Criterion-Referenced Measurement for Educational Evaluation and Selection

Christina Wikström

Department of Educational Measurement
Umeå University
No. 1

Department of Educational Measurement
Umeå University
Thesis 2005

Printed by Umeå University
March 2005

ISSN 1652-9650
ISBN 91-7305-865-3

Abstract

In recent years, Sweden has adopted a criterion-referenced grading system, where the grade outcome is used for several purposes, but foremost for educational evaluation on student- and school levels as well as for selection to higher education. This thesis investigates the consequences of using criterion-referenced measurement for both educational evaluation and selection purposes. The thesis comprises an introduction and four papers that empirically investigate school grades and grading practices in Swedish upper secondary schools.

The first paper investigates the effect of school competition on the school grades. The analysis focuses on how students in schools with and without competition are ranked, based on their grades and SweSAT scores. The results show that schools that are exposed to competition tend to grade their students higher than other schools. This effect is found to be related to the use of grades as quality indicators for the schools, which means that schools that compete for their students tend to be more lenient, hence inflating the grades. The second paper investigates grade averages over a six-year period, starting with the first cohort who graduated from upper secondary school with a GPA based on criterion-referenced grades. The results show that grades have increased every year since the new grading system was introduced, which cannot be explained by improved performances, selection effects or strategic course choices. The conclusion is that the increasing pressure for high grading has led to grade inflation over time. The third paper investigates if grading practices are related to school size. The study is based on a similar model as paper I, but with data from graduates over a six-year period, and with school size as the main focus. The results show small but significant size effects, suggesting that the smallest schools (<300 students) are higher grading than other schools, and that the largest schools (>1000 students) are lower grading than other schools. This is assumed to be an effect of varying assessment practices, in combination with external and internal pressure for high grading. The fourth and final paper investigates if grading practices differ among upper secondary programmes, and how the course compositions in the programmes affect how students are ranked in the process of selection to higher education. The results show that students in vocationally oriented programmes are higher graded than other students, and also favoured by their programmes' course compositions, which have a positive effect on their competitive strength in the selection to higher education.

In the introductory part of the thesis, these results are discussed from the perspective of a theoretical framework, with special attention to validity issues in a broad perspective. The conclusion is that the criterion-referenced grades, both in terms of being used for educational evaluation, and as an instrument for selection to higher education, are wanting both in reliability and in validity. This is related to the conflicting purposes of the instruments, in combination with few control mechanisms, which affects how grades are interpreted and used, hence leading to consequences for students, schools and society in general.

Utan tvivel är man inte klok

Tage Danielsson

Acknowledgements

I would like to start by expressing my gratitude to the Department of Educational Measurement for giving me the opportunity to complete a doctoral degree, and for treating me so well during the years of my doctoral studies.

I entered the doctoral programme with a very confused idea of what the doctoral studies really implied. “How hard can it be?” I said to my former colleagues in overweening confidence, when I (with a great deal of regret) told them I was leaving and returning to school. They just laughed at me and threatened to pin a banner with these words on the door to my study (but fortunately they never did).

Very soon after arriving at the Department of Educational Measurement, I was offered the opportunity to participate in the VALUTA project, a large research project, with project members representing the front edge in Swedish educational research, and with access to an impressive database any researcher would dream of. With embarrassment, I now remember that I actually hesitated, because my mind was set on other things, and validating a selection system was very far off from my ideas of what educational measurement is really about. Fortunately, for once I listened to reason, and eventually joined this project, a decision turning out to be a good one. For this reason and many more, I would like to express my gratitude to my supervisor Widar Henriksson, who has patiently guided me through this process, no matter how muddled I have been at times. I also owe my gratitude to the Bank of Sweden Tercentenary Foundation and the participants in the VALUTA project, for giving me the opportunity to work in such an interesting research project.

It is difficult not to direct my thanks to everyone I know, since so many are important to me. However, there are a few people that have been especially important for my work. I want to thank Tomas Tjäderborn, for leading me into the area of educational measurement in the first place. It has been an adventure! I would also like to thank my friends in the AEA-

Europe council, who have given me great support. Furthermore, my thanks are due to Simon Wolming and Lotta Jarl, for pampering me during my doctoral years and to Björn Sigurdsson for helping me with the practical issues connected to the printing of my thesis. Without his help, I would probably have ended up with a pile of stencils! I would also like to send a happy thought to Per-Erik Lyrén, for good advice and for having the mental strength to share his study with me for so long. My thanks are also due to Gunnar Persson, who has patiently proofread my manuscripts. Any mistakes that can be found are my sole responsibility, since I cannot keep from making major changes at the last minute.

Most of all, I want to thank my husband Magnus, who has been invaluable for many reasons. Even though his own workload has often been overwhelming, he has taken the time to guide me both practically and theoretically. He even took on the hazardous task of conducting a study with me, a project turning out to be more successful than we both had probably anticipated.

I would also like to thank my children; Elliot, Ludwig and Anton, who have generously accepted having a mother who spends most of her time in front of the computer rather than doing "what other mothers do".

Thank you colleagues, friends and family, for being there, and for being you!

Well, how hard was it? - Pretty hard, but a lot of fun as well!

Criterion-Referenced Measurement for Educational Evaluation and Selection

This thesis is based on the following articles:

I. 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.

II. Wikström, C. (in press). Grade stability in a criterion-referenced grading system: the Swedish example. Forthcoming in *Assessment in Education - Principles, Policy and Practice*.

III. Wikström, C. (2005). Does school size affect teachers' grading practices? Submitted for publication.

IV. Wikström, C. (2005). The reliability and validity of the grade point average as instrument for selection to higher education. Submitted for publication.

All referencing to these articles will follow the enumeration used above.

Table of Content

1. Introduction

The overall purpose of this thesis is to study the implications of using a criterion-referenced instrument for educational evaluation and student selection. Four empirical studies have been conducted that illuminate aspects connected to this purpose. These studies are presented in papers I-IV. The papers are joined together by an introductory section, which has the purpose of putting these research issues and the results that have been found into the perspective of a theoretical framework, in the hope of making the findings even more relevant and useful for the reader.

Terminology is always a difficult issue, and in my field of research, there are a number of terms that can mean similar things to some, but have different meanings to others. I will therefore start by clarifying some of the key concepts that will be used in this thesis. The first regards the terms *measurement* and *assessment*. For me, there is not much difference between these terms, even though I appreciate that, for others, there may be a world of difference. In *Collins COBUILD English Language Dictionary* (1987) *assess* is described as "to estimate the value of […]" "to determine the amount of […]" and "to evaluate". *Measurement* means, among other things, "the act or process of measuring" and "a system of measures based on a particular standard". This does not make me much wiser. The title of the thesis contains the word *measurement*, because most often this word is used in connection with the criterion-referenced approach (CRM). However, I like the word *assessment*. It is generally used in a wider sense, which I find suitable here. Nevertheless, since it is difficult to avoid using only one of them, these terms will be used inter-changeably in this thesis, even though *assessment* will be the first choice.

