Abstract

Title: Grades and grade assignment: effects of student and school characteristics

Language: English

Keywords: compulsory school, grades, grade assignment, national tests, student background

ISBN: 978-91-7346-636-3

The main aim of the thesis is to explore the dimensionality of grades and how different student and school characteristics influence grades at individual and school levels, by the use of multilevel multivariate techniques. The thesis comprises three empirical studies based on a single large-scale set of data. The participants were 99 070 ninth grade students born in 1987 who left compulsory school in 2003. Grades, national test scores, a student questionnaire and different school characteristic variables have been used.

Previous research indicates that grades are influenced by different student and school characteristics. At the same time there are widely held assumptions that grades form a one-dimensional and objective measure of student knowledge and skills.

The first study focuses on identifying and separating different dimensions in grades, which on the one hand, might be interpreted as expressing variance in knowledge and skills or, on the other, different systematic factors. Another purpose is to examine differences related to gender and family background. The second study focuses on the influence of different student characteristics, such as motivation, interest and parental engagement, on the identified dimensions of grades found in the first study. A further purpose is to investigate how different student characteristics mediate the effect of gender on grades. The purpose of the third study is to investigate the relations between different school characteristic variables, such as size and location of schools, the educational provider and different teacher characteristics, and the identified dimensions of grades.

The results showed that grades are multidimensional and a structure was found that separated the variance in grades into subject-specific dimensions in Swedish, English and mathematics, and into a single common grade dimension. At both the individual and school levels, the largest part of the variance in grades was due to achievement in the different subject areas, measured by the national tests. At both levels, the common grade dimension cut across the three subject grades, which suggests that grades are influenced by factors other than just cognitive abilities. Gender differences were discovered in the Swedish, English and common grade dimensions, with girls having a higher value on these three dimensions at the individual level. Analyses of mediating relations showed that student motivation fully explained the gender differences in the English and common grade dimensions and a major part of the variance in the Swedish dimension. Thus, one explanation why girls receive higher grades is that they have a higher motivation for schoolwork and learning. The results also showed a strong positive relation between parental education and the Swedish, English and mathematics dimensions, whereas at the school level there was a negative relation between parental education and the common grade dimension. The substantial negative relation between parental education and the common grade dimension may be due to compensatory grading practices where schools with a large proportion of students with less well-educated parents have a higher value on this dimension. School level analyses showed that some school characteristics relate highly to grades but, when controlling for parental education, all the relations decreased and, in most cases, became non-significant. A strong positive effect of independently-operated schools on grades was shown to be primarily due to independently-operated schools having students with a higher level of parental education.

Acknowledgements

Thank you to my supervisor Cristina Cliffordson! I have enjoyed every minute of our work and it has been such great fun working with you. You have always had the time for me and my hundreds of questions. I am still fascinated by how quickly you pick up the phone! Of course, I have experienced some difficulties as well but I am grateful for how much I have achieved and all the things I have learned. I am deeply grateful for your help and your generosity in sharing your knowledge with me and, above all, your support and belief in me and my work.

Thank you to my other supervisor Jan-Eric Gustafsson. You have shown a great interest in my work and spent countless hours of your time reading my manuscripts in minute detail and offering critical, positive and encouraging comments. You have also encouraged me to discuss methodological issues, which I have found particularly interesting. I am very pleased to have two such acknowledged experts in their fields as my supervisors.

Thanks FUR! I can only say that I have had such great luck coming to the FUR research group with all these generous, hard-working, competent and sharing people; thank you all of you. Special thanks too to Åsa and Bo for your understanding and gentle approach when I was lost in SPSS. Of course, I enjoyed and appreciated the nice bed in your guestroom, Åsa!

I must also thank Lisbeth Åberg-Bengtsson and Horst Löfgren for their critical and constructive comments on my manuscript.

Thanks Jonas and Malin! Thank you so very much for your delicious dinners, luxurious wine and good company and of course our discussions and of course “my” very, very comfy bed in your house, perfect! My friends Milla and Anders, you have both supported me all through my PhD studies and I am so grateful that you have always been there for me.

Thanks Eva and Gert! My supportive parents who have always believed in everything I do and encouraged me all through my PhD studies and life!

Thanks Aino and Peter Milner! Thank you for always being there for me and my family. A special thank you to Peter for reading my manuscript and for giving me useful comments at a late stage.

Thanks also to Rasmus, Thea, Elvira, Wiggo and Märta! My lovely kids, you are the best! And, for taking such a great responsibility with the horses and dogs when I have been away, you are all amazing!

Thanks Calle, for supporting me and for your love!

Helgagården, November 2008

Alli Klapp Lekholm

Table of contents

The studies I-III

I Klapp Lekholm, A., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: effects of gender and family background. *Educational Research and Evaluation*, *14*(2), 181-199.

II Klapp Lekholm, A., & Cliffordson, C. (in press). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation*.

III Klapp Lekholm, A. (submitted). Effects of school characteristics on grades in compulsory school.

INTRODUCTION

The overall purpose of this thesis is to explore grades and their dimensionality in order to better understand and explain the meaning of grades in terms of what they measure. Three large-scale empirical studies have been conducted, each of which highlights and focuses on the validity of grades. The studies are linked theoretically and methodologically by the use of validation theories and multivariate, multilevel analyses on a single common dataset.

Grades are a controversial issue with which most people have some form of experience; either good or bad. Grades affect individuals' life-chances and concern issues of fairness and individual rights and are thus of vital importance for both individuals as well as society at large. The legitimacy of grades is dependent on the quality of the grading process; i.e. if grades function as intended and whether they affect students in the ways that are assumed. It is thus of great importance that grades are perceived as reliable, valid, comparable and fair. The issue of what grades measure in terms of student knowledge, skills and characteristics forms the main focus of this work and encompasses issues such as validity, fairness, transparency and the comparability of grades.

The assumption among different stakeholders in the educational system and society at large is that grades are an objective measure of student knowledge and skills (in this thesis, the concepts 'student academic knowledge and skills' and 'student achievement' are used interchangeably). This assumption implies that, primarily, grades measure student academic knowledge and skills, and that they do not capture irrelevant variance such as student and school characteristics. This assumption also implies that teachers employ objective and technical measurement principles when carrying out assessment and making grading decisions (McMillan, 2003). However, research has indicated that grade assignment is an ambiguous, complex and problematic exercise where a plethora of factors have an impact on grades (Andersson, 1998; Brookhart, 1991, 1993, 1994; Cliffordson, 2004; Hidi, Renninger & Krapp, 2004; McMillan, 2003; Pilcher, 1994; Wentzel, 1991).

Within the area of research on assessment and grading, a large number of perspectives and concepts are discussed in the research literature. Therefore, it is necessary to conceptualise and define the perspectives and conceptions that are relevant for the purposes of this thesis.

First, teachers' assessment decisions and grading practices constitute an important set of issues since it is teachers who assess, evaluate and grade their students. Other factors that have indicated to have an influence on grades and offer an important perspective concern different student characteristics such as gender and parents' educational background (Rosén, 1998; Yang, 2003). Additionally, school characteristics have also been shown to have an impact on grades. For example, school type has been demonstrated to have an effect on grades (Wikström, 2005). The interpretation of grades is also affected by the characteristics of the grading and educational systems. The purposes, functions and uses of these systems form a fundament of which grades are the outcome.

It has been stressed that grades have both explicit and implicit functions. Bergenhenegouwen (1987) argues that there exist unspoken demands and traditions in the actions of teachers and students. Whilst the concept of the "hidden curriculum" has been widely used in educational research, it has also been criticised due to the limitation of the concept in that it only concerns the curriculum (Bergenhenegouwen, 1987) and that it primarily relates to the implicit aspect in the educational setting. Bergenhenegouwen claims that the concept of "implicit education" is more appropriate since this concept concerns the implicit aspects in evidence throughout the educational system. Implicit education denotes what is implied in the organization of education, the interaction and communication patterns present in schools, as well as the informal rules concerning student behaviour and the tacit demands of students for a certain achievement level and in the exercise of the grading process. It is thus argued here that the implicit functions may confound the validity of grades in such a way that we do not really know or understand what it is that grades actually measure.

The fundamental assumption of the function of grades is that human knowledge and skills vary, and are therefore measurable properties. Grades can be conceptualized as a summary measure of student knowledge and skills, against some point of reference, often in relation to curriculum documents and educational policy. Typically, grades work within a particular grading system which is built upon different theories of science and epistemological beliefs which incorporate assumptions concerning the technical and measurement characteristics of the grading system as well as perceptions of and attitudes towards knowledge. The different grading systems used in different countries are not isolated phenomena but related to the context of the educational system in which they work. In order to explore the dimensionality and variability of grades it is also necessary to take into account the context of the educational system, as well as the construction and purposes of the grading system, of which grades themselves are the outcome.

A large number of studies have highlighted the fact that, in many industrial countries during the 1990s, a paradigm shift concerning the views, perspectives and theories of assessment and grading practices in school took place. Gipps (2001) describes the previous prevailing paradigm as being based on behaviouristic theories of learning and psychometric measurement principles, whereas the new paradigm is based upon contemporary educational and sociological perspectives related to cognitive, constructivist and interpretive perspectives on learning and assessment (Gipps, 2001; McMillan, 2003). However, there is always a risk of simplification when conceptualizing educational science into simple dichotomies. Furthermore, whilst in certain educational systems there is a high degree of teacher autonomy, such that teachers have the main responsibility for assessing their students, evaluating their knowledge and awarding grades, in others, standardized tests and external assessors are used in the grading process.

In the current thesis, the research topic is the assigning of grades in Sweden that takes place in a highly decentralized grading system, itself within the context of a decentralized educational system. The context of grade allocation in Sweden is thus one that is characterized by far-reaching autonomy for teachers. Teachers are responsible for interpreting the curriculum documents, for developing locally-defined criteria for levels of educational achievement, for the assessment of student performances and, finally, for assessing their results and awarding grades. Grades are used as an instrument for selection to the next level in the educational system and have high-stake implications for students. From a historical perspective, teachers in Sweden have, to a large extent, enjoyed far-reaching autonomy in assessing and grading their own students, irrespective of the different grading systems that have been employed (Wedman, 1983).

However, grading systems differ substantially, both over time and in terms of national and educational contexts, with respect to both construction and function, which also affect the conceptualization and meaning of grades. Certain functions have been emphasised during certain periods, such as, for instance, that grades are primarily used for selection or evaluation purposes, or that their primary function is to give information about student attainment.

Two main grading systems which differ substantially with respect to their construction, function and choice of reference points, namely the norm-referenced and the criterion-referenced grading systems, can be identified. These grading systems are connected to different legal, economic and ideological forms of steering. The norm-referenced grading system was constructed based upon principles grounded in the psychometric measurement tradition. The purpose was primarily to construct a system which would function for selection purposes built upon comparisons between the performances of students within a group. The criterion-referenced grading system, on the other hand, was constructed with the purpose of delivering information about student attainment measured against the centrally formulated goals and locally defined criteria for specific subject domains, as well as to function as a diagnostic instrument of student abilities. In the criterion-referenced grading system, the purpose of selection is not the primary focus.

Grades are intended, first and foremost, to be one-dimensional (i.e. that primarily they measure student knowledge) and should not be infected with irrelevant variance. This means that they should be comparable between teachers, schools and over time, thus legitimizing their function both as an instrument of selection and for evaluating the quality of the educational system. However, indications from research on the ambiguity and complex nature of grades suggest that a main concern relates to the issue of validity. The assumption that grades are a fair, reliable and valid measure of student knowledge and skills can be questioned (Brookhart, 1991; Pilcher, 1994; Cizek, Fitzgerald & Rachor, 1995). Thus, there is gap in knowledge between that which it is assumed that grades measure and that which research has indicated grades actually measure. However, it is difficult to generalize much of the research findings since studies are often based on limited sets of data and are sometimes afflicted by problems relating to methodological issues. Therefore, in this research, population data is used in order to conduct a series of large-scale studies using powerful methods in order to investigate the objects of measurement and to identify patterns that are possible to generalize.

Against this background the main purpose of the thesis is to explore the dimensionality of grades in order to better understand, but also to attempt to explain what grades actually measure in terms of student cognitive and non-cognitive abilities. Sources of variability in grades are sought within and between schools and by investigating different student and school characteristics. The three studies that are included in this thesis will be referred to later as:

Study I	Discrepancies between school grades and national tests scores at individual and school level: effects of gender and family background.
Study II	Effects of student characteristics on grades in compulsory school.
Study III	Effects of school characteristics on grades in compulsory school.

The thesis consists of two parts. The first part is an integrative essay and the second part includes the three empirical studies (I-III). The integrative essay contains some main sections which are described below.

The section **Theoretical framework** presents and discusses the research context and problems, definitions, purposes and functions of grades and grade assignment. This section also discusses theories concerning cognitive and non-cognitive abilities, the influence of student and school characteristics on grades and grade assignment, and the validity of grades. The theories are used in order to construct a theoretical framework for the thesis. In the section **Research questions and the purposes of the studies** the research questions are discussed and the specific purposes of each of the three studies are presented. In the **Method** section the analytical techniques used in the studies are described. In the **Summary of the studies** section, the subjects, variables and the findings of the three empirical studies are presented. Thereafter, in the section **Discussion and conclusions** the results of the empirical studies are discussed and the implications of the findings considered. Finally, methodological issues are discussed and suggestions for further research are proposed.

THEORETICAL FRAMEWORK

In this section, the purposes and functions of grades, the grading system and the national tests are discussed since the development of and implications inherent in these systems constitute an important and necessary perspective and starting point for investigating grades. The next area to be covered in this section concerns research and theories regarding different aspects or factors that might have an influence and an effect on grades. First, the conceptualisation of assessment is discussed, which is followed by a discussion of the linkage between assessment and grades. This is followed by a discussion of teachers' perspectives and their decision-making. A further perspective concerns student and school characteristics which have a potential influence on grades and grading. All of these different factors may have an influence on grades and have the potential to explain variability in grades. Encapsulation theory (Gustafsson & Carlstedt, 2006) has also been used in order to discuss ways in which different student abilities are related to each other and reflected in grades.

A major perspective concerns validity theory which offers conceptions of substantial value for understanding the implications of grades and assists in clarifying the relevance of the different factors that might influence grades. The validity concept is of major importance since it focuses the core of the questions being asked in this thesis, namely what it is that grades measure. The use of the validity concept also helps to clarify the discussion of the complexity of grades and what grades measure in terms of student cognitive and non-cognitive abilities.

Grading systems, grades and national tests

One part of this research area concerns the characteristics of the different educational and grading systems and whether the different systems have different purposes, functions and effects, and the nature of their implications for grades. This section highlights the "educational and grading system perspective" which takes into account the importance of the context of which grades are a part. In order to understand and analyse the meaning of grades, this perspective requires some attention.

Grading systems and grades

Over time changes to the grading system in Sweden have taken place and the different grading systems have developed out of the need for grades to fulfil certain functions. Not only have the functions differed, but so too have the points of references for grade assignment. In the two main grading systems, the norm-referenced and the criterion-referenced systems, the points of reference differ substantially. In the norm-referenced system the point of reference is defined by the distribution of grades at the population level, while in the criterion-referenced system the points of reference are defined by goals and criteria. These grading systems are also connected to different legal, economic and ideological forms of control. However, before these two grading systems were developed, a system of subjective and absolute grades was used, where teachers had the entire responsibility for measuring students' knowledge, skills, characteristics and behaviour. Since no external point of reference was used in this system, it suffered from differences in teachers' grading practices between schools and over time. The absolute grades were constructed in terms of a grading scale with different levels assumed to represent absolute knowledge. These grades functioned primarily as a guarantee for a certain requisite level of knowledge that allowed students to continue to the next level in the educational system. However, because of the criticisms of the use of subjective and absolute grades as an instrument for selection to the next level in the educational system, a relative, norm-referenced grading system was developed out of the need for a grading system that was primarily constructed for selection purposes (e.g. SOU 1942:11).

The basic assumption of the norm-referenced system is that individuals' knowledge and skills vary and are hence relative in relation to the knowledge and skills of other students in the same group. The point of reference used in this system was the mean performance of all the students who are in the same group on a national level (subject, year). The norm-referenced grading system was centrally-controlled and based upon the assumption that the distribution of grades followed a normal distribution pattern and that the grades for a class in the core subjects were to be based upon the results achieved on centrally-created standardized tests (Jarl, Kjellgren & Quennerstedt, 2007; Lindensjö & Lundgren, 2000).

It was assumed that since the norm-referenced grading system was based upon measurement properties it was a reliable and valid measure of student knowledge and skills. The basic principle of the norm-referenced grading system was to compare the performances of individuals in a group in order to rank them, and hence so as to function as a reliable and valid instrument for selection purposes. The grading scale used in this system ranged from one to five, with a mean of three and a standard deviation of one.

The norm-referenced grades were used within the Swedish centrally-controlled educational system, where the curricula and syllabus were detailed with respect to content and methods, with some directions being imperative, and others advisory. Apart from the legal control of schools, the distribution of the funding was also an important component in the centrally-controlled system where the Government had the overall control of, and responsibility for, both funding and policy decisions (Jarl et al., 2007). The ideological dimension of control was characterized by detailed directions such as the determination of teaching and advisory materials that were directed towards teachers. The principals and teachers were, in this system, not primarily regarded as professionals but more as civil servants with the task of following the centrally-created regulations. The municipalities had no power to influence the control of the schools in their localities.

The norm-referenced grading system was based on measurement theories relating to the curve of normal distribution, at the population level. However, this system was widely misunderstood among teachers who employed the principles of the normal distribution on their own group of students within the classroom, which, as a consequence, made the system unfair and unreliable (National Agency for Education, 2005; Tholin, 2006). It was also heavily criticised due to its competitive characteristics where students competed for the highest grades in their group and, due to the misunderstanding of how to use the curve of normal distribution, only a small group of students in any class could receive the highest grades. The norm-referenced grading system was also criticised on the grounds that the level of knowledge for a certain grade was unknown since students were only ranked in relation to the national distribution (Cliffordson, 2004; Tholin, 2006).

Glaser and Klaus (1962) introduced the concept of criterion-referenced measurements and several researchers have followed their lead and developed the principles of this system. The basic principle is that a student's level of knowledge is judged in relation to given standards for a particular domain or subject area. The main function of a criterion-referenced grading system is to establish whether the student masters a particular area of knowledge and not to compare inter-group performance in order to rank individual students. The main purpose of this system is to ascertain the level of knowledge a student has mastered, as well as for diagnostic purposes. Another function is to be a normative system in order to implement the epistemological beliefs and attitudes towards knowledge among teachers. This means that the principles of selection were not in focus when the criterion-referenced system was constructed. Theoretically, all students within a criterion-referenced grading system can receive the highest grades, hence making the system useless as an instrument for selection to the next level in the educational system (Wedman, 1983). However, research has established that the criterion-referenced grading system can indeed be used for selection purposes (Cliffordson, 2004, 2008). Criterion-referenced grades have been shown to function at least as well as the norm-referenced grades and indeed considerably better than the Swedish Scholastic Aptitude Test (SweSAT) as a means of predicting student academic success in higher education. The explanation for this is that the grades are a summary measure of several assessments and observations made over extended periods of time, and when used as a selection instrument, the grades are transformed into a grade point average (GPA) which has good measurement characteristics (Cliffordson, 2004, 2008).

In Sweden, the criterion-referenced grading system is a part of the decentralised result- and goal-steered educational system. Nevertheless, whilst the Government and the Swedish Parliament remain the ultimate authorities for policy decisions, the instruments for steering the educational system have changed radically (Jarl et al., 2007). Whilst the goals in the curriculum are steered centrally by legislation, it is the municipalities, schools and individual professionals who decide how the goals should be reached. The transfer of economic responsibility has also provided the municipalities with the overall operative responsibility for running schools.

Whilst on the one hand, teacher autonomy is highly valued by several stakeholders in the Swedish educational system, on the other, problems of subjectivity and reliability can arise and, in a highly decentralized grading system with a great amount of teacher autonomy, the validity of grades may be open to question. Peterson and Woessmann (2007) argue that in a highly decentralized educational system, there is a need for extensive evaluations of students' performances and that these evaluations should be closely connected to the subject domain in question.

