Educational Research and Evaluation
Vol. 14, No. 2, April 2008, 181–199

Routledge
Taylor & Francis Group

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

Alli Klapp Lekholm[a]* and Christina Cliffordson[a,b]

[a]*Department of Education, University of Gothenburg, Sweden;* [b]*University West, Sweden*

(*Received 22 May 2007; final version received 11 October 2007*)

The purpose of the study was to use multivariate multilevel techniques to investigate whether it was possible to separate different dimensions in grades that relate to subject-matter achievement and to other factors. Data were derived from The Gothenburg Educational Longitudinal Database (GOLD), and the subjects were 99,070 ninth-grade students born in 1987. The analyses were based on subject grades and scores on national tests in Swedish, English, and mathematics. The results showed that, at both individual and school levels, the greatest part of the variance in grades was due to achievement in the different subject areas. At both levels, it was possible to identify a dimension that cut across the grades in all 3 subjects, which suggests that grading is influenced by factors other than achievement. One of the most interesting results concerns the relation between parental education and the common grade dimension at the school level.

Keywords: grades; grade assignment; compulsory school; national tests; gender; family background

Introduction

In many educational systems, grades are the main instrument used to measure student success in school. Grades may have a number of functions, depending on the grading system used and its purposes. In Sweden, the main functions of the grading system are to provide information about students' attainment in relation to required standards, to evaluate the educational system, and to be an instrument for selection to the next level in the educational system. The recently introduced system of criterion-referenced grading is based on the evaluation of student attainment measured against defined goals. According to policy documents on the grading system and recommendations in the National Curriculum, the purpose of grades is to measure students' subject knowledge (National Agency for Education, 2004). Since the Swedish grading system is highly decentralized, leaving teachers with the primary responsibility for instruction, assessment, and the award of grades, there may be elements of subjectivity in the grading process. Research indicates that teachers' evaluations of student performance may be related to students' individual characteristics and that evaluations may differ between teachers, which, among other things, can be a cause of grade inflation and differences between schools (Cliffordson, 2004; Wikström, 2005).

*Corresponding author. Email: alli.klapp.lekholm@ped.gu.se

ISSN 1380-3611 print/ISSN 1744-4187 online

DOI: 10.1080/13803610801956663
http://www.informaworld.com

Several studies have indicated that grades not only reflect academic subject knowledge and skills but also student characteristics, such as temperament, effort, motivation, and communicative abilities (Alexander, 1935; Andersson, 1998; Brookhart, 1991; Cizek, Fitzgerald & Rachor, 1995; Cross & Frary, 1999; McMillan, Myran, & Workman, 2002; Selghed, 2004; Stiggins & Bridgeford, 1985). Student characteristics may influence grades both directly and indirectly. Thus, some characteristics, such as motivation, influence student learning and therefore have an indirect influence on grades. Other characteristics, such as politeness, may influence grades in spite of the fact that such characteristics do not influence student learning. Such direct effects may be seen as contributing construct-irrelevant variance (Messick, 1994) to grades, since they influence the grading process but are irrelevant in relation to the criteria defined in the curriculum.

The main aim of the present study is to investigate whether there are different dimensions in grades which can be associated with achievement and non-achievement.

Previous research

Achievement and non-achievement

Research indicates that grades reflect factors other than just achievement. Parsons (1959), for example, identified both cognitive and moral dimensions in grades in elementary school:

> In the elementary grades these two components are not clearly differentiated from each other. Rather, the pupil is evaluated in diffusely general terms; a good student is defined in terms of a fusion of the cognitive and the moral components, in which varying weights are given to one or the other. Broadly speaking, then we may say that the "high achievers" of the elementary school are both the "bright" pupils, who catch on more easily to their more strict intellectual tasks, and the more "responsible" pupils, who, "behave well" and on whom the teacher can "count" in her difficult problem of managing her class. (p. 304)

Alexander (1935) investigated relations between students' abilities and the influence these abilities have on student achievement. Of particular interest in Alexander's study was the appearance of a certain factor which influenced all student achievement, which he labelled the "X" factor. This dimension was interpreted as being a personality characteristic defined by determination and persistence and which would be of importance for student success in all school subjects. According to Alexander (1935), different student characteristics played different parts in different subject-related areas, the complexity of the problems involved being one determinant. However, the distinction between direct and indirect influences was not stressed by Alexander, who primarily emphasized student characteristics that had an influence on learning and which, in turn, had an influence on achievement and grades.

In order to understand what things influence grades, teachers' grading practices form a central area of enquiry. Teachers interpret grades and grading recommendations, as well as assessing and grading students. Manke and Loyd (1990) showed that different student characteristics, such as effort, behaviour, personality, and homework completion, are commonly taken into consideration when teachers assign grades. According to Nava and Loyd (1992), student classroom behaviour, non-content academic skills, and student personal characteristics may also be of importance when teachers assign grades. Additional aspects may be punctual submission, completion of assignments, and working hard both in and out of class (e.g., Cox, 1995; Pilcher-Carlton & Oosterhof, 1993).

According to Pilcher (1994), teachers in mathematics graded students in a more concrete "right or wrong" manner and they generally described their grading practice as objective. Effort was used as a grading criterion when adjusting grades, for example taking into consideration homework, which mirrors student effort. In contrast, English teachers graded their students without making a distinction between student efforts and writing ability. Compared to their mathematics colleagues, they described their grading practices as being subjective. Additionally, students' levels of achievement also seem to influence grading practices (e.g., Korp, 2006; Stiggins, Frisbie, & Griswold, 1989), with low-ability students being graded both on the basis of achievement and non-achievement, whereas high-ability students are, to a greater extent, graded on achievement alone.