Furthermore, the reader will also find the term *instrument* frequently used. The aim is not do discuss a specific test mode, but to focus on general issues in educational assessment, while exemplifying by investing empirical evidence. It is the assessment process that is interesting, and the decisions made on the basis of this process. Therefore, the term *instrument* is found to be the most suitable. It

refers to any kind of mode used for assessment purposes; it can be a standardized instrument, like a test, but also, for instance, the process of assessing practical or academic performances made by a judge or a teacher. In other words: the instrument represents anything that is used for retrieving information of which some kind of decision in an educational context will be made.

In measurement theory, instruments are generally categorised as norm-referenced or criterion-referenced. This distinction has more to do with the decisions the instrument will lead to than the instrument itself (see, for instance, Black & Dockrell, 1987; Ghiselli, Campbell and Zedeck, 1981; Gregory, 2004). This is also appreciated by Messick (1989), who uses the distinction *selection decisions* and *classification decisions* (p.71). In norm-referenced measurement, the performance of an individual is compared with the performance of other individuals, the norm group. This type of instrument is very common in psychological testing, education, sports etc, or whenever it is considered important to distinguish between the performances of individuals (the selection decision). In criterion-referenced measurement, the purpose is to determine how an individual is performing in relation to a criterion, or educational objectives (the classification decision) (see, for instance, Hambleton, 1994; Linn, 1994). Criterion-referenced measurement is generally used for decision-making (masters–non-masters), for bench-marking (educational evaluation) and for formative purposes, for instance when individualising education. There is an extensive amount of literature and research concerning both approaches, and how such instruments are developed, used and interpreted. Sometimes, these two approaches are combined; i.e. the instrument is expected to yield information about the performance of an individual, but also to be used for ranking the students relative to each other (Gregory, 2004). Still, even though these types of instruments are not uncommon, very limited research can be found concerning their qualities, limitations and consequences, leading to unfortunate knowledge gaps in this area.

The Swedish upper secondary school grades constitute a good example an assessment instrument with several purposes. The grades are criterion-referenced, and their main purpose is to give information concerning how well

the students have reached educational goals (Andersson, 1999). They are also used for educational evaluation on higher levels, for instance when evaluating school performance (Lyrén, 2005). Furthermore, they constitute the main instrument for student selection to higher education, by ranking the students according to their achievements (Wedman, 2000; Wolming, 1999). The Swedish upper secondary school grades will therefore constitute the empirical evidence in this thesis.

1.1. Disposition of the thesis

This thesis consists of four papers (I, II, III and IV), all focusing on the characteristics of the Swedish upper secondary school grades. The results from these studies will be discussed from a validity perspective, presented in the introductory section.

Two main issues are addressed: First, can the criterion-referenced grades meet the demands of a criterion-referenced instrument – does a certain grade mean the same, irrespective of when and where it was achieved, and if not: what are the causes? Second, is this type of grade system suitable for selection purposes, and what are the consequences of using criterion-referenced grades for ranking students?

The thesis is structured as follows: First, the theoretical framework of the thesis will be addressed. This includes a brief introduction to measurement theory and a description of the characteristics and history of norm-referenced and, chiefly, criterion-referenced measurement. This will be followed by a section about reliability and validity in educational measurement and a chapter describing the Swedish assessment system. Then the four papers representing the empirical studies will be summarised. In the last sections, the results will be discussed, chiefly from the perspective of validity. Some conclusions will be made, as well as suggestions for future research. Finally, the papers will follow in numerical order.

2. Instruments in educational assessment

This chapter gives a short introduction to measurement theory and the types of instruments used in educational assessment. The purpose is to give the reader a theoretical background to the issues investigated in this thesis and the discussion that will follow.

2.1. Norm-referenced measurement

As previously discussed, assessment instruments are generally categorised as absolute or relative, or in measurement terminology: criterion-referenced or norm-referenced.

The history of norm-referenced measurement goes back to the early days of intelligence testing, and when, for instance, Binet developed his famous instruments for measuring "mental age" (Gregory, 2004). The transition from intelligence testing to educational testing started in the 1940s, when Educational Testing Services developed the SAT, which is used for college admission in the USA. The majority of all instruments used in educational measurement have a norm-referenced approach. The purpose of a norm-referenced instrument is to compare the performance of an individual with the performances of other individuals (the norm group). Since the purpose is to distinguish between individuals, there is no point in having an instrument where everyone receives the same outcome, which means that variability in the outcome of the instrument is very important (Shavelson, Baxter & Gao, 1993; Yen, 2003). Therefore, a norm-referenced instrument generally consists of tasks or items that vary in difficulty. The general assumption is that student performances differ, and by receiving information about these differences, various types of decisions can be made. Sometimes the assumption of variability is taken even further, by also assuming that the performances (and hence the outcome of the instrument) is distributed according to a normal (Gaussian) curve (Gregory, 2004). The score is often expressed as a percentile, a grade equivalent score, or a stanine (Yen, 2003).

Norm-referenced instruments are common in competitive surroundings, when an instrument is needed for retrieving information about how individuals are performing relative to each other. They are also common in education, chiefly in different forms of examinations, but are generally not suitable for educational evaluation. One reason is that the score is based on the norm group, and the norm group is likely to differ among administrations. Furthermore, since the area being measured is not necessarily based on a clearly defined criterion, it may be difficult to link the outcome to learning objectives. A criterion is here defined as the domain from which inferences about individuals' performance can be made (see also Hambleton, 1994).

The long history of the norm-referenced tests has given them a strong position in the area of assessment as well as an established theoretical foundation. Their potentials and limitations are known, and methods for research well documented. There are a large number of approaches to how to ensure reliability and validity, and most statistical methods in both classical and modern test theory have been developed with norm-referenced tests in mind (Hambleton, 1994; Wiberg, 2004).

2.2. Criterion-referenced measurement

The term *criterion-referenced measurement* was introduced by Glaser and Klaus more than forty years ago (Glaser & Klaus, 1962). In 1963, Glaser published the paper "Instructional technology and the measurement of learning outcomes. Some questions", which is considered the official starting point of the criterion-referenced measurement era, even though various forms of absolute measurement (without the proper label attached) at least to some extent had existed before. This approach to assessment was welcomed chiefly by practitioners in the educational field, since it is found to be especially useful for formative purposes (see for instance, Hambleton, Swaminathan, Algina & Coulson, 1978).

Here too, the terminology is somewhat confusing. Forms of criterion-referenced measurement are sometimes referred to as objective-referenced,

domain-referenced or curriculum-oriented measurement (Linn, 1994). Most likely, there are also an additional number of other similar names. Some use only one of these names, some distinguish between them and give them separate meanings and some use them interchangeably. In this thesis, I will not make any distinction between these terms. I will use the term criterion-referenced measurement for any measurement that has the purpose of measuring skills, performance or knowledge, has been defined by a criterion and is used for finding out performance levels of the individuals that are being assessed.