Grades are a summative measure derived from several assessment occasions and which, typically, have several purposes and functions, some of which are explicit and some implicit. The explicit functions of grades are, primarily, to provide information of a student's attainment, to increase the motivation to learn, to function as a selection instrument to the next level in the educational system, and to function as an instrument of evaluation. However, the implicit functions of grades and their assignment may have a substantial impact on outcomes in school. Whilst grades are often regarded as having an explicit function in motivating students to learn, research has however consistently indicated that grades may function in ways that motivate some students but leave others demotivated and, indeed, even excluding some students from learning altogether (Ainley, Hidi & Berdorff, 2002; Brookhart & Durkin, 2003; Hidi & Renninger, 2006; Pintrich, 2002). Grades may also be an instrument of power for the teachers and schools and an instrument of control for schools and society at large. The implicit functions of grades incorporate aspects of a hidden curriculum or implicit education (Bergenhenegouwen, 1987) where different aspects of student knowledge, characteristics and behaviour are being measured in the form of grades.

However, the problem with several explicit purposes in one grading system may be that whilst the system may fit very well for one purpose it may not do so for another. A system with several purposes may end up as a system where each discrete purpose can be compromised by trade-offs, thus not really fulfilling any of the purposes. Newton (2007) suggests that an explicit prioritization of the purposes in a grading system is of major importance in order to enable different stakeholders to draw valid inferences from the results. Additionally, the implicit purposes and functions of grades may cause the legitimacy of the grading system to deteriorate.

One important issue that concerns the validity of grades is the point of reference that is used. In the systems of subjective judgements and in the absolute grading system, the point of reference was the teachers themselves; they had the full authority to assess their students without any external control. In a way, it is possible to view these systems as providing the teachers with a high professional status. Later, the norm-referenced grading system was criticized due to the belief that the teachers were deprived of their professional status by the implementation of the measurement principles and the standardized tests that guided the awarding of grades (Carlgren and Marton, 2000). However, during periods in which the norm-referenced grading system was used, teachers had far-reaching autonomy due, amongst other things, to the fact that standardized tests were only available in a few subjects and they were seldom administered. In line with the view that teachers in the norm-referenced grading system were deprived of their professional status, it has been argued by some researchers (Carlgren & Marton, 2000) that through the use of a criterion-referenced grading system, the teachers would regain some of their professional status by means of the creation and application of local criteria and classroom assessment.

The issue of the point of reference for grades, and hence too in the grading practices, concerns the overall explicit assumption that teachers only assess and award grades on the basis of student subject knowledge and skills. A large amount of research has indicated that grades are influenced by subjectivity and that factors such as different student characteristics, teachers' grading practices and systematic differences within and between schools, exert an influence on the assignment of grades (Alexander, 1935; Brookhart, 1991, 1993, 1994; Cliffordson, 2004; Pilcher, 1994; Wikström, 2005). Several studies have indicated that, when carrying out assessment and awarding grades, teachers take account of student effort, personality and willingness to cooperate, and that they expect students to listen attentively, to follow instructions and to control their behaviour (Lane, Givner & Pierson, 2004). It seems thus as if cognitive as well as non-cognitive factors are of importance for understanding the grading processes in school (Gipps & Murphy, 1994; Peterson & Woessmann, 2007).

The issue of the kind of student knowledge and skills that grades are intended to measure is of vital importance irrespective of the grading system that is employed. Despite the substantial differences concerning the construction of the different grading systems, and hence the functions of grades, these differences may not influence the teachers' grading practices to any substantial degree. As has already been observed, grades from the norm-referenced and criterion-referenced grading systems have a similar prognostic validity (Cliffordson, 2004, 2008). Previous research has also shown that grades from the preceding stage in the educational system are the instrument with the highest validity for predicting academic success (Carroll, 1982; Gustafsson & Carlstedt, 2006).

Most educational systems use some sort of national tests as well as internationally developed tests such as PIRLS and TIMSS. The national tests are often used in order to calibrate grading, to monitor the educational system and to hold schools accountable for their results. In most educational systems, there is a relation between the tests and the grades, in that the tests are curriculum-based which means that they are based on the focal constructions, learning criteria and goals articulated in the curriculum. The history, functions and effects of the national tests are of importance for the current research since the results from the Swedish national tests are used as indicators of student academic or subject knowledge throughout the thesis.

The National tests

The first systematically developed national assessment tests for compulsory school in Sweden were introduced in the late 1940s to support use of the norm-referenced grading system. The increasing number of applications to higher education placed higher demands on the calibration of grades since the grade levels were found to vary substantially. The tests were used in order to function as an instrument to calibrate the assessment and awarding of grades (National Agency for Education, 2005; SOU 1942:11).

In the norm-referenced grading system, teachers were supposed to grade their students in relation to the normal distribution on a national level, which made it necessary to develop standardized tests, in order to calibrate grades. The tests functioned as a benchmark for the class average and distribution of grades and teachers' mean final grades were only permitted to diverge 0.2 units of a standard deviation from the mean grade score the class obtained on the standardized test. If there was a larger difference the teachers had to provide a written explanation of the reason for this to their principal. This test system was developed on the basis of strict statistical principles and it was heavily standardized in terms of strict rules concerning the routines of the tests and that similar conditions existed in all test situations at every school. The tests only ranked the students in relation to the results from the same year, which means that an equivalent grade did not necessarily represent the same level of student knowledge over time. Standardized tests were used only in three subjects, Swedish, English and mathematics and their sole purpose was to calibrate grading.

The present system of national tests in Sweden has several purposes including contributing to increased goal attainment, exemplifying course goals and grading criteria, assisting in the process of setting fair and reliable grades, identifying students' strengths and weaknesses and for monitoring the educational system. However, the national tests should not influence the teachers to choose any particular teaching method or function as final examination tests (National Agency for Education, 2005). In Sweden, the national tests are used in order to support equity, reliability and fairness in grading, and as such provide support for teachers' assessment and grading (National Agency for Education, 2005).

A characteristic of the current Swedish national tests is that the marking of the tests is carried out by the teachers themselves. It has been concluded that the advantages of decentralized marking outweigh the disadvantages (National Agency for Education, 2005). For instance, when teachers mark the tests of their own students, the tests function as a form of competence enhancement and the professionalism of teachers is strengthened. This system is also conceptualized as a more cost-effective system since no external marking procedure is necessary. Some argue that the workload for the teachers increases and that the time spent on marking could be used more efficiently, for instance in direct classroom teaching. Nevertheless, and not withstanding these criticisms, a large majority of teachers support the decentralized marking system (National Agency for Education, 2005).

It is argued that the national tests are multidimensional which implies that they measure several student abilities which, together, form a hierarchical structure (Gustafsson, 2001; Gustafsson & Carlstedt, 2006; Åberg-Bengtsson & Erickson, 2006). Using confirmatory factor analysis Carlstedt and Gustafsson (2006) developed a three level model where the highest level generates a general factor influencing overall performances and achievement. The second order factors, which are less general, are related to more specific student abilities, and the factors at the lowest level are related to even narrower and highly specialized student abilities. Åberg-Bengtsson and Erickson (2006) investigated the Swedish national tests in Swedish, English and mathematics using two-level structural equation modelling. They concluded that the national tests have a hierarchical structure both according to the structure of student abilities, divided into more and less general and specific abilities, but also according to the hierarchical structure of educational data where students in the same school are more similar to one another compared to students in other schools. Åberg-Bengtsson and Erickson (2006) also demonstrated that about 12 per cent of the variance in the national tests was due to school differences. The multidimensionality of the tests was associated with both the content and formats of the different subjects.

Another issue concerns gender differences in the national tests. Whilst the national tests in Sweden reveal certain gender differences, these are not of the magnitude of those in evidence in teacher-awarded grades. Whilst these gender-related differences are most apparent in the national test for Swedish, the differences in the national test for English are only moderate whereas in mathematics they are very small (National Agency for Education, 2003, 2005).

Analyses of the Swedish national tests conducted by the National Agency for Education (2005) show that the largest proportion of students reaches the goals in English and Swedish, whilst in mathematics it is the least proportion of students who do so. Students in independently-operated schools score higher levels of achievement on the national tests compared to students in municipally-operated schools.

The Swedish national tests are developed by the National Agency for Education in cooperation with several universities. The tests are distributed to the students by the teachers who also assess and grade the tests. It could be argued that the tests are teacher-owned and therefore afflicted with irrelevant variance and, consequently, not reliable. However, teachers have access to extensive guidelines each with a series of annotated benchmarks, as well as websites where a large number of examples of student attainment on the tests are offered. Teachers are also strongly encouraged to cooperate in the assessment procedures coupled to these tests. The tests have several different formats such as oral, written and listening tasks which are administered both individually and in groups, and there are several assessment occasions which make it hard for students to develop "test wiseness" skills (Åberg-Bengtsson & Erickson, 2006). The multiple assessment occasions and the varied formats of the tests, as well as rigorous pre-testing procedures, ensure that reliability is high.

Assessments and teachers' decision making

In school, students are assessed more or less continually, in assessment events prior, during and after instruction (McMillan, 2003) and teachers regularly make many formative and summative assessment decisions. Assessments are often defined as having a "formative purpose" or being a "summative judgement" (Newton, 2007). The definition of formative assessment means that assessments are primarily used in order to give students relevant feedback to enhance their learning and an important feature of formative assessment is that students need an opportunity to put the feedback they receive into practice (Segers, 2008). Summative assessment is a summary of many different sub-assessments which often result in a judgement and a summary score or grade, which can be used for selection purposes or as an evaluation of the school system.

There are also different kinds of assessments, such as classroom assessment, with a vast variety of assessment events that take place on a daily basis in the classroom, where teachers observe student performances over an extended period of time. Performance assessment is another type of assessment where students are assessed during actual performance which is thus argued to be a form of direct assessment (Kane, Crooks, & Cohen, 1999). Other types of assessment involve high-stakes achievement tests with varying levels of standardization, which may or may not be judged by external assessors.

Grades in Sweden are primarily based on classroom assessments. Classroom assessment is an overall definition of several types of sub-assessment such as performance assessments, portfolios, self-assessments but also different kinds of tests. McMillan (2003) has developed a theory of classroom assessment and teachers' assessment decisions in which he contrasts the traditional measurement principles with recent theories about learning. He argues that whilst the use of traditional measurement principles for developing large-scale objective tests is necessary for that specific purpose, the same principles however may be hard to apply in the classroom context. In the classroom many factors, such as teacher beliefs and values, classroom realities and external aspects, all impact upon the assessment decisions that teachers make. These different 'themes' of teacher beliefs, values, classroom realities, external factors, decision making rationale, assessment practices and grading practices, together form the different perspectives in this theory (McMillan, 2003).

The context of the classroom contains many interacting and competing forces that teachers must address (Airasian & Jones, 1993). It has been emphasized that measurement principles must be modified and related to the classroom assessment processes, so that there is a consistency between assessment, learning and teaching in order to develop a sound measurement theory for classroom assessment (McMillan, 2003). See Figure 1.

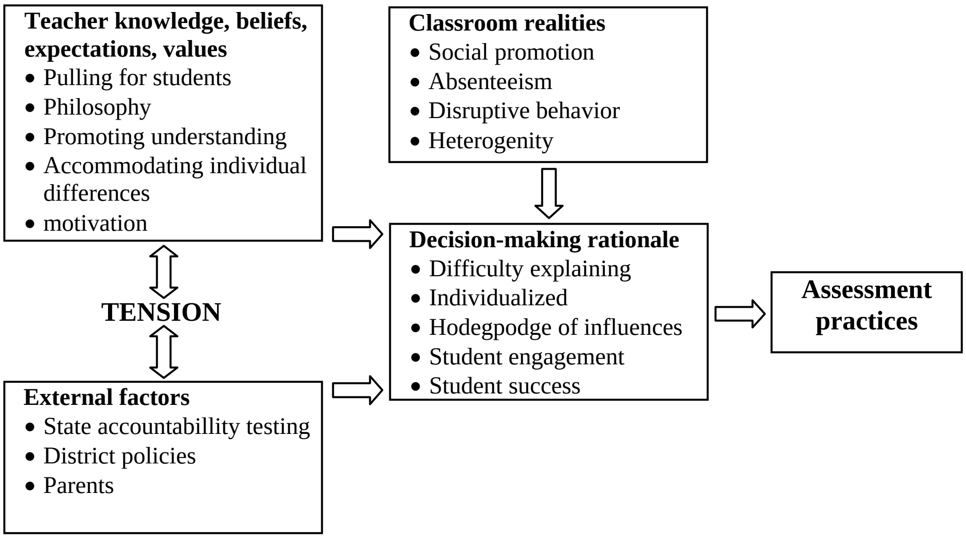


Figure 1. A modified figure of teachers' classroom assessment decision-making (McMillan, 2003).

Since grades in Sweden are based on classroom assessment, classroom assessment theory seems relevant as a means of highlighting the many factors that influence grades. The different factors that McMillan (2003) has found in his research, and which he claims have an impact on the assessment and decision-making in the classroom, concern teachers' internal beliefs and values about student learning and motivation, the classroom realities and external factors such as high-stake tests. These different factors lead teachers to make decisions that are prefaced on different rationales. Teachers want their students to succeed and therefore they adjust and modify their assessment events to provide students with good opportunities to be successful. Another perspective includes teachers' overall philosophical beliefs about education and goals for students, which also include non-cognitive abilities. According to McMillan, teachers were also found to use constructed response assessment in order to check whether students had reached a deeper understanding. Teachers were found to change the assessments in line with different student characteristics and to base their assessments and grading decisions on these different factors so that students would be more motivated and engaged in their learning.

Classroom realities concern all those different factors that influence the work in the classroom. The classroom environment, the students' home backgrounds, student attitudes and behaviour all interact with teachers' beliefs and goals in their procedures of instruction and assessment. Placing a value on these factors in assessment and decision-making processes creates problems for teachers. According to McMillan (2003), the external factors that have an impact on teachers' assessment and decision-making practices are high-stake tests, policies and parents. In line with several studies, McMillan argues that high-stake tests have a strong influence on both what is assessed and how classroom assessments are conducted.

Several researchers (Brookhart & Durkin, 2003; McMillan, 2003; Moss, 1994) have emphasised that the definitions of assessment and assessment decisions must be reconceptualised in order to better align with the contemporary view of learning and teaching. They claim that this implies that it is not appropriate to use traditional measurement principles when validating classroom assessments and identify the need for a validity model appropriate for classroom assessment decisions. Collecting, interpreting and evaluating information is an ongoing process of classroom assessment and McMillan (2003) argues that the assessment decision-making conducted by teachers is a subjective and intuitive process where teachers need to have the competence and ability to evaluate their interpretations and the effects that these have on assessment decisions. This theory focuses on ways in which to validate classroom assessment and teachers' assessment decisions in the classroom, not, as traditionally been the case, by measurement principles, but by using a wider definition of validity which recognizes that assessment decisions are subjective and intuitive and which also incorporates the consequences of such decisions. However, it is emphasized that it is necessary to develop teachers' analytical skills in order to reach valid assessments (Shepard, 2000). Consequently, assessment decisions should be regarded as closely related to validity, in that they involve an ongoing process of gathering evidence to support or reject a certain interpretation and to ascertain whether this decision is sound, trustworthy and legitimate.

Messick (1994) emphasizes that different kinds of assessments must be evaluated using the same validity criteria, both evidential and consequential. In educational settings, the enthusiastic embracing of the consequential aspect of validity may weaken the evidential aspects. Further, basic assessment issues such as reliability, validity, comparability and fairness must be uniformly addressed, irrespective of the assessment measures that are selected. Messick (1994) argues that:

> This is so because validity, reliability, comparability, and fairness are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgements and decisions are made (p.7).

Because teachers are involved in processes of assessing and grading their students, this makes the rationales and objectives of their decisions of great importance and interest when investigating variability in grades. It has been found that teachers take account of students' levels of achievement when grading, in that high-ability students are, to a larger extent, graded on the basis of achievement only, whereas low-ability students are graded on the basis of both achievement and non-achievement (Stiggins, Frisbie & Griswold, 1989). It has also been claimed that there exist different grading practices among teachers due to the subject they teach. For example, teachers in non-academic subjects tend to attach greater weight to non-achievement factors than their colleagues in academic subjects (Agnew, 1985; Tholin, 2006). Pilcher (1994) found that teachers in different subjects graded their students based on different rationales and took varying account of different student characteristics when grading. For example, on the one hand, teachers in mathematics graded their students in a "right or wrong" manner and they described their own grading practice as objective, making a clear distinction between students' knowledge in mathematics and their effort. They also took account of student effort when students were said to be on the borderline of achieving a higher grade. On the other hand, teachers of English graded their students without making a clear distinction between subject knowledge and effort, and they described their grading practices as subjective. Pilcher (1994) also argues that the interpretation of grades differs among the different stakeholders in the educational system. Teachers assign grades by taking account of cognitive, affective and motivational behaviours, which suggests that cognitive abilities, effort and attitudes are all factors of importance in grades. Pilcher also argues that teachers' grading rationales fluctuate as student ability levels increase and decrease, and that teachers use different student characteristics in order to adjust student achievement.

Some researchers have indicated that teachers, by taking non-achievement into consideration in grade setting practices, do so to protect students, teachers and schools from negative consequences (Cross & Frary, 1999). One example that highlights this is that, according to certain research, independently-operated schools in Sweden award higher grades and that this is explained by teachers' grading practice in that teachers experience a sense of pressure to award higher grades in order to attract students and that it is the competition for numbers that is influencing these practices (Wikström, 2005). However, this has been questioned by the National Agency for Education (2005) which believes that differences between schools seem to be related to other school-related characteristics, such as, for example, school culture. School culture may determine the extent to which teachers take account of different aspects of student abilities, skills and characteristics when assigning grades (Agnew, 1985; Cizek et al., 1995).

In Sweden, classroom assessment is the basis for judging student performance and for awarding grades. No final exams are used. Nevertheless grades have a high-stake status for students since they are used for selection purposes.

It is possible to view the different kinds of assessment along a continuum where strictly defined and short periods of assessment form one end whilst, at the other, there are loosely defined assessments which take place over extensive periods of time. On the one hand, assessments may be conceptualized narrowly where a student performance is valued in strict relation to certain criteria and during a short period of time. On the other, assessment may be understood as including the entire process of collecting information, evaluating that information, assessing it and, finally, making a decision as to the score or grade to award (Airaisian, 1993). This implies that the whole process takes place over an extended period of time and includes many assessment events on many occasions. However, even though all of these different kinds of assessment can be high-stakes for the individual students involved, the number of occasions and the formats of the testing affect the reliability of the overall assessment. Despite the reliability of the assessments, an overall and fundamental issue concerns the type of knowledge, skills and abilities that form the subject of assessment.

Whilst formative assessment has the explicit purpose of enhancing student learning, summative assessment is based on a summary of several different assessments and is primarily used as an evaluation and selection instrument. Grades are a summary score and an overall assessment and do not have a primary function of enhancing student learning.

Student characteristics

A large number of studies have indicated that grades not only reflect student subject knowledge and skills, but also different student characteristics such as motivation, interest, effort, classroom behaviour and homework completion (Cizek et al., 1995; Cross & Frary, 1999; Manke & Loyd, 1990; McMillan, Myran & Workman, 2002).

The influence of different student characteristics, such as gender and socio-economic background on grades, has been investigated in several studies which have shown that student background exerts an influence on educational outcomes and grades (Coleman et al., 1966; Rosén, 1998; Yang, 2003). In most industrialized countries girls perform better in school and obtain better grades than boys (National Agency for Education, 2003) and indeed the same is true for students with high socio-economic status (SES) (Hanushek & Luque, 2003; Yang, 2003). A substantial body of studies has consistently found a positive relationship between students' family background and educational outcomes (Coleman et al., 1966; Yang, 2003). The measure of family socio-economic status used in research is often a composite of different measures such as family income, level of parental education and parents' occupation. Even though the definition of family SES differs, these basic dimensions of SES are generally accepted. These dimensions are often seen as a unitary concept and as different kinds of capital (e.g., economic, cultural, educational, and social) that influence grade outcomes.

