between pupils of different SES.

Introducing SES at the between classroom level, it was found, when achievement was kept under control, that high SES classes received substantially higher judgements from their teachers. An interpretation of this result could be that high SES classrooms were rated in an unfair manner, too high relative to test results. However, this interpretation does not accord well, for example, with the results of Klapp Lekholm (2008) who found that, rather than teachers giving advantages to schools with a large share of high SES pupils, a compensatory grading strategy was in evidence. One plausible explanation for the result in the present thesis is that the SES variable holds more information at the between level than is the case for the test results. Teachers have at least some information about their pupils' SES, for example the school area or pupil intake. For the teacher, the achievement variable does not contain any information about the performance of one class relative to another since teachers are neither informed about their own pupil achievement, nor about the performance of other classes on the PIRLS test. The mediating SES variable did fully account for the effect of class average achievement and on average judgements, which however, was initially quite modest. The effect of average SES on average teacher judgement was about .40, which may be considered as quite low, given the large

achievement differences between different SES groups. This relationship was indicated by the direct effect of classroom SES on a class average test-result, which was about .80. About 20% of the variance in achievement differences was between classes, and these differences were thus mostly accounted for by pupil SES.

The influence of SES and gender on pupil self-assessment within classrooms

Differences in pupil self-assessments were investigated with respect to pupil gender, SES and attitudes in Study IV (not included in model above). If there are differences in how well different groups of pupils can assess their knowledge and skills, this may mean that it is problematic to use self-assessments in school. For example, if a pupil with lower SES inaccurately assesses his/her reading skills, the learning conditions will be not as good as for a pupil who accurately assesses their reading skills. This may be particularly true if the teacher presence in school is low and the share of individualized learning is high.

The first analysis showed gender and SES differences in the self-assessments, indicating that girls and pupils with higher SES had higher levels of self-assessment. This was expected since achievement differences between these groups exist. However, when controlling for achievement differences (test-results and teacher judgements) the differences between girls and boys and different SES groups' self-assessments disappeared. Pupil attitudes towards reading were positively related to all of the other variables.

Thus the present investigation did not reveal any differences in the self-assessments on the basis of gender or SES. However, as schools become increasingly segregated with respect to SES (Myrberg & Rosén, 2006) it will be of importance to follow up these results with more recent pupil cohorts. Moreover, it is likely that individual responsibility will be less pronounced in the primary school. The data in the current thesis comes from 2001 and it is plausible that individual responsibility and use of self-assessments have increased in the past few years.

Exploring the relationship between teacher competence, teacher judgements and pupil test results

Despite teachers in different classes did not assess their pupils' knowledge and skills in a similar way, the correlation within a class was high in most classes (.65 in average). This indicates that teachers are largely able to rank order their pupil achievement in the Swedish language domain. However, certain variability for

the within classroom relations can also be expected, especially since significant differences across grades have previously been found. One hypothesis that would strengthen the validity of teacher judgements is that there would be that, for more competent teachers, there would be a higher correspondence between their judgements and their pupil achievement. Put another way, the higher the competence levels of the teacher, the better able they may be in identifying pupil knowledge and skills. The hypothesized rebuttal here is thus that no differences will be found in the judgements between teachers with different competence levels. However, before investigating this hypothesis, teacher competence must be operationalized and its role in relation to pupil achievement needs to be determined. This was also the main objective of Study II.

In order to explore the impact of teacher competence on pupil achievement, two achievement measures were selected; 1) test-results from PIRLS 2001, and 2) teachers' judgements of pupils' reading and writing skills.

To study the effects of teacher competence, the grade 3 sample was selected. At least two reasons supported this choice. First, 70% of the pupils had the same teacher over the first three years in school, typically first changing teacher in grade 4. This implies that grade 4 teachers had not taught their pupils for more than a few months, thus making the investigation of teacher effects difficult. Second, the teachers in grade 3 had different training and experience than most of the 4th grade teachers. For example, their teacher education would have had a greater emphasis on reading pedagogy. Teachers holding primary-school training (småskollärarutbildning och lågstadielärarutbildning) were typically educated in the 60s-80s. Teachers with this kind of training have been shown to achieve significantly better results on reading tests measuring grammar and phonological awareness than their more recently educated colleagues (Alatalo, 2011). Among the grade 3 teachers, several other kinds of training existed, as well as great variability in experience and further training.

In order to operationalize 'teacher competence' a latent variable was formulated. The measurement model of teacher competence was based on the 4 manifest variables; 1) appropriate education (coded according to a categorization made by Frank (2009)), 2) number of years of experience in 3rd grade, 3) whether or not the teacher held a certification, and 4) the amount of reading pedagogy included in their basic education. An excellent fit to the data was achieved for this measurement model.

In order to examine the impact of teacher competence, this variable was related to both achievement measures at the between level part. The results showed that teacher competence had a significant impact on pupil achievement,

regardless of whether it was measured by teacher judgements or PIRLS test results. This implies that teachers with higher competence levels worked in classrooms with higher achieving pupils. However, potential selection effects could have biased the analyses. Teachers with high competence are likely to "select" schools with high achieving pupils, or alternatively, parents with high SES may choose schools where more competent teachers work. As in Study I pupil SES was shown to correlate highly to pupil achievement, SES was chosen as a variable controlling for selection effects. The effect of teacher competence on teacher judgements and pupil test results decreased slightly. However the SES variable was also found to be uncorrelated to the teacher competence, which is in accord with the findings of Frank (2009). Pupil SES did not have as large effect on teacher judgements as on pupil test results, where SES explained 64% of the variance. One reason for this may be that teacher judgements were associated with ceiling effects; performance differences across classrooms which actually existed were not captured by the teacher judgements. For example, two high achieving classes would probably get similarly high average ratings, even though there might have been actual differences in levels of achievement. Thus, there was no variance for the SES variable to explain in the judgement variable. Another explanation may be that selection mechanisms other than SES were associated with teacher judgements, such as for example motivation or effort. This forms an important question for further research.

In summary, the results of this investigation showed that it was possible to operationalize a latent teacher competence variable which had a substantial effect on both achievement measures. Next step was to investigate the role of teacher competence for the accuracy of teacher judgements.

Although the influence of teacher competence on assessment practice has rarely been investigated, it may provide support for the validation process of teachers' judgements – if those with higher competence levels are more accurate in their judgements of their pupil reading skills. This was the main focus of Study III. More accurate judgements basically mean, in this case, that there is a higher correspondence with the pupil reading test results. The results of Study I form the starting-point for this hypothesis, since it was shown that the relationship between teacher judgements and pupil achievement in the different grades differed significantly. The judgements of teachers in grade 3 had higher correspondence, which is reasonable considering that they had taught their pupils for a longer period of time, thus having a fuller picture of their pupil abilities. A high correlation between judgements and test results could therefore be an indication of high quality judgements.

The first step in the analysis was to investigate whether the relationship between teacher judgements and pupil test-results varied between classes. Basically, this involved testing whether or not the correlation was close to .65 in most classes. To examine this, a multilevel model with random slopes was formulated. The significance of the variation of the slope between classrooms, i.e. whether the relationship between teachers' judgements and pupil reading achievement varied between classes, was determined. The estimated variance in the slope (.33) was significant, which implied that the relationship between teacher judgements and pupil achievement varied significantly between classes. The variance of this relationship justifies further analyses and will define a new variable at the between level part of the model.

In the next step, the latent teacher competence variable and average achievement for classes were related to the slope variable as explanatory variables. In order to estimate the effect of teacher competence on the slope, it is necessary to control for group achievement so the level of achievement is equal for all classes. The results revealed that teacher competence had a significant positive effect on the slope, implying that the effect of achievement on judgements was greater for more competent teachers. In other words, for teachers with a higher level of competence there was a higher correspondence between their pupils' test-results and their own judgements of their pupils' reading achievement. These results speak to the importance of an appropriate teacher education for teachers' ability to assess pupil achievement. The results also indicate that the test-results represent a valid measure of pupil achievement since the correspondence between these and teachers' judgements was higher for those pupils taught by more competent teachers.