Criterion-referenced measurement has been an intensely discussed issue for a long period of time (Hambleton, 1994; Millman, 1994). The common argument for criterion-referenced measurement has been, and still is, that it gives the educator better information about educational progress for individual students and groups of students, and is hence useful for individualising instruction, which is an area of special interest in education today. However, criterion-referenced measurement has also led to new research issues, such as how to ensure the validity and the reliability of such an instrument. Classical test-theoretical methods were initially found inadequate, due to the fundamental differences between norm-referenced instruments and criterion-referenced instruments in their construction and use. In the main, the question was how to analyse the outcome without score variability, which is necessary in a norm-referenced instrument, but often irrelevant in a criterion-referenced instrument. Another issue, which has perhaps been the most intensely discussed topic, was to establish what could be measured with the criterion-referenced instruments (Hambleton, 1994; Nitko, 1996). This has more to do with the type of knowledge from a cognitive perspective, than with subject matter areas in general. One argument is that if the objectives to be measured are on a high abstraction level, the scope of the area as well as the performance levels will be interpreted differently by educators, test developers and test takers. The concern was that such situations would lead to serious reliability and validity problems (see, for instance, Hambleton et al, 1978).

Originally, the general opinion was that criterion-referenced measurement generally is most suitable for subjects and subject matters with a simple structure and with objectives corresponding on the lower levels of Bloom's cognitive taxonomy, which is described by Anderson & Krathwohl (2001) and originally by Bloom (1956). The opinion that the criterion had to be on a rather detailed level and that criterion-referenced measurement was limited to non-complex areas and low cognitive levels, was also interpreted as a strengthened position for the multiple-choice item format (see, for instance, Popham & Husek, 1969; Wedman 1973). This view was not necessarily supported by Glaser who "founded" the concept of criterion-referenced measurement, however, who merely had addressed the idea and usefulness of criterion-referenced measurement in an overall perspective.

Later on, the general view of the formerly described limitations of the criterion-referenced instrument was revised, and most researchers agreed that criterion-referenced measurement is also suitable for measuring more complex areas, performance and higher order thinking, even though the problem of how to define the test specifications for the instrument when working with complex areas and high cognitive levels of knowledge is difficult to solve (Millman, 1994; Skolöverstyrelsen, 1983).

In the 1980s, the discussion about criterion-referenced measurement seemed to lose intensity, which is interpreted by some as if the concept lost in popularity. However, this is contradicted by Hambleton (1994), who argues that this is more of a terminology issue. He thinks that the term *criterion-referenced measurement* became less frequently used because other terms, such as *authentic assessment* and *performance assessment* and the use of portfolios, became popular towards the end of the 1980s and onwards. He also points out that these terms are in fact forms of criterion-referenced measurement and hence have made the overall denomination unnecessary.

Most likely, the discussions described above will continue for some time. It is evident, however, that criterion-referenced measurement is useful and suitable for making decisions about performance standards, as long the necessary

prerequisites and quality issues are ensured (see chapter 3). Today, criterion-referenced instruments are very common outside the traditional educational field, for instance in licensing and different types of certification programmes. Within education, criterion-referenced instruments are common for formative purposes, but still norm-referenced instruments are in domination in examination and selection.

3. Quality issues

The reliability and validity of assessment instruments, and the types of decisions these instruments will result in, are important issues that must be considered at all times. In some situations it is especially important to ensure the quality of the assessment process. An instrument is often referred to as "high stakes" or "low stakes". This has to do with the consequences the instrument will have for those who take the test, or those who for any other reason are affected by the decisions that the outcome will lead to (Lane, 2004; Roos, 2005). The higher the stakes, the higher must the accuracy, i.e. the reliability and validity, be. Generally, when talking about high stakes instruments, various forms of examination tests come to mind, but the stakes can vary with the instrument, the examinees, and the circumstances connected to the assessment situation. For some, the stakes are high, for others, the outcome does not matter as much.

3.1. Reliability

Assessment instruments must be reliable in order to be useful. Lack of reliability means that the instrument will yield different outcomes, even though the skills, knowledge or performance of the examinee have not changed. In theory, reliability and validity are generally treated as two separate issues. In practice, reliability and validity issues are strongly connected. In the main, reliability is necessary in order to achieve validity (even though the opposite is not necessarily the case).

The reliability of an instrument can be evaluated in several ways. Methods and approaches are related to the type of instrument, its construction, content,

interpretation and use (see, for instance, Ghiselli et al. 1981; Linn, 1989; Nitko, 1996; Wiberg, 2004). In measurement theory the evaluation of reliability is directed towards tests. How to evaluate the reliability of decisions that are based on performance assessment and classroom assessment is seldom discussed, most likely because of the disparity of these types of assessments. The most common approaches are to evaluate the temporal stability, the equivalence of several forms and the internal consistency. The stability is often evaluated by administrating the same test form to the same groups of testees on two occasions (test–retest), by comparing the results from several forms (alternate forms), or by investigating the consistency between items (internal consistency) (see, for instance, Gregory, 2004; Wiberg, 2004). There are of course practical problems connected to such evaluations. For instance, administrating the same instrument twice to the same group of examinees can be difficult, because there are often repeated test effects that need to be considered (see, for instance, Henriksson & Bränberg, 1994; Törnkvist & Henriksson 2004).

It is even more difficult to transfer these methods to a grade system where the instrument is the grading process determined by different teachers' judgements. For instance, since assessment methods are diverse, and data from the ongoing assessment generally not collected or reported, the reliability is very difficult to investigate, unless the number of observations can be scaled up, as argued by Shavelson, Baxter & Gao (1993) as well as Yen (1993). Such analyses have been discouraging according to Parkes (2000), who claims that both performance assessment and portfolio assessment, which are common elements in criterion-referenced assessment and especially within the framework of classroom assessment, are often wanting in reliability. The reliability problem is referred to low agreement between tasks and methods, as well as the varying assessment skills and experiences of those who are responsible for the assessment.

3.2. Validity

All researchers, test developers and test users, irrespective of field and orientation, are likely to have come across the concept of validity, some more often than others. The original view on validity addressed the question "Does the test measure what it purports to measure?" (Shepard, 1993). Validity is, however, not a one-dimensional concept but much more complicated. How to define what it comprises and what it affects has been a matter of theoretical discussions for decades.

The traditional way of defining validity has been to divide the concept into three sub areas: content validity, criterion-related validity, and construct validity. Originally, there were four aspects, since criterion-related validity was divided into two: concurrent and predictive validity (Wolming, 2001). These aspects are tightly connected to the instrument, its construction and how the outcome correlates with some external criteria. These are not independent of each other, however. What affects one of them will also be likely to affect the others. Within these categories there are also other, subordinate, validity aspects (Gregory, 2004; Messick, 1989).

Content validity has to do with how well an instrument meets the specifications for that instrument. It addresses the question of whether the elements or samples in the instrument are representative of all the aspects and facets of the definitions of the trait that is being assessed (Ghiselli et al., 1981), for instance, by investigating how the items in a test correspond to the detailed objectives described in the specifications for the test.