On the individual level, measures of SES explain about ten per cent of the variance of academic achievement, regardless of how SES and school outcomes are measured (Yang, 2003). However, there may be reason to look upon the SES definition as a more diffuse, non-unitary concept, since it has been shown that different dimensions of SES relate differently to student achievement (Bloom, 1976; Yang, 2003). Yang (2003) argues that different dimensions of SES, such as economic and cultural capital, influence student achievement in different ways and hence that the SES concept is multidimensional. Yang also argues that SES is a hierarchical measure that works at different levels, for example individual and school levels. She suggests that it could be more fruitful to separately investigate the different aspects of SES in relation to outcome variables/student achievement. The results of Yang's studies (2003) reveal that although students from families with ample cultural and educational capital are more likely to be higher achievers, the economic status of the family did not have any positive impact on school achievement. It is therefore for this reason that parental education is used as a family background characteristic in this thesis.

Research has found that the influence of gender on grades is significant and that gender differences are much greater in teacher-awarded grades in comparison to results on achievement tests (National Agency for Education, 2005, 2007). However, the gender differences found in the achievement tests are most apparent in the language domain (National Agency for Education, 2005). In a study of grades in compulsory school, Andersson (1998) found a strong general school-achievement factor, which influenced the grades in all subjects and where girls had a higher mean on this factor. Gustafsson and Balke (1993) had, in an earlier study, also found a strong general school-achievement factor, which correlated .60 with earlier (three years) measures of student general cognitive ability. This implies that even though general school-achievement overlaps with general cognitive ability, a large amount of variance in achievement is however independent of cognitive ability.

The gender differences in grades, it has been suggested by several researchers, stem from the different learning approaches that girls and boys develop during their upbringing (Gipps & Murphy, 1994; Murphy, 2000; Rosén, 1998). Not only do girls seem to be better prepared, but they have also developed interests that are more in line with school activities (Murphy, 2000). The different learning approaches that girls and boys develop may also influence other student characteristics such as motivation, self-perception of competence and social skills, all of which have an influence on grades. Murphy (2000) suggests that the different behaviours that girls and boys develop may be due to the different kinds of experiences and expectations they meet, thus leading to different self-perceptions, perceptions of others and of the social environment. The different expectations from the environment will often lead girls and boys towards different experiences, which influences their perceptions of, for example, male and female domains and of their relative competencies. Girls are expected by students, teachers, parents and society at large to do better in language domains, whereas boys are expected to do better in quantitative domains. These socially-derived presumptions are likely to be more related to behaviours and social skills than to cognitive abilities. One reason for the better performances of girls in school may be due to other characteristics such as motivation, interest, effort and adjustment, and that some of these characteristics are of importance for students' learning which, in turn, affects performance.

Manke and Loyd (1990) found that different student characteristics, such as effort, behaviour, personality, and homework completion, are commonly taken into consideration when teachers assign grades. Wentzel (1989, 1991) found that students' grade point averages were positively influenced by student motivation and effort. Student effort also seems to be a key criterion for students on the borderline to achieving higher grades that is taken into account by teachers in their assigning of grades (McMillan, 2003). Moreover, student interest also seems to influence the awarding of grades. Hidi and Renninger (2006) claim that student interest is characterized by varying amounts of affect, knowledge and value and hence includes cognitive as well as non-cognitive aspects. According to some researchers, students have specific individual interests, some of which may be in line with the goals of classroom learning, while other individual interests are not (Ainley et al., 2002). Pintrich (2002) argues that students may have domain-specific interests as well as a more general individual interest in learning. A distinction is made between individual and situational interest both of which are perceived as motivational since they concern perception and the range of actions that a student considers possible (Bergin, 1999; Brookhart & Durkin, 2003). Students with a general interest for learning seem to develop both general and specific goals which they try to achieve with an attitude and approach that helps them understand new phenomena. It is also argued that student general interest for learning is closely related to motivation (Brookhart & Durkin, 2003). For instance, the overall concept of interest may be perceived as having two aspects, the first of which concerns specific interest in some specific issue or subject, whilst the second relates to a general interest in learning and, primarily, concerns attitudes towards learning as a whole. The concept of general interest for learning may thus be argued to be closely related to the concept of motivation.

However, it is not clear from previous research whether student characteristics are related to grades because they influence the development of student knowledge and skills, which in turn affect grades, and thus is an indirect effect, or whether different student characteristics have a direct effect on grades. The distinction between indirect and direct effects encompasses several aspects of importance. The question that needs to be asked is thus which characteristics have a direct effect on grades and which have an indirect effect. Furthermore, questions about the nature of the variance in grades, i.e. whether it is relevant or irrelevant, also need to be addressed.

School characteristics

The effects of different school characteristics on students' school achievement have been of major interest for stakeholders, researchers and society at large. A large amount of research has indicated that certain demographic variables exert an influence on student performances and grades (Coleman et al., 1966; Darling-Hammond, 1999, 2000; Darling-Hammond, Ross & Milliken, 2006; Gustafsson & Myrberg, 2002; Hanushek, 1989, 1997). Yang (2003) found that school-level SES had a major impact on the variance in school performance, and that the relation between school-level achievement and school SES varied substantially between the different countries in her studies. It has also been indicated that different factors related to schooling, such as teacher competence and quality, teacher density, school-size, location and the educational provider (e.g. municipally- or independently-operated schools) also influences student performances and grades.

Peterson and Woessmann (2007) have argued that in the US, the decentralized educational system and school policy, coupled with extensive residential segregation has had the effect of reinforcing existing social patterns. In a similar vein Machin (2007) has demonstrated that in Britain, rather than mitigating the effects of social distinctions, the expansion of higher education has in fact reinforced and enhanced social inequality. Children from families with well-educated parents are more likely to pursue advanced studies than children from less well-educated families (Peterson & Woessmann, 2007). Thus, family background still has a powerful influence on schooling, even in societies that have expanded their educational systems. One explicit reason for expanding the educational sector is to offer students from less well-educated families improved opportunities. However, Peterson and Woessmann (2007) argue that educational expansion in itself does not necessarily enhance human capital or facilitate social mobility. Indeed, this pattern does not seem to vary significantly between countries (Peterson & Woessmann, 2007). Additionally, Hanushek (2007) has argued that decentralized educational systems where school-choice voucher systems are operated improves the competition among schools thus enhancing the quality of education and resulting in a more effective school and educational system.

The widely cited Coleman et al. (1966) report demonstrated that student learning and performance in school was heavily influenced by demographic characteristics such as mother's and father's education and family income. The study also showed that, even when controlling for the demographic characteristics, different school characteristics had only marginal effects on student performance. The result indicated that school characteristics such as class- and school-size, teachers' qualifications and ability, and classroom equipment did not contribute substantially to student performance and learning outcomes. Ever since the Coleman report was published, the results of the research conducted into the effects of different school characteristics and their influence on school performances have been inconsistent. Hanushek (1997, 2007) has conducted several meta-analyses of the relations between economic resources and educational results, leading him to the conclusion that there is no relation between economic resources and school outcomes. Similarly, a number of European researchers have found that policies and resources focused on disadvantaged students have not improved performance or in any sense ameliorated individual inequalities (Leuven & Oosterbeek, 2007).

However, several researchers argue that the research outlined above suffers from problems relating to a lack of control variables and models and methods that are excessively parsimonious thus causing misleading inferences to be drawn. Hedges, Laine and Greenwald (1994) conducted a meta-analysis on the same data material as Hanushek and found positive relations between resources and results and, furthermore, that an increase in expenditure per students would increase the results quite substantially. In a later study, Greenwald, Hedges and Laine (1996) conclude that school resources are systematically related to student achievement and that global resource variables show a strong and consistent relation with achievement. They also argue that different teacher quality characteristics, such as teacher ability, education and experience, had a very strong relation to achievement. Additionally, Wenglinsky (1998) argues that different kinds of resources have different implications for different groups of students and that the combination of different resources may enhance results.

An important distinction when discussing the influence of different school characteristics on student achievement and grades is whether they influence achievement, or whether they in fact impact on grading practices so that, at any given achievement level, higher or lower grades are obtained. For example, Wikström (2005) found that independently-operated schools award higher grades relative to students' performance on the Swedish Scholastic Assessment Test (SweSAT) in comparison to municipally-operated schools, even after controlling for parental education. Wikström suggests that the result shows that independently-operated schools award higher grades than municipally-operated schools given the same achievement levels. However, an alternative interpretation of this result is that student achievement was not properly measured by the SweSAT.

Research on the 'independent school effect' is somewhat inconsistent and whilst some studies have found a strong, positive effect of independently-operated schools on grades, the results of other studies have only found small and often insignificant relations (Somers, McEwan, & Willms, 2001). In Sweden, Bergström and Sandström (2001) investigated differences between independently- and municipally-operated schools and found positive effects of independent schools on grades. Using the grade point average (GPA) and controlling for background variables they found that teachers in independently-operated schools award higher grades. They also claimed that independently-operated schools increase the competitive allocation of resources in municipally-operated schools and that the competition therefore produces higher standards (Bergström & Sandström, 2001).

Somers et al. (2001) agree that the relationship between independently-operated schools and grades is significant and quite substantially positive but, when controlling for the effects of student background and for peer-group characteristics, the positive relation between independently-operated schools and grades decreases and, in some instances, is even negative (Gustafsson & Myrberg, 2002; McEwan, 2000; Figlio & Stone, 1999; Somers et al., 2001).

In Sweden, Wikström (2005) found that, given the same achievement levels on the SweSAT test, there was a school size effect on grades, such that teachers in small schools (<300 students) award higher grades in comparison to second smallest (300-499 students) and very large (>1000 students) schools, the second largest schools (500-1000) was the reference category. Wikström (2005) concluded that this result may be due to a pressure for higher grading, in combination with different grade assignment practices among schools, due to the fact that grades function as an instrument for selection to the next level in the educational system. Another suggested explanation is that, as a result of the voucher system, small schools are more economically vulnerable to the loss of students and therefore award higher grades.

Darling-Hammond et al. (2006) reviewed studies on the relation between school-size and achievement and found that students in small schools have higher levels of achievement. It is suggested that this effect may be due to small schools either being "small by design" or "small by default". Schools in rural areas are often found to be "small by default" whereas independently-operated schools are more likely to be "small by design". These different characteristics may have implications for the relation between school size and results. School size may influence students' results indirectly through other school characteristic factors, such as a strong academic curriculum, a shared school mission, or differences in teachers' grading practices. There seem to exist confounding relations between school size, location and different demographic variables (Darling-Hammond et al., 2006; Ready & Lee, 2006).

Studies which investigate the relation between the location of schools and grades are rare. There are some studies that have investigated the relation between the location of schools and different achievement tests. Using two-level modelling, Åberg-Bengtsson (2004) investigated the relation between reading achievement among Swedish nine-year-olds and school size and location. The result showed small and mostly insignificant relations between rurality, school size and reading achievement. According to the National Agency for Education (2000) and statistics on the leaving certificates from compulsory school, students in sparsely populated municipalities attained better results, probably benefiting from higher teacher density and better working conditions in school. However, the results also showed that, compared to schools located in urban areas, rural schools had a larger numbers of unqualified teachers.

Analyses from the International Education Association (IEA) show that, in a global perspective, students in rural areas have lower achievement levels on literacy tests in comparison to students living in urban areas. However, it was also found that, in highly developed countries, students in rural areas achieved as well as, or even better than students in urban areas (Åberg-Bengtsson, 2004).

Some studies have investigated the effects of rural and urban communities on different school-related variables (Stanley, Comello, Edwards & Marquart, 2008) and found mediating effects of parental education, income and rurality on school adjustment. Roscingo and Crowely (2001) and Williams (2001) found, for example, that the relation between rurality and achievement disappeared when socio-economic status was taken into account.

Teachers' formal competence in the form of appropriate teaching qualifications has been found to have a positive effect on students' results (Wayne & Youngs, 2003). A few Swedish studies also found a strong positive effect of teacher qualifications (Andersson & Waldenström, 2007; Myrberg, 2007). According to Hattie, Biggs and Purdie (1996) it is primarily the 'teacher factor' that has a strong effect on students' results. They argue that what matters most are different teacher characteristics such as, for example, giving relevant feedback (i.e. where students receive feedback information on a task and how to improve it), direct instruction and reciprocal teaching.

Research also indicates that teachers' evaluation methods and grading practices may differ between individuals as well as of a result of school characteristics. A central area of inquiry is teachers' grading practices and, as Sadler (1989) states "The focus is on judgements about the quality of student work; who makes the judgements, how they are made, how they may be refined, and how they may be put to use in bringing about improvement" (p.119). Black and Wiliam (1998) claim that assessment processes are, at heart, social processes, "taking place in social settings, conducted by, on and for social actors". The classroom assessment and evaluation culture in the classroom has been described by Stiggins and Bridgeford (1985) and Stiggins and Conklin (1992) as being more related to teacher practices than student perceptions. Teachers' general approach to assessment and grades creates a certain assessment and grade environment in the classroom.

Research also suggests that the consideration of non-cognitive abilities when awarding grades may serve to protect students, teachers and schools from negative consequences (Cross & Frary, 1999; Wikström, 2005 Agnew, 1985). Cizek et al. (1995) also suggest that differences in grade setting practices may be related to school characteristics such as the nature of the educational provider, and that school culture may play a part in determining the extent to which teachers take account of student knowledge, skills and characteristics when awarding grades. Such behaviour can differ in terms of its correspondence with the goals and criteria in the grading system. Brookhart (1991) suggests that, by taking both cognitive and non-cognitive abilities into account in the grade setting practices, teachers maximize student outcomes which may be perceived as a win-win situation in which students, teachers and schools all benefit. An issue of importance concerns the influence of the different school characteristics given a certain achievement level, and the extent to which schools with different characteristics take account of cognitive and non-cognitive abilities in the award of grades.

The model that has been developed by McMillan (2003) indicates that assessment practices and decision-making are constituted by different perspectives where the agents in the educational system are in focus. External factors such as policies, grading systems and different school characteristics, such as the educational provider, form one perspective whilst teachers' interpretations, knowledge and beliefs concerning assessment and decision-making form another, with classroom realities such as different student characteristics and behaviour creating an additional perspective. These different perspectives influence assessment practices, decision-making and the awarding of grades and are thus of importance when exploring the dimensionality and meaning of grades.

Different factors and perspectives have been discussed in this section and, in order to highlight how we can understand the meaning, implications and interpretation of grades, theories of encapsulation and validity are discussed in the following two sections.

Encapsulation Theory

Encapsulation Theory (Gustafsson & Carlstedt, 2006) takes as its point of departure the theory of Investment developed by Catell (1971, 1987), who argues that general fluid abilities (*Gf*) predict general crystallized abilities (*Gc*). Investment theory suggests that *Gf* is a single, general ability that has an overall influence on the development of knowledge and skills, which are required by children through the experiences and practices they face during their upbringing. Cattell refers to such knowledge and skills as crystallized abilities (*Gc*). *Gf* is thus being invested in all types of situations and educational settings and is a basic ability that influences the development of *Gc*, which suggests that *Gc* is partially a function of the level of *Gf* at earlier points in time.

Lohman (2004) argues that it is possible to view different measures or instruments of abilities along a continuum from a fluid end to a crystallized end. At the fluid end, we find measures of primarily cognitive abilities (*Gf*) such as intelligence tests, whereas at the crystallized end we find measures such as aptitude tests and school grades (*Gc*). School and course grades are thus on the outer end of the crystallized continuum and reflect a broad range of student cognitive and non-cognitive abilities such as motivation, self-discipline and interest.

Much research has shown that measures of *Gc* are good predictors of achievement and one interpretation of this is that *Gc* includes both *Gf* and individual differences in knowledge and skills which are of importance for further learning and achievement (Gustafsson & Carlstedt, 2006). Research also has shown that the grade point average (GPA) is a better predictor of achievement in higher education than different college entrance tests (Cliffordson, 2004, 2008). Thus, it is suggested that grades reflect student knowledge and skills which represent both cognitive and non-cognitive abilities and that the different non-cognitive variables determine success in the educational system and for achievement in higher education. Thus, *Gf* predicts general crystallized abilities (*Gc*) which in turn predicts grades. This explains why *Gf* does not have any direct effect on grades. Encapsulation theory suggests that the *Gf* information is encapsulated in the *Gc* measure, and that the *Gc* information is encapsulated in grades (Gustafsson & Carlstedt, 2006). See Figure 2.

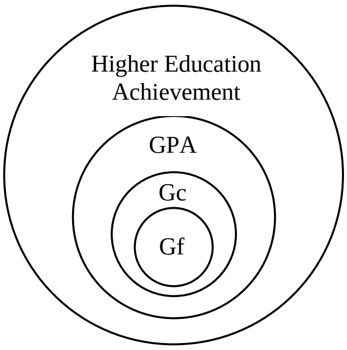


Figure 2. A modified figure of the theory of Encapsulation (Gustafsson & Carlstedt, 2006).

The theory of Encapsulation suggests that grades are a measure of different dimensions of abilities, and that this explains why grades are the best predictive instrument for achievement at the next level of education. Encapsulation theory highlights the issue of validity in grades and the types of variance that are relevant components of variance in grades. In order to clarify the relevance of different sources, the issue of relevant variance in grades may be clarified by the help of validity theories.

Validity

In order to understand and discuss the nature and quality of the interpretations of grades, Messick's unified validity theory appears to be particularly useful in that it includes several important aspects for validating assessment in schools and, in particular, the consequences of the assessments which seem to be of major importance for educational assessment. However, Messick's theory is primarily theoretical and somewhat abstract and gives little guidance in terms of its practical application in educational settings. In the section that follows I will not only discuss Messick's validity theory but also a number of other relevant theories which offer further understanding about the complex theoretical concepts involved in validity theory.

Grades are a high-stake summative assessment and score, especially in educational systems where they are used as an instrument for selection. The consequences for those being assessed are serious, often affecting their chances of entering further education. The fairness, validity and reliability of grades are thus of fundamental concern and of major importance in order to legitimize their use. It could be argued that classroom assessment does not require high levels of reliability since it is not a high-stake situation and that the inferences made may be corrected later on. However, if classroom assessment is the primary assessment method for awarding grades, this argument does not hold. On the contrary, the individual judgements made by teachers in the classrooms, affect the level of quality and validity in terms of the nature of the student knowledge and skills that are being assessed and whether these constructs are in line with the purpose of the assessment and curriculum demands.

The question of reliability concerns, among other things, the level of independence of the individual conducting the assessment and awarding the grades. The assessment of students' performances and attainment should, in a perfect world, be the same irrespective of the individual carrying out the assessment and when and where it is done. Thus, the quality of the individual judgements is of fundamental importance in grade-setting practice since such evaluations have long-term consequences for individuals, as well as for society at large. In a grading system involving multiple assessment occasions and where a large number of varied assessment methods are used, the reliability may be perceived to be high. Many assessment occasions, with different teachers carrying out evaluations of student performances, enhance the reliability of grades.

However, the issue of the object of assessment still poses a problem since high reliability doesn't necessarily guarantee that the assessment is measuring or focusing the intended focal constructs. If the wrong constructs are measured, validity will be low. There is thus sometimes a trade-off between the levels of reliability and validity, in that a higher reliability can imply lower validity (Messick, 1994).

Grades are a summarization, or a summative score, extrapolated from a large number of evaluations - for example tests and performances in the classroom – of different observed performances. Multiple sources of evidence are gathered by teachers in different situations over extended periods of time and by means of different types of assessment. This process provides a basis for the validation of the interpretations made where each source of evidence contributes to making the most reasonable case (Messick, 1989, 1994). Different modes of assessment, such as multiple choice testing and performance assessment in the classroom, may seem to render different validation processes due to their different nature, but as Messick (1989) points out, it is not the observation device as such that is evaluated but rather the inferences and interpretations derived from the test scores or the performances. The use of a score is to be understood in a broad sense, where the score is used as an encapsulation of different kinds of observations of student performances and products. Validity is about evidence for or against a certain interpretation which means that an assessment procedure or an assessment score is neither valid nor invalid of itself, but rather the interpretation of these scores that are validated (Kane, Crooks & Cohen, 1999).