It has been suggested that taking non-achievement into consideration when assigning grades may serve to protect students, teachers, and schools from negative consequences (Cross & Frary, 1999; Wikström, 2005). By regarding both achievement and non-achievement when assigning grades, it is possible to maximize student grade outcomes, which may be perceived as a win-win situation where students, teachers, and schools all benefit (Brookhart, 1991; Wikström, 2005). Differences regarding grade setting practices may be related to the type of school (e.g., independent or public schools), and school culture may play a part in determining the extent to which teachers take achievement and non-achievement into account when grading students (Agnew, 1985; Cizek et al., 1995; Wikström & Wikström, 2005). A few studies (Agnew, 1985; Tholin, 2006) also indicate the existence of differences in grading practices of teachers in academic and non-academic subjects. Teachers of non-academic subjects tend to weigh non-achievement more heavily than their colleagues in academic subjects. The behaviour of teachers may more or less be in correspondence with the goals of the educational system. The relation between different stakeholders in the educational system (e.g., teachers and principal) may suffer from information gaps and communication difficulties when monitoring the educational process (Bishop, 2001; Woessmann, 2002). Indeed, informational asymmetries and different interests among stakeholders affect teachers and school culture, which may influence the assessment and grading practices.

Student background

Grades are also influenced by the students' gender and their social and national backgrounds. It has been found that these different background characteristics may influence grades, both directly and indirectly. The direct influence of student characteristics can include social behaviour or politeness, whilst indirect influences could be communicative skills or the degree of effort expended (National Agency for Education, 2006; Nycander, 2006; Svensson, 1971).

Several studies have shown that students from a higher socioeconomic background are more likely to perform well in school and are more often on academic track (Hanushek & Luque, 2003; Wikström, 2005). Even though the definition of family socioeconomic status (SES) differs, some basic dimensions of SES, such as family income, parents' level of education, and occupation, 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. However, there may be reason to look upon these SES definitions as more diffuse, non-unitary concepts, since it has been shown that different dimensions of SES relate differently to student achievement (Bloom, 1976; Yang, 2003).

An additional factor of importance to be taken into consideration is the fact that female students seem to perform better in school and also obtain better grades (National Agency for Education, 2003). A recent Swedish study showed that girls get better grades in relation to their performance on national tests than boys (Nycander, 2006). At the end of compulsory school, girls have better grades in all subjects, with the exception of sports/athletics. The differences in grades are smallest in mathematics, physics, and technology and greatest in art, home economics, religion, and Swedish (National Agency for Education, 2005; Nycander, 2006). The results of national tests in Swedish, English, and mathematics reveal certain gender differences but not, however, to the same extent as teacher-awarded grades. These gender-related differences are most apparent in the national test for Swedish, whereas differences in the national test for mathematics are very small. In English, they are moderate (Nycander, 2006). Additionally, girls and boys seem to develop different approaches towards the learning environment; whilst girls seem to nurture their general cognitive abilities, boys nurture their specific cognitive abilities (Rosén, 1998; Wernersson, 1989, 1992).

Validity of grades

In the construct-centred or "unified" concept of validity, construct validity subsumes and integrates different aspects of validity (Messick, 1989). In order to justify interpretation and the use of scores (scores in a broad sense, both qualitative and quantitative summaries, where grades may be seen as a quantitative summary), multiple sources of evidence provide a framework for the validation of assessment where each source of evidence (content relevance and representation, substantive, internal, external, generalizability, and consequences) contributes to make the most reasonable case (Downing, 2003; Messick, 1991, 1994). Different modes of assessment, for example, performance assessment and multiple-choice testing, may seem to render different validation processes due to their different nature, but, as Messick points out, in the light of the unified validity concept "it is not the test or observation device as such that is being evaluated but rather the inferences derived from test scores or other indicators" (Messick, 1991, p. 2).

Among other things, evidence is supposed to function as a mean to seek arguments to avoid two major threats to construct validity, namely construct-irrelevant variance and construct underrepresentation. The former implies that the measure or assessment contains variance connected to constructs other than the one in focus, whereas the latter implies that the measure is too narrow, not covering important aspects of the focal construct.

The unified validity approach focuses on the complexities of knowledge, skills, or other attributes to be assessed, which highlights questions concerning the types of behaviours and performances the construct should reveal. Scores or grades were interpreted in terms of the current construct, where some attributes are consistent with the scores or grades and correlates to some characteristics of individuals or other objects of measurement. Given this interpretation, when used in the admissions process to the next level in the educational system, grades could, for example, 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. Messick (1994) suggests that "any negative impact on individuals or groups, especially gender and racial groups, should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance" (p. 34). For example, a student's low score ought not to be low because the measurement is missing some relevant construct which, if present, would have revealed the student's real

ability. Obviously, a central issue concerns the identification of factors that constitute the construct-irrelevant variance in grades.

According to Dylan (1996), the subject domains are in focus in most educational assessment situations. A domain is defined by experts as well as different assessments, the curriculum, and grading criteria. Certain aspects of a domain and different constructs related to the domain are emphasized, while others are de-emphasized. Dylan (1996) argues that the inferences that are made are related to the domain that is being assessed (within-domain), but also related to other domains (beyond-domain). Dylan has modified and simplified Messick's four-facet framework for validity arguments and argues that "any process of validation must address inferences from, and consequences of, assessments, both within and beyond the domains that they address" (p. 142).

Research in this field indicates that teachers consider consequences of grades when grading students and that they interpret grades in different ways (Brookhart, 1993, 1994). Additionally, distributive justice of grades, a definition focusing the use and consequences of grades which affect teacher's grading practices (Deutsch, 1979; Pilcher, 1994), is also a factor affecting the award of grades. Indeed, Pilcher (1994) suggests that "teachers realize that students use grades to judge their own merit and others use grades to make decisions about students ... perhaps teachers are more concerned with the distributive justice of a grade (how it is used and consequences for its use) than with how a grade is interpreted" (p. 71).