Chapter Seven: Concluding remarks

The purpose of this thesis has been to explore validity issues in different assessment forms. The research questions have concerned the validity of the inferences made about teacher judgement, test results and pupil self-assessments of reading literacy in grades 3 and 4 in Swedish primary school. In particular focus has been directed to teacher judgements of reading literacy and their trustworthiness. From this research it is reasonable to conclude that teachers are largely able to rank order their own pupils in terms of their knowledge and skills. However, the correspondence between teacher judgement and pupil test results on PIRLS varied between teachers. A higher correlation between these variables was found for teachers with higher competence. The role of competent teachers for pupil achievement as well as for valid judgements is something which needs to be addressed more comprehensively in future research. Further, pupil self-assessments corresponded relatively well to both teacher judgements and PIRLS test results.

To calibrate teachers' judgements so that pupils in different classrooms are judged similarly when they have similar knowledge and skills, remains one of the most problematic issues in a school system aiming at educational equality. Pupils are likely to get different feedback or grades depending on which teacher they have and which school they attend. Differences in assessment can also lead to an imbalance in the allocation of resources to schools. This is a serious concern for equality in education. In spite of these results it can nevertheless be concluded that highly competent teachers are in a good position to identify pupil achievement levels, which in turn can mean that, resources permitting, they are able to provide assistance to those who need it most.

Informing teachers about the knowledge levels of pupils in other classes and other schools could be beneficial for fairness and equity in the Swedish school system. Further, this would also relate to the formative aspects of assessment, since teacher would be able to get feedback on their assessment practice.

Although teachers' judgements were not consistent across different classes, this does not imply that teachers were not aware of fairness and equality in assessment. Indeed, they seem to anchor their assessments in the performances within their own class, ranking pupils in terms of relative performance rather well. Generally, people tend to look for a starting point, an origin, upon which

they can base further estimations. Tversky and Kahneman (1974) talk about the psychological mechanism of 'anchoring'. Anchoring implies that all estimations are biased towards an initial value. In classrooms with similar performances, estimations of pupils' knowledge and skills could be very different. Starting with a high value, the other observations will be quite close to that value. If the starting point is a lower value, then the estimations that follow will also be lower. Thus teachers' anchoring of performances will depend on the starting point. Whether this is a low, average or high performing pupil, this will in all likelihood have consequences for their assessment practice.

Methodological issues

Pupils in a class will influence the group climate. They also share common experiences and have the same teachers. Individual observations are therefore not independent of each other since the group level influences the individual level. Research questions that have to take account of influences from different levels are multilevel problems and have to be examined using a multilevel framework. This approach is quite rare in educational research, due in no small part to the complexity of the techniques and that many items and observations are needed. However, great advances in the development of multilevel techniques have been made in recent years and several dissertations in Sweden have made use of advanced multilevel methods (e.g., Frank, 2009; Hansson, 2011; Holfve-Sabel, 2006; Klapp-Lekholm, 2008; Yang, 2003). Most of the variables used in the current thesis also have between class effects, which were accounted for using multilevel modeling.

An approach using latent variables is also an advantage compared to an approach with single indicators or sum scores of directly observable measures. The correspondence between teacher judgements and the PIRLS test results, for example, is expressed by a correlation coefficient which is error free in the sense that the judgement variable is latent. This is not the case in much of the previous research (see for example, Hoge & Coladarci, 1989; Südkamp et al., 2012).

Further, it should be noted that the correlation expresses the relationship between the assessment instrument teachers used, and the PIRLS results. The correspondence does not express the "true" correspondence between teacher judgements and PIRLS test results. In reality, the relationship could be stronger, especially for those teachers who had the most relevant competence for assessing pupil reading achievement. In this thesis, teacher ratings including 12 aspects of reading and writing have been adapted from a diagnostic instrument. The diagnostic instrument has been used as a tool to provide a measure for

teacher judgements. The ratings have been extrapolated to form an overall measure of teacher judgements. This is important to bear in mind when interpreting the results of this thesis. The "fairly high" correspondence between judgements and test-results nevertheless shows that it is possible to use the diagnostic instrument to document teachers' judgements of pupil reading achievement.

Future research

The current investigation brought a validity perspective to bear on the PIRLS 2001 international study. International studies are a hotly debated phenomenon and the validity of such studies is currently being examined from several different angles and at different levels. The items used in international studies are under constant scrutiny. Reliability of single items is a necessary condition for validity and there a several ways to secure validity at this level. Further, the constructs being measured (e.g., reading, mathematics, science) have to be aligned to external criteria that aim to measure the same constructs. Different ways of validating the constructs can be achieved through the use of content analyses of the framework used in the studies, and the steering documents of different countries. Another way is to examine the correspondence between the pupil achievement in the studies and another external achievement measure, as was carried out in this thesis.

The impact of international studies results has, in particular, been investigated at policy level. At the same time as the international studies provide valuable results of knowledge trends in different countries, unintended consequences of these studies have been put forward in literature. Lending and borrowing of educational polices in a globalised world are hypothesized to cause international educational differences to disappear and that there will be convergence towards a world curriculum (e.g., Baker & Le Tendre, 2005; Pettersson, 2008). If countries try to find solutions for undesirable results by scrutinizing the educational curricula of successful countries, isomorphism may be a possible result. The longitudinal design of the international studies facilitates such investigation. Rutkowski & Rutkowski (2009) attempted to address the question of global processes on education. Using data from TIMSS 1995, 1999 and 2003 they examined whether pupil responses had become more similar over time. The results did not provide any support for a trend towards isomorphism in educational policy and practice. One reason for this might have been that the data represented a time range too narrow to study the effects of globalization. Further research is therefore needed in order to shed more light on this issue.

Taking advantage of the older IEA studies, it would be possible to study trends in educational policies in a long-term perspective. Another way to address questions of validity in international studies is to examine the correspondence between the international studies and curricula. The use of the "opportunity to learn" (OTL) information that is available for all IEA studies can facilitate such analyses. The OTL concept represents the degree to which the content of the tests is aligned with curricula and actual teaching content. OTL-information has been systematically collected from teachers and through analyses of curricula and text-books in many of the IEA studies. However, few attempts have been made to connect the OTL-information to analyses of the achievement data.

In the current thesis I used the information about the knowledge and skills measured in the Swedish PIRLS 2001 assessment and compared it with teacher judgements and pupils' self-assessment of the same construct. Even though support for the PIRLS results could be found, it is necessary to explore the results at different levels and across different countries when addressing validity issues in international studies. The results of the international studies are goldmines for further research on these matters.

Swedish summary

Abstract

Många aktörer i skolan och samhället i stort har intresse av validiteten i de slutsatser som dras på grundval av olika bedömningsformer. I den här avhandlingen undersöks validitetsaspekter i olika bedömningsformer, närmare bestämt lärares bedömningar, resultat på ett standardiserat läsprov och elevers självskattningar i årskurs 3 och 4. Data inhämtades från den internationella undersökningen PIRLS 2001 (Progress in International Reading Literacy Study, 2001). Strukturell ekvationsmodellering med latenta variabler utgjorde den huvudsakliga metoden för analys. Ett av de viktigaste resultaten var att lärare väl kunde skatta elevernas språkliga kunskaper inom den egna klassen, medan de däremot har svårare att göra detta på ett samstämmigt och likvärdigt sätt över klassrum. Resultaten tyder också på att faktorer på elevnivå (kön och SES) påverkade lärarens skattningar av elevernas färdigheter. Lärarens kompetensnivå var viktig för såväl elevernas resultat i skolan som hur väl läraren bedömde elevens kunskaper. Vidare visade resultaten att elevers självskattningar stämmer relativt väl överens med såväl lärarens bedömning som elevens provresultat.

Inledning

Huvudsyftet med denna avhandling är att undersöka validiteten i de tolkningar och slutsatser som dras av lärares bedömningar, elevers testresultat och elevers självskattningar av läsförmåga. Avhandlingen består av två delar, dels en kappa som belyser validitetsaspekter när det gäller olika bedömningsformer, dels fyra artiklar som var och en diskuterar relationen mellan olika bedömningsformer och faktorer som kan påverka denna relation.

Tidigare forskning har givit skiftande stöd för olika bedömningsformer. Lärares bedömningar har ansetts vara giltiga eftersom överensstämmelsen mellan dem och andra mått på prestation varit hög (ex., Hoge & Coladarci, 1989; Meisels et al., 2001). Gipps (1994) menade exempelvis att eftersom lärare observerat sina elever under lång tid kunde deras bedömningar anses vara giltiga mått. Andra har argumenterat för att det inte går att betrakta lärares bedömningar som särskilt trovärdiga, då det har visat sig finnas liten överensstämmelse mellan dessa och exempelvis provresultat (Harlen, 2005).