Criterion-related validity evidence deals with how an instrument relates to some external criteria, of which the instrument is expected to give information used for further inferences. The criterion-related validity is generally evaluated by correlating the outcome of the instrument with another instrument. In the traditional perspective, criterion-related validity can be divided into concurrent validity and predictive validity. Concurrent validity deals with how well the instrument reflects another, external but *coincident* measure, and predictive

validity deals with how well the instrument reflects another, external, but *future* measure. Messick describes the distinction as follows:

> *Predictive evidence bears on the extent to which the test scores forecast an individual's future level on the criterion, whereas concurrent evidence relates to the test scores to an individual's present standing on the criterion.* (Messick, 1989, p.71)

Construct validity is a wider concept than the other validity aspects. The most common definition of construct validity is that it refers to how well the instrument measures higher-order objectives, usually some sort of unobservable trait. Intelligence, motivation, creativity or some sort of aptitude or ability can represent such constructs. Some argue that all other validity aspects are in some respects subordinate to construct validity (Messick, 1995; Shepard, 1993).

This definition of validity has been frequently discussed. There are many different views on what a certain category involves, what aspects that are subordinate other aspects, and so forth. For instance, some treat predictive validity as a sub-category of construct validity, some treat construct validity and predictive validity as two entirely separate concepts (Brualdi, 1999; Wolming, 2001). The implications of construct validity are most intensely debated, and is often claimed to be superior to all other aspects, especially in more recent interpretations (Messick, 1998; Shepard, 1993).

The traditional view on validity has a strong fundament in educational assessment. Most researchers agree that all the validity aspects discussed here are important to consider when validating an instrument, even though some aspects may be more important, considering the purpose of the instrument.

However, the validity discussion took a new turn when Samuel Messick proposed a somewhat different view on validity. He argued that the traditional view on validity is too limited. In his opinion, validity is not necessarily connected to the instrument itself, but more to how the outcome is interpreted and used and the consequences this will lead to from a wider and more long-term perspective (Messick, 1989; see also Shepard, 1993).

In Messick's model "The facets of validity" (1989), he identifies a number of validity aspects, and reduces them to a system. His model takes the purpose of the test score, grade, or other form of assessment outcome, into consideration. He appreciates how it is interpreted and used, and what type of evidence and consequences will be the result. Even though Messick makes a clear distinction between the aspects of the model, he does not deny that they are also intertwined and emphasises the importance of taking all these aspects into consideration when validating an instrument (Messick, 1989; 1995; 1998).

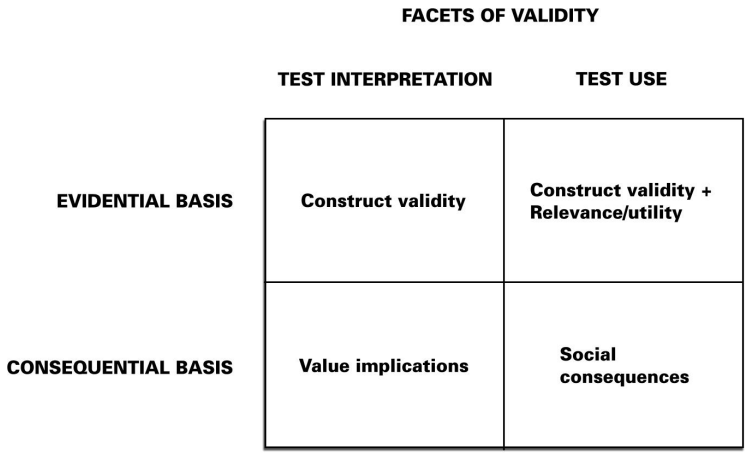

Figure 1. The facets of validity, according to Messick (1989, p.20)

The validity discussion has been intense ever since Messick's broadened view of validity was introduced. There are a number of researchers, for instance Popham (1997) and Mehrens (1997) who criticise this view on validity, with the argument that the focus is too distant from the instrument (see also Shepard, 1993). They agree that the consequences are important to consider, but that such issues should be kept separate from the traditional validity aspects. Sireci (1998), for instance, argues that content validity has often been overlooked by researchers focusing too much on other aspects, which he regards as a serious problem for ensuring the quality of the instruments.

However, while the debates are focusing on what is to be included in the concept of validity, and what is not, most researchers agree that it is necessary to widen the perspective when validating an instrument, since one instrument can be used for more than one purpose, and the validity is known to vary with how the outcome is used, a fact that was well known decades before Messick's model was introduced (see, for instance, Ghiselli et al, 1981).

In my opinion, the important contribution of Messick's model is not only the broadened view on validity, but also that it provides the researcher with a useful tool for validation processes (see Figure 1). This is especially valuable when aiming to systematically investigate the effects of assessment instruments or systems that are complex and have an impact on many levels. Since the Swedish grade system and the instruments for selection to higher education are examples of such complex instruments, I find Messick's model most useful, and especially appreciate the focus on social consequences, since the decisions based on such instruments will affect not only the examinees, but also schools, educational systems, and society in general.

4. Assessment in Sweden

This chapter will give a short presentation of the assessment and grading system in Sweden. More details may be found in the papers in this thesis, chiefly in papers I and II.

4.1. The change of system

Until the mid-1990s, the Swedish grading system was norm-referenced. Students were graded on a scale from 1 to 5, where 3 represented average achievement. This was based on the assumption that students' abilities and performances are distributed along an achievement scale, following a normal distribution (Andersson, 1999; Wedman, 2000). This distribution, or the Gaussian normal distribution curve, was the basis for the grading process. Standardised tests were to help the teachers to place the students along the scale. There was, however, a lot of criticism of this approach. The criticism

came chiefly from professionals in the educational field, but also from students and parents, and concerned the grades' lack of knowledge orientation, since the students (and teachers) were more focused on how a student performed relative to other students than on what they actually learned. This system was regarded as very competitive and hence contributing to a negative climate in schools (Andersson, 1999, 2002). However, it should be noted that the criticism of the norm-referenced system concerned how it was used in schools, and not its role of being an instrument for selection to higher education (Skolöverstyrelsen, 1983; Wedman, 2000).