Purposes

The review of the literature indicates that there may be reason to suspect that the award of grades is influenced by factors other than achievement, such as different student characteristics. Other sources of influence seem to operate at the school level. If grades are systematically affected by other factors than achievement, this should manifest itself as a common grade dimension which cuts across different subjects and academic fields. This dimension may reflect variance due both to individual and school-level factors, which makes it necessary to adopt a multilevel analytical approach.

The main purpose of this study is to use multivariate multilevel techniques to ascertain whether it is possible to identify and separate different dimensions in grades, which, on the one hand, may be interpreted as expressing variance in knowledge and skills and, on the other, different systematic factors. Another purpose is to examine differences related to gender and family background.

Method

Subjects

The subjects were 99,070 ninth-grade students born in 1987, who left compulsory school in 2003. This is the whole population, and full information is available for subject grades, national test results, gender, and educational background. However, two reductions have been made, namely that, first, individuals for whom information on subject grades and national test result is lacking and, secondly, schools with 14 students or fewer, have been excluded from the analyses. In sum, 1,246 schools are included in the analyses. Data used in this study come from The Gothenburg Educational Longitudinal Database (GOLD), which contains register data compiled by Statistics Sweden for all individuals born between 1972–1987 and where information, such as student background, parental educational attainment, native country, grades from compulsory and secondary

education, and results on national tests is available. In this study, only data concerning subject grades and scores on national tests, together with parental education and gender, are used.

Instruments

In the current study, two measures were used: subject grades from the end of compulsory school and test scores from national tests, similarly undertaken at the end of compulsory school. The national tests are designed to support the grading practice in three core subjects: Swedish, English, and mathematics. The subject grades and national test scores are hypothesized to measure different aspects of achievement. Hence, the subject grades are hypothesized to measure both a subject-specific dimension and a common grade dimension, while the national test scores are hypothesized to primarily measure the subject-specific dimension.

The Swedish grading system

In the beginning of the 1990s, the Swedish educational system changed profoundly in several ways. The school system became decentralized through major school reforms and deregulations, leaving the responsibility for primary and secondary education to the local municipalities. The reforms also introduced a voucher system, which permitted the establishment of independently run schools within the Swedish educational system. During this period of reform, a new grading system was implemented. This new system introduced criterion-referenced grades, where grading is primarily based on classroom assessment, and the emphasis is on sets of grading criteria based on an underlying assumption that teachers will all interpret these criteria in a similar way. The criterion-referenced system has multiple purposes. First, the aim is to provide information about the individual student's acquisition of required standards. Secondly, it is intended to evaluate the educational system, and finally, it can be used as an instrument for selection to the next educational level in the school system. Since the intention is that the Swedish upper secondary school system should provide education to all students, the issue of selection was not emphasized during the development and implementation of the grading system. However, since the issue of selection is a very important factor, the criterion-referenced grading system has become a quantitative system where grades are used for ranking students for selection purposes, as well as for information (Wikström, 2005).

In Sweden, grading is decentralized. Grades are based primarily on classroom assessment, where individual teachers judge the performance of their students. The different subject domains are defined by the curriculum, the criteria, the national tests, and the grading. In the Swedish curriculum, the grading criteria in Swedish and English often reach beyond the respective subject domain. The criteria emphasize, for example, active communication, to develop thinking and argumentation, to state opinions, while a few criteria deal with subject content, such as grammar and vocabulary. However, in mathematics, the criteria primarily emphasize subject content, for example, that students are to demonstrate knowledge in area, geometric, diagrams, and equations. (National Agency for Education, 2007).

There are no standardized tests or external referees involved in the process of grading and assessment. The grading scale used in schools consists of four levels: not pass (IG), pass (G), pass with distinction (VG), and pass with special distinction (MVG). In order to use grades for selection to the next level in the educational system, they are converted into

numbers, and a weighted mean is computed. The scale ranges from 0–20, where not pass (IG) = 0, pass (G) = 10, pass with distinction (VG) = 15, and pass with special distinction (MVG) = 20. 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. As regards the standards for the grade levels, G, VG, and MVG, for the semester grades in Year 8 and the autumn grades in Year 9, it is the teachers themselves who define criteria for each grade level. In this study, grades from the end of Year 9 in three core subjects will be used, namely grades for Swedish (SGSW), English (SGEN), and mathematics (SGMA).

National tests

The main purpose of the national tests used in Sweden is to help teachers calibrate their grading, in order to support an equivalent and fair assessment and grading, and to concretize grading criteria. In Grade 9, national tests are used in three core subjects, Swedish, English, and mathematics. The tests are comprised of different subtests in each core subject, and there are oral as well as written tests. The curriculum and the syllabus are the starting point for the tests, although not all of the centrally defined criteria are tested, which implies that the respective subject domain is not fully covered in the tests. In Swedish, there are three subtests, the first test being a reading comprehension test, the second an oral test conducted in pairs, and the final test 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 done in a group, and a test of arithmetic where use of a calculator is not permitted, a test with more extensive tasks, and, finally, a test which demands problem-solving and for students to account for the calculations they make (National Agency for Education, 2006). The tests are produced centrally and their contents are not revealed in advance. As regards the grade-setting process, there are no external referees involved. The system relies on the teacher's ability to interpret grade criteria and to make judgements and evaluations that are fair and relevant in relation to the syllabus goals. However, in order to help teachers calibrate their grading, authentic examples of student test performances are available, where teachers have the possibility to compare their grading with these centrally evaluated examples. The National Agency for Education (2004) recommends that both individual teachers and schools should collaborate in the grade setting process.

The scales for the national tests correspond with the scale for the subject grades and range from 0–20, where not pass (IG) = 0, pass (G) = 10, pass with distinction (VG) = 15, and pass with special distinction (MVG) = 20. In this paper, the following abbreviations are used for the national tests: NTSW1, NTSW2, and NTSW3 for the test scores in Swedish; for English NTEN1, NTEN2, and NTEN3 are used; and for mathematics, only one summarized test score is available, NTMA.