Samma resultatmönster har noterats för sambandet mellan elevers självbedömningar och t.ex., lärares bedömningar. Dock är detta samband inte studerats särskilt mycket i lägre åldrar. Vidare finns det svensk skolforskning som tyder på att bedömningar av elevers kunskaper och färdigheter varierar mellan lärare. Skolverket (2007, 2009) har jämfört slutbetyg med provbetyg på nationella prov i årskurs 9 och i gymnasiet och kommit fram till att variationen är stor när det gäller lärares betygsättning. Det kan finnas rimliga anledningar till skillnader mellan provresultat och betyg; lärare har följt eleven under lång tid och kan därför justera bedömningen om resultatet på provet inte överensstämmer med de observationer som läraren gjort under undervisningen och vid tidigare bedömningssituationer. Emellertid kan det också vara så att likvärdiga kunskaper och färdigheter inte bedöms likvärdigt av lärare. Eftersom summativa bedömningar har setts variera mellan klassrum (se t.ex., Skolverket, 2007, 2009) finns det skäl att tro att också bedömningar med formativa syften kan variera mellan klasser. Om lärare inte rättvisande kan identifiera elevers kunskapsnivåer, påverkas elevers möjligheter att få adekvat hjälp och stöd samt deras förmåga att lära. Att lärare kan identifiera elevers kunskapsläge och kan ge återkoppling som kan stödja lärande är bland det viktigaste i lärares uppdrag (Hattie, 2009).

Att skolresultat påverkas av elevers och lärares olika egenskaper är sedan länge känt. Det finns också forskning som visar att lärare väger in olika personliga egenskaper i bedömningen. Elevers skolprestationer och kunskaper är det som starkast påverkar lärarens bedömning men forskning visar att även andra faktorer än kunskaper och färdigheter blir bedömda, till exempel elevens ansträngningar och motivation (Klapp-Lekholm, 2008; Thorsen & Cliffordson, 2011).

Det är viktigt att resultaten av olika bedömningar tolkas på ett valitt sätt. Olika bedömningsmetoder har olika styrkor och svagheter och det gäller att kunna identifiera dessa på ett systematiskt sätt. Validering är en verksamhet som aldrig når perfektion i alla avseenden och därför kan inte det som utvärderas någonsin anses helt valitt (Kane, 2006). Flera modeller finns för att validera prestationer och i denna avhandling har bland annat Bachmans (2005) användning av så kallade argument bidragit till ett exempel på validering. Bachman har utgått från Toulmins (1958/2003) användning av logiskt strukturerade argument. Modellen utgår från att det finns ett påstående; exempelvis att ”eleven klarar av att läsa de texter av den svårighetsgrad som krävs i årskurs 3” som är baserad på data, t.ex., lärarbedömning. Data kan vara att läraren bedömer att eleven kan det som påståendet anger samt att det stoff läraren har bedömt finns specificerat i kursplaner, mål, etc. Påståendet att elevens

läskunskaper är goda, kan försvagas av så kallade 'motbevis'. Exempel på sådana motbevis skulle kunna vara 1) att lärare bedömer olika saker och att elevens prestationer kunde bedömts annorlunda av en annan lärare, 2) lärare bedömer inte enbart läsprestationen utan även personlig karaktäristika. För att kunna eliminera dessa motbevis och bekräfta påståendet måste stöd inhämtas. Stödet kan till exempel vara en undersökning som visar att lärares bedömningar inte reflekterar personliga egenskaper eller en undersökning som visar att lärares bedömningar speglar samma kunskaper som mäts på ett annat sätt, såsom genom ett skriftligt prov bedömt av andra.

Syfte

Det övergripande syftet i avhandlingen är att bidra till kunskaper om styrkor och svagheter hos olika bedömningsformer. Genom att ömsesidigt belysa validitetsaspekter i lärares bedömningar, standardiserade prov och elevers självbedömningar kan slutsatser om styrkor och svagheter dras. Följande frågeställningar har varit till ledning i de fyra studierna som har genomförts:

1. Hur ser sambandet ut mellan lärares bedömningar och PIRLS läsprovsresultat inom klassrum och hur fungerar lärares bedömningar för jämförelser mellan klassrum i årskurs 3 och 4?
2. Påverkar elevens kön eller sociala bakgrund lärarens bedömningar?
3. Hur ser elevernas kunskaper och färdigheter i läsning ut för lärare med högre kompetens, och påverkar kompetensnivån bedömningspraktiken?
4. Hur väl kan elever skatta sina egna kunskaper och färdigheter i läsning? Finns det skillnader i denna förmåga beroende på kön och social bakgrund?

Data och metod

För att besvara syften och frågeställningar har data från den storskaliga undersökningen Progress in International Reading Literacy Study 2001 (PIRLS) inhämtats. 2001 ingick 35 länder i undersökningen och data finns tillgängligt från elever, lärare och skolledare. PIRLS design finns beskriven såväl i den internationella rapporten (Gonzalez & Kennedy, 2003) som i den svenska (Rosén et al., 2005). Till skillnad från många andra länder deltog Sverige med ett urval från årskurs 4 och ett från årskurs 3. Eftersom 70 % av lärarna i årskurs 3 har undervisat sina elever sedan årskurs 1 finns goda möjligheter att studera effekter av lärarens undervisning. Som ett tillägg till den internationella undersökningsdesignen hade Sverige en nationell utökning där lärarna ombads

att skatta sina elevers kunskaper i relation till flera olika aspekter av svenska språket. Detta formulär finns som bilaga 1. Bedömningsaspekterna är sprungna ur diagnosmaterialet ”Språket lyfter” (Skolverket, 2002) och istället för att formulera sina bedömningar i ord fick lärarna kvantifiera sina bedömningar på en tiogradig skala i anslutning till PIRLS-undersökningen 2001. Lärarna som har ingått i delstudierna fick ta ställning till 12 olika påståenden om elevernas kunskaper i läsning och skrivning och bedöma deras kunskapsnivåer på en skala från 1-10. Vidare har information om lärarnas utbildning, erfarenhet och fortbildning använts i denna avhandling, liksom information om elevernas sociala bakgrund, kön samt självskattningar. Till sist har också elevernas resultat på kunskapsprovet i PIRLS utnyttjas.

Analysmetoder

Data i samhällsvetenskaplig forskning och speciellt inom utbildning är ofta av hierarkisk natur (Gustafsson, 2009; Hox, 2002). Detta innebär att individer är klustrade inom ett klassrum, att klassrummen är klustrade inom en skola och så vidare. Individerna inom ett kluster tenderar att vara mer lika varandra än individer i andra kluster. De delar liknande erfarenheter, men också lärare och kamrater. Dessa beroenden måste hanteras statistiskt för att tolkningarna av analyserna ska bli korrekta. Många statistiska test tillämpar ett antagande om oberoende mellan de observationer som analyserna baseras på och bryts detta antagande kommer standardfelen att bli för små, vilket kan leda till flera felaktiga resultat (Hox, 2002).

Flernivåmodellering hanterar problemet med beroenden inom nivåer. Genom att dela upp variansen i komponenter inom och mellan grupper går det att separera ut den variation som går att härleda till individuella respektive gruppskillnader. Den så kallade intraklass-korrelationen ger signaler om huruvida flernivåmodellering bör användas eller inte; den ger information om den andel av variationen som kan härledas till skillnader mellan grupper. Ett exempel kan utgöras av olikheter i läsprestationer mellan svenska elever i årskurs 4, vilka finns både inom och mellan klasser. Ofta skiljer sig prestationer inom en klass sig åt ganska mycket, medan klassernas genomsnittliga prestationer är mer lika. Om klassernas resultat däremot varierar kraftigt kommer intraklass-korrelationen att vara hög, vilket indikerar heterogena prestationer bland klassrummen.

I delstudierna användes strukturell ekvationsmodellering (SEM) på flera nivåer. SEM har flera viktiga fördelar gentemot till exempel multipel regressionsanalys, bland annat genom användningen av latenta variabler. En latent variabel bygger på samvariationen mellan flera olika indikatorer och gör

det möjligt att bättre operationalisera ett begrepp än vad som vore möjligt med endast enstaka indikatorer eller manifesta variabler. En annan fördel med latenta variabler är att de kan sägas vara fria från mätfel. Indikatorer är alltid behäftade med ett visst mått av mätfel, men genom att konstruera en latent variabel sorteras mätfelet ut till en så kallad residualfaktor.