In the unified validity approach formulated by Messick (1989, 1994), construct validity subsumes and integrates different aspects of validity (content relevance and representation, substantive, internal, external, generalizability and consequences). In particular, Messick emphasizes that these aspects or evidences are complements to one another, not alternatives. He also argues that the evidence is supposed to function as a way of seeking arguments for and against an interpretation and to avoid two major threats to construct validity; namely construct irrelevant variance and construct underrepresentation. Construct irrelevant variance implies that a score contains variance related to other constructs, or method variance, not relevant for the construct in focus. It may also be that the construct irrelevant variance advantages or disadvantages particular groups of students. Messick (1994, 1995) is also at pains to stress that the interpretations and conclusions that are made must not be influenced by factors such as students' prior knowledge, background and gender, and hence should not derive from measurement invalidity due to construct irrelevant variance or construct underrepresentation.

Underrepresentation implies that the measure is too narrow and does not cover important aspects of the construct at hand. Construct irrelevant variance and underrepresentation influence all forms of assessment. In educational systems that emphasize higher-order skills, reasoning and complex skills and abilities, the question of what constitutes construct irrelevant variance is a problematic issue. The complex issues about the relevance of different student characteristics in grades and, in the long term, whether grades are fair, concerns how we define construct irrelevant variance (Messick, 1994).

An important question raised by Kane et al. (1999) concerns how we can validate the inferences drawn from performance assessments. In Sweden, the educational and grading system is highly decentralized which implies that the most commonly used assessment procedure is classroom assessment, which includes performance assessments. This means that teachers assess their students on a daily basis, with the use of a wide range of different assessments, such as different teacher-made tests, written assignments, classroom observations, performance assessments and portfolio assessments. The definition of performance assessment provided by Kane et al. (1999) is that performance assessment involves a sample of performance from a particular domain of performance. The concept of performance assessment may be used interchangeably with authentic and direct assessment since these conceptions all concern the closeness of the type of performance that is observed and the type of performance that is of interest. Put another way, performance assessment includes a sample of student performances, which is of importance for the interpretations and inferences that teachers make.

Kane et al. (1999) argue that performance assessments have certain advantages or strengths (in comparison to other assessments such as for example multiple-choice tests) which concern evidences and consequences, due to the direct line between the inferences from an observed performance and the conclusions being made. This directness or authenticity is thus a strong foundation for validity. Kane et al. (1999) emphasize that the interpretive argument, which is a basic and overall concept, incorporates several stages of scrutiny of the arguments for and against a certain interpretation. This chain or line of arguments stretches from the observed student performances to the final assessment of student attainment.

In order to visualize the different aspects of this chain of validity, but also to clarify the different threats to the interpretive arguments and validity, a bridge analogy may be helpful. (See Figure 3). This bridge analogy is concerned, from the left to the right, with observed performances, observed scores, universe scores and target scores.

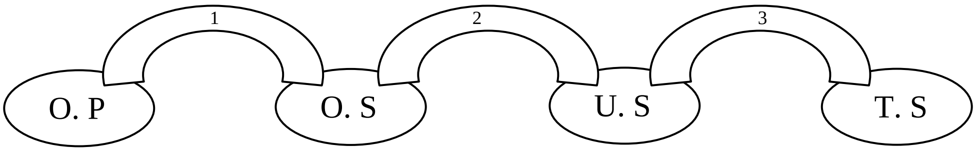


Figure 3. The bridge analogy: Observed Performances (O. P), Observed Score (O. S), Universe Score (U. S), Target Score (T. S). This is a modified figure from Kane et al. (1999).

In this figure, the bridges (one to three) are seen as the inferences or interpretations in the validation process. The first inference or bridge deals with the scoring of the observed performances of the students and the appropriateness of that scoring. It rests on assumptions that the criteria which have been used in the scoring procedure have been used in an appropriate way and that the performances were assigned under conditions compatible with the intended score interpretation. The threats to the scoring of the observed performances are numerous, from the exclusion of some important criteria and tasks not clearly communicated to the students, to the lack of appropriate equipment needed for solving a certain task, and may distort the interpretations and inferences that are drawn.

The second inference or bridge deals with the generalization from the observed scores to a universe scores. Universe scores are understood as a generalized score, in that the observed score leads to a conclusion about the expected performances on similar tasks. Scores obtained on a sample of tasks are thus generalized beyond that specific set of tasks. However, this assumes that the set of tasks is a representative sample of the domain in focus. A major threat to the generalizability is the lack of representativeness of the set of tasks that has been chosen from that domain. In performance assessment, the number of tasks is relatively small which affects the reliability and increases the risk of sampling errors. In particular, the variability over tasks has been shown to be considerable in performance assessment and therefore a threat to generalizability. Kane et al. (1999) argue that the consistency and generalizability can be improved by standardizing the task characteristics and administration procedures.

The third inference or bridge is the extrapolation from the universe scores to a target score. This implies that the expected score from the universe of generalization (e.g a sub-domain within the target domain) is a summarization or conclusion generalized and applied to the target domain. The target domain is, in education, permitted to be less-well defined and to have fuzzier boundaries in accordance with the goals and criteria in the different steering documents and curricula.

By using these definitions and concepts, grades are understood as a target score extrapolated from the universe of generalization or, in other words, generalised from the observed score. In educational settings, the interpretive argument thus functions in the following way. First, teachers observe student performances and, in their assessment processes, crystallise this into some observed score (scoring). Next, this observed score is generalized into a universe score (generalization) which in turn evaluates the level of knowledge of the student from the universe score (extrapolation). Finally, a decision about what grade to award on the basis of the student's level of knowledge (decision) is made.

Kane (2006) emphasizes that the fundamental issue at stake is to make the interpretive argument as explicit as possible, and, in particular, to ensure that all assumptions are made explicit, such as, for example, assumptions about what is relevant variance in grades. For low-stake situations, for example when a teacher gives feedback to a student, the evidence from the development stage is probably enough, whereas in high-stake situations more extensive appraisal is demanded. Messick (1994) also emphasizes that the validation process demands that the criteria and goals in the curricula are made explicit. Cronbach (1988, 1989) acknowledges that the validity argument is of major importance in high-stake situations, and has to be persuasive to all different stakeholders.

However, classroom assessment is a qualitative form of assessment, which focuses on the evolving interpretations of many different performances in a context, as opposed to simply focusing on scores. In classroom assessment, teachers adapt instruction to the needs of their students, which means that 'standard consistency' over classrooms is not relevant. Thus, when interpretations and decisions made in individual classrooms are applied in other contexts, across classes and schools and over time, problems may occur. Grades may be a high-stake measure for students, particularly if grades are used as an instrument for selection to the next level in the educational system. Indeed Kane (2006) explicitly acknowledges that "The extension of the teacher's conclusions about students to contexts beyond the classroom may be higher stakes than classroom assessment because they may not be easily corrected" (p. 49-50). Kane thus makes a distinction between quantitative and qualitative models of validation where the quantitative model tries to eliminate the influence of different conditions such as social and physical context and student characteristics (e.g. achievement tests) when assessing students, while the qualitative model tries to integrate all kinds of evidences in order to reach a coherent view of the student (e.g. grades).

One of the main issues of the interpretive argument concerns alternative interpretations and inferences. Discriminant evidence discounts alternative inferences which may disguise the nature of the knowledge being assessed, for example that the national tests in mathematics demand a certain level of language knowledge in order for students to fully understand the questions. Discriminant evidence may be of particular interest in schools with, for example, students from minority backgrounds, since their specific situation may require alternative modes of assessment. The interpretation of results or evidence must be made in a cautious manner, and reasonable rival inferences must be highlighted, so that valid inferences can be drawn.

During the validity process, the focus is on the complexities of student knowledge, skills, or other attributes to be assessed, which highlights issues concerning the types of performances and behaviours that are in line with the interpretive argument. According to Wiliam (1996) certain aspects of a domain and different constructs related to the domain are emphasized, while others are de-emphasized. Scores and grades are interpreted in terms of the current construct, where some attributes are consistent with the scores and grades and correlate with certain student characteristics. Given this interpretation, when used in the admission process to the next level in the educational system, grades could, for instance, predict success and indicate that some skills are relevant for learning and that no adverse impact to any group is due to construct irrelevant variance. A student's low score or grade ought not to be low because the measurement is missing some relevant construct which, if present, would have revealed the student's real ability.

In educational research, Messick's model of validity would appear to be particularly fruitful, since this theory takes into account the consequences of a judgement. According to research, teachers seem to take into account the consequences of their decisions and also take into account the distributive justice of a grade, how it is used and the consequences of its use (Pilcher, 1994). It also has been found that teachers interpret grades in different ways (Brookhart, 1993, 1994) in that teachers emphasize certain constructs whilst other constructs are de-emphasized. It is also suggested that the inferences that teachers make are related both to the domain in focus and to other domains. Wiliam (1996) has modified Messick's four-facet framework for validity arguments and argues that the validation process must address a broad range of inferences from, and consequences of assessments, both within and beyond the domains in focus.

Reflections on the theoretical framework

This conceptual framework has been chosen out of the need, usefulness and appropriateness of the different perspectives given above, in order to understand and explain the meaning and dimensionality of grades. Teachers' assessment decisions and rationales for grading decisions form one perspective which concerns ways in which grades are affected by explicit and implicit assumptions in the grading procedure. The perspective of student characteristics and their influence on grades further highlights the implicit assumptions that concern grades. The school characteristics perspective also places an emphasis on different systematic factors that have an effect on grades. These three perspectives have been developed in this research over a long period of time.

Furthermore, Encapsulation Theory provides distinct assistance in understanding and unfolding the different cognitive dimensions and their importance for understanding the nature of the student abilities that are measured by grades. Finally, validation theories add a number of additional useful concepts which highlight how to think about the quality and validity of grades in terms of the distinction between relevant variance and irrelevant variance in grades. The unified validity theory and conceptions such as construct validity, construct-irrelevant variance, underrepresentation, discriminant evidences, as well as the bridge analogy, incorporates and provides a foundation of different perspectives, linking them with the research problem in order to understand the meaning of grades and what it is that they measure. Of course, several other perspectives may have added further understanding of the meaning of grades but the selection of theories accounted for in this section functions as a comprehensive and overarching basis for understanding the complexity of grades and what it is that grades measure.

RESEARCH QUESTIONS AND THE PURPOSES OF THE STUDIES

Based on previous research reported in the literature and its relation to the present research problem, a conceptual framework has been created which incorporates a number of the main perspectives and conceptions described in the previous section.

The literature suggests that there may be reasons to suspect that grades are multidimensional in that they measure both cognitive and non-cognitive abilities. Accordingly, a hypothesis was developed, that if grades are systematically influenced by factors other than student knowledge and skills then this should manifest itself as a common grade dimension which cuts across different academic fields. This means that a part of the variance in grades is common to all grades, irrespective of the subjects in which they are assigned, which suggests that this has to do with another dimension than achievement in the different academic domains. Consequently, different student and school characteristics were incorporated in the model developed in order to investigate their relations to grades and whether any of them could explain some of the variability in grades.

Three empirical studies were carried out in order to investigate the dimensionality of grades and the effects of student and school characteristics on grades. These studies were driven by the following specific purposes:

- To identify and separate different dimensions in grades, which on the one hand, might be interpreted as expressing variance in knowledge and skills or, on the other, different systematic factors. Another aim is to examine differences related to gender and family background. Since there are reasons to suspect that grades are influenced by factors other than achievement - some of which operate at the individual level, and others at the school level - the main purpose was to identify and separate different dimensions of grades by the use of multilevel multivariate techniques;

- To investigate the influence of different student characteristics on the two identified dimensions of grades found in study I; i.e. the subject-specific and common grade dimensions. A further aim was to investigate how gender influence grades and, specifically, how different student characteristics mediate the effect of gender on grades; and

- To explore how school characteristics influence grades and how these characteristics influence the subject-specific and common grade dimensions. An additional purpose was to investigate the extent to which parental education confounds the relations between school characteristics and the subject-specific and common grade dimension in grades at the school-level.

METHOD

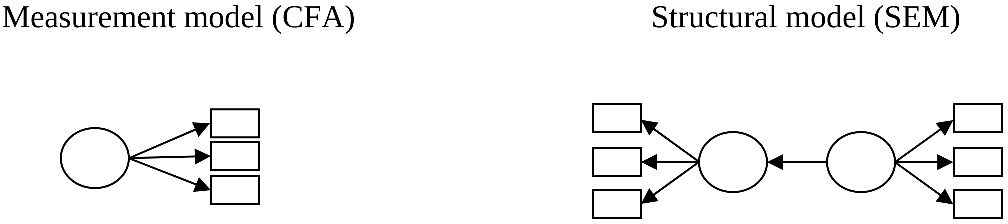
In educational research and in research which focuses on measurement of student achievement, the relations between different background variables and student outcome variables often are of major importance and interest. These different background variables typically are unobservable and hard to measure. However, the use of latent variable modelling offers a solution to these problems. For example, in order to measure student motivation, different observed variables such as how interested the student is in learning or whether the student wants to continue to further education may be used. Of course, there are several ways to measure the different unobserved (or latent) variables by using a variety of observed measures or indicators.

However, the correct measure of a construct is dependent on the indicators and their ability to represent the construct which is the major focus of construct validity. "Construct validity can be defined as "the extent to which a latent construct can be inferred by a set of observed variables" (Yang, 2003, p. 26). The validity and reliability in the process of creating constructs must be ascertained and properly handled. Of course, in different times and different contexts indicators may have different meanings and different implications. When measuring student motivation the meaning of the construct may be the same but the indicators may vary over time, place and context. The indicators are effects of the construct and the indicators reflect the construct and thus create a factor model (Loehlin, 2004).

Confirmatory factor analysis and structural equation modelling

Factor analysis is an analytic approach designed to identify underlying dimensions by using multiple measures of a construct. In a factor analysis, a latent variable is identified by several observed (manifest) variables or indicators. Thus, latent variables or factors represent theoretical constructs measured by several observable variables or indicators. For example, the measurement of student achievement is achieved indirectly through several observable indicators, for example test scores and grades. The relationship between observed indicators and unobserved factors are defined in terms of weights or factor loadings, where a factor loading is the weight of an indicator on the factor. With this method, the constructs are isolated in terms of the observed variance being partitioned into common, specific and error variance (Hox, 2002; Loehlin, 2004; Maruyama, 1998).

The first specifications of a confirmatory factor analysis (CFA) model should be made on the basis of theoretical assumptions and thorough considerations. A CFA model can be seen as a sub-model of a more general structural equation model (SEM). CFA is a method that "…begins with a theoretical model that has to be identified and must attempt to see whether or not data are consistent with the theoretical model" (Maruyama, 1998, s. 139). The SEM model is to be regarded as an extension of the CFA model, because the SEM model also accounts for specified relations between factors or constructs, as outlined in Figure 4 (Jöreskog, 1993). Hence, the SEM model is comprised of both CFA models (relations between a factor and its indicators) and a structural model (relations between factors).



Measurement model (CFA)

Structural model (SEM)

Figure 4. Example of a measurement model (CFA) and a structural model (SEM)

The hierarchical structure of educational data must also be handled adequately by using multilevel analysis (Gustafsson & Stahl, 2005; Hox, 2002). Multilevel analysis refers to the hierarchical data structure in the population where relationships between individuals and society are studied. Both the data in itself and the different sampling procedures involve multilevel structures, such as for example cross-sectional studies where students are nested within schools, but also longitudinal research where different distinct observations are nested within individuals, and meta-analysis where individuals are nested within different studies (Hox, 2002). In order to handle the multilevel problems, the traditional use of aggregation and disaggregation of variables to one level, and a standard analysis method, such as for example ordinary multiple regression analysis, leads to both statistical and conceptual problems. On the one hand, using aggregated data, where information from small units is combined into fewer values for higher-level units, much information is lost and the statistical analysis loses power. Disaggregated data, on the other hand, blows up values from a small number of units, for example schools, into values for the smaller units (individuals), which may lead to incorrect calculations of significance levels. The conceptual problems of aggregation and disaggregation may lead to difficulties of interpretations, such as the ecological fallacy, when interpretation are made of the results from aggregated data on the individual level, or the atomistic fallacy when disaggregated data are interpreted on, for example, the school level (Hox, 2002; Yang, 2003).

When groups of students, such as classes or schools, are sampled and observed, the individual observations are in general not completely independent, which violates the assumption of independence of the observations which standard statistical tests typically rely on. As a consequence, the design effect or the clustering effect may thus lead to estimates of standard errors that are too small. To take this into account, multilevel CFA and SEM have been used in the studies reported on in this thesis.

In order to take into account the clustering effects in the data the complex option offered by the Mplus program provides an alternative to the full multilevel analysis approach. This method, whilst compensating for the disturbances in the χ^2 and standard errors due to clustering effects, does not affect the estimates. In the complex analysis, the standard errors become larger and the t-values become smaller due to losses in information caused by the clustering. The extent of the information loss due to clustering effects is a function of the intraclass correlation and the cluster size (Muthén & Muthén, 2004).

A common problem in educational research is that information is missing on different variables. A distinction is made between data that is missing completely at random (MCAR) and data that is missing at random (MAR), whilst both assume that the failure of observing a certain data is independent of the missing values. MCAR is however, a much more restrictive assumption, because data must be completely independent of all other variables. With MAR data, on the other hand, the missing may depend on other variables in the model and correlate with the unobserved values. The MAR assumption implies that the procedure yields unbiased estimates when the missingness is random given the information in the data. This is a much less restrictive assumption than the assumption that the data is ‘missing completely at random’. High interrelations among the observed variables provide good possibilities to satisfy the MAR assumption (Schafer & Graham, 2002).

In CFA and SEM, the hypothesized model may be statistically tested against data, which implies that the discrepancies between the sample covariance matrix and the fitted covariance matrix are determined and an overall goodness-of-fit test χ^2 is computed. This measure of model fit is relatively sensitive to sample size, where large samples tend to produce high χ^2 values. The Root Mean Square Error of Approximation (RMSEA) measures the amount of discrepancy between the model and the data, taking the number of parameters and sample size into account, making the RMSEA measure relatively independent of large sample sizes. The RMSEA is strongly recommended as a tool when evaluating model fit. It should be below .08 to be acceptable and below .05 to indicate good fit (Jöreskog, 1993). The Standardized Root Mean Square Residual (SRMR), which is a measure of residuals, compared separately for within and between levels, was also used. This measure should be below .08.

When investigating complex and underlying relations among variables, for instance how different variables account for differences in student behaviours, typically there are both direct and indirect relations. Baron and Kenny (1986) make a distinction between variables which function as moderators and variables which function as mediators, emphasizing that the failure to recognize the distinction between moderators and mediators has conceptual implications. The difference between moderator and mediator variables is that while moderators “specify when certain effects will hold” mediators “speak to how or why such effects occur” (p. 1176).

In the current research there is a particular focus on the mediating effects of certain variables, which involves determination of whether different factors have a direct or indirect effect on grades. In a mediator type of model, the mediating variable accounts for the relation between the independent and dependent variables. First, there is a relation between the independent and dependent variable. Then it is necessary to establish a relation between the independent and the mediating variable, and also between the mediating variable and the dependent variable, see Figure 5.

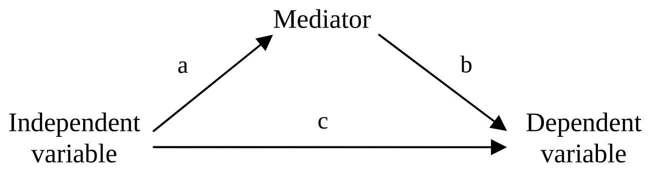


Figure 5. A modified figure of a mediator-type of model (Baron & Kenny, 1986).