Methods of analysis

In order to investigate the dimensions of grades, one-level and two-level confirmatory factor analysis (CFA) was used. First, a latent variable was created (*Sw*), designed to reflect achievement in Swedish, and which was related to the manifest variables subject grades (SGSW) and national test scores in Swedish (NTSW1–3) in a measurement model.

In the same way, a measurement model was estimated for English, where a latent variable was created (*En*) and designed to reflect achievement in English and related to the manifest variables for subject grades (SGEN) and national tests scores (NTEN1–3) in English. Finally, for mathematics, the measurement model was non-identified since only two indicators were available, but, nevertheless, a latent variable (*Ma*) was assumed.

Several steps then followed in the modelling process. In the first model (A), the factors for the different subjects (*Sw, En, Ma*) were each hypothesized to reflect a subject-specific dimension and were related to all their respective manifest variables or indicators, with covariance between the factors. In order to separate a common grade dimension, an additional latent variable was then specified, a common grade factor (*ComGr*), which was related to the three subject grades; Swedish, English, and mathematics (SGSW, SGEN, SGMA). In this baseline model (B), covariances between *Sw, En*, and *Ma* were estimated (Figure 1).

The next step was to add gender and family educational background variables into the baseline model. Dummy variables were created to represent parental education (0 = upper secondary education or lower and 1 = higher than upper secondary education) and gender (0 = boys, 1 = girls). The gender and educational variables were then related to all the factors (*Sw, En, Ma, ComGr*) (Model C).

Since previous research indicates that there are differences between schools regarding achievement and grades, the next step was to investigate the school differences. A two-level model (D) was estimated from the previous model (B), with the same relations for the within and between levels. In order to control for differences between girls and boys, the creation of two two-level models for girls (E-girls), and boys (E-boys), respectively, formed the next step. The final models (F and G) were estimated with the parental education variable related to all the factors on both levels. The F model was estimated for the entire sample, whereas the G model was estimated for girls and boys separately.

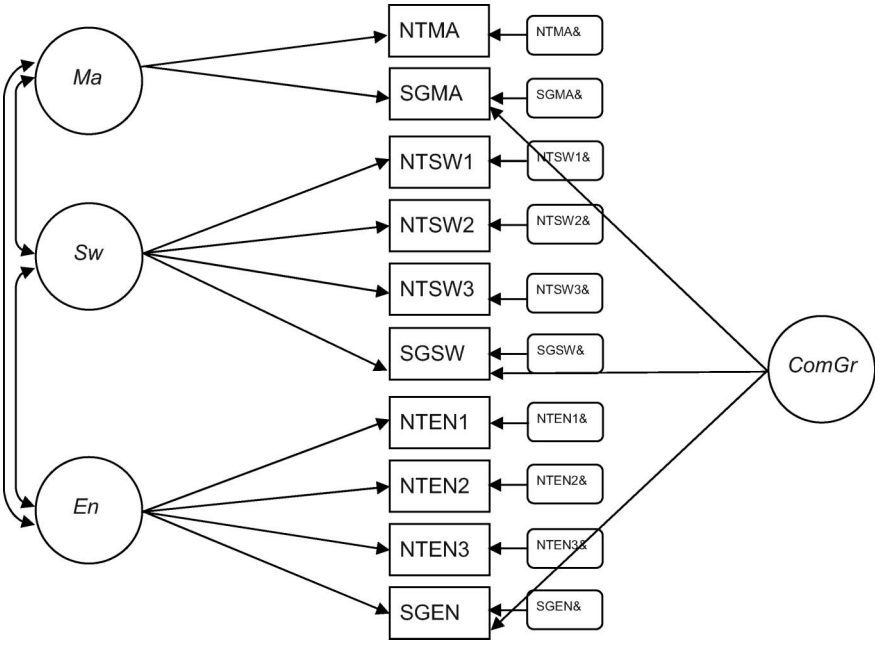

Figure 1. The basic four-factor model (C) with covariance between the subject factors.

As measures of model fit, the χ^2 goodness-of-fit test and Root Mean Square Error of Approximation (RMSEA) were used. In order for a model to be acceptable, the RMSEA should be below .07, whilst to be good, the RMSEA should be below .05. The RMSEA is strongly recommended as a tool when evaluating model fit, since it takes both the number of observations and free parameters into account (Jöreskog, 1993). The CFI goodness-of-fit measure was also used. This index should be as close to 1.0 as possible, and values below .95 are not acceptable (Bentler, 1990). The Standardized Root Mean Square Residual (SRMR), which is a measure of residuals compared separately for within and between levels, was also used and should be below .08. However, while the model fit is of fundamental importance, it should also be stressed that a model should be meaningful and the parameters interpretable.

For the national tests, there was some missing information, which was handled using missing data modelling (Muthén, Kaplan, & Hollis, 1987), which, under relatively mild assumptions, produces unbiased estimates. The large amount of missing data in the national test in mathematics was due to the test leaking out in advance in some areas in Sweden, and there is no reason to assume that this missing is biased due to achievement in mathematics. Furthermore, the missing data analysis applied makes the assumption that the data are "missing at random" (MAR), which implies that the procedure yields unbiased estimates when the missing is random given the information in the data. This is a much less restrictive assumption than the assumption that the data are "missing completely at random". The fact that there are high interrelations among the observed variables provides good possibilities to satisfy the MAR assumption.

Mplus, version 3 (Muthén & Muthén, 2004) was used for the purposes of estimation and for testing all of the models. STREAMS (Gustafsson & Stahl, 2005) is a modelling front-end environment, which was used to execute these analyses.

Results

The number of individuals, means, and standard deviations for subject grades and national test scores are shown in Table 1.