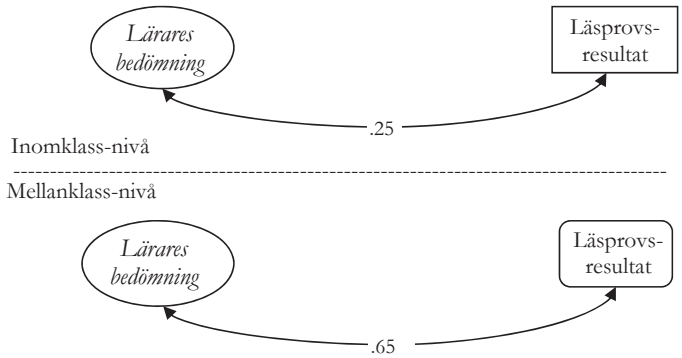
Analyserna gjordes med hjälp av programmet Mplus 6.1 (Muthén & Muthén, 2007-2012) som användes i programmet STREAMS analysmiljö (Gustafsson & Stahl, 2005). STREAMS låter användaren sätta upp modellerna i ett enkelt språk, vilket bland annat förstås av Mplus.

I de fyra delstudierna skapades latenta variabler för lärares bedömningar, lärares kompetens, elevers självskattningar och för elevers sociala bakgrund. För en mer detaljerad beskrivning av tillvägagångssättet och de statistiska egenskaperna hänvisas till delstudierna.

Sammanfattande resultat och diskussion

Den första delstudien fokuserade relationen mellan lärares bedömningar och elevers provresultat i årskurs 3 och 4. Indikatorerna som bildade den latenta variabeln lärares bedömning delades först in i en två-nivå modell med två faktorer, en läs- och en skrivfaktor. Emellertid passade denna modell data dåligt och indikerade att de två faktorerna var högt korrelerade. Faktorerna visade sig också korrelera högt på både individ- och gruppnivå (ca .96). Det verkade således svårt att separera bedömningen av läsning och skrivning empiriskt eftersom läraren tycktes göra en global bedömning av den språkliga förmågan hos individuella elever. Vidare analyser ledde fram till en latent variabel definierad genom fyra paketsummor av lärares bedömning. En utförligare diskussion av paketerings-förfarandet finns framför allt i den första delstudien. I nästa steg analyserades variationen i bedömningarna och det visade sig att den större delen av variationen hänförde sig till individnivå, alltså inom klasser. Ungefär 70 % av variationen i bedömningarna kunde härledas till inomklassvariation medan ungefär 30 % utgjordes av variation mellan klasser. Genom att introducera elevernas provresultat i PIRLS som en oberoende variabel, förklarades ca 45-50 % av den totala variationen i den beroende variabeln lärares bedömning på individnivå. På gruppnivå var motsvarande procentsats ca 3-4%. Dessa resultat innebär i praktiken att lärares bedömningar följer elevernas prestationer på PIRLS-testet relativt väl inom klassrum. Däremot varierade olika lärares bedömningar kraftigt trots att klassernas medelresultat hölls konstant. På individnivå motsvarar den förklarade variationen en korrelation på ungefär 0.65, vilket är i linje med tidigare forskning som gjorts på

området (Hoge & Coladarci, 1989; Südkamp et al., 2012). En modell över relationerna presenteras i Figur 1.



Figur 1. Relationen mellan lärares bedömning och elevernas resultat på PIRLS-provet

I den följande analysen relaterades elevens sociala bakgrund och kön till lärares bedömningar och elevers resultat, främst för att undersöka om dessa faktorer var något som lärare tog med i sin bedömning av elevens läs- och skrivkunskaper. Vad som kan noteras på individnivå är att både kön och social bakgrund har en effekt på lärares bedömning. Denna effekt betyder att flickor och elever från högre social bakgrund får en något högre bedömning av lärarna, givet att deras prestationsnivå på provet är samma som för pojkar och elever från lägre social bakgrund. Både SES och kön är också positivt relaterade till prestation, vilket visar att flickor och elever med högre SES har bättre resultat på provet. Att flickor tenderar att bedömas något högre än vad prov resultat visar är i linje med tidigare forskning (Reuterberg & Svensson, 2000). En förklaring kan vara att flickorna visar kunskaper som inte avspeglas i provresultatet (t.ex., verbal förmåga) och som alltså läraren väger in i sin bedömning (Wernersson i SOU, 2010:15). En annan förklaring skulle kunna vara att flickorna bedöms anstränga sig mer och därför får en lite bättre bedömning på grund av detta. Ett liknande resonemang kan föras när det gäller effekten av social bakgrund.

När det gäller klassnivån visar det sig att social bakgrund har en mycket stark effekt på klassens resultatnivå. Klassrum med ett högt genomsnittligt SES visar betydligt högre prestationer. Klassers SES har också en relativt stark effekt på lärares bedömningar, samtidigt som man kan notera att effekten av provresultat på lärares bedömningar inte längre är signifikant när SES introduceras i modellen. Det ser sålunda ut som att relationen mellan provresultat och lärares bedömningar helt medieras av social bakgrund. Då lärarna skattade elevernas kunskaper och färdigheter kände de inte till den egna klassens, eller andra

klassers testresultat, medan de hade information om elevernas sociala bakgrund och denna variabel är högt korrelerad med testresultat.

I den andra delstudien undersöktes betydelsen av lärares kompetens för elevers läsförmåga. Eftersom effekter av lärare troligtvis uppstår efter en tid användes endast informationen från årskurs 3. Dessa lärare hade i regel haft sina elever sedan årskurs 1, medan eleverna i årskurs 4 bara haft sin lärare i ungefär en termin. Lärarkompetens definierades med hjälp av en latent variabelmodell baserad på fyra olika indikatorer på lärarkompetens. Dessa var 1) om läraren var behörig, 2) typ av lärarutbildning 3) antal års erfarenhet av undervisning i årskurs 3, samt 4) om särskilt fokus lades vid läspedagogik under utbildningen. I urvalet finns flera olika lärarutbildningar och en kategorisering gjordes för att kunna rangordna lärarna; från de med mest adekvat utbildning till de med minst adekvat utbildning för att lära ut läsning till yngre elever. Denna kodning validerades på två sätt. Dels användes information från Franks (2009) beskrivning av hur mycket betoning på läsning de olika lärarutbildningarna i Sverige har haft sedan 1969, vilket senare låg till grund för kategoriseringen. Dels användes resultaten från Alatalos (2011) avhandling som visar att småskollärarna och lågstadielärarna hade de högsta kunskapsnivåerna, vad det gäller stavningsregler och fonologisk medvetenhet. Alatalos resultat låg också i linje med den kategorisering som Frank (2009) gjort med avseende på vilka lärare som har haft mest tyngdpunkt på läsinlärning i utbildningen.

Ett flertal studier har visat att lärarkompetensen är viktig för elevernas prestationer i skolan. I delstudie II relaterades lärarkompetens till elevernas resultat, definierad genom både lärarbedömningar och elevers provresultat. Resultaten visade att klasser med mer kompetenta lärare hade klart högre resultat på läsprovet i PIRLS 2001 och bedömdes högre av lärarna. Eventuella selektionseffekter kontrollerades genom att ta hänsyn till elevernas sociala bakgrund. Mer kompetenta lärare skulle kunna söka sig till skolor och klasser där eleverna presterar på en hög nivå. Det kan också vara så att föräldrar med hög socioekonomisk status väljer att placera sina barn i skolor där lärarna är kompetenta. Emellertid visade det sig att SES inte hade något samband med lärarkompetens. SES visade sig fånga större delen av variationen mellan klassernas resultat (R^2=0.64) men lärarbedömningarna hade inte så högt samband med SES, vilket indikerar att oförklarad varians kan hänföras till andra variabler. Anledningen till att lärarbedömningarna korrelerade lägre med SES än vad PIRLS-testresultaten gjorde kan också bero på takeffekter i lärarbedömningsvariabeln vilka kan bidra till att skillnader som faktiskt existerar mellan olika SES grupper inte går att uppfatta. Om lärare 1 och 2 bedömer sina

respektive klasser med i genomsnitt nio på den tio-gradiga skalan trots att klassmedelvärdet skiljer sig avsevärt när det gäller provresultaten, går det inte heller att förklara variation som kan finnas, men som inte manifesteras genom lärarnas bedömning. Andra tänkbara förklaringar till variationen i lärares bedömningar kan vara andra variabler, exempelvis motivation. Detta är en fråga av stort intresse för framtida forskning inom området.