From the 1970s onwards, there were discussions of changing to a criterion-referenced system (Skolöverstyrelsen, 1983; Wedman, 2000). A number of reports were produced, and most of them were agreed that a criterion-referenced system had many advantages, even though a change of system would also lead to many practical problems. Wedman, for instance, expresses strong concerns about how to define the objectives, and how to achieve fairness and comparability in a system based on subjective interpretations, and argues:

> *What type of grade system should be chosen is related to what purpose the grades should have. If the grades are to be used for assessing if and when the student has reached a certain performance level, the criterion-referenced grading is sufficient and also possible to carry out, in spite of the comparability problems. If the grades are to be used to rank students after their graduation, some sort of relative grading is necessary from a fairness and comparability point of view. [...] A true criterion-referenced grading procedure that will ease the stress and competitiveness can only be achieved if schools are relieved of participating in the selection to closed study positions and professions.* (my translation, in Skolöverstyrelsen, 1983, p.105-106)

In spite of these concerns, the Swedish grade system was changed to criterion-referenced in 1994. The grades were to be used for multiple purposes: performance assessment on an individual basis, educational evaluation in general and for the selection to higher education. Standardised curricula and grading criteria were issued, with overall objectives for each subject, which were left to the schools to interpret and define on a practical level (Andersson, 2002; Wedman, 2000). So-called national tests were made available for some subjects, for the purpose of guiding the teachers to the grading scale

(Lindström, 2003: Nyström, 2004). When the school system was decentralised, a new level of competition was also introduced, opening up for private alternatives, and making the schools compete for their students.

The introduction of the criterion-referenced system was a turbulent period of time. For a while three fundamentally different systems were practised at the same time: the old norm-referenced system, the new criterion-referenced system, and a mixture of both. There were also problems with the criteria for grading, since they did not apply to all grade levels. General guidelines were issued centrally, but still a large part of the responsibilities for formulating criteria for each subject were left to the individual schools.

4.2. Grading

Sweden differs from most other countries by leaving the entire responsibility for assessing and grading the students to the teacher (the National Agency for Education, 2004). The grading is a result of an extensive process centred round standard documents: the curricula with the learning objectives and the grading criteria for each course and subject (see Figure 2). The grading criteria are supposed to describe the objectives and performance levels that are needed for each grade level. The general idea is to base the performance standards on a general knowledge domain, but differentiated by cognitive demands, in line with the taxonomy defined by Bloom (Anderson & Krathwohl, 2001; Bloom, 1956). The more advanced the student proves to be when approaching the particular subject, the higher grade will be awarded.

The teachers base their decisions on information gathered in the classroom. Classroom assessment is however not only one practice, but many, and the grading process based on such assessment comprises several steps, as illustrated in Figure 2 (see also McMillan, 1997). Nevertheless, the Swedish system is based on the assumption that teachers are proficient in assessment and grading, and hence will do a good job in assigning their students the accurate grade. The modes of assessment are optional. In theory, the focus is expected to be on performance assessment and portfolios, which is agreed to be appropriate for

formative purposes (Linn, 1995). To some extent this is true, but traditional teacher constructed tests are very common. The teachers evaluate the evidence, and relate the findings to the grading criteria. In Figure 2, the process from the curricula with the learning objectives to grade outcome is illustrated.

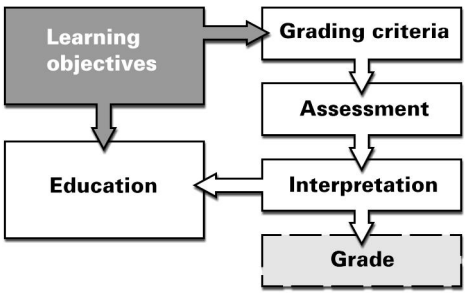

Figure 2. The grading process, from learning objectives to grade outcome.

4.3. Selection to higher education

As previously discussed, school grades serve a number of purposes. Apart from evaluating educational performance, grades are also used for selecting students to higher education (Wedman, 2000, Wolming, 1999).

The purpose of a selection instrument is to distinguish between individuals for the purpose of deciding who are to be selected (in this case, admitted) and those who are not (Gregory, 2004; Nitko, 1996). When developing and using such an instrument, a number of issues should be considered. First of all, what does the instrument intend to measure? In some cases, the purpose of the instrument is to ensure that only those with the necessary requirements in terms of knowledge and skills will be accepted. In such a case, there must be a pass score, or standard, and the students that will be considered are those who scores higher. In other cases, ensuring basic requirements is not the main issue, but the purpose is to rank the applicants according to their performance, and to select as many as there are openings (starting from the top in the distribution). The construct here is more likely a trait, such as study success, work performance, etc. The purpose and use of the instrument will then be what determines the approach in the validation process. In most cases decisions are made on whom

to accept and whom to reject. An instrument for selection to higher education should therefore be categorised as high-stakes. The consequences of incorrect decisions are generally serious, since they affect not only the students but also the universities and society in general.

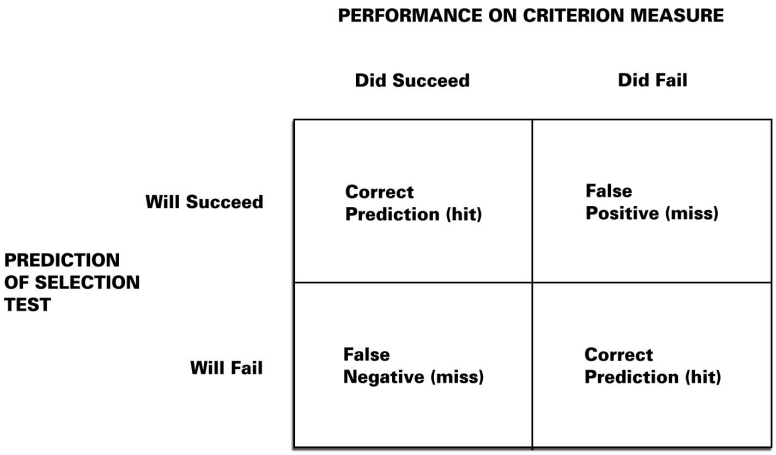

Figure 3. Decision theory model according to Gregory: "Possible outcomes when a selection test is used to predict performance on a criterion measure" (2004, p. 104)

According to the model illustrated in Figure 3, the consequences of correct-incorrect decisions depend on perspective and outcome. If there is a false negative, the student who were not admitted (but would have succeeded) will be affected. If there is a false positive, the school who admitted the student (who failed) will be affected.

4.3.1. The SweSAT

In Sweden there are two main instruments used for selection to higher education: school grades and the SweSAT (Stage, 1994). It should be noted that these instrument are not used together, as is common in other countries. All applicants are expected to have graduated from upper secondary school (or education regarded as equal), which means that they will have a GPA. The

SweSAT is optional, and a score from this test can be used as an alternative to the GPA, if it is regarded as more beneficial for the student.

The SweSAT is a test battery similar to the American SAT, and is expected to predict study success, by including objectives considered important for successful studies in higher education (Stage & Ögren, 2004). Such areas are Swedish word comprehension, reading comprehension (both in Swedish and English), mathematical reasoning and interpretation of diagrams, maps and tables. All subtests are based on multiple-choice items. The SweSAT is open to anyone, administered twice a year, and scored centrally. There is no pass score, however. The test is norm-referenced, and scores are given on a standardised scale in order to make test scores comparable over time. The SweSAT is described more thoroughly in the papers I-III.

4.3.2. The GPA

Upper secondary school grades constitute the main instrument for selection to higher education. The grades are used for two purposes in the selection process. First of all, they are to ensure that students fulfil the necessary requirements for admission to higher education (these differ among university courses). Second, they are used for ranking the students when the number of applicants exceeds the number of openings (SOU 2004:29, Wedman, 2000).