The descriptive statistics showed that there were missing observations for all variables, particularly in the case of the national tests in mathematics where there was a considerable proportion of missing data, 43.1%. For the national tests in Swedish and English, the proportion of missing data ranged from 17.8% for NTEN2 to 21.3% for NTSW2. For the subject grades, Swedish, English, and mathematics, the missing data proportion was low, 0.7%. The high rate of missing data for the national test in mathematics was due to problems when collecting the test data.

The pooled-within models

The first step was to estimate a model (A) with three factors, *Sw, En*, and *Ma*, specified by subject grades and national test scores in each subject (SGSW, SGEN, SGMA, NTSW1–3, NTEN1–3, and NTMA) and with covariance between the factors. The standardized factor loadings for this model (see Table 2) were substantial and the fit indices were acceptable ($\chi^2(32, 99070) = 7260.54$; RMSEA = .048).

In the next step, a baseline model (B) was developed with a common grade factor, *ComGr*, added to the model. The *ComGr* factor was related to subject grades, SGSW, SGEN, and SGMA, which also had relations from their respective subject factor. The goodness-of-fit indices for model B (Table 5) were better than for model A (χ^2(29,

Table 1. Descriptive statistics for the manifest variables; subject grades, national tests, gender and educational background.

Variables	*N*	*% missing*	*M*	*SD*
Subject Grades				
SGSW	98353	0.7	12.98	4.34
SGEN	98353	0.7	13.14	4.68
SGMA	98353	0.7	12.09	4.52
National tests				
NTSW1	81391	17.8	11.79	4.82
NTSW2	77979	21.3	12.91	4.00
NTSW3	81131	18.1	12.17	4.14
NTEN1	79832	19.4	13.15	4.25
NTEN2	81426	17.8	13.84	5.00
NTEN3	81221	18.0	12.73	4.40
NTMA	56325	43.1	11.90	4.44
Gender				
Girls	48660	0.0		
Boys	50410	0.0		
Educational backgr.				
Higher education[1]	44008	0.0		
Lower education[2]	54569	0.0		
Unspecified[3]	493			

Note: [1] = Higher than upper secondary education; [2] = Upper secondary education or lower; [3] = An unspecified group, to a large proportion composed by immigrants.

N=99070) = 5061.25; RMSEA = .042). Table 2 presents the standardized factor loadings for this model, which showed that the subject factors (*Sw, En, Ma*) accounted for a considerable proportion of the variance in both subject grades and in the national tests.

The *ComGr* factor explained variance in all subject grades, the factor loadings being highest for Swedish and lowest for English. These results thus support the hypothesis that there is a common grade dimension.

In order to investigate gender differences and differences related to the level of parents' education, gender, and family educational background, variables were related to the latent variables (*Sw, En, Ma*, and *ComGr*) in model B. Table 2 presents the standardized factor loadings and regression coefficients for this model (C). The estimated factor loadings were similar to those of the baseline model (B). The educational background variable seemed primarily to be of importance in the subject factors (*Sw, En, Ma*) and accounted for between 9.0% to 9.6% of the variance in each respective factor. The educational variable had a small, but nevertheless significant, negative relation (–.04) with *ComGr*. The gender variable primarily influenced *Sw* and *ComGr*, favouring girls, whereas there was a low but nevertheless significant relation with the English factor (*En*) in favour of girls. The goodness-of-fit indices were even better (χ^2(49, 99070) = 6408.26; RMSEA = .036).

The two-level models

In the next step, model B was developed into a two-level model (model D) in order to investigate differences between schools. This model was estimated with the same relations for within and between levels. The goodness-of-fit indices are presented in Table 5 and they showed a further improvement of fit (χ^2(58, 99070) = 4397.43; RMSEA = .027).

Table 2. Standardized coefficients for model A, B, and C and *t* values for gender and educational background.

Model	*A*			*B*				*C*							
Latent Variables	*Sw*	*En*	*Ma*	*Sw*	*En*	*Ma*	*Com Gr*	*Sw*		*En*		*Ma*		*ComGr*	
								t value		*t* value		*t* value		*t* value	
Manifest variables:															
SGSW	.91			.90			.23	.90						.20	
NTSW1	.76			.76				.77							
NTSW2	.74			.74				.74							
NTSW3	.75			.76				.76							
SGEN		.94			.93		.17			.93				.18	
NTEN1		.85			.85					.85					
NTEN2		.80			.80					.80					
NTEN3		.84			.85					.85					
SGMA			.93			.92	.19					.91		.21	
NTMA			.87			.89						.89			
Gender[1]								.26	79.16	.08	25.08	.00	.63	.14	20.44
Educational backgr.[2]								.31	96.44	.30	92.87	.31	90.32	−.04	−5.90

Note: [1]Boys are the reference; [2]Upper secondary education or lower is the reference. All estimates are significant when no *t* values are presented.

Table 3. Standardized coefficients for the population (model D) and for boys and girls separately (models E-girls and E-boys), within and between levels.

	pop	*g*	*b*	*pop*	*g*	*b*	*pop*	*g*	*b*	*pop*		*g*		*b*	
										t value		*t value*		*t value*	
Within level		*SwW*			*EnW*			*MaW*				*ComGrW*			
SGSW	.90	.90	.88							.23	19.26	.18	13.11	.24	14.05
NTSW1	.75	.76	.73												
NTSW2	.73	.72	.71												
NTSW3	.75	.74	.73												
SGEN				.93	.93	.92				.17	19.26	.18	13.11	.16	14.05
NTEN1				.84	.84	.84									
NTEN2				.80	.81	.80									
NTEN3				.84	.84	.83									
SGMA							.92	.91	.91	.19	19.34	.19	13.26	.20	15.21
NTMA							.89	.89	.90						
Between level		*SwB*			*EnB*			*MaB*				*ComGrB*			
SGSW	.94	.93	.95							.26	3.95	.30	2.75	.30	3.43
NTSW1	.83	.83	.84												
NTSW2	.83	.81	.85												
NTSW3	.86	.87	.85												
SGEN				.95	.94	.96				.27	4.00	.26	2.76	.22	3.51
NTEN1				.95	.94	.95									
NTEN2				.83	.82	.84									
NTEN3				.92	.93	.92									
SGMA							.94	.91	.95	.19	4.20	.17	2.77	.18	3.51
NTMA							.86	.86	.87						

Note: pop = population; g = girls; b = boys. All estimates are significant when no *t* values are presented.