Den tredje delstudien bygger på resultaten från studie II. Huvudsyftet i studien var att undersöka om lärarkompetens påverkade sambandet mellan lärarnas bedömningar och elevernas provresultat. Hypotesen var att lärare med högre kompetens i högre grad skattar elevernas kunskaper i samstämmighet med provresultatet. För att undersöka detta genomfördes en flernivåanalys, vilken innebar att den latenta lärarkompetensvariabeln som formulerades på klassnivå relaterades till relationen mellan bedömning och provresultat på inomnivå. En så kallad random slope-teknik applicerades i syfte att undersöka sambandet mellan lärarkompetens, lärarbedömning och elevers provresultat. I korthet innebär det här tillvägagångssättet att relationen mellan provresultat och lärarbedömning antas variera över klasser. I det första steget undersöks om detta är ett korrekt antagande genom att undersöka variansen i. I och med att variansen var signifikant, kan slutsatsen dras att relationen mellan prov och bedömning varierade över klasser. Därför går det att försöka förklara detta med hjälp av variabler på klassnivå. När lärarkompetens relaterades till relationen mellan provresultat och lärarbedömning visade det sig att sambandet mellan prov och bedömning var något högre för lärare med högre kompetens. Det kanske viktigaste resultatet från den här delstudien är således att lärare med högre kompetens bedömer elevernas kunskaper mer i enlighet med deras provresultat, vilket också ger fog för tolkningen att provet mäter läsförmåga väl. Eftersom provet väl kan anses spegla de mål som finns i svenska kursplaner, är en rimlig slutsats att lärare med högre kompetens är bättre på att bedöma sina elevers kunskapsnivåer. Implikationerna av den här studien är således att högre lärarkompetens i klassrummen inte bara har betydelse för elevernas prestationer, utan även för förmågan att bedöma elevernas kunskaper på ett korrekt sätt.

I den fjärde och avslutande studien studerades sambandet mellan elevers självbedömningar, lärares bedömningar och elevers provresultat i årskurs 3 Eftersom elever tolkar information från läraren då de gör självbedömningar, verkade de lämpligt att välja ut de elever som haft sina lärare en längre period. Som tidigare nämnts var detta fallet i årskurs 3 men inte i årskurs 4. I tidigare forskning är ofta elevers bedömningar jämförda med lärares, som i så måtto oftast är det initiala kriteriet (Ross, 2006). Elevers självbedömningar är ibland

också jämförda med provresultat, men dessa studier är färre. I delstudie IV finns möjligheten att jämföra elevers bedömningar med både provresultat och lärares bedömningar, i syfte att undersöka hur väl elever lyckas skatta sina kunskaper. Dessutom undersöktes om det fanns någon skillnad mellan hur pojkar och flickor skattar sina kunskaper samt om elever med olika socioekonomisk status skattar sina kunskaper på olika sätt.

De bedömningar som eleverna fick göra handlade om hur goda de ansåg sina läskunskaper vara, dels relativt andra elever, dels med sig själva som utgångspunkt (T.ex. ”Läsning är väldigt lätt för mig”). Eleverna fick ta ställning till fyra påståenden på en skala som gick från Instämmer inte alls – Instämmer helt (1-4).

Ett av huvudresultaten i den här studien var att elevers självbedömningar inte varierade mellan klassrum i särskilt stor utsträckning, till skillnad från vad som gällde för lärarbedömningar och läsprovsresultat. Detta betyder att eleverna i olika klassrum bedömer sina kunskaper och färdigheter på en likvärdig nivå trots att prestationsskillnader föreligger. De små skillnaderna i självbedömning mellan klasser möjliggjorde inte någon undersökning av självskattningarna på gruppnivå. En anledning till detta resultat kan vara att eleverna tog sin utgångspunkt i sina klasskamraters kunskaper och färdigheter då de uppskattade sina egna. En annan anledning kan ha varit att de indikatorer som definierade den latenta variabeln *elevers självbedömning* endast innehöll fyra svarsalternativ, vilket kan ha lett till att variationen i dessa indikatorer inte var så stor.

I nästa steg studerades relationen mellan elevers självbedömning, lärares bedömning och läsprovsresultat. Det visade sig att sambandet mellan självbedömning och läsprovsresultaten var lite lägre (.58) än sambandet mellan lärarnas bedömning och läsprovsresultat (.65). När det gällde relationen mellan lärarnas bedömningar och elevers självbedömningar var denna ungefär densamma som relationen mellan självbedömningar och läsprovsresultat (.59). Resultaten indikerar att både provresultaten och lärarbedömningarna har relativt god överensstämmelse med elevers självbedömningar. Vidare undersöktes om skillnader förelåg beträffande hur pojkar och flickor skattade sin läsförmåga. Givet samma prestation på provet visade det sig att det inte fanns några signifikanta skillnader mellan könen vad gäller att bedöma sina egna kunskaper i läsning. Inte heller verkade det som att elevens socioekonomiska status hade någon större betydelse för hur eleven skattade sina kunskaper. Resultaten i dessa studier visade alltså att eleverna i årskurs tre skattade sina generella läskunskaper relativt väl, och sambandet var nästan identiskt oavsett om självbedömningarna relaterades till lärarbedömningar eller läsprovsresultat.

Slutsatser

I avhandlingen har tre mått på elevers läsförmåga problematiserats utifrån validitetsaspekter. Dessa är lärarbedömningar, provresultat och elevers självbedömningar av läsförmåga. I Sverige har lärares bedömningar varit det klart dominerande måttet på elevers kunskaper och även om prov använts så har det varit i relativt liten omfattning i de lägre årskurserna. Avhandlingens resultat kan ge formativ feedback gällande vilka styrkor och svagheter som finns i olika bedömningsinstrument.

När förhållandet mellan lärares bedömningar och provresultat ömsesidigt belystes, indikerade resultaten att båda instrumenten kan vara lämpliga mått på läskunskaper inom klassrum. Dock verkar lärare ha olika referensramar vid bedömning och att samma kunskaper skattas på skilda sätt av olika lärare. Detta gör att lärarbedömningar kan vara problematiska att använda då jämförelser ska göras mellan elever som går i olika klasser. Framför allt kanske detta är problematiskt vid summativa bedömningar som ligger till grund för individuella utlåtanden och betyg. Summativa utlåtanden kan emellertid också ligga till grund för insatser som görs för att främja elevers lärande, och detta innebär att villkoren för lärande inte blir likvärdiga för eleverna i de tidiga åren i grundskolan. Även om inte undervisningen behöver ske på samma sätt mellan skolor föreskriver Skollagen (2010) att utbildningen i grundskolan ska vara likvärdig och att hänsyn ska tas till elevernas olika förutsättningar och behov. Om bedömningen inte är likvärdig kan det få konsekvensen att en del skolor ger stöd och feedback till elever som behöver det, medan andra inte gör det.

Ytterligare feedback om validiteten i de slutsatser som dras av lärares bedömningar gavs i studierna. Det visades att kön och social bakgrund tenderar att påverkar lärarens bedömning. Frågor om rättvisa gör att dessa faktorer inte kan förbises, även om det finns anledning att tro att lärare skattat dessa elevers prestationer högre eftersom de kan ha visat prov på kunskaper som inte undersökts i provet. Vidare kan en lärare med högre formell kompetens vara en viktig del i vägen till en framgångsrik bedömningspraktik. Då lärare med högre kompetensnivå hade högre samband mellan provresultat och lärarbedömning ger detta vid handen att PIRLS provresultat också är ett bra mått på elevernas kunskaper och färdigheter inom området läsning.

I denna avhandling har några svar givits om validiteten i olika bedömningsformer, men i ett föränderligt samhälle gäller det påminna sig om att validering är en ständigt pågående aktivitet.