Grades used for selection to higher education have some basic requirements attached: they should be fair and they should make it possible to compare and distinguish the applicants from each other on the basis of their performance and assumed ability. In order to function as selection instruments, they should also measure a construct relevant for the selection process. However, after the educational reform in 1994, the selection aspect has been toned down, while the other aspects of the criterion-referenced grades have been emphasised (Wolming, 1999).

It has been argued that criterion-referenced instruments are not always suitable for ranking students (see Paper I as well as Wedman, 2000). However, measurement theory does not always contradict the possibility of using a

criterion-referenced instrument for ranking purposes, if some necessary prerequisites can be met (Black and Dockrell, 1987; Hambleton et al. 1978). One such prerequisite is score variability (Shavelson et al., 1993). Without variability it is not possible to rank the students. The problem with a criterion-referenced system is chiefly that all students may receive the same grade (= no variability). One way of solving this problem has been to calculate the grade point average, consisting of a large number of grades. The more grades that are included, the greater is the chance of variability. The way the GPA is calculated is an important issue that unfortunately has not been paid much attention. There are likely to be a number of complications associated with transforming the qualitative scale into a quantitative one, and, in order to complicate matters further, also with multiplying the quantitative grades by the number of credits in the various courses. Calculating an average based on this model will most likely lead to a fair share of problems with regard to comparability.

Students who apply for higher education are ranked on the basis of their GPA, which is calculated by the quantified upper secondary school grades (*fail*=0, *pass*=10, *pass with distinction*=15, *pass with special distinction*=20). The GPA consists of a large number of grades, of which all are considered equal. What determines the weight is the length of the course, since the calculation of the average takes the number of credits (which are based on the number of hours assigned to each course) into consideration. The GPA is calculated as follows:

$$GPA = \frac{\sum_{j=1}^{h} c_j \cdot g_j}{\sum_{j=1}^{h} c_j}$$

In the equation, h denotes the number of courses included in the GPA, g_j denotes the awarded grade level for course j, and c_j the length, in terms of the number of credits awarded for course j (see also Lyrén, 2005).

4.4. The selection instruments

The aims of the selection process to higher education is to have a broad recruitment to higher education, rank students according to their achievements, while also being fair, by avoiding giving some groups of students advantages that others do not have (SOU 2004:29; Wedman, 2000).

The selection to higher education has for a long time been subject to an intense debate among educational researchers in Sweden. During the last couple of years, this issue has also received considerable media attention. One of the reasons is that an increasing proportion of students are leaving upper secondary school with a maximum GPA. The medical faculties, for instance, have now a larger number of applicants with a maximum grade average than there are available study positions to offer (SOU 2004:29).

The Swedish situation is unique, by having two instruments for selection to higher education, very different in format and purpose, that can be used more or less interchangeably. While one instrument is a standardised test, the other is the result of teachers' decisions based on evidence of classroom assessment. The test is norm-referenced and the grades are aiming to be criterion-referenced. Both instruments have their problems and advantages. Grades are the result of evidence collected over a long period, while the test only represents a snapshot of a person's knowledge and skills on a certain occasion. On the other hand, the test can be taken an (almost) unlimited number of times. The grades includes a varying degree of subjectivity, while the test is objectively administered and scored. Even though most students show a similar degree of success on both instruments, there are also categories of students who are advantaged by one and disadvantaged by the other. Furthermore, both instruments offer some degree of strategic behaviour. There is a repeated-test effect, which generally makes re-taking the test worthwhile (Henriksson & Bränberg, 1994; Törnkvist & Henriksson, 2004). It is also claimed that the GPA can be manipulated by course choices and supplementation of grades (Cliffordson, 2004a; Löfgren, 2004a; SOU 2004:29).

Since the instruments are so different in their characteristics, opinions are divided on the issue of which instrument is the most suitable. Some are in favour of using the school grades, while others propose standardised tests or other forms of assessments (SOU 2004:29). Research shows that the predictive validity of the GPA is higher than that of the SweSAT, when the external criterion is the production of the expected number of university credits per year (Cliffordson, 2004b; Gustafsson, 2003; Henriksson & Wolming, 1998). According to Gustafsson, (2003), one reason is that the GPA to a higher degree than the SweSAT captures factors like industriousness, endurance and motivation. However, a relevant question regarding selection instruments and their validity is whether we should be content with ensuring that the predictive validity is on an acceptable level? Because even if the predictive validity is high, it does not necessarily mean that the instrument is fair. Furthermore, if the system is open for strategies, "clever" students will find these openings and may very well be successful when once admitted to higher education.

5. Summary of the papers

In this chapter the four papers that together represent the core content of this thesis, will be summarised. These papers all investigate empirically the Swedish upper secondary grades, how they vary and plausible explanations of this variation. The first paper (I) was co-written with Magnus Wikström and in the other three papers (II-IV) I am the sole author. The papers are presented in chronological order, in terms of when the studies were conducted.

The studies are, to some extent, linked together. While focusing on separate but related issues, they are also slightly overlapping. One reason is that the results of one study have given the idea for the next study, and so forth. Another reason is that they were all developed within the framework of a subproject of the VALUTA project, aiming to illuminate the characteristics of the criterion-referenced upper secondary school grades (Löfgren, 2004b).

The idea behind the first analysis, described in paper I, was based on an ongoing debate in Sweden concerning the increasing number of independent schools and how the competition among schools affects educational outcome. Grades were commonly used as an instrument for capturing performance differences, and our idea was to investigate if the grading is comparable, and how it is affected by the competition among schools. Another discussion during that period concerned the increasing grade averages, and if the increase was related to improved performances or not, which initiated study number two (paper II). In the analyses in these studies, some of the variables included in the models indicated that school size and programme characteristics could be related to how students are graded and ranked by their GPA. This led me to conduct two more studies, one investigating the effect of school size, as described in paper III, and the other focusing on the effect of programme enrolment (paper IV).

5.1. Paper I

Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools.

As previously mentioned, the idea behind this study originates from a public debate about whether school competition affects school quality or not. Since school grades were used as evidence in this debate, it felt important to investigate their reliability. The hypothesis was that grading varies between schools, and that competition puts pressure on the grading.

The study investigates how competition between schools affects how students are graded. The analysis is based on data from students enrolled in theoretically oriented programmes (natural science orientation and social science orientation) in Swedish upper secondary schools. The selected students graduated in 1997, and were the first cohort that graduated with a GPA based on the criterion-referenced grades. In order to analyse how grading varies among schools, a separate measure was created, based on the students' final grades (their GPA) and their scores on the SweSAT test. Schools in municipalities without

competition (only one upper secondary school) were compared with schools in municipalities with competition, where a distinction was made on the basis of whether the competition was from public or private schools (or both). The results show that public schools exposed to competition from other public schools in the same municipality grade their students slightly higher than other schools. However, the most conspicuous result is related to the private schools, who prove to inflate their grades considerably. This is interpreted as an effect of grades being used as a quality indicator for the schools, which in combination with competition leads to grade inflation.