Table 4. Standardized coefficients for educational background, for the population (model F), and for boys and girls separately (models G-girls and G-boys), within and between levels. *T* values are presented for educational background on the *ComGr* factor.

	pop	*g*	*b*	*pop*	*g*	*b*	*pop*	*g*	*b*	*pop*		*g*		*b*	
											t value		*t value*		*t value*
Within level		*SwW*			*EnW*			*MaW*				*ComGrW*			
Educational backgr.[1]	.28	.30	.29	.26	.26	.27	.28	.28	.28	−.02	−2.61	−.01	−.51	−.01	−1.04
Between level		*SwB*			*EnB*			*MaB*				*ComGrB*			
Educational backgr.[1]	.72	.65	.74	.80	.76	.81	.74	.71	.77	−.33	−2.95	−.26	−.2.39	−.54	−3.96

Note: pop = population; g = girls; b = boys; [1]Upper secondary education or lower is the reference. All estimates are significant when no *t* values are presented.

The intra-class correlations showed that 7.9% to 8.5% of the variance in the national tests in Swedish (NTSW1–3) was explained by school differences. For the national tests in English (NTEN1–3), the intra-class correlations ranged from 7.3% to 8.1%. For the national test in mathematics (NTMA), 7.3% was explained by school differences. For subject grades in Swedish, English, and mathematics, the intra-class correlations were 6.2%, 6.8%, and 6.3%, respectively. The intra-class correlations indicate that there are school effects, thus making it worthwhile to conduct a two-level analysis.

The standardized factor loadings (Table 3) at the school level were substantial, both for the grades and for the national tests, but slightly higher for the subject grades in Swedish, English, and mathematics, (SGSW, SGEN, SGMA). The *ComGr* factor explained more variance in subject grades in Swedish (6.8%) and English (7.3%) at the school level, whereas in mathematics the amount of explained variance was the same for the individual and school levels, 3.6%.

The next step was to estimate a two-level model for girls and boys separately (E-girls, E-boys). Model fit indices are presented in Table 5 and showed a good fit for both girls and boys; ($\chi^2(58,\ 48660) = 2129.50$; RMSEA = .027) and ($\chi^2(58,\ 50410) = 2378.43$; RMSEA = .028), respectively.

The standardized factor loadings showed that differences between girls and boys primarily concerned the loadings for English on *ComGr* at the school level. The *ComGr* factor accounted for 6.8% and 4.8% of the variance in subject grades in English for girls and boys respectively. The *ComGr* factor explained 9% of the variance in the Swedish subject grades and about 3% of the variance in subject grades in mathematics, for both girls and boys at the school level.

In the final step, the parental educational variable was related to the latent variables on both the within and between levels and estimated for the whole sample (Model F), as well as for girls and boys separately (models G-girls and G-boys). The two-level model on the whole sample (F) showed a good fit ($\chi^2(77,\ 99070) = 4735.24$; RMSEA = .025), and similarly good indices were observed for both girls and boys; ($\chi^2(77,\ 48660) = 2339.63$; RMSEA = .025) for girls and ($\chi^2(77,\ 50410) = 2524.43$; RMSEA = .025) for boys (Table 5).

The parental educational variable accounted for a considerable amount of variance (Table 4) in the subject factors, *Sw, En*, and *Ma*, namely 51.8% to 64.0% at the school level, and also had significantly negative relations to the *ComGr* factor for the F model

Table 5. Goodness-of-fit indices for the models A–G.

					SRMR	
Model	*Df*	χ^2	*RMSEA*	*CFI*	*W*	*B*
Model A	32	7260.54	.048	.982		
Model B	29	5061.25	.042	.987		
Model C	49	6408.26	.036	.995		
Model D	58	4397.43	.027	.985	.019	.036
Model E-girls	58	2129.50	.027	.988	.018	.040
Model E-boys	58	2378.43	.028	.986	.020	.035
Model F	77	4735.24	.025	.986	.017	.034
Model G-girls	77	2339.63	.025	.988	.015	.039
Model G-boys	77	2524.43	.025	.987	.017	.033

(−.33). The relation between parental education and *ComGr* was more strongly negative for boys (−.54) than for girls (−.26).

Discussion and conclusions

The modelling results showed that, at both the individual level and at the school level, the largest part of the variance in grades was due to achievement in the different subject areas, as measured by the national tests. However, the results also showed that, at both levels, it was possible to identify a dimension that cut across the grades in the three subject areas. This common grade dimension suggests that grading is influenced by factors other than achievement. The factor model does not, however, reveal whether the common grade dimension should be interpreted as construct-irrelevant variance, or whether it is part of the construct. One way to obtain further insight into the common grade dimensions is to investigate the pattern of association with other variables.

At the individual level, the common grade dimension accounted for between 3%–5% of the variance in the different subject grades. Girls had a higher level on the common grade dimension within schools, a result which is in agreement with previous findings that the grades achieved by girls are higher than their test scores (Nycander, 2006). One interpretation of this finding is that girls and boys develop different rationales in school; whilst girls nurture their general abilities, evidenced, for example, by a desire to become ''responsible students'' and to ''please their teachers'', boys seem to nurture their specific abilities (Rosén, 1998; Wernersson, 1989, 1992).