References

Alatalo, T. (2011). *Skicklig läs- och skrivundervisning i åk 1-3. Om lärares möjligheter och hinder.* [Proficient teaching of reading and writing in grades 1-3. About teachers' opportunities and barriers.] (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/25658

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity*. (pp.19-32) Hillsdale, NJ: Lawrence Erlbaum.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly, 2*, 1-34. doi: 10.1207/s15434311laq0201_1

Bagozzi, R.P., & Heatherton, T.F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling, 1*, 35-67.

Baker, B., & LeTendre, G. (2005). *Global Similarities: World Culture and the Future of Schooling.* Stanford, Calif: Stanford University Press.

Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy, 18*, 215-228. doi: 10.1080/0268093022000043065

Ball, S. J. (2010). New Voices, New Knowledges and the New Politics of Education Research: the gathering of a perfect storm?, *European Educational Research Journal, 9*(2), 124-137. http://dx.doi.org/10.2304/eerj.2010.9.2.124

Bates, C., & Nettelbeck, T. (2001). Primary School Teachers' Judgements of Reading Achievement. *Educational Psychology, 21*, 177-187.

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences, 42*, 825-829.

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A Comparative Study of Teacher Ratings of Emergent Literacy Skills and Student Performance on a Standardized Measure. *Education*, *126*, 116-137.

Binet. A., & Simon, T. (1916). *The development of intelligence in children.* Baltimore, Williams & Wilkins. (Reprinted 1973, New York: Arno Press; 1983, Salem, NH: Ayer Company)

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*, 7-73.

Black, P., William, D. (2011). The Reliability of Assessments. In J. Gardner (Eds.), *Assessment and learning.* (Second edition; pp. 243-263). Gateshead, UK: Sage.

Bong, M., Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist, 34*, 139-154.

Borsboom, D., Mellenbergh, G. J., van Heerden, J. (2004). The concept of validity. *Psychological review, 111*, 1061-1071.

Borsboom, D., Cramer, A.O.J., Kievit, R.A., Zand Scholten, A., Franić, S. (2009). The end of construct validity. In R. Lissnitz (Ed.). *The Concept of Validity. Revisions, New Directions and Applications.* (pp. 135-170). Charlotte, NC: IAP.

Broadfoot, P. (1996). *Education, assessment and society – a sociological analysis.* Buckingham: Open University Press.

Brookhart, S. M. (2012). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, 1-22. doi:10.1080/0969594x.2012.703170

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: The Guilford Press.

Brunner, M., Ludtke, O., & Trautwein, U. (2008). The Internal/External Frame of Reference Model Revisited: Incorporating General Cognitive Ability and General Academic Self-Concept. *Multivariate Behavioral Research, 43*, 137-172.

Butler, Y. G., & Lee, J. (2006). On-Task versus Off-Task Self-Assessments among Korean Elementary School Students Studying English. *Modern Language Journal, 90*, 506-518.

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.

Coladarci, T. (1986). Accuracy of Teacher Judgments of Student Responses to Standardized Test Items. *Journal of Educational Psychology*, *78*, 141-146.

Cronbach, L. J. (1963). Course Improvement Through Evaluation. *Teachers College Record*, *64*, 672-683.

Cronbach, L. J. (1971). Test Validation. In Robert L. Thorndike (Ed.), *Educational Measurement* (Second edition, pp. 443-507). Washington, D.C: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (Eds.), *Test validity*. (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin 52*, 281-302.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge Grading: Endorsed by Students and Teachers Alike. *Applied Measurement in Education, 12*, 53-72.

Cunningham, A. E. & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 934-945.

Cureton, E. (1951) Validity. In E. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington, D.C: American Council on Education.

D'Agostino, J. V., Welsh, M. E., & Corson, N. M. (2007). Instructional Sensitivity of a State's Standards-Based Assessment. *Educational Assessment, 12*, 1-22.

Darling-Hammond, L. (2000). Teacher Quality and Student Achievement: A Review of State Policy Evidence. *Education Policy Analysis Archives, 8*(1), 1-50.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does Teacher Certification Matter? Evaluating the Evidence. *Educational Evaluation and Policy Analysis, 23*(1), 57-77.

Darling-Hammond, L., Bransford, J. (2005). (Eds.), *Preparing teachers for a changing world. What teachers do and should be able to do*. San Francisco, CA: Jossey-Bass.

DeVellis, R. F. (2003). *Scale Development. Theory and Applications*. Thousand Oaks, CA: Sage.

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

Emanuelsson, I. & Fischbein, S. (1986). Vive la difference? A Study on Sex and Schooling. *Scandinavian Journal of Educational Research, 30*, 71-84.

Englund, T., Forsberg, E., & Sundberg, D. (2012). *Vad räknas som kunskap? Läroplansteoretiska utsikter och inblickar i lärarutbildning och skola*. Stockholm: Liber.

Falchikov, N., & Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of Educational Research, 59*, 395-430.

Feinberg, A. B., & Shapiro, E. S. (2009). Teacher Accuracy: An Examination of Teacher-Based Judgments of Students' Reading With Differing Achievement Levels. *Journal of Educational Research, 102*, 453-462.

Forsberg, E., Lindberg, V. (2010). *Svensk forskning om bedömnings – en kartläggning.* [Swedish research about assessment – an overview]. Stockholm: Vetenskapsrådet.

Frank, E. (2009). *Läsförmågan bland 9-10-åringar. Betydelsen av skolklimat, hem- och skolsamverkan, lärarkompetens och elevers hembakgrund.* [Reading literacy among 9-10 yeaar olds. The importance of school climate, teacher competence and pupil home background] (Doctoral thesis. University of Gothenburg, Göteborg). Retrieved from http://hdl.handle.net/2077/20083.

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. doi: 10.1080/02702711003608279

Geske, A., & Ozola, A. (2009). Different Influence of Contextual Educational Factors on Boys' and Girls' Reading Achievement. *US-China Education Review, 6*, 38-44.

Gijsel, M. A. R., Bosman, A. M. T., & Verhoeven, L. (2006). Kindergarten risk factors, cognitive factors, and teacher judgments as predictors of early reading in Dutch. *Journal of Learning Disabilities, 39*, 558– 571.

Gipps, C. (1994). *Beyond Testing. Towards a theory of educational assessment*. London: The Falmer Press.

Gipps, C. (2001). Sociocultural Aspects of Assessment. In G. Svingby & S. Svingby (Eds.) *Bedömning av kunskap och kompetens.* [Assessment of knowledge and competence] (pp. 15-67). Stockholm: Lärarhögskolan i Stockholm, PRIM-gruppen.

Gipps, C., Brown, M., McCallum, B., & McAlister, S. (1995). *Intuition Or Evidence?: Teachers and National Assessment of Seven Year Olds.* Buckingham: Open University Press.

Goffin, R. D. (2007). Assessing the adequacy of structural equation models: Golden rules and editorial policies. *Personality and Individual Differences, 42*(5), 831-839. doi: 10.1016/j.paid.2006.09.019

Goldhaber, D. D., & Brewer, D. J. (2000). Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement. *Educational Evaluation and Policy Analysis, 22*, 129-145.

Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 User Guide for the International Database.* Chestnut hill, MA: Boston College.

Guilford, J (1946) New standards for test evaluation. Educational and Psychological Measurement, 6, 427-438.

Gustafsson, J-E. (2009). Strukturell Ekvationsmodellering. [Structural Equation Modeling]. In G. Djurfeldt & M. Barmark (Eds.), *Statistisk verktygslåda 2 Multivariat analys* (pp. 269-321). Lund, Sweden: Studentlitteratur.

Gustafsson, J-E., & Erickson, G. (in press). To trust or not to trust? –Teacher marking versus external marking of national tests.

Gustafsson, J-E., & Myrberg, E. (2009). Resursers betydelse för elevers resultat. [The importance of resources for pupil achievement]. In Skolverket (Ed.), (pp. 160-206). *Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer.* [What influences the achievement in Swedish compulsory school? Review of the importance of different factors]. Stockholm: Skolverket.

Gustafsson, J.-E., & Rosén, M. (2005). *Förändringar i läskompetens 1991-2001. En Jämförelse över tid och länder.* [Changes in reading literacy 1991-2001. A comparison over time and across countries]. Göteborg: Institutionen för pedagogik och didaktik.

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

Gustafsson, J-E., & Yang-Hansen, K. (2009). *Resultatförändringar i Svensk grundskola*. [Achievement changes in Swedish compulsory school] In Skolverket (Ed.), (pp. 40-83).Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer. [What influences the achievement in Swedish compulsory school? Review of the importance of different factors]. Stockholm.