In a mediator-type of model, the independent variable (a) accounts for variations in the mediator (path a), and the variations in the mediator account for variations in the dependent variable (path b). When paths a and b are introduced, a previous significant relation between the independent and dependent variable decreases or become zero (path c). As an example of this model, let us assume that there exists a relation between gender and grades (path c), in favour of girls. However, there may be other variables affecting this relation; for example girls may work harder in school than boys. As a mediator variable we choose to investigate student effort to see whether student effort explains some of the gender differences found in grades. If the relation between gender and grades decreases or even becomes zero when the mediator is included in the model, student effort mediates and explains the gender effect on grades.

SUMMARY OF THE STUDIES

This thesis comprises three large-scale empirical studies. In this section, the data used in the three studies are presented, along with the participants, the variables, the measurement of the constructs and the results.

Subjects

All three studies are based upon one dataset which derived from The Gothenburg Educational Longitudinal Database (GOLD) which contains register data compiled by Statistics Sweden for all individuals born between 1972 and 1987 and where a large amount of information is available. For example the GOLD database contains information about gender, family background and grades from compulsory and post-compulsory education, results from national tests and questionnaire data from students, parents and teachers. The database also contains school characteristic data such as school-size and location, the management of schools (e.g. municipally- or independently-operated schools), teacher density, teacher average age and sex, experience and qualifications. The student questionnaire data used in the studies derives from the ETF-project (Evaluation through Follow-up) where a 10 per cent national representative sample has been collected for certain specific cohorts (Härnqvist, 2000). In all three studies, the subjects were 99 070 students born in 1987, and who left compulsory school in 2003 (9th grade). The number of participants in the questionnaire data was 8 717. In sum, 1246 schools were included in the analyses.

Two reductions of the dataset have been made. Initially schools with 14 students and fewer have been excluded from the analyses since schools with very few students are often schools for students with special needs or other specific groups that are in need of specialised support and education. Secondly, individuals for whom both information on subject grades and national test result are lacking have also been excluded. In all, 1782 individuals have been excluded from the analyses.

Variables

Throughout the studies variables on grades and scores on the national tests have been used. These variables provide the foundation for exploring the dimensionality of grades and are used to create the conceptual model. In order to further investigate the nature of the influences on the different dimensions in grades, and to explain the variability in these dimensions, different student and school characteristic variables were used.

Individual characteristic variables

Grades and national tests

In all three studies, two measures have been used: the final subject grades and test scores from the national tests. Both of these measures are derived from the period at the end of compulsory school (9th grade) in three core subjects; Swedish, English and mathematics.

The subject grades are awarded by the teachers themselves and no external referees are involved in the grade setting practice. The grading process is highly decentralized and is based primarily on classroom assessments where the individual teacher evaluates the performances of her/his students. In the curriculum, there is an emphasis on student abilities such as active communication, the development of thought, argumentation skills and the ability to state opinions. The grading scale consists of four levels which, in order to function as an instrument of selection to the next level in the educational system, are converted into numbers: not pass (IG) = 0, pass (G) = 10, pass with distinction (VG) = 15, pass with special distinction (MVG) = 20. No intermediate numbers are used. These levels reflect student attainment of the objectives or criteria for each subject. Overall standards for the final semester of secondary education, that is, the spring of the ninth year of school, are defined centrally for all the grade levels in the curriculum. In this thesis, the following abbreviations are used for the subject grades: SGSW for subject grades in Swedish, SGEN for subject grades in English and SGMA for subject grades in mathematics.

The national tests are composed of different subtests in each subject, and include both written and oral performances, as well as group and individual tasks. The tests are produced centrally and their contents are not revealed in advance. Whilst the curriculum and the syllabus provide the starting-point for the construction of the tests, not all of the centrally defined goals are however covered by the tests, implying that the respective subject domains are not fully covered by the tests. In Swedish there are three subtests, the first of which is a reading comprehension test, which is followed by an oral test conducted in pairs and, finally a written assignment. In English, the three subtests consist of oral interaction and production, usually conducted in a group, reading and listening comprehension tasks, and a short essay. In mathematics there are four subtests: an oral task carried out in a group, a test of arithmetic where the use of a calculator is not permitted, a test with more extensive tasks, and finally, a test which demands problem-solving where students have to account for the calculations they make (National Agency for Education, 2005, 2007, 2008).

As regards the grade setting process for the national tests, the teachers themselves assess and grade the results of their own students and no external referring procedure is used. However, teachers and schools are recommended to collaborate in the grade setting practice for the national tests and encouraged to carry out comparisons of student results. Furthermore, teachers have access to extensive guidelines each with a series of annotated benchmarks, as well as websites where a large number of examples of student attainment on the tests are offered.

The scale for the national tests corresponds with the scale for the subject grades and ranges from 0-20, where not pass (IG) = 0, pass (G) = 10, pass with distinction (VG) = 15 and pass with special distinction (MVG) = 20. No intermediate numbers are used. In the thesis, the following abbreviations are used for the national tests: NTSW1, NTSW2, NTSW3 for the test scores in the Swedish national test; NTEN1, NTEN2, NTEN3 for English and in the case of mathematics, where only one summarized test score is available, the abbreviation NTMA is used.

Student background variables

In all three studies two student background variables were used, namely student gender and the level of parental education within the student's family. Parental education was indicated by a dummy variable (PAREDU) where education higher than upper secondary education was assigned the value 1. Gender was also a dummy variable (GENDER: boys = 0, girls = 1).

In Study II, the purpose was to ascertain which student characteristics influence grades. The student questionnaire data were included in order to create factors of different student characteristics. The questionnaire data were collected during spring 2003 when the students were finishing the 9th grade of compulsory school.

One overall purpose of the questionnaire is to gather information to be used for longitudinal studies in order to investigate changes over time. Because of that purpose, whilst the questions have remained almost unchanged since the start of the project, the questionnaire items were not primarily created for advanced analyses methods like factor analysis. This implies that extensive consideration needed to be taken as to whether or not the questions and specific items were possible to use in the creation of constructs or factors of different student characteristics. Initially, this meant carefully exploring these variables and testing the items in different models. Whilst of course, the use of this type of data brings with it certain limitations due to the nature of the items, it is nevertheless argued here that these data are of good quality and possess the potential to contribute to the understanding of the types and nature of student characteristics that are of importance for grades.

From a large number of items on the student questionnaire, 43 items were selected in order to create eight theoretical factors, hypothesized to represent certain student characteristics, see Table 1 in study II. One difficulty in creating factors concerns the necessity of labelling them. The name or term can "take over" the meaning of the construct with the result that the underlying items and meaning of the factor become somewhat limited and shadowed. It is problematic to find a suitable and short enough term suitable for the construct at hand. It is emphasized here that the name of the factors must be dealt with in a broad sense, so that it is clear that the construct is of primary importance in order to understand these factors. Therefore, whilst the presentation of the factors below will be thorough, in order to get a more detailed description, the reader is referred to study II.

Three self-perceptions of competence factors *SpSw*, *SpEn* and *SpMa* hypothesized to reflect students' self perceptions of competence, interest and self-esteem in the three subjects, Swedish, English and mathematics, were created. The items used for these three factors were similar to each other, with only the subject itself differing. By means of example, the Swedish items are now presented. The question "How well do you think you are able to accomplish the following in Swedish?" was followed by six items all focusing on the student's self-perception of different competencies such as "reading and understanding a text", "reading the subtitles", "writing a story" and "participating in a discussion". There was also a general question "How good do you think you are in Swedish" and one question concerning student interest in the same subject, namely "How interested are you in learning more in Swedish?" These items concern the self-perceived ability and interest of the student in the particular subject and it is hypothesized that students who believe they are good in accomplishing the different tasks in a subject also are more interested in learning in that particular subject. The interest item in this factor was thus hypothesized to reflect the student's subject-specific interest.

A parental involvement factor, *Parent*, measured by four indicators and hypothesized to reflect the level of parents' involvement or engagement in students' school work was created. By means of example some of the items were "How often do you tell your parents about how it has been or how you have felt in school?", "How often do you show your parents your test results or other things you have done in school?" and "How often do your parents participate in school?" Whilst it could be argued that these items only show that the student talks to her/his parents, not that the parents are involved or engaged in their child's school work, it is however argued here that students who tell their parents about how they feel in school and what results they get, have parents who listen and are interested and involved in their child's school work. Conversely, parents who do not listen to their child about school work are not regarded as very involved or engaged and the child will be likely to stop telling them about experiences at school.

A *Coping* factor, measured by four indicators, was hypothesized to reflect student anxiety about school work and how the student is coping in school. Items such as "Do you find it difficult to keep up in lessons?" and "Do you often give up if you get a difficult task to do in school?" were hypothesized to indicate whether the student is anxious about school work and whether she/he feels able and cope with the tasks in the classroom.

An *Interest* factor was hypothesized to reflect students' overall or general interest in learning. This factor concerns the student's interest in learning more in all the three subjects and was measured by three indicators about interest in each of the respective subjects (Swedish, English and mathematics) together with one indicator relating to future university studies. This factor was hypothesized to reflect student general interest and motivation for learning and thus reflects a positive attitude towards learning, school and higher education. Interest and motivation are very similar concepts, including both cognitive and non-cognitive aspects and are thus closely related (Brookhart & Durkin, 2003; Hidi & Renninger, 2006).

A student adjustment factor, (*Adjust*) hypothesized to measure student adjustment and comfort in school was created out of five indicators. Items, such as "How content are you with other pupils?" and "How content are you with the teachers?" were hypothesized to reflect students' relationship with class-mates and teachers, attitudes to school work and general wellbeing in school.

A factor (*Cooperate*) which was hypothesized to measure student cooperation with classmates and teachers was created from three items including "How often do you work in groups in your class?" and "How often do the teacher and students discuss in your class?" It is argued here that these items demonstrate the existence, or otherwise, of a cooperative climate in the classroom and, as such provide an indication of a teaching approach where the students are empowered, have opportunities to state their own opinions and can cooperate with one another.

School characteristic variables

In study III, different school characteristic manifest variables were used. School authority has been categorized into a dummy variable (INDEPSCH: municipally-operated schools = 0, independently-operated schools = 1) School size (SIZE) has been categorized into four sizes and, hence three dummy variables have been created. In this categorization small schools have up to 299 students, the group of second smallest schools has 300 to 499 students, the group of third largest schools has 500 to 1000 students, while the largest schools have more than 1000 students. The small schools are used as the reference category. The schools are also divided into three location categories relating to the level of rurality of the location of the school, and two dummy variables have been created (LOCATION). Schools which are located in 'rural areas' are situated more than 45 minutes from the nearest town with more than 3000 inhabitants, schools in 'suburban areas' are between 5 to 45 minutes from the nearest town (3000 inhabitants) whilst schools in 'urban areas' are located in a town with a maximum of 5 minutes travel time to a town with more than 3000 inhabitants. Urban schools form the reference category. Different teacher characteristics on the school level such as the proportion of female teacher at the school (TEAWOM), the average age of the teachers at the school (TEAAGE), the average proportion of qualified teachers at the school (TEACER), average years of teaching experience among the teachers at the school (TEAEXP) and teacher density at the school (TEADENS) have also been included in Study III.

The manifest school characteristic variables are somewhat limited due to the level of detailed information. In particular, the teacher qualification variable is rather imprecise and does not reveal whether or not the teachers have qualifications other than a degree in education that might be of importance for their work as a teacher. Neither does this variable reveal whether the teachers are "close" to obtaining a qualification or have a long way to go before becoming qualified.

Treatment of missing data

In the subject grades the amount of missing data was negligible, 0.7 per cent (717 students). The national test contained a larger amount of missing values, ranging from between 17.8 to 21.3 per cent for the different subtests in Swedish and English. For the national test in mathematics there was a considerable amount of missing data, 43.1 per cent. The large amount of missing data in the national tests in mathematics was due to the unfortunate fact that, in some areas in Sweden, an unauthorised prior divulgence of content of the test had taken place. There is however no reason to assume that this lack of sufficient information is biased due to achievement in mathematics.

Since the questionnaire data was drawn from the population data, there are two kinds of missing data. The first is that the questionnaire is a 10 per cent sample drawn from the population represented in GOLD, whilst the second relates to the fact that some students have missing values on certain items. The questionnaire data contained missing values, ranging from 26.4 to 27.9 per cent for the different items.

For the school characteristics data, there were only negligible amounts of missing data, ranging from 0.1 to 0.7 per cent for the different teacher characteristics. School characteristics, such as the educational provider (municipal or independent), school size and parental education had no missing values. The variable for school location had missing values of 0.1 per cent.

In all of the studies, the missing data was handled by the use of missing data modelling (Muthén, Kaplan & Hollis, 1987), and the missing data option was used in the Mplus program (Muthén & Muthén, 2004). The main reason for applying missing data modelling was the fact that both population and sample data were used in the studies and, that this option also handles the missing information of the different items. This modelling is based on the assumption that the data is MAR and the basic principle is that subsets of cases with a particular pattern of missing observations each having a separate covariance matrix and where the matrices are combined into one total matrix (Gustafsson & Stahl, 2005; Muthén, Kaplan & Hollis, 1987).

Intraclass correlations of the variables

One important step before estimating a full multilevel analysis is to estimate the amount of variance accounted for by the between level (here school level) for the variables in the analysis. This is done in order to check whether a multilevel approach is necessary. The intraclass correlation (ICC) indicates the variation on the between level, and is a measure of the homogeneity of the observed variables within each cluster (Muthén, 1994; Yang, 2003). If the ICC measure is large and of a magnitude in the region of 0.05 it is necessary to conduct a multilevel analysis. The ICC showed that between 6.2 and 9.5 per cent of the variance was accounted for by the between level, with the consequence that the multilevel approach was conducted in most of the analyses. In the second study (Study II), in order to account for the clustering effects which correct the fit indices (standard errors) without changing the estimates, the complex option in the Mplus program was used.

The data analysis was carried out in Mplus (Muthén & Muthén, 1998-2004) using the STREAMS (Gustafsson & Stahl, 2005) system as a front-end.

Results

The purposes of the three studies were: (1) to explore the dimensionality of grades and the influences of gender and parental education on individual level and grades at individual and school level; (2) to investigate the influences of different student characteristics mediating the effects of gender on grades; and (3) to explore the influence of different school characteristics on grades and the confounding effects of parental education on grades.

Study I

In study I, the dimensionality of grades was explored at the individual and school level. In this study the effects of student gender and the level of parental education on grades were also investigated. The total variation of grades can be decomposed according both to the dimensions (student abilities) and the levels (student and school). This initial study laid the foundation for the two subsequent studies and the conceptual model that was developed in this study was used as a baseline model throughout the series of investigations.

The main purpose was to explore whether it was possible to identify different dimensions in grades where the variance was decomposed, on the one hand into subject-specific variance and, on the other hand, into variance common to the three subject grades. Factors were created in order to develop the conceptual model. A factor for each subject (S*w*, *En*, and *Ma*) was created and hypothesized to reflect three subject-specific dimensions; Swedish, English and mathematics. The *Sw* factor had four indicators; the subject grade in Swedish (SGSW) and three scores from the subtests on the Swedish national tests (NTSW1-3). Similarly, the *En* factor also had four indicators; the subject grade in English (SGEN) and three scores from the subtests on the English national tests (NTEN1-3). The *Ma* factor however had only two indicators; the subject grade in mathematics (SGMA) and one combined score from the four subtests on the national test in mathematics (NTMA). Thus, a three-factor CFA model comprising the three subject-specific factors and with covariances among the factors was developed which was considered theoretically and interpretable sound. In this model, the factors (*Sw*, *En*, and *Ma*) were hypothesized to reflect subject-specific dimensions and measure variance that is common for the grades and scores on the national tests in Swedish, English and mathematics and thus to reflect student achievement in each subject.

In order to identify a possible additional dimension in grades a fourth factor (*ComGr),* was created and included in the model and only related to the three subject grades (SGSW, SGEN, and SGMA), reflecting a common grade dimension which cuts across the three subject grades. Consequently, the baseline model included four factors; three subject-specific factors related to grades and scores on national tests in Swedish, English and mathematics, and one common factor related only to the subject grades. The path diagram of this baseline four-factor model is shown in Figure 6. The result of this four-factor model supports the hypothesis that grades are multidimensional in that subject-specific dimensions and a common grade dimension were identified, at both individual and school levels.

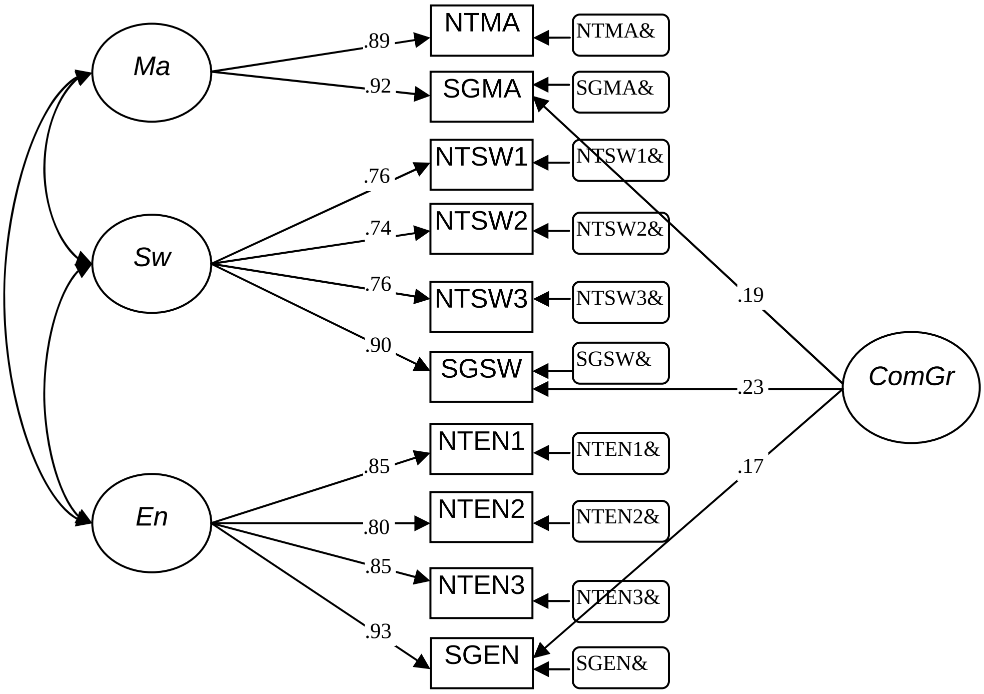


Figure 6. The baseline four-factor model with covariances among the three subject factors.

The individual level. The result from the four-factor model showed that the standardized factor loadings were substantial for the subject-specific factors indicating that the largest amount of variance in grades was accounted for by the three subject-specific dimensions (*Sw*, *En* and *Ma*). The estimates are shown in Figure 6. However, the relations between the *ComGr* factor and the three subject grades were also positive and significant. This model shows that the relation between the subject grade in Swedish (SGSW) and the *ComGr* factor is the strongest one of the relations between the *ComGr* factor and the different subject grades, which supports previous findings that the grade in Swedish is the one most strongly influenced by factors not related to a subject-specific dimension (National Agency for Education, 2005, 2008). The result from the four-factor model suggests that there exists a common grade dimension which accounts for variance that is common for the three subject grades and which has no relation to the national tests (SGSW, SGEN, and SGMA).

Another purpose in the first study was to investigate the effects of gender and parental education on the subject-specific and common grade dimensions. Gender and parental education were thus included as dummy variables in the baseline four-factor model, related to the four factors (*Sw*, *En*, *Ma* and *ComGr*). Gender primarily influences the *Sw* (.25) and the *ComGr* (.15) factors, where girls have a higher value. Thus, in comparison to boys, girls register higher achievement in the Swedish and the common grade dimensions. The effect of gender on the *En* factor although small was nevertheless significant (.08) whilst no significant influence of gender on the *Ma* factor could be discerned. These results correspond with previous research indicating that girls have higher achievement in the language domain (National Agency for Education, 2005). That girls also have higher achievement levels in the common grade dimension is a finding that requires further investigation in order to be adequately explained.