5.2. Paper II

Grade stability in a criterion-referenced grading system: the Swedish example.

Ever since the criterion-referenced grading system was introduced, grade averages in upper secondary schools have been rising. This has been interpreted as improved achievements, at least from a political point of view. Others have offered other explanations of this effect, for instance that teachers are slow in learning to use the scale, and so forth. The purpose of this study is therefore to investigate the causes of the grade increase. Four hypotheses are presented as plausible explanations: better student achievements, student selection effects, strategic behaviour in course choices, and lowering of grading standards. The analysis is based on data including all students who graduated from Swedish upper secondary schools over a six-year period, starting with the year 1997, which was the first year when all upper secondary students graduated with criterion-referenced grades. The result shows that all student groups increased their grades over time, and high performers gained the most. An analysis of students' performances on the SweSAT does not support the idea of increased achievements, but the contrary. Even though important explanatory variables are held constant, there is no effect on grade averages, which is taken to indicate that selection effects cannot be the cause. The effect of course choices is investigated by studying grades from various

courses separately. The results show that all grades are increasing in line with the GPA. When adding up the various results, the only plausible explanation is that grading standards have been lowered over time, which is interpreted here as grade inflation. The grade inflation is assumed to be an effect of the leniency of the grading system in combination with pressure for high grading, chiefly related to the upper secondary school grades' function as an instrument for selection to higher education.

5.3. Paper III

Does school size affect teachers' grading practices?

This paper is based on the hypothesis that school size is relevant to how students are being assessed and graded. The analysis is empirically based, and focuses on grade outcome from students enrolled in the natural science programme who graduated from upper secondary schools between the years 1997 and 2002. The approach is to analyse the rank differences between school grades and test scores from the Swedish Scholastic Assessment Test (the SweSAT) among schools in different size categories. The results show that there is a grading effect related to school size. Students enrolled in large schools (>1000 students) are systematically graded slightly lower than other students. Students in small schools (<300 students) have higher grade averages but are also graded higher than other students, and students enrolled in private schools (which generally are small) are even higher graded. The exception is small schools in large cities. These schools have a lower grade average but are also lower grading. The reasons for the size effect cannot be determined by the analysis, but is assumed to be related to the use of different assessment methods, or a combination of such differences with internal and external pressure for high grading. This is supported by the fact that both small schools and private schools behave similarly when grading their students. These schools attract high performing students, who are known to press for high grading and are also more likely to be more sensitive to competition than other schools. The conclusion is that grades are not strictly comparable among schools, which,

among other things, negatively affects the grades' function as instruments for educational evaluation on school level, as well as student ranking.

5.4. Paper IV

The reliability and validity of the grade point average as instrument for selection to higher education.

This study follows the findings in previous studies, indicating that programme enrolment is of importance for how students are graded. The study analyses the upper secondary GPA as an instrument for selection to higher education in relation to the consequences of the educational reform in 1994, which led to fundamental changes in the educational system in general, but also changed the prerequisites connected to assessment and grading. One such change is that all upper secondary programmes are to provide the students with the basic requirements for university eligibility, irrespective of orientation. This means that if no additional requirements are set, grades and GPA from all programmes are treated as equal in the process of selection to higher education. The purpose of the study is to find out if some programmes are systematically higher or lower grading than other programmes, and also how the course compositions of the various programmes affect the students' GPA and hence their ranks in a selection process. The analysis is based on data from all students graduating from Swedish upper secondary schools in 2002. Students are sorted into categories depending on their affiliation, and their grades from compulsory courses as well as their GPA are investigated. Variables indicating rank differences are created for both the first research question (the programme's effect on the grading) and the second (the course composition's effect). The results show that students in the vocationally orientated programmes are slightly higher graded, and that the course composition in different programmes affects the students' GPA considerably. Students in vocationally oriented programmes are advantaged, while especially the students in the natural science programme are disadvantaged. This will have impact on the students' competitive strength in the process of selection to higher education.

6. Discussion

In this chapter, the results from the empirical studies (Paper I-IV) will be discussed from a validity perspective.

As previously mentioned, the validation process is always related to the purpose of the instrument or assessment system that is to be analysed. Furthermore, it has been concluded that validity is not a fixed quality connected to an instrument, since it will change with the different circumstances of the measurement situation. The validity issues of scores (or grades) and meanings, their relevance, utility and social consequences are faceted and intertwined. This complicated matter of validation is supported by Messick, who concludes that there is a reason why validity has come to be viewed as a unified concept, and that the different aspects cannot stand alone in validity arguments (Messick, 1995, 1998).

However, since the concept of validity is complex and the Swedish grade system has multiple purposes and functions, there is no limit to the number of issues that could be raised, all more or less affecting the others. It is not practical or even possible to address all these issues. The discussion here will therefore be limited to the two research issues presented as parts of the purpose of this thesis. The first addresses the characteristics of the assessment instrument (the grades) when used for educational evaluation (the criterion-referenced approach), and the second aspect focuses on the implications of using the instrument (the GPA) for ranking students (the norm-referenced approach). Due to these research issues and the nature of the research included in this thesis, Messick's broad validity perspective is found suitable and useful. The discussion will chiefly be based on the consequential basis for how assessment outcomes are interpreted and used.

6.1. The results

In this section I will summarise the main results from the studies described in Paper I-IV in relation to traditional reliability and validity aspects. In the following sections, validity issues will be more thoroughly discussed from a broader perspective.

First of all, an instrument used for educational evaluation must be reliable. The reliability is necessary for ensuring that the outcome is comparable among administrations. Lacking reliability is known to affect validity negatively in general, but perhaps chiefly the content validity (Nyström, 2004), which, in turn, affects other validity aspects. Secondly, neither the instruments for selection to higher education should be affected by random or systematic bias. Here, the predictive validity is central, but in order to achieve predictive validity, the instrument must be construct valid (which also incorporates content and criterion validity).

All studies described in Papers I-IV focus on grades, but foremost on the GPA, which should be considered a different instrument than the separate grades, since its purpose differs. The separate grades are foremost investigated in Paper II and IV. The evidence provided by all these studies show that the reliability of the grades is lacking. Grade levels varies between students, groups of students, schools, their characteristics and location. This is assumed to be related to the grading criteria (see discussion in section 2.2.) and its openness for subjective interpretation, as discussed in section 4.2., in combination with few control mechanisms and a strong pressure for high grading, both from students, parents and schools, due to the grades' functions as quality indicators and selection instruments. The GPA is directly affected by the problems with the separate grades, but also proves to include other sources of error. Foremost, the results show that the GPA is affected by the course composition the included grades are representing, which is related to upper secondary programmes and the students' course choices. These choices may relate to personal interests but also from ambitions to maximise the GPA.