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] (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/26669

Hanushek, E. A. (1989). The impact of differential expenditures on school performance, *Educational Researcher*, *18*, 45–65.

Hanushek, E. A. (2003). The Failure of Input-Based Schooling Policies. *The Economic Journal*, 113, 64-98.

Harlen, W. (2005). Trusting Teachers' Judgement: Research Evidence of the Reliability and Validity of Teachers' Assessment Used for Summative Purposes. *Research Papers in Education, 20*, 245-270.

Harlen, W. (2011). On the Relationship between Assessment for Formative and Summative Purposes. In J. Gardner (Eds.), *Assessment and learning*. (Second edition; pp. 61-80). Gateshead, UK: Sage.

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*, 81-112. doi: 10.3102/003465430298487

Hau, K.-T., & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology, 57*, 327-351.

Hecht, S. A., & Greenfield, D. B. (2002). Explaining the predictive accuracy of teacher judgments of their students' reading achievement: The role of gender, classroom behavior, and emergent literacy skills in a longitudinal sample of children exposed to poverty. *Reading and Writing: An Interdisciplinary Journal, 15*, 789– 809.

Hoge, R. D., & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Review of Educational Research, 59*, 297-313.

Holfve-Sabel, M-A., (2006). *Attitudes towards Swedish comprehensive school.* (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/10035

Hox, J. (2002). *Multilevel Analysis - Techniques and Applications*. Mawah, NJ: Lawrence.

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

James, W. (1890/1998). *The principles of psychology*. Bristol, UK: Thoemmes.

Jönsson, A. (2008). *Educative assessment for/of teacher competency: a study of assessment and learning in the "interactive examination" for student teachers.* (Doctoral Thesis, Malmö University, Malmö, Sweden). Malmö studies in educational sciences.

Jönsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*. 130-144.

Kane, M. (1992). An argument-based approach to validity. *Applied Psychological Bulletin, 112*, 527-535.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, D.C: American Council on Education and National Council on Measurement in Education.

Kane, M (2009). Validating the Interpretations and Uses of Test Scores. In R. W. Lissnitz (Ed.), *The Concept of Validity. Revisions, New Directions and Applications*. (pp. 39-64) Charlotte, NC: IAP.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice, 18*, 5-17.

Kenny, D. T., & Chekaluk, E. (1993). Early Reading Performance: A Comparison of Teacher-Based and Test-Based Assessments. *Journal of Learning Disabilities, 26*, 227-236.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54*, 757–765.

Klapp Lekholm, A. (2008). *Grades and grade assignment: effects of student and school characteristics*. (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/18673

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*, 181-199.

Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education, 2*, 145-163.

Klenowski, V., & Wyatt-Smith, C. (2010). Standards-Driven Reform Years 1-10: Moderation an Optional Extra? *Australian Educational Researcher, 37*, 21-39.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage.

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

Liberg. C. (2010). Texters, textuppgifters och undervisningens betydelse för elevers läsförståelse. [Texts, text tasks and the teachings importance for pupil reading literacy]. In Skolverket (Ed.), *Texters, textuppgifters och undervisningens betydelse för elevers läsförståelse. Fördjupad analys av PIRL S 2006*. (pp. 8-130). Stockholm: Skolverket.

Lissitz, R. W. (2009). Introduction. In R. W. Lissnitz (Ed.). *The Concept of Validity. Revisions, New Directions and Applications*. (pp. 1-15) Charlotte, NC: IAP.

Lissitz, R. W., & Samuelsten, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*, 437-448.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To Parcel or Not To Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling, 9*, 151-173.

Llosa, L. (2007). Validating a Standards-Based Classroom Assessment of English Proficiency: A Multitrait-Multimethod Approach. *Language Testing, 24*, 489-515.

Llosa, L. (2008). Building and Supporting a Validity Argument for a Standards-Based Classroom Assessment of English Proficiency Based on Teacher Judgments. *Educational Measurement: Issues and Practice, 27*, 32-42.

Loehlin, J. C. (2004). *Latent variable models* (4th edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological reports*, *3*, 635-694.

Lundahl, C. (2006). *Viljan att veta vad andra vet. Kunskapsbedömning i tidigmodern modern och senmodern skola*. [To know what others know: Assessment in education in pre-modern, modern and late-modern times] Uppsala Universitet, Uppsala.

Lundahl, C. (2011). *Bedömning för lärande*. [Assessment for learning]. Stockholm: Norstedt.

Madaus, G. F., & O'Dwyer, L. M. (1999). A Short History Of Performance Assessment. *Phi Delta Kappan, 80*, 688-695.

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences, 42*, 851-858.

Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external fram of reference model. *American Educational Reasearch Journal*, *23*, 129-149.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling, 11*, 320-341.

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

Martínez, J. F., Stecher, B., & Borko, H. (2009). Classroom Assessment Practices, Teacher Judgments, and Student Achievement in Mathematics: Evidence from the ECLS. *Educational assessment 14*, 78 - 102.

McCoach, D. B. (2010). Hierarchical Linear Modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 123-140). New York; NY: Routledge.

Mehrens, W A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice 16*, 16-18.

Meisels, S. J., Bickel, D. D., Nicholson, J., Yange, X., & Atkins-Burnett, S. (2001). Trusting Teachers' Judgments: A Validity Study of a Curriculum-Embedded Performance Assessment in Kindergarten to Grade 3. *American Educational Research Journal, 38*, 73-95.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (Third edition, pp. 13-103). New York: American Council on Education/Macmillian.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

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.

Muthén, B. O. (1994). Multilevel Covariance Structure-Analysis. *Sociological Methods & Research, 22*, 376-398.

Muthén, L. K., & Muthén, B. O. (2007-2012). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Myrberg, E. (2006). *Fristående skolor i Sverige. Effekter på 9-10-åriga elevers läsförmåga.* [Independent schools in Sweden. Effects on 9-10 year old pupil reading achievement]. (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/16820

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*, 145-162.

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*, 185-205.

Newton, P. E. (2007). Clarifying the Purposes of Educational Assessment. *Assessment in Education: Principles, Policy & Practice, 14*, 149-170.

Nusche, D., Halász, G., Looney, J., Santiago, P., & Shewbridge, C. (2011). *OECD Reviews of Evaluation and Assessment in Education, Sweden*. Paris: OECD.

Nyström, P. (2004). *Rätt mätt på prov*. [Validation of educational assessments]. Umeå University.

Näsström, G. (2005). *Lärares skattningar av sina elevers provresultat*. [Teachers' ratings of their students' test results]. Umeå: Institutionen för beteendevetenskapliga mätningar.

Pedhazur, E. J. & Pedhazur-Schmelkin, L. (1991). *Measurement, Design and Data analysis. An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associate.

Pehrsson, A. & Sahlström, F. (1999). *Kartläggning av läsning och skrivning ur ett deltagarperspektiv: Analysverktyg för alla*. (Specialpedagogiska rapporter, nr 14) Göteborg: Göteborgs universitet, Institutionen för specialpedagogik.

Perry, N. E., & Meisels, S. J. (1996). How Accurate Are Teacher Judgments of Students' Academic Performance? Working Paper Series: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 555 New Jersey Avenue, N.W., Room 400, Washington, DC 20208-5652.

Pettersson, A. (2011). Bedömning – vad, varför och varthän? [Assessment – what, why and where to?] In L. Lindström & V. Lindberg (Eds.), *Pedagogisk bedömning: att dokumentera, bedöma och utveckla kunskap*. [Educational assessment: to document, assess and develop knowledge]. (pp. 32-41). Stockholm: HLS Förlag.

Pettersson, D. (2008). *Internationell kunskapsbedömning som inslag i nationell styrning av skolan*. Acta Universitatis Upsaliensis. Uppsala Studies in Education No 120. Uppsala: Uppsala universitet.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4-13.

Reeves, D. J., Boyle, W. F., & Christie, T. (2001). The Relationship between Teacher Assessments and Pupil Attainments in Standard Test Tasks at Key Stage 2, 1996-98. *British Educational Research Journal, 27*, 141-160.

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: University of Gothenburg.