The level of parental education in terms of students with well-educated parents, influenced the three subject-specific factors *Sw*, *En* and *Ma,* to an almost equal extent, around .30. The parental educational variable had a small, but nevertheless significant, negative relation (-.04) with the *ComGr* factor at the individual level. These results show that whilst a higher level of parental education has quite a considerable positive effect on the subject-specific dimensions, there is only a negligible influence of parental education on the common grade dimension. This result supports previous research that has indicated that students with well-educated parents achieve better in school and achieve higher grades (Yang, 2003). However, the findings also show that a high level of parental education is not of importance in the common grade dimension which suggests that students with well-educated parents do not have benefits in this dimension or benefit in the same way as in the subject-specific dimensions.

In previous research, the multidimensionality of grades, in terms of subject-specific and common grade dimensions, has not been properly recognized. Instead, grades have been seen as a one-dimensional measure, ignoring the fact that grades measure several dimensions of student knowledge and skills. The modelling results obtained in the first study indicate that it is possible to separate the variance in grades into both specific and common dimensions and that gender and parental education influence these dimensions in different ways.

The school level. Intraclass correlations showed that between 7.3 and 8.5 per cent of the variance in the national tests in Swedish, English and mathematics was due to school differences. For subject grades in Swedish, English and mathematics, the intraclass correlations ranged from 6.2 to 6.8 per cent, which also indicated the presence of school effects, thus making it necessary to conduct two-level analyses.

In order to take account of systematic differences in grades related to the school level, multilevel CFA and SEM analyses were conducted. The baseline four-factor model was thus estimated with the same relations and structure among the variables for within and between levels. The standardized factor loadings showed a similar pattern on the school level as for the individual level, the only difference being that they were slightly higher for grades as well as for the national tests at the school level. See Figure I in Study III.

Since previous analyses and research has shown that the level of parental education has a substantial influence on grades at the school level, the parental education dummy variable (PAREDU) was included in the model and related to the four factors (*Sw*, *En*, *Ma* and *ComGr*) at both levels. At the school level this variable thus represents the proportion of students who have parents with education above upper secondary education. The standardized regression coefficients were rather high for *Sw*, *En* and *Ma* (.72, .80, and .74, respectively) whereas the relation to the *ComGr* factor was substantially negative and significant (-.34), indicating that schools with a large proportion of students with less well-educated parents have a higher level on the *ComGr* factor.

The school-level effect of parental education on the common grade dimensions indicates the presence of a compensatory grading practice. A compensatory grading practice implies that teachers compensate or take into account different rationales or objectives when awarding grades to students.

To sum up, the results from Study I show that grades are multidimensional, and that the variance can be decomposed into subject-specific dimensions as well as a common grade dimension. Furthermore, this pattern was similar at both the individual and school level. The study also showed gender differences in the Swedish and common grade dimensions, with girls having a higher value in these dimensions. It could also be concluded that students with parents with a high level of education achieve higher on the subject-specific dimensions, whereas no such effect was found for the common grade dimension. However, at the school level the effects of parental education on the subject-specific dimensions were even stronger, whilst the effect on the common grade dimension, which was also strong, was negative. This is interpreted as a compensatory grading effect, such that students in schools with a high proportion of students with low-educated parents are assigned higher grades than is motivated by their level of performance on the national tests.

Study II

In this study the baseline four-factor model was extended with different student characteristics. The student questionnaire was used in order to construct factors from the different questions and items. For an overview of the items and the latent constructs, see Table 2 in Study II. From a large number of items, 43 were selected for use in creating factors, based on theoretical assumptions, hypothesized to measure different student characteristics such as interest, parental support and engagement, cooperation, adjustment, coping in school and self-perception of competence. Hence, a student characteristic model (CFA) comprised of eight factors (*SpSw*, *SpEn*, *SpMa*, *Parent*, *Coping*, *Interest*, *Adjust* and *Cooperate*) with relations to all their respective indicators and with covariances among the factors was estimated. The standardized factor loadings were generally high and significant and are presented in Table 3 in Study II. The results show that the three self-perception of competence factors in Swedish, English and mathematics (*SpSw*, S*pEn* and *SpMa*) accounted for a substantial degree of variance in almost all of their respective indicators. However, some indicators, such as, for instance, “interest in learning more in Swedish” (*SpSw*; .36) and “read subtitles” (*SpEn*; .39) had relatively low loadings, they were nevertheless significant. The *SpEn* and *SpMa* factors had generally higher loadings in comparison to the *SpSw* factor. The standardized factor loadings for the remainder of the factors (*Parent*, *Coping*, *Interest*, *Adjust* and *Cooperate*) also showed a pattern where a few items, such as for instance “if parents participate in school” (*Parent*; .29) and “are you considering going to university” (*Interest*; .20), although having relatively low loadings but were nevertheless significant and theoretically reasonable.

The eight student characteristic factors were then added to the baseline four-factor model and, in order to investigate the relations between the different student characteristics and the subject-specific and common grade dimensions, a twelve-factor model was estimated. Covariances were estimated, on the one hand, between the eight student characteristic factors and the common grade factor (*ComGr*) and, on the other, between the eight student characteristic factors and the subject-specific factors (*Sw*, *En* and *Ma*) and also between the residuals of *Sw*, *En* and *Ma*.

This twelve-factor model was applied to data and the standardized covariances revealed (Table 4, study II) that the relations between the three self-perception of competence factors (S*pSw*, *SpEn* and *SpMa)* and their respective subject-specific factors, *Sw*, *En* and *Ma* were substantial (.63, .70, .65, respectively) suggesting that students’ perceptions of their own competence in each respective subject corresponds to a large extent with their national test scores and their grades. The result also showed that the relations between the other student characteristic factors, *Parent*, *Coping*, *Interest*, *Adjust* and *Cooperate* and the subject-specific factors *Sw*, *En* and *Ma* factors were, with the exception of the relation between the *Interest* factor and the *En* and *Ma* factors which was non-significant, all positive. Thus, general interest in learning and motivation would appear to be of minor importance for high achievement in the English and mathematics dimensions.

The result of this model also suggests that parental engagement and support (*Parent*) has a higher influence on the *Sw* factor (.25) compared to the *En* and *Ma* factors (.16). The *Coping* factor also seems to be of major importance in the subject-specific dimensions (*Sw*, *En* and *Ma*; .42, .41, .48, respectively). Thus, students who report that they feel competent and believe that they manage well with school-work and in the classroom, have high values in all three subject-specific achievement dimensions.

Self-perception of competence in Swedish and English (*SpSw* and *SpEn)* was significantly negatively related to *ComGr* (-.08, -.12), whereas self-perception of competence in mathematics (*SpMa*) was significantly positively related to *ComGr* (.09). Thus students, who have negative self-perceptions of competence in Swedish and English, have slightly higher values in the common grade factor. This may be due to a compensatory grading practice, where teachers, to a larger extent, take into account student characteristics and non-achievement when assigning grades for certain groups of students. Students' general interest and motivation (*Interest*) had the strongest relation to *ComGr* (.26), whilst parental engagement and support (*Parent*), student adjustment (*Adjust*) and student coping (*Coping*) all related positively but weaker to the common grade factor (*ComGr*). These findings show that student characteristics influence the subject-specific and common grade dimensions in different ways. The factors that concern students' perceptions of their competence in the different subjects and whether they believe they are coping with their school work in the classroom, primarily affect the subject-specific dimensions. For the common grade dimension it is primarily student general interest for learning and motivation that are of primary importance.

Because the result obtained from Study I coupled with the results from previous research indicate that gender influences grades, an additional purpose of Study II was therefore to investigate whether the gender difference is mediated by the different student characteristics (Baron & Kenny, 1986). Student characteristic factors were therefore added to the baseline four-factor model with gender as a dummy variable related to all the four factors in the model, each individually in succession, see Figure 7. This resulted in eight models, all of which estimated the same structure but with different composite student characteristic factors.

.

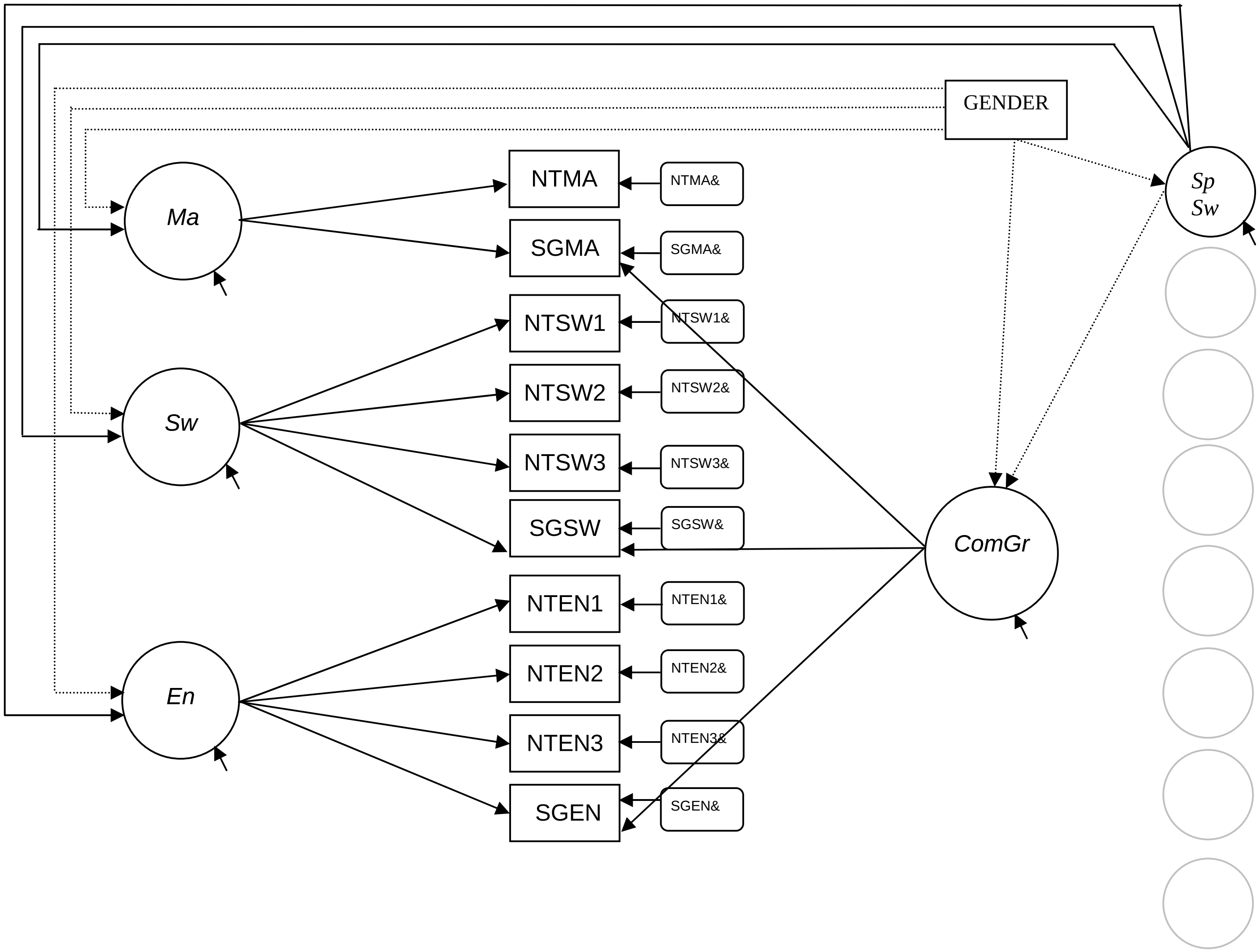


Figure 7. The structural model with relations between one student characteristic factor and the subject-specific and common grade factors and with covariances between the residuals for *Sw*, *En* and *Ma*. Gender as a dummy variable with a direct and mediating effect on *Sw*, *En*, *Ma* and *ComGr* factors.

The standardized regression coefficients decreased substantially for the direct relation between gender and *ComGr* when the indirect relation of gender with *ComGr* via *Interest* was introduced. In fact, the *Interest* factor explained almost all of the gender differences in the *ComGr* factor. The *Parent* factor also explained some of the gender differences in *ComGr*. The models with the *Coping*, *Adjust* and C*ooperate* factors, mediated no gender effect on the *ComGr* factor. These results show that the gender difference found in the common grade dimension was almost fully explained by student general interest for learning and motivation. This means that girls show greater general interest for learning and have stronger motivation, thus giving them a higher value in this dimension.

To sum up, the results from the analyses in Study II reveal that the subject-specific and common grade dimensions are both affected by student characteristics such as general interest for learning, parental engagement, self-perceptions of coping with school work and contentment in school. Gender differences, primarily in the Swedish subject-specific dimension (*Sw*) and in the common grade dimension (*ComGr*), where girls achieve higher values, are also evident. In order to investigate the underlying relations between student characteristics, gender and grades, a mediator-type of model was estimated (Baron & Kenny, 1986). The results reveal that together general interest for learning and motivation mediates almost all of the gender differences in the common grade dimension and the English dimension. The *Interest* factor also mediates a proportion of the gender difference in the *Sw* dimension.

Study III

Study I revealed the existence of systematic differences between schools and the substantial influence of parental education on the subject-specific dimensions (positively) and on the common grade dimension (negatively). Previous research has also indicated that different school characteristics such as demographic and teacher quality variables exert an influence on grades (Darling-Hammond, 1999; Gustafsson & Myrberg, 2002; Hanushek, 1997). Eight school characteristic manifest variables (INDEPSCH, SIZE, LOCATION, TEAAGE, TEAWOM, TEAEXP, TEACER, TEADENS) have been included, one at a time, in the baseline four-factor model, at the school level. As is shown in Table 2 in Study III, the strongest standardized regression coefficients concern the relation between INDEPSCH and the *Sw*, *En* and *Ma* factors (.35, .42, .39, respectively), whereas no significant relation was found with the *ComGr* factor. This result shows that independently-operated schools have a higher level of achievement on the subject-specific factors (*Sw*, *En*, and *Ma*) in comparison to municipally-operated schools.

The SIZE variable was related with *ComGr* which primarily was due to a significantly lower value on the *ComGr* factor (-.16) for the group of the second largest schools (500-1000 students) than for the reference category of small schools (<300 students). The largest schools (> 1000 students), had a higher level on the *Sw*, *En* and *Ma* factors (.12, .12, .07, respectively). All the other relations concerning size were non-significant.

The model with the LOCATION variable showed no significant differences between schools located in rural areas and the *Sw*, *En*, *Ma*, *ComGr* factors in comparison to the reference category of urban schools. However, schools located in suburban areas had significantly lower values on the *Sw*, *En* and *Ma* factors (-.11, -.17, -.07, respectively) and a significantly higher value on the *ComGr* factor (.25) in comparison to the reference category of urban schools. This finding suggest that suburban schools had lower values of achievement on the subject-specific factors, but that greater account of the common grade dimension was taken when awarding grades, in comparison to urban schools.

Some of the different teacher characteristic variables had significant relations to the subject-specific factors. The standardized regression coefficients showed significant relations between teacher age (TEAAGE) and teacher experience (TEAEXP) and the *Sw* factor (.-08, -.11, respectively) and *En* factor (-.12, -.15, respectively). Teacher certification (TEACER) was only significantly related to the *Ma* factor (.12). However, no significant relations were found between the teacher characteristic variables (TEAAGE, TEAEXP, TEAWOM, TEACER, and TEADENS) and the *ComGr* factor. These findings show that younger and less experienced teachers award higher grades and that teacher qualifications were positively related to the *Ma* factor. Schools with a larger proportion of female teachers and with a higher density of teachers also showed significant and positive relations to the subject-specific factors. The non-significant relations between the teacher characteristics and the common grade dimension may be due to either methodological and/or conceptual problems, both of which will be discussed later.

In order to investigate to the extent in which parental education confounds the relations between the different school characteristics and the subject-specific and common grade dimensions, covariances were estimated between parental education and the different school characteristic variables and added to the previously estimated models. Eight models were estimated with one school characteristic variable at a time, as illustrated in Figure 7 above and Figure 2 in Study III.

The result from the model with covariance between INDEPSCH and PAREDU shows that there is a strong, positive covariance between PAREDU and INDEPSCH (.44) indicating that, in comparison to municipal schools, independent schools have students with well-educated parents to a greater extent. The strength of the relation between INDEPSCH and the subject-specific factors decreases substantially, and for the *Sw* factor becomes non-significant.

The model with relations between SIZE and PAREDU revealed that there was a significant covariance between these variables for both the second largest and the largest categories of schools in comparison to small schools. All of the regression coefficients between the SIZE and the subject-specific and common grade factors became non-significant when parental education was controlled for which suggests that it is not school size per se that influence the grades but, rather, the proportion of students with well-educated parents in the school.

The model with LOCATION and PAREDU revealed that, when parental education was controlled for, students in rural schools have, to a greater extent, parents with lower education levels in comparison to students in the suburban and urban schools. In this model, it was also shown that the relations between the *Sw* and *En* factors and LOCATION became significantly negative for schools located in rural areas (-.11, -.07, respectively) and significantly positive for schools located in suburban areas (.12, .06, respectively). For the *Ma* factor, schools located in rural as well as suburban areas had significantly positive values (.11, .16, respectively), in comparison to the reference category urban schools. The relation between suburban schools and C*omGr* decreased somewhat but still remained strong when PAREDU was controlled for (.19). This could be interpreted in the sense that, to a greater extent than schools in rural areas, suburban schools have a large proportion of students with well-educated parents.

The results from the models with relations between teacher characteristic variables and PAREDU (Table 3 in Study III), show that in the models with TEAAGE and TEAEXP there were negative covariances between TEAAGE, TEAEXP and PAREDU indicating that schools with older and more experienced teachers have students with less well-educated parents. The models with TEAWOM, TEADENS and PAREDU revealed significant and positive covariances. These results indicate that schools with a large proportion of female teachers and with a high teacher density have, to greater extent, students with well-educated parents. No significant relations were found in the model with TEACER when PAREDU was controlled for.

To sum up, when parental education was controlled for the result showed that the strong influence of independently-run schools on the subject-specific dimensions decreased substantially and, for the Swedish dimension (*Sw*) became non-significant. In the same way, the associations between the size of schools and the subject-specific and common grade dimensions became non-significant when parental education was controlled for. Interestingly, the non-significant association between rural schools and the subject-specific factors became significant when parental education was controlled for. This result may suggest that in rural schools the proportion of students with less well-educated parents is higher in comparison to urban schools and this explains the non-significant relation between rural schools and the subject-specific factors.

All the associations between the teacher characteristic variables and the subject-specific factors became non-significant when parental education was controlled for with just one exception. This relates to the association between TEAAGE and the common grade factor which instead became significantly positive, suggesting that in schools with a large proportion of students with well-educated parents, the teachers are older.

Summary of the results

The results from the three studies show that grades are multidimensional in the sense that there are subject-specific dimensions as well as a common grade dimension. The pattern of these dimensions was similar at both individual and school levels. Different student characteristics influenced the subject-specific and common grade dimension in different ways, sometimes in reversed directions.

Gender differences were primarily found, at the individual level, in the common grade dimension and in the Swedish and English dimensions, where girls had higher levels of achievement. When analysing mediating models, the results showed that students' general interest for learning and motivation mediated all of the gender differences found in the common grade and English dimension and some of the differences in the Swedish dimension.

Parental education was also found to have a strong effect but in reverse directions for the subject-specific and common grade dimensions at both individual and school levels. At the school level, parental education had a significant and substantial negative relation to the common grade dimension. This may be due to compensatory grading practices where schools with a large proportion of students with less well-educated parents have a higher value on this dimension. Different school characteristics were also found to influence the subject-specific and common grade dimensions. However, when parental education was controlled for, these relations became, in most cases, non-significant. This result shows that there exist confounding relations between different school characteristics, parental education and grades.

DISCUSSION AND CONCLUSIONS

My research interest focuses on three main themes, namely the dimensionality of grades, the effects of different student and school characteristics on grades and the ways in which systematic differences in grades operate within and between schools. These different aspects of grades were explored in the three studies.