Compared to boys, girls also had a higher level on the achievement dimensions, particularly so in Swedish. The student's family background was quite strongly related to the subject achievement dimensions, a result which is also in agreement with previous research (Hanushek & Luque, 2003; Wikström, 2005; Yang, 2003). However, parental education did not associate with the common grade dimension within schools either for boys or for girls. For the total group, there was a weakly negative association, indicating that students with a lower educational background received somewhat higher grades.

In order to further clarify the meaning of the common grade dimension at the individual level, it would be interesting to incorporate measures of individual characteristics into the model, such as motivation, effort, work completion, classroom behaviour, and personality, which in previous research have been shown to influence grading (Alexander, 1935; Andersson, 1998; Brookhart, 1991; Cizek et al., 1995; Cross & Frary, 1999; Manke & Loyd, 1990; McMillan, Myran, & Workman, 2002; Selghed, 2004; Stiggins & Bridgeford, 1985). Since the model separates achievement from the common grade dimension, this would make it possible to determine whether such individual characteristics influence grades indirectly, via the effect on achievement, or directly, through effects on the common grade dimension. The nature of the gender difference in the common grade dimension can also be clarified in models which investigate which individual characteristics mediate the gender effect on the common grade dimension.

At the school level, the common grade dimension explained between 3% (mathematics) and 9% (Swedish) of the variance in grades. Interestingly, the results also showed that the proportion of students with a high level of parental education was negatively related to the common grade dimension at the school level. This effect was stronger for the proportion of boys (–.54), with a high level of parental education, than for the proportion of girls (–.26). This result reveals, at the school level, the presence of a compensatory grading practice, in that at schools where there are many students from families with lower educational backgrounds, the grades assigned are higher than the test

scores. One reason for this may be that teachers try to get as many students as possible up to the "pass" level because of the negative consequences both for the individual student and for the school that are associated with a "not pass" grade. The competition between schools for students may also cause schools to assign higher grades in order to compensate for poor results on the national tests. Such practices may result in grade inflation and differences between schools (Cliffordson, 2004; Cross & Frary, 1999; Wikström, 2005).

Issues of validity in grades

The fact that there is a dimension which cuts across the grades does not, of itself, demonstrate that this variance is construct irrelevant. It may, of course, be that teachers can capture aspects of achievement that cannot be captured by the national tests and which cut across different subject areas. However, the fact that there is a common grade dimension at both individual and school levels also suggests that there may be sources of construct-irrelevant variance in the grades at both these levels.

The most obvious kind of construct-irrelevant variance is due to factors which directly influence the outcomes of the grading process and which are irrelevant to the goals and criteria established for the subject area. In the Swedish grading system, an example would be if the grading were to be affected by the politeness of the student.

Other student characteristics may have an indirect influence on grades, since they influence the amount of knowledge and skills acquired. Examples of such factors are motivation and effort. Such indirect effects of grades would not contribute to construct-irrelevant variance; on the contrary, one of the expressed functions of grading systems is to enhance motivation and effort. Within the framework of the current study, it has not been possible to separate the nature of different sources of variance in the common grade dimension, and this therefore remains an important task for future research.

The common grade dimension found in this study may thus be fundamental in order to understand what construct-irrelevant variance implies for grades. Brookhart's study (1991) demonstrated that the use and consequences of grades constituted the meaning of grades to teachers, sometimes at the expense of their interpretation. Other research has shown that low and high achievers may be graded differently due to moral considerations among teachers (Cox, 1995; Parsons, 1959; Pilcher, 1994). Construct-irrelevant variance in grades may thus vary according to the level of student achievement, the student's background, the current subject domain, and different teacher grading strategies. The evidence or assessment of data collected in order to support a decision "are more or less valid for some very specific purpose, meaning or interpretation, at a given point in time and only for some well-defined population" (Downing, 2003, p. 830).

One of the most interesting results to emerge from the present study concerns the relation between parental education and the common grade dimension at the school level. One possible interpretation of this finding is that teachers are concerned with the distributive justice of grades (Deutsch, 1979; Pilcher, 1994). Systematic differences between schools in the interpretation of the grading criteria may be another source of construct-irrelevant variance in grades. Such systematic differences may be related to 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.

At a more general level, the role and type of accountability system in the educational system may affect the behaviour of the agents in the educational system (Bishop, 2001; Woessmann, 2002). It has been suggested that in a highly decentralized educational system, the accountability system needs to focus on the domains and constructs in the

curriculum and to primarily use curriculum-based rigorous exams in order to decrease the level of construct-irrelevant variance in grades (Bishop, 2001; Woessmann, 2002).

When grades are used as an instrument for selection to the next level in the educational system, grades are assumed to be comparable between schools and over time. In a goal-referenced grading system without a rigorous accountability system, grades are often assumed not being reliable, hence not suitable as an instrument of selection. In the Swedish educational system, goal-referenced grades are used as a selection instrument, and they have been shown to function as well as or even better than norm-referenced grades (Cliffordson, in press). Grades also have been shown to be better predictors than scholastic aptitude tests (Cliffordson, in press). Different noncognitive skills may be an important part of grades, and this may contribute to their predictive validity.

Limitations

One limitation of the current study is that the information available has not made it possible to clearly differentiate between sources of variance in the common grade dimension, and thus this remains a task for future research. Another limitation concerns the national tests used as instruments for measuring achievement. The national tests are assumed to measure achievement or subject-specific knowledge, but it seems reasonable to believe that the tests also measure some amount of non-achievement, since the national tests are administered and marked locally by the teachers themselves.

Future research

The result of the present study can be seen as a starting point for further analyses. In order to understand what it is that grades actually measure, both the common grade dimension and the nature of construct-irrelevant variance in grades must be investigated more closely. The use of questionnaire data that included students' responses to questions about ambitions, effort, interest, and the like would enable answers to be found to this question.

Acknowledgements

The Swedish Research Council has financially supported the research reported in this article. The work is a part of the GRAM (Grades and grade assignment: Functions and effects) project.