Rosén, M. (1998). *Gender differences in patterns of knowledge*. (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/13820

Rosén, M. (2012). Förändringar i läsvanor och läsförmåga bland 9- till 10-åringar. Resultat från internationella studier. [Changes in reading habits and reading achievement among 9-10 year olds. Results from international studies]. In Swedish Government Official Reports (2012, report no 10), (Ed.), *Läsarnas marknad, marknadens läsare – en forskningsantologi*. [The readers market, the market readers - a research anthology]. (pp. 111-139). Stockholm. Department of Education.

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

Ross, John A. (2006). The Reliability, Validity, and Utility of Self-Assessment. *Practical Assessment Research & Evaluation, 11*(10). 1-13. Available online: http://pareonline.net/getvn.asp?v=11&n=10

Rutkowski, L., & Rutkowski, D. (2009). Trends in TIMSS responses over time: evidence of global forces in education? 1. *Educational Research and Evaluation, 15*(2), 137-152.

Sadler, D. R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science, 18*, 119-144.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods, 7*, 147-177.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation*, (pp. 39-83), Chicago, IL: Rand McNally.

Selghed, B. (2004). *Ännu icke godkänt - Lärares sätt att erfara betygssystemet och dess tillämpning i yrkesutövningen.* [Not yet passed: How teachers' experience a criterion-referenced grading system and what they say about its use in Swedish secondary school]. (Doctoral thesis, Malmö University, Malmö, Sweden)

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: The interplay of theory and methods. *Journal of Educational Psychology, 46*, 407-441.

Shepard, L. A. (1993). Evaluating Test Validity. *Review of Research in Education, 19*, 405-450.

Shrauger, J. S., & Osberg, T. M. (1981). The Relative Accuracy of Self-Predictions and Judgments by Others in Psychological Assessment. *Psychological Bulletin, 90*, 322-351.

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research, 75*, 417-453.

Sjöberg, L. (2010). *Bäst i klassen? Lärare och elever i svenska och europeiska policytexter* [Top of the Class? Teachers and pupils in Swedish and European policy texts] http://hdl.handle.net/2077/24101

Skaalvik, E. M. (1997). Self-enhancing and self-defeating ego-orientation: Relation with task and avoidance orientation, achievement, self-perceptions, and anxiety. *Journal of Educational Psychology, 89*, 71-81.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

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*, 51-79.

Stobart, G. (2011). Validity in Formative Assessment. In John Gardner (Ed.), *Assessment and Learning* (Second edition; pp. 233-242). London: Sage Publications.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762.

Svensson, A. (1971). *Relative achievement. School performance in relation to intelligence, sex and home environment*. Stockholm: Almqvist & Wiksell.

Swalander, L. (2006). *Reading achievement: Its relation to home literacy, self-regulation, academic self-concept, and goal oreintation i n children and adolescents*. Doctoral thesis, Lund University.

Swalander, L., & Taube, K. (2007). Influences of family based prerequisites, reading attitude, and self-regulation on reading ability. *Contemporary Educational Psychology, 32*, 206-230. doi: 10.1016/j.cedpsych.2006.01.002

Swedish Government Official Reports. [SOU] (1942). *Betänkande med utredning och förslag angående betygssättningen i Folkskolan.* [Clear goals and knowledge demands in the compulsory school. Proposal for a new goal and follow-up system]. (Report No. 11). Stockholm, Sweden: Department of Education.

Swedish Government Official Reports. [SOU] (2010). *Könsskillnader i skolprestationer – idéer om orsaker*. [Gender differences in school performances – ideas about causes]. (Report No. 51). Stockholm, Sweden: Department of Education.

Swedish Schools Inspectorate (2010). *Kontrollrättning av nationella prov i grundskolan och gymnasieskolan.* [Control marking of national tests for comprehensive school and upper secondary education]. Retrieved 20 July 2012 from http://www.skolinspektionen.se > Publikationer.

Swedish Schools Inspectorate (2011). *Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan.* [Remarking of national tests for comprehensive school and upper secondary education]. Retrieved 20 July 2012 from http://www.skolinspektionen.se > Publikationer.

Taras, M. (2005). Assessment – Summative and Formative – Some Theoretical Reflections. *British Journal of Educational Studies, 53*, 466-478.

Taras, M. (2009). Summative Assessment: The Missing Link for Formative Assessment. *Journal of Further and Higher Education, 33*, 57-69.

Taylor, H. G., Anselmo, M., Foreman, A.L., Schatschneider, C., & Angelopoulos, J. (2000). Utility of kindergarten teacher judgments in identifying early learning problems. *Journal of Learning Disabilities, 33*, 200–210.

The Swedish Education Act. [Skollagen]. (2010). Report no, 800. Retrieved 15 December from:http://www.riksdagen.se>Dokument-Lagar

The Swedish National Agency for Education [Skolverket] (1994). *Läroplan för grundskolan.* (LPO94) [Curriculum for the Compulsory School System, the Pre-School Class and the Leisure-time Centre]. Stockholm: Author.

The Swedish National Agency for Education [Skolverket] (2000). *Kursplaner för grundskolan.* [Syllabi for the compulsory school]. Stockholm: Author.

The Swedish National Agency for Education [Skolverket] (2001). *PISA 2000. Svenska femtonåringars läsförmåga och kunnande i matematik och naturvetenskap i ett internationellt perspektiv.* [PISA 2000. Swedish 15 year-olds reading literacy and proficiency in mathematics and science in an international perspective.] Stockholm, Sweden: Author.

The Swedish National Agency for Education [Skolverket] (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 second language for grades 2-5] Stockholm, Sweden: Author.

The Swedish National Agency for Education [Skolverket] (2006). *Med fokus på läsförståelse. En analys av skillnader och likheter mellan internationella jämförande studier och nationella kursplaner.* [Reading Literacy in Focus. An analysis of differences and similarities between international comparative studies and national syllabuses] Stockholm, Sweden: Author.

The Swedish National Agency for Education. [Skolverket] (2007). *Provbetyg - Slutbetyg - Likvärdig bedömning? En statistisk analys av sambandet mellan nationella prov och slutbetyg i grundskolans årskurs 9, 1998-2006.* [Test grades - Final grades – Equality in assessment? A statistical analysis of the relationship between the national tests and the final grade of compulsory school year 9, 1998-2006]. Stockholm, Sweden: Skolverket.

The Swedish National Agency for Education [Skolverket] (2009). *Likvärdig betygssättning i gymnasieskolan? En analys av sambandet mellan nationella prov och kursbetyg.* [Equality in grading in upper secondary school? An analysis of the relationship between national tests and course grades.]. Stockholm: Author.

The Swedish National Agency for Education. [Skolverket] (2011). *Läroplan för grundskolan, föreskoleklassen och fritidshemmet 2011.* (Lgr11) [Curriculum for the Compulsory School System, the Pre-School Class and the Leisure-time Centre]. Stockholm: Author.

Tholin, J. (2006). *Att Kunna Klara Sig i Okänd Natur. En studie av betyg och betygskriterier - historiska betingelser och implementering av ett nytt system.* [Being able to survive in an unknown

environment. A study of grades and grading criteria – historical factors and implementation of a new system]. (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/16892

Thorsen, C., & Cliffordson, C. (2012). Teachers' Grade Assignment and the Predictive Validity of Criterion-Referenced Grades. *Educational Research and Evaluation, 18*, 153-172.

Toulmin, S. (1958/2003). *The Uses of Argument*. Cambridge, UK: Cambridge University Press.

Tversky, A., Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *4157*, 1124-1131.

Wayne, A. J., & Youngs, P. (2003). Teacher Characteristics and Student Achievement Gains: A Review. *Review of Educational Research*, *73*, 89-122.

Wernersson, I. (1989). *Olika kön samma skola? - En kunskapsöversikt om hur elevernas könstillhörighet påverkar deras skolsituation*. [Different gender same school? – A review of research about how pupil gender influence their school situation]. Stockholm.

Wolming, S. (1998). Validitet - Ett traditionellt begrepp i modern tillämpning. [Validity. A modern approach to a traditional concept]. *Pedagogisk Forskning i Sverige, 3*, 81-103.

Wolming, S., & Wikstrom, C. (2010). The Concept of Validity in Theory and Practice. *Assessment in Education: Principles, Policy & Practice, 17*, 117-132.

Yang, Y. (2003). *Measuring socioeconomic status and its effects at individual and collective levels: A cross-country comparision*. (Doctoral Thesis, University of Gothenburg, Göteborg, Sweden). Retrieved from http://hdl.handle.net/2077/15895