The dimensionality of grades

In order to achieve an improved understanding of the dimensionality of grades, the structure of grades was investigated by the use of multilevel CFA and SEM. The results show that grades are multidimensional (Klapp Lekholm & Cliffordson, 2008), there being both subject-specific dimensions, which account for the largest part of the variance in grades, as well as a common grade dimension, which cuts across different subject grades. Whilst this holds true on both the individual and school levels the strength of the associations are however somewhat stronger on the school level.

Whereas the three subject-specific dimensions (*Sw*, *En*, *and Ma*) thus accounted for a considerable amount of variance in the different subject grades and national tests, the common grade dimension (C*omGr*) was found to account for between three to five percent of the variance in the three subject grades at the individual level. At the school level, the common grade dimension explained between three per cent (mathematics) and nine per cent (Swedish) of the variance in grades (Klapp Lekholm & Cliffordson, 2008). Grades are thus subject to variance that cuts across different subject domains. Furthermore, this variance is related neither to what is common for the national tests and grades, nor to achievement in these subject domains.

The common grade dimension may seem to explain a rather low part of the variance in each subject grade whilst the measure of explained variance tends to give an expression of small effects. However, even a low level of explained variance may be of theoretical and practical importance. The fact that the common grade dimension influences all subject grades at the individual and school level also implies that its effects are magnified when grades are summarized.

Whilst the common grade dimension could potentially be a source of construct irrelevant variance, this need not necessarily be the case. The main reason for this is the presence of curricular goals that are not measured by the national tests, but which may be captured by grades. In the curriculum there are, on the one hand, syllabuses for different subjects with subject-specific constructs that specify what students should achieve, in terms of, for example, knowledge in mathematics or Swedish required for grade nine. On the other hand, the curriculum also includes overall goals, such as, for example, that students should develop motivation and a positive attitude towards learning and school. The variance in grades that cuts across the different subject grades - as the common grade dimension does - implies that there is variance of importance for all of these grades. One explanation is that certain skills and characteristics that are of importance in most subjects, such as, for example, motivation for learning, are not captured by the subject-specific dimensions. A difficult issue might then relate to whether motivation is a relevant or irrelevant variance in for example mathematics or social science.

Both Gustafsson and Balke (1993) and Andersson (1998) found a general school-achievement dimension in grades. Gustafsson and Balke (1993) demonstrated that a general school-achievement factor correlated around .60 with general cognitive ability measured a couple of years previously, which suggests that there exists a large amount of variance that is not related to cognitive ability. Andersson (1998) has suggested that the general school-achievement factor she found might be related to different student characteristics, and proposed that motivation for school work is one of these. These findings as well as those of several other researchers (Brookhart, 1991, 1993, 1994; Pilcher, 1994; Cizek et al., 1995; Cross & Frary, 1999; Manke & Loyd, 1990; McMillan et al., 2002) support the notion that motivation is indeed important for achieving high grades.

The dimensionality of grades is also well in line with the Encapsulation Theory, which suggests that grades are a measure which contains culturally valued knowledge (Gustafsson & Carlstedt, 2006). According to this theory, grades are a broad measure which reflects a large amount of different student cognitive and non-cognitive abilities. The common grade dimension may thus capture a broad range of different student abilities and characteristics, valued both by teachers and by society at large.

The construct of implicit education (Bergenhenegouwen, 1987) is also relevant in this context. Students who want to obtain high grades need to meet the formal as well as the informal demands of school (Bergenhenegouwen, 1987; Brookhart, 1994). For example, while characteristics such as a general interest in learning and motivation may be basic characteristics or attitudes, necessary for school achievement, they are not explicit criteria in the curriculum for the subject grades. However, this is problematic due to the nature of the curriculum and how it is interpreted. Whilst the curriculum has explicit criteria for each subject grade and for the different grade levels, it also contains stipulations concerning the overall attitudes students should develop. The criteria for the different subject grades are related to student knowledge and skills in the particular subject (National Agency for Education, 2008). Therefore, the issue of whether student motivation is a relevant source of variance in grades relates to the interpretations of the curriculum and the criteria for each subject grade. In the criteria for Swedish, English and mathematics motivation is not an expressed and explicit skill that should be a part of the variance in grades, and therefore it could be perceived as construct irrelevant variance. However, as suggested previously, this is primarily a question of interpretation of the overall curriculum. Nevertheless, one of the main purposes of the grading system and grades is to enhance students' motivation to learn. So, if motivation is an underlying construct that is of major importance for students' learning processes, goal-attainment and for success within the educational system, motivation seems to be a construct of overall importance, relevant for grades, and something which all students ought to be aware of.

Bergenhenegouwen (1987) argues that the implicit demands from teachers and schools are made partly consciously and partly unconsciously and furthermore, that students are not explicitly taught how to meet informal and implicit demands. However, if informal and implicit demands and rules do exist but are not explicitly taught, this can create a lack of transparency, inequality, uncertainty, unfairness as well as poor comparability between schools. Indeed, it has been stressed by several researchers that the transparency of assessment and grading systems is fundamental in order to evaluate and validate the quality, fairness and comparability of grades (McMillan, 2003; Kane, 2006). It is thus argued that the finding that grades are multidimensional contributes to making the grades and decision making processes more transparent in order to understand what grades measure and to discuss the validity and quality of grades.

Student characteristics of importance for grades

Another purpose of the present research was to investigate more closely the nature of the different grade dimensions, in order to discover which student characteristics they relate to. The results show that student characteristics such as, for example, if students believe they are competent in a particular subject, if they are happy in school, whether their parents are engaged and involved in their school work, and if students cooperate with each other, all exert a positive influence on the subject-specific dimensions in grades. These results also show that the subject-specific dimensions are differently influenced by the different student characteristics.

As indicated by Pilcher (1994) teachers in different subjects take differing student characteristics into account in the grading process. For instance, whereas teachers in mathematics emphasized effort when students where on the borderline to achieving a higher grade, teachers in the language domain emphasized effort without making a distinction between effort and other abilities in the particular subject. However, generally speaking, the differences among the subject-specific dimensions were only small. Students who are happy in school believe that they are coping with their schoolwork and those who have positive beliefs in their self-competence in the subjects achieve higher grades in all three subject dimensions. In the Swedish dimension the importance of having engaged and supportive parents was slightly higher in comparison to the English and mathematics dimensions.

Whilst students' specific interest in learning more in a certain subject influences the respective subject-specific dimension, general interest for learning and motivation influence the common grade dimension (Klapp Lekholm & Cliffordson, in press). The relation between student self-perception of competence in Swedish and English and the common grade dimension is negative, which suggests that the importance of the common grade dimension is greater for students who believe they are not very good in Swedish and English. One interpretation of this result is that the common grade dimension has a 'compensatory function' where low-achieving students may receive some form of compensation. As demonstrated by Stiggins et al. (1989), teachers take into account students' level of achievement, in the sense that whilst high-ability student are to a larger degree, graded on achievement only, low-ability students are graded on the basis of both achievement and non-achievement. The present result seems to support this claim, as well as providing support for McMillan's (2003) contention that teachers develop rationales for decision making that take into account the teachers' own moral considerations and beliefs, such as for instance to accommodate individual differences and pulling students. The common grade dimension may be a manifestation of a compensatory grading practice where teachers adjust their grade setting practice to the students' knowledge, skills and characteristics in order to take into account a much wider range of different aspects of the students' abilities but also in terms of a consideration of the consequences on students, teachers and schools of their decision.

The results also demonstrate that gender differences exist in the Swedish, English and the common grade dimensions, where girls' values are higher. This result confirms previous findings relating to the existence of gender differences in teacher-awarded grades, primarily in the language domains (National Agency for education, 2005). Additionally, this result confirms, and indeed extends, that of Andersson (1998) who found a strong and robust general school-achievement dimension which affected all the subject grades, where girls had higher values. Whilst Andersson could not explain this general school-achievement factor, she suggested that it was rooted in different student characteristics, a speculation which is supported by the present results.

When grading their students, teachers have to take account of a complexity of factors, some of which derive from underlying and tacit relations in a structure where some factors operate directly, whilst others do so indirectly. Whilst some student abilities or characteristics affect learning and achievement and are relevant for the construct at hand and thus should be a part of grades, others do not affect learning and may therefore be regarded as an irrelevant variance in grades (Messick, 1994).

One of the most interesting findings concerns the mediating relations between student characteristics, gender and grades. The result show that students' general interest for learning and motivation mediates all of the gender differences found in the common grade dimension and a large part of the difference in the language domains. This means that girls have a greater general interest for learning and motivation in comparison to boys, and therefore achieve higher grades. Thus, one important mechanism behind the advantage for girls in grades seems to be that they express a greater degree of interest and motivation in schoolwork than boys do. As discussed by Murphy (2000), the better school performances of girls may be due to girls coming to school better equipped for educational activities in that they have developed a learning approach that is more closely related to school practices and which is better rewarded by teachers in the grade setting practices. A learning approach, characterised by interest and motivation, is one that is culturally valued by teachers and by the curriculum. Motivation is here an indirect effect which seems to influence learning and is thus perceived as relevant variance in grades. However, if motivation for learning in the curriculum is identified as a goal, even direct effects of such a characteristic should be regarded as a source of valid variance.

The findings discussed above were reached in analyses with several different student characteristics. The use of large-scale data makes it possible to generalize the findings but limitations exist due to the particular variables used and to the latent variables or constructs that have been created. Of course, these variables have their limitations and further research could fruitfully make use of additional student characteristic variables in order to elaborate on these findings.

School characteristics of importance for grades

The results of studies I and III indicate that there are systematic differences between schools in the awarding of grades (Klapp Lekholm & Cliffordson, 2008; Klapp Lekholm, 2008). Research has indicated that different school characteristics such as school size, location, the local control of schools and teacher characteristics all have effects on student achievement and grades. There are, however, contradictory findings about the importance of these variables for school outcomes in the research literature (Hanushek, 1997; Gustafsson & Myrberg, 2002).

Parental education (study I) was found to have a strong effect on grades at the school level, in the sense that schools with a large proportion of students with well-educated parents achieved higher grades as a result of the operation of the subject-specific dimensions whereas schools with a large proportion of students with less well-educated parents achieved higher grades as a result of the operation of the common grade dimension. The results from the present studies show that while a number of separate school characteristics have effects on grades, when however parental education is controlled for most of these effects are reduced and become non-significant (Klapp Lekholm, 2008). Thus, these results show that there are confounding associations between parental education and the different school characteristics and grades. It would appear that parental education is an underlying and powerful factor influencing achievement and grades in different confounding ways, and that it is not, for instance, the local control of the school, school size or school location that per se influence the subject-specific and common grade dimensions in grades.

The initial finding that students in independently-operated schools gain higher grades is, to a large extent, due to the peer characteristics of the schools. Independent schools have a higher proportion of students with well-educated parents. Thus the independent school effect is primarily a parental educational effect. This finding is in line with the results of several researchers who have shown that an extensive control for student composition decreases the relation between independent schools and grades and in most cases, rendering it non-significant (Myrberg & Rosén, 2006, Figlio & Stone, 1999; McEwan, 2000; Somers et al., 2001). It may be noted, however, that these results contradict the conclusion drawn by Wikström (2005) that independent schools award higher grades than is motivated by the actual level of achievement of the students. This is probably because Wikström (2005) used a poor achievement measure, namely the SweSAT.

Whilst previous research has indicated that small schools award higher grades (Darling-Hammond et al., 2006; Wikström, 2005) the results of the present study do not however confirm these findings. On the contrary, the result of the analyses conducted here suggests that the effect of school size may be due to confounding influences of parental education and school location. The first analyses indicated that, in comparison to small schools, large schools awarded higher grades. However, when the effects of parental education were controlled for, the effect of large schools on the subject-specific dimension in grades became non-significant. This result supports a number of previous findings which suggest that it is not school size per se that affect achievement and grades but, rather, different demographic factors (Darling-Hammond et al., 2006; Ready & Lee, 2006).

The result also showed that students in schools located in suburban areas, in comparison to students in schools located in urban- and rural areas, are awarded higher grades. When parental education was controlled for, while suburban schools were still found to award higher grades, the influence of the common grade dimension in grades however decreased. This result may be due to confounding relations between school size, school location and parental education. Large schools are primarily located in suburban and urban areas where parents to a greater extent are well educated. Schools located in rural areas may be small schools and, in these areas, parents tend to be less well educated. Another influence concerns segregation effects where peer characteristics may differ substantially among schools located in urban areas. According to the current voucher system for compulsory education in Sweden, parents have a free choice of school for their children, which results in a segregated school system which tends to function as a tracked educational system. However, in Sweden formal tracks do not exist in compulsory school. These results are in line with Stanley et al. (2008) and Williams (2005) who claim that it is primarily the characteristics of the neighbourhood and the underlying structures in society that influence student achievement and grades.

The results also showed that certain teacher characteristics had positive associations with the subject-specific dimensions in grades. However, when parental education was controlled for, almost all of the effects decreased and became non-significant. The initial analyses showed that a high density of teachers (or small classes) and a large proportion of female teachers had positive relations to the subject-specific dimensions. Additionally, while teacher qualifications influenced the mathematics dimension, no such relations were found for the Swedish and English dimensions. When parental education was controlled for these relations also became non-significant, indicating that the initial results were due to confounding effects caused for example by compensatory allocation of resources and selection effects among teachers.

Another result revealed that younger and less experienced teachers award higher grades in comparison to older and more experienced colleagues. However, when parental education was controlled for these relations too became non-significant.

One of the main findings concerns the relations between the common grade dimension and the level of parental education at the school level. Schools with a high proportion of students whose parents have a low level of education have a substantially higher level on the common grade dimension and this effect was strongest for boys. Thus, in schools with many students whose parents have a low level of education, the importance of non-cognitive aspects (the common grade dimension) for their grades is higher.

One possible explanation may be systematic differences between schools in terms of how well established the grading criteria are at different schools, and the degree of support for grading that is in fact offered by the national tests and by the educational system generally. Another explanation for the strong effect of parental education on the common grade dimension may be that specific grading practices exist at the school level, where some schools have developed an assessment and grading culture, which takes account of a broader range of constructs not explicitly defined in the curricula and not directly relevant to the goals established for the subject area. This suggests that culturally valued knowledge and implicit education (Bergenhenegouwen, 1987) to a greater amount develop in certain types of schools, namely schools with a large proportion of students with less well-educated parents.

Stiggins et al. (1989) suggest that teachers apply different rationales in the grade setting practice as a result of students' achievement levels. As suggested by McMillan (2003) teachers' decision-making includes a broad range of factors which, sometimes, can be contradictory. The compensatory grading practice is possibly a way for teachers to handle the external pressure for objective grading and the consequences of their decisions, that teachers' decision making primarily concerns how to adjust students' knowledge, skills and characteristics in order to motivate the students to learn.

Additionally, it could also be that teachers try to get as many students as possible up to the "pass" level because of the negative consequences for students, teachers and schools that are associated with a "not pass" grade. It has also been suggested that the competition between schools for attracting students may also cause schools to assign higher grades and hence compensate for poor results on the national tests which may, it has been suggested, lead to grade inflation and differences between schools (Cliffordson, 2004; Cross & Frary, 1999; Wikström, 2005).

According to Deutsh (1979) and Pilcher (1994) teachers appear to be concerned with the distributive justice of grades and with the consequences that grades have for their students. Accordingly, several researchers emphasize that teachers need to develop their analytic skills in order to make reliable and valid interpretations of student performances (Shepard, 1993; Messick, 1994).

Issues of validity

Classroom assessment, which is the basic grading procedure in the Swedish educational system, has been subjected to heavy criticism due to the risk for subjectivity in assigning grades. Since the criteria for the different levels of achievement are formulated locally based on interpretations of the national syllabus and curriculum goals, and there are no standardized examination tests, the reliability, fairness and validity of the grades have all been questioned (National Agency for Education, 2007; Tholin, 2006; Wikström, 2005).

However, McMillan (2003) claims that in interpretations of student performances, subjectivity is always present and indeed has to be so since the nature of assessments in the classroom and the awarding of grades are characterized by subjectivity and intuition. Teachers' internal and moral beliefs are therefore not congruent with the external pressure of objective grading (McMillan, 2003; Kane, 2006; Wiliam, 1996). The multidimensionality of grades found at both the individual and school level suggests that teachers take account of both cognitive and non-cognitive student abilities when awarding grades. The distinction between quantitative and qualitative models of validation that Kane (2006) suggests seems valuable when discussing the multidimensionality of grades. While the quantitative model aims to eliminate the influence of different conditions, such as social and physical context and student characteristics, when assessing students, the qualitative model tries to integrate all kinds of evidence in order to reach a coherent view of the student and student abilities. The existence of the common grade dimension in grades over and above the subject-specific dimensions indicates that teachers use the qualitative model when evaluating the performances of their students. In this practice, teachers integrate student knowledge, skills and characteristics such as interest, motivation and parental support as well as student subject knowledge into the grades that they assign. It is also argued that this is one reason why grades function as a good predictive instrument of student achievement for subsequent levels in the educational system (Cliffordson, 2008; Gustafsson & Carlstedt, 2006).

Different student characteristics were found to influence the different dimensions in grades. General interest for learning and motivation had the greatest effect on the common grade dimension. This can be interpreted in different ways. A student characteristic that does not affect learning, but affects the grades directly may be perceived as construct irrelevant variance in grades. However, student characteristics which affect learning and achievement, and hence influence grades indirectly, may thus be perceived as construct relevant variance. Another example of these types of relations concerns the gender differences found in this research. For example, if students are awarded higher grades due to their sex, this is irrelevant to the goals and criteria in the curriculum and indeed is an infringement on the rights of the individual. However, the result from the mediating models showed that the gender difference in the Swedish, English and common grade dimensions, where girls are favoured, were almost fully explained by general interest for learning and motivation. This indicates that girls have a greater interest in learning and are therefore awarded higher grades. Hence it could be argued that what seems to be construct irrelevant variance caused by gender is instead mediated by certain student characteristics that are relevant for student learning and achievement, thus contributing construct relevant variance in grades.

When applying the bridge analogy (Kane, 2006) to the grade setting practice in Sweden, where grades are based primarily on classroom assessments, it becomes evident that there are threats to the construct validity all through the three bridges of the assessment and grading process. The teachers themselves decide what to teach, they choose content and formats, they observe students' classroom performances, they evaluate these performances and award grades and in so doing, they value their own interpretations and validate their own inferences. So the issue of construct validity is highly relevant and must be properly recognized. The dimensionality of grades suggests that tacit, implicit and underlying structures constitute a part of the variance in grades. According to Kane's (2006) theory one of the major threats to construct validity is, in particular, the first step in the bridge analogy, where teachers' observations may include construct irrelevant variance as well as the underrepresentation of constructs. However, in the Swedish educational system, the national tests function as an instrument to exemplify the goals and criteria in the curriculum, as well as to assist in the process of setting fair and reliable grades in order to support teachers' decision-making processes (National Agency for Education, 2005; Åberg-Bengtsson & Erickson, 2006). If threats to construct validity are present in the observation phase, the whole assessment procedure and the interpretations of students' performances, that teachers make, may be distorted, thus resulting in invalid inferences.

The dimensionality of grades indicates that teachers' grading practices are tacit and implicit, as evidenced, for example, by the influence of student adjustment and parental engagement influence on grades. This may lead to non-transparent grading processes which influence the possibility to evaluate the arguments for a certain grade, and hence influence the validity, comparability and fairness of the grades. Kane (2006) and Messick (1994) emphasize that the interpretive argument needs to be as explicit as possible, and indeed that all assumptions in the grading process must be made explicit.