Notes on contributors

Alli Klapp Lekholm is a Ph.D. student at Gothenburg University, Sweden. She is working with research on grades and grade assignment within the GRAM (Grades and grade assignment: Functions and effects) project.

Christina Cliffordson, associate professor in education, is working with research on assessment and measurement based on large-scale longitudinal data.

References

Agnew, E.J. (1985, March/April). *The grading policies and practices of high school teachers*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

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.

Bentler, P.M. (1990). Comparative indexes in structural models. *Psychology Bulletin*, *107*, 238–246.

Bishop, J.H. (2001). A steeper, better road to graduation. *Education next*, *1*(4), 56–61.

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.
Cizek, G.J., Fitzgerald, S., & Rachor, R.E. (1995). Teachers' assessment practices: Preparation, isolation and the kitchen sink. *Educational assessment*, *3*(2), 159–179.
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. (in press). 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*.
Cox, K.B. (1995, March/April). *What counts in English class? Selected findings from a state-wide study of California high school teachers*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.
Cross, L.H., & Frary, R.B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education*, *12*(1), 53–72.
Deutsch, M. (1979). Education and distributive justice. *American Psychologists*, *34*(5), 391–401.
Downing, S.M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*, 830–837.
Dylan, W. (1996). National curriculum assessments and programmes of study: Validity and impact. *British Educational Research Journal*, *22*(1), 129–142.
Gustafsson, J.-E., & Stahl, P.-A. (2005). *STREAMS 3.0 User's Guide*. Mölndal, Sweden: Multivariateware.
Hanushek, E.A., & Luque, J.A. (2003). Efficiency and equity in schools around the world. *Economics of Education Review*, *22*, 481–502.
Jöreskog, K.G. (1993). Testing structural equation models. In K.A. Bollen & J. Scott Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.
Korp, H. (2006). *Lika chanser på Gymnasiet? En studie om betyg, nationella prov och social reproduktion* [Equal chances at the upper secondary level? A study on grades, the national test and social reproduction]. (Malmö studies in educational science, Vol. 24). Malmö, Sweden: Department of Education.
Manke, M.P., & Loyd, H. (1990, April). *An investigation of nonachivement related factors influencing teachers grading practices*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston.
McMillan, J.H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203–213.
Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Messick, S. (1991). Validity of test interpretation and use. In M.C. Alkin (Ed.), *Encyclopedia of educational research* (6th ed., pp. 1–29). New York: Macmillan.
Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher*, *23*(2), 13–23.
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, L.K., & Muthén, B.O. (2004). *Mplus user's guide* (3rd ed.). Los Angeles: Author.
National Agency for Education. (2003). *Skolverkets lägesbedömning 2003* [The National Agency for Educatioṅs determination of position 2003]. Stockholm: Author.
National Agency for Education. (2004). *Likvärdig bedömning och betygssättning. Skolverkets allmänna råd 2004* [Equal assessment and grading. General advice issued by the National Agency for Education]. Stockholm: Author.
National Agency for Education. (2005). *Skolverkets lägesbedömning 2005* [The National Agency for educatioṅs determination of position 2005]. Stockholm: Author.
National Agency for Education. (2006). *National subject tests in compulsory school*. Retrieved May 2, 2007, from http://www.skolverket.se/sb/d/276

National Agency for Education. (2007). *Steering documents*. Retrieved October 2, 2007, from http://www.skolverket.se/sb/d/287

Nava, F.J.G., & Loyd, B.H. (1992, April). *An investigation of achievement and non-achievement criteria in elementary and secondary school grading*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Nycander, M. (2006). *Pojkars och flickors betyg* [Grades received by boys and girls]. Uppsala, Sweden: Uppsala University, Department of Education.

Parsons, T. (1959). The school class as a social system: Some of its functions in American society. *Harvard Educational Review*, *29*, 297–318.

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

Pilcher-Carlton, J., & Oosterhof, A. (1993, April). *A case study analysis of parents', teachers', and students perception of the meaning of grades: Identification of discrepancies, their consequences, and obstacles to their resolution*. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

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

Selghed, B. (2004). *Ännu icke godkänt. Lärares sätt att erfara betygssystemet och dess tillämpning I yrkesutövningen* [Not yet a pass. Teachers' ways of experiencing the grading system and its application in professional practice]. (Malmö studies in educational science, Vol. 15). Malmö, Sweden: Department of Education.

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

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.

Svensson, A. (1971). *Relative achievement. School performance in relation to intelligence, sex and home environment* (Göteborg studies in educational sciences, Vol. 6). Göteborg, Sweden: Acta Universitatis Gothoburgensis.

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]. Borås, Sweden: Department of Education, The University college of Borås.

Wernersson, I. (1989). *Olika kön samma skola? En kunskapsöversikt om hur elevernas könstillhörighet påverkar deras skolsituation* [Different gender same school? A research overview about how the student's gender affects their school situation]. Stockholm: National Board of Education.

Wernersson, I. (1992, June). *Gender differences in social interaction in the classroom setting: Alternative explanations*. Paper presented at "Alice in Wonderland" – First international Conference on Girls and Girlhood: Transitions and Dilemmas, Amsterdam.

Wikström, C. (2005). *Criterion-referenced measurement for educational evaluation and selection. No. 1*. Umeå, Sweden: Umeå University, Department of Educational measurement.

Wikström, C., & Wikström, M. (2005). Grade inflation and school competition: An empirical analysis based on the Swedish upper secondary schools. *Economics of Education Review*, *24*, 309–322.

Woessmann, L. (2002). *How central exams affect educational achievement: International evidence from TIMSS and TIMMS-Repeat* (Report PEPG/02–10). Cambridge, MA: Harvard University, Kennedy school of Government.

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, Sweden: Acta Universitatis Gothoburgensis.