Perhaps this is one reason why Messick's unified validity theory seems most relevant when validating grades. Whilst the different evidences or arguments are complementary to one another, they are not alternatives, and consequences are also included in the validity arguments. As Brookhart (1991), Pilcher (1994) and Deutsch (1979) have all argued, teachers can focus more on the use and consequences of grades which also constitutes the meaning of grades to teachers, sometimes at the expense of their interpretation. The compensatory grading practice, as discussed in this research, suggests that the consequences of grades for the students are of major importance for teachers. The consequences may concern different aspects, such as, for example, that students need a certain grade in order to continue to the next level in the educational system or that teachers feel a moral responsibility to reward students for the interest and motivation they have shown in their schoolwork. The distributive justice and compensatory grading practice may thus be fundamental in order to understand the rationales teachers employ in their grading practice (Deutsch, 1979; Pilcher, 1994). Messick's unified validity approach (1989; 1994) thus incorporates teachers' conceptual and moral beliefs and the fact that teachers bring with them their own beliefs, values and moral considerations to the decision making process, and in drawing inferences about what grade to assign. The validation process, as suggested by certain researchers (Messick, 1989; Wiliam, 1996), needs to be based upon broad conceptions of validity that take into account a variety of evidence which cuts across the domains in focus, which, in the context of the research carried out here, can be found reflected in the dimensionality of grades. It has also been emphasized that traditional measurement principles are inappropriate when validating assessment decisions in school. In line with Messick and Wiliam, McMillan (2003) also argues, that the subjective and intuitive characteristics of classroom assessments need to be broader in terms of the definition of validity. However, Messick (1994) emphasizes that both evidential and consequential evidences of student performances are of importance in the decision making process and claims that, in educational settings, the consequential aspects of validity are often emphasized at the expense of the evidential aspects.

So, if we recognize that teachers' decision making and grade awarding practices in school are a complex business involving a degree of subjectivity, teachers' rationales for these decisions must be in focus. McMillan's (2003) model of teachers' decision-making highlights the complexities of the many contradicting forces that teachers have to face. If the educational system and the stakeholders in school recognize this hodgepodge of factors that teachers have to handle in the grade setting practice, the transparency of these practices would be highlighted. The subjectivity in grades may be better understood and indeed perceived differently if teachers' rationales for decision-making were transparent and that some of the factors taken into account by teachers in their grading were recognized as important for learning and relevant variance in grades.

Conclusions

The findings from the present research indicate that grades are multidimensional as evidenced by the subject-specific and common grade dimensions. This dimensionality, which was robust, was found at both the individual and school level. Whilst the subject-specific dimensions are perceived as cognitive dimensions which reflect knowledge and skills in different subject areas, the common grade dimension is perceived as a non-cognitive dimension which reflects other aspects in the grading practice. The findings from the analyses of the composition of the common grade dimension suggest that the common grade dimension is related to different student characteristics of which some can be perceived as construct relevant variance whereas others can be perceived as construct irrelevant variance in grades. This is due to whether there is a direct or an indirect influence of student characteristics on grades. However, whilst the influence of students' general interest for learning on the common grade dimension may be perceived as construct irrelevant variance for a certain grade, it may nevertheless be relevant for student achievement and with respect to the overall goals in the curriculum, demonstrating that students need to develop a positive attitude towards learning and school. The common grade dimension is suggested as being reflective of the moral dilemmas teachers face when assessing and awarding grades to their students. The conflict between the pressure for objective grading and the subjective relations between teachers and students in the classroom may lead to opaque grading practices which compensate low-achieving students to a greater extent by taking student characteristics into account in the grades.

At the individual level, gender differences were found in the Swedish, English and the common grade dimension but, when mediating models were analysed, it was found that the general interest for learning mediated all of the gender effect in the English and the common grade dimension and some of the gender effect in the Swedish dimension. This is explained by girls showing a greater interest and motivation for learning which is a characteristic valued by the teachers when awarding grades.

At the school level, parental education was found to have a strong negative effect on the common grade dimension, indicating that in schools with a large proportion of students with parents with low levels of education, students benefit in the common grade dimension. This is interpreted as reflecting a compensatory grading practice where teachers take into account a different amount of cognitive and non-cognitive student abilities when awarding grades. Furthermore, the analyses of the school characteristics showed that there were many confounding relations between the different school characteristics, grades and parental education.

Methodological issues

The result of this research highlights the issue of using appropriate methods when conducting studies involving educational data. In order to take account of both the hierarchical and complex characteristics of educational data, multivariate multilevel approaches are necessary in order to obtain adequate research results. The results from the studies reported here are indicative of the considerable analytic power offered by these techniques when applied to large-scale data. To unravel the complexities it seems necessary both to apply latent variable and multilevel modelling.

Suggestions for further research

The results of the present research show that the investigated student and school characteristics are somewhat limited, and there may be other characteristics of importance for grades. Additionally student and school characteristics may provide a more thorough and detailed perspective on the meaning of grades. Primarily, it would appear that it is the instrument measuring teacher competence that seems to be rather broad and insufficiently precise in that it may not capture important and relevant aspects of teacher competence, education and qualifications. Therefore, it seems of major importance to develop instruments that measure teacher abilities and skills more adequately as a means of investigating the effects of teacher competence on the dimensionality of grades.

A suggestion for further research concerns the investigation of the comparability of grades; for example whether grades are comparable over time, teachers and schools. Another area of interest concerns the effect of grades on student motivation and learning, as well as the importance of the dimensionality of grades and the role the cognitive and non-cognitive dimensions have in predicting success in the educational system.

REFERENCES

Agnew, E. J. (1985, April). The grading policies and practices of high school teachers. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology*, 3, 545-561.

Airasian, W. P. (1993). Policy-driven assessment or assessment-driven policy? *Measurement and Evaluation in Counseling and Development*, *26*(1), 22-31.

Airasian, W. P., & Jones, M. A. (1993). The teacher as applied measurer: Realities of classroom measurement and assessment. *Applied Measurement in Education*, *6*(3), 241-254.

Alexander, W. P. (1935). Intelligence concrete and abstract. *British Journal of Psychology*. Monograph Supplement, 19, 177.

Andersson, A. (1998). The dimensionality of the leaving certificate in the Swedish compulsory school. *Scandinavian Journal of Educational Research, 42(*1), 25-40.

Andersson, C., & Waldenström, N. (2007). Teacher certification and student achievement in Swedish compulsory schools. Working paper 2007:6. Uppsala, Sweden: The Institute for Labour Market Policy Evaluation.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.

Bergenhenegouwen, G. (1987). Hidden curriculum in the university. *Higher Education*, 16, 535-543.

Bergin, D. A. (1999). Influences on classroom interest. *Educational Psychologist*, 34, 87-98.

Bergström, F., & Sandström, M. (2001). *Konkurrens bildar skola – en ESO-rapport om friskolornas betydelse för de kommunala skolorna.* [Competition and the Quality of Municipal Schools] Ds 2001:12. Stockholm, Sweden: Ministry of Finance.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Bloom, B. S. (1976). *Human Characteristics and School Learning.* New York: McGraw Hill.

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice, 10*(1), 35-36.

Brookhart, S. M. (1993). Teachers´ grading practices: Meaning and values. *Journal of Educational Measurement, 30*(2), 123-142.

Brookhart, S. M. (1994). Teachers´ grading: Practice and theory. *Applied Measurement in Education, 7*(4), 279-301.

Brookhart, S. M., & Durkin, D. T. (2003). Classroom assessment, student motivation and achievement in high school social studies classes. *Applied Measurement in Education*, *16*(1), 27-54.

Carlgren, I., & Marton, F. (2000). *Lärare av imorgon* [Teachers of tomorrow]. Stockholm: Lärarförbundets Förlag.

Carroll, J. B. (1982). The measurement of intelligence. In R. J. Sternberg (Ed.) *Handbook of Human Intelligence*. New York: Cambridge University Press.

Cattell, R. B. (1971) *Abilities: Their structure, growth and action*. Boston: Houghton-Mifflin.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. New York: North-Holland.

Cliffordson, C. (2004). Betygsinflation i de målrelaterade gymnasiebetygen. [Inflation in goal-referenced grades from upper secondary school]. *Pedagogisk forskning i Sverige, 9*(1), 1-14.

Cliffordson, C. (2008). Differential prediction of study success across academic programs in the Swedish context: The validity of grades and tests as selection instruments for higher education. *Educational Assessment, 13*(1), 56-75.

Cizek, G. J., Fitzgerald, S., & Rachor, R. E. (1995). Teachers´ assessment practices: Preparation, isolation and the kitchen sink. *Educational assessment*, *3*(2), 159-179.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartlant, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U. S. Government Printing Office.

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

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

Darling-Hammond, L. (1999). *Teacher quality and student achievement: A review of state of policy evidence.* Washington, DC: U. S: Center for the Study of Teaching and Policy, University of Washington.

Darling-Hammond, L. (2000). How teacher education matters. *Journal of Teacher Education*, *51*(3), 166-173.

Darling-Hammond, L., Ross, P., & Milliken, M. (2006). High school size, organization, and content: What matters for student success? Brookings Papers on Educational Policy: 2006/2007.

Deutsch, M. (1979). Education and distributive justice. *American Psychologists*, *34*(5), 391-401.

Figlio, D. N., & Stone, J. A. (1999). School choice and student performance: Are private schools really better? Working paper 1141-97, California: Institute for Research on Poverty, University of Oregon.

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., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity.* Buckinghamshire: Open University Press.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. Gagné (Ed.). *Psychological principles in system development*, (pp. 421-427). New York: Holt, Rinehart & Winston,

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *66*(3), 361-396.

Gustafsson, J.-E. (2001). On the hierarchical structure of personality and ability. In J. Collis & S. Messick (Eds.). Intelligence and personality: Bridging the gap in theory and measurement (pp. 25-42). Mahwah, NJ: Lawrence Erlbaum Associates.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioural Research*, *28*(4), 407-434.

Gustafsson, J.-E., & Carlstedt, B. (2006, August). Abilities and grades as predictors of achievement: The encapsulation theory. Paper presented at the symposium "The investment theory of intelligence: New evidence, new challenges" at the annual meeting of the American Psychological Associations, New Orleans.

Gustafsson, J.-E., & Myrberg, E. (2002). *Ekonomiska resursers betydelse för pedagogiska resultat* [The importance of economical resources on pedagogical results]. Stockholm: Liber.

Gustafsson, J.-E., & Stahl, P.-A. (2005). *STREAMS 3.0 User's Guide.* Mölndal, Sweden: Multivariateware.

Hattie, J. A., Biggs, J., & Purdie, N. (1996). Effects of learning skills intervention on student learning: A meta-analysis. *Review of Research in Education*, 66, 99-136.

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher, 18*(4), 45-65.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An Update. *Educational Evaluation and Policy Analysis*, *19*(2), 141-164.

Hanushek, E. A., & Luque, J. A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review*, 22, 481-502.

Hanushek, E. A. (2007). Some U.S. evidence on how the distribution of educational outcomes can be changed. In L. Woessmann & P. E. Peterson (Eds.), *Schools and the equal opportunity problem* (pp. 159-190). Cambridge, Mass.: MIT Press, cop.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5-14.

Hidi, S., Renninger, K. A., & Krapp, A. (2004). Interest, a motivational variable that combines affective and cognitive functioning. In D. Y. Dai & R. J. Sternberg (Eds.), *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 89-115). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 4, 111-127.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* New Jersey: Lawrence Erlbaum Associates, Inc.

Härnqvist, K. (2000). Evaluation through follow-up. A longitudinal program for studying education and career development. In C.-G. Janson (Ed.), *Seven Swedish Longitudinal Studies in Behavioral Science.* Stockholm: Forskningsrådsnämnden.

Jarl, M., Kjellgren, H., & Quennerstedt, A. (2007). Förändringar i skolans organisation. [Changes in the organisation of school] In J. Pierre (Ed.), *Skolan som politisk organisation* [School as a political organisation] (pp. 23-48). Lund: Gleerups Utbildning AB.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen, & J. Scott Long (Eds.), *Testing Structural Equation Models* (pp. 294-316). California: Sage Publications.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport, CT: Praeger Publishers.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performances. *Educational Measurement: Issues and practice,* Summer, 5-17.

Klapp Lekholm, A., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school levels: Effects of gender and family background. *Educational Research and Evaluation*, *14*(2), 181-199.

Klapp Lekholm, A., & Cliffordson, C. (in press). Effects of student characteristics on grades in compulsory school. *Educational Research and Evaluation.*

Klapp Lekholm, A. (submitted). Effects of school characteristics on grades in compulsory school.

Lane, K. L., Givner, C. C., & Pierson, M. (2004). Secondary teachers´ view on social competence: Skills essential for success. *The Journal of Special Education*, 3, 174-186.

Leuven, E., & Oosterbeek, H. (2007). The effectiveness of human-capital policies for disadvantaged groups in the Netherlands. In L. Woessmann & P. E. Peterson (Eds.), *Schools and the equal opportunity problem* (pp. 191- 208). Cambridge, Mass.: MIT Press, cop.

Lindensjö, B., & Lundgren, U. P. (2000). *Utbildningsreformer och politisk styrning* [Educational reforms and policy steering]. Stockholm: HLS Förlag.

Loehlin, J. C. (2004). *Latent variable models.* London: Lawrence Erlbaum Associates, Publishers.

Lohman, D. (2004). Aptitude for college: The importance of reasoning tests for minority admissions. In R. Zwick (Ed.). *Rethinking the SAT. The future of standardized testing in university admissions* (pp. 41-56). New York and London: RoutledgeFalmer.

Machin, S. (2007). Education expansion and intergenerational mobility in Britain. In L. Woessmann & P. E. Peterson (Eds.), *Schools and the equal opportunity* problem (pp. 29-50). Cambridge, Mass: MIT Press, cop.

Manke, M. P., & Loyd, H. (1990, April). An investigation of nonachievement related factor influencing teachers' grading practices. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

McEwan, J. P. (2000, June). Comparing the effectiveness of public and private schools: A review of evidence and interpretation. Occasional paper presented at teachers college, Columbia University, New York, National Centre for the study of privatization in education.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary Teachers´ Classroom Assessment and Grading Practices. *The Journal of Educational Research*, *95*(4), 203-213.

McMillan, J. H. (2003). The relationship between instructional and classroom assessment practices of elementary teachers and student scores on high-stakes tests. Manuscript submitted for publication in *ERIC, Clearinghouse on Assessment and Evaluation*, January, 2003.

Maruyama, G. M. (1998). *Basics of structural equation modelling.* Thousand Oaks: Sage Publications, Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition) (pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-8.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12.

Murphy, P. (2000). Equity, assessment and gender. In J. Salisbury & S. Riddell (Eds.), *Gender, Policy & Educational Change* (pp. 134-152). London: Routledge.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modelling with data that are not missing completely at random. *Psychometrica, 52*(3), 431-462.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, 22, 376-398.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus User's Guide* (3rd edition). Los Angeles, CA: Muthén & Muthén.

Myrberg, E. (2007). The effect of formal teacher education on reading achievement of 3rd-grade students in public and independent schools in Sweden. *Educational Studies*, *33*(2), 145-162.

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

National Agency for Education. (2000). Utbildningsvillkor i glesbygd [Conditions of education in sparsely populated areas]. Stockholm: National Agency for Education.

National Agency for Education. (2003). *Skolverkets lägesbedömning 2003* [The National Agency for Education's determination of position 2003]. Stockholm: National Agency for Education.

National Agency for Education. (2005). National assessment and grading in the Swedish school system. Retrieved August 4, 2008 from http://www.skolverket.se/publikationer?id=1524

National Agency for Education. (2007). National subject tests in compulsory school. Retrieved May 2, 2007 from http://www.skolverket.se/sb/d/276

National Agency for Education. (2008). Ämnesprovet 2007 i grundskolans årskurs 9 [The national subject tests in compulsory school, year nine] Retrieved August 4, 2008 from http://www.skolverket.se/publikationer?id =1779

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education*, *14*(2), 149-170.

Peterson, P. E., & Woessmann, L. (2007). Introduction: Schools and the equal opportunity problem. In L. Woessmann & P. E. Peterson (Eds.), *Schools and the equal opportunity problem* (pp. 3-27). Cambridge, Mass.: MIT Press, cop.

Pilcher, J. K. (1994). The value-driven meaning of grades. *Educational Assessment, 2*(1), 69-88.

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice*, 41, 219-225.

Ready, D. D., & Lee, V. E. (2006). Optimal Context size in elementary schools: Disentangling the effects of class size and school size. *Brookings Papers on Education Policy*, 2006/2007, 99-135.

Roscingo, V. J., & Crowley, M. L. (2001). Rurality, institutional disadvantage, and achievement/attainment. *Rural Sociology*, *66*(20), 268–292.

Rosén, M. (1998). Gender differences in patterns of knowledge. *Göteborg studies in educational science, Vol. 124*. Göteborg: Acta Universitatis Gothoburgensis.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, 18, 119-144.

Schafer, L. J., & Graham, W. J. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147-177.

Segers, M. (2008, July). The relations between assessment and student learning: Research evidence, challenges and future directions. Paper presented at the JURE conference of EARLI: Innovative and creative perspectives: New directions in educational research.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Somers, M.-A., McEwan, P. J., & Willms, D. J. (2001). How effective are private schools in Latin America? Occasional paper, New York: National Center for the Study of Privatization in Education, Columbia University.

SOU 1942:11. Betygsättning i folkskolan. Betänkande med utredning och förslag angående betygsättningen i folkskolan avgivet inom Ecklesiastikdepartementet tillkallade sakkunniga [Grade setting in primary school. Considerations of the investigation and proposal regarding the setting of grades in the primary school organized within the Ecclesiastical department]. Stockholm: Ecklesiastikdepartementet.

Stanley, R. L., Comello, M. L. G., Edwards, W. R., & Marquart, S. B. (2008). School adjustment in rural and urban communities: Do students from “Timbuktu” differ from their “City slicker” peers? *Youth Adolescence*, 37, 225-238.

Stiggins, R. J., & Bridgeford, R. J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, *22*(4), 271-286.

Stiggins, R.J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* Albany, NY: SUNY Press.

Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice*, *8*(2), 5-14.

Swedish National Rural Development Agency. Rural Sweden. Retrieved March 4, 2008 from http://www.glesbygdsverket.se/site/default.aspx.

Tholin, J. (2006). *Att kunna klara sig i ökänd 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 the implementation of a new system]. (Doctoral thesis). Borås: The University college of Borås, Department of Education.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*(1), 89-122.

Wedman, I. (1983). *Den eviga betygsfrågan. Historiskt och aktuellt om betygssättningen i skolan* [The everlasting issue about grades. History and present about grade assignment in school]. Utbildning och forskning, FoU Rapport 48. Stockholm: Skolöverstyrelsen.

Wenglinsky, H. (1998). Finance equalization and within-school equity: The relationship between education spending and the social distribution of achievement. *Educational Evaluation and Policy Analysis*, *20*(4), 269-283.

Wentzel, K. R. (1989). Adolescent classroom goal, standards for performance, and academic achievement: An interactionist perspective on primary prevention. *Journal of Consulting and Clinical Psychology,* 59, 830-841.

Wentzel, K. R. (1991). Relations between social competence and academic achievement in early adolescence. *Child Development*, 62, 1066-1078.

Wikström, C. (2005). Criterion-Referenced Measurement for Educational Evaluation and *Selection. No. 1.* Umeå: Umeå University, Department of Educational measurement.

Wiliam, D. (1996). National curriculum assessments and programmes of study: Validity and impact. *British Educational Research Journal, 22*(1), 129-142.

Williams, S. L. (2001). City kids and country cousins: Rural and urban youths, deviance, and Labor market ties. In R. T. Michael (Ed.), *Social awakening* (pp. 379–414). New York: Russell Sage Foundation.

Yang, Y. (2003). Measuring socioeconomic status and its effects at individual and collective levels: A cross country comparison. *Göteborg studies in educational science, Vol. 193.* Göteborg: Acta Universitatis Gothoburgensis.

Åberg-Bengtsson, L. (2004). Do small rural schools differ? A comparative two-level model of reading achievement among Swedish 9-year-olds. *Scandinavian Journal of Educational Research*, *48*(1), 19-33.

Åberg-Bengtsson, L., & Erickson, G. (2006). Dimensions of national test performance: A two-level approach. *Educational Research and Evaluation*, *12*(5), 469-488.

Study I

Discrepancies between school grades and test scores at individual and school level: effects of gender and family background.

Educational Research and Evaluation, 14(2), 181-199, 2008

Study II

Effects of student characteristics on grades in compulsory school.

Educational Research and Evaluation (in press)

Study III

Effects of school characteristics on grades in compulsory school

(Submitted)