The general findings in the papers are therefore that there are problems with both the reliability and the validity of the grades, both in terms of being instruments for educational evaluation and selection instruments to higher education.

6.2. Evidence of, and consequences for interpretation and use

I will now move on from the specific results provided by the papers to focus more on validity aspects in a wider perspective.

6.2.1. Construct validity – evidence, relevance and utility

Construct validity is often claimed to be an overarching validity issue (Messick, 1998; Shepard, 1993). Due to the type of analysis that has been conducted in this thesis, the criterion validity, content validity and construct validity of the criterion-referenced grades cannot be directly investigated. There is no information concerning how students are assessed and how the instruments being used correlate with the criteria. However, inferences from the findings in papers I-IV all point in the same direction. The Swedish school grades are most likely not absolute measures of certain performance standards, since they incorporate a certain degree of relative grading, as well as different sources of error. A grade can differ considerably, depending on how, where and when the student was graded.

The GPA complicates matters further, since the purpose of the GPA is not to give information about a certain construct. There is no pass score that will tell if the student is a master or non-master. It is not expected (or formulated) to be a relation between a certain GPA and how this student will perform in higher education. The assumption is "the higher, the better". The GPA can therefore be investigated in terms of content, but not in terms of the construct it intends to represent. Furthermore, there are a number of additional complications built into the system (see Figure 4), such as the composition of grades included in the GPA, which will vary among students. The lengths of these courses are also important, since a longer course is weighted higher. This supports the view that content validity is problematic for the grades and especially for the GPA, which most likely affects the criterion validity and, consequently, the construct validity.

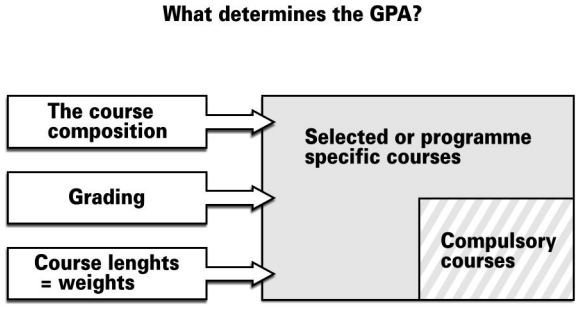

Figure 4. Threats to validity in the GPA model.

When the construct validity is problematic, both the relevance and the utility of the instrument will be affected. Students may be more or less advantaged in the process of selection to higher education, for instance due to the grade inflation that can be noticed among schools and over time. Even a small variation can be crucial for the student, especially if s/he is aiming for an attractive study position, for instance where students are selected on the basis of the second decimal of the GPA (SOU 2004:29). Furthermore, an inflated grade will no longer have an acceptable relation to the performance criteria for that grade level. This will affect the construct validity, and may lead to students (or teachers) misinterpreting the students' previous knowledge, which will negatively affect her/his chances of being successful in education where such performance levels are necessary.

6.2.2. Value implications

Construct validity, in terms of utility and relevance, affects the value implications of an instrument. The value of assessment outcomes is dependent on a number of issues, but chiefly on how the outcome can be used. Hence, a disadvantage of using the criterion-referenced grades for selection is that they decrease in value as instruments for educational confirmation. The grade level that corresponds to a certain performance standard in a specific subject turns into a figure when included in the GPA, which means that the courses and

what they represent become less important than finding ways to maximise the rank position, relative to other students.

6.2.3. Social consequences

As previously mentioned, the major advantages of an absolute approach is that the outcome can be used for giving feedback to students and parents concerning performance in relation to standardised objectives, and for monitoring educational progress in general. In the perspective of educational feedback, the variation in grading generally has fewer consequences for students than for teachers and schools. A student may be underrated or overrated, which may seem unfair and also be misguiding for future educational planning. However, a high-grading school is easily mistaken for providing better education than a low-grading school. This is an important issue, since school output is regularly reported both in official reports for policy purposes and by the media (Lyrén, 2005). This information is used by students and their parents in school choices, and low grading schools are likely to be disadvantaged when competing for the students. Furthermore, in the absence of information from standardised tests, grades are commonly used by researchers who focus on educational evaluation. Failure to recognise the systematic variations may lead the researcher and consequently policy-makers to incorrect conclusions.

The consequences for the students are much more serious when grades are used as a selection instrument. In the analysis of how students are ranked by their grades (I-IV), the results confirm that some student categories are advantaged, both in the grading, but also in the GPA model. Here a rather uncommon effect can be found, related to student gender. Traditionally, male students have been overrepresented in higher education, and especially in attractive programmes, such as medicine and law, which has been considered a serious problem from an equality point of view. However, the increasing gender difference that can be noticed in the grades is shifting this proportion (see paper II). Male students are becoming underrepresented in many university programmes and are also disadvantaged in the selection process (SOU 2004:29), unless they apply with SweSAT scores, where male

students are generally more successful (and the problem is reversed) (Stage and Ögren, 2004).

A matter of serious concern is related to social segregation. The results in papers I-IV all show that that students with a high socio-economic background and students born in Sweden are ranked higher than other students, everything else equal, both when it comes to the separate grades and the GPA. This means that if comparing "similar" students, in terms of, for instance, previous performance, students with smaller economic means and immigrants are disfavoured. These findings are highly worrying, since they imply not only that there is a fairness issue for the students who are affected, but also that students who are already underrepresented in higher education are disfavoured by the system. There are also other mechanisms that may lead to increased social segregation. High performing students seem more successful in pressuring for high grading (papers II and III), and schools are also becoming increasingly selective (paper I, Arnman, Järnek & Lindskog, 2004). There are also a number of openings for strategic behaviour, which usually already advantaged students exploit (Löfgren, 2004a; Törnkvist & Henriksson, 2004). Furthermore, inflated grades affect not only the students' competitive strength in the selection process, but also society in general, and especially from an economical point of view. To be able to compete for attractive study positions, students often need to supplement their grades, which makes them extend their upper secondary education longer than necessary (Löfgren, 2004a; SOU 2004:29). Furthermore, the GPA model leads to extended consequences in terms of course choices. For instance, a basic level course may be worth as much as a deeper level course, and a basic level course in, for instance, cooking is regarded as equal to an advanced course in, for instance, language or natural science, if assuming that they are equal in length. Hence, advanced courses in theoretical subjects are losing in popularity (SOU 2004:29). This affects not only university students' previous knowledge, but it also diminishes the student basis for university programmes where such courses are necessary. This will consequently affect professions where there are special requirements for advanced education based on such subjects.

7. Conclusions

The overall purpose of this thesis is to investigate the implications of using a criterion-referenced instrument for both educational evaluation and student selection to higher education. The analyses focuses on the Swedish upper secondary school grades, since they have these characteristics. In papers I-IV, grade outcome has been studied from a number of perspectives. The results have then been discussed from a perspective of validity, where the validity model according to Messick (1989) was found to be most useful.

The conclusion is that the Swedish grading model involves several sources of error, which causes problems with both reliability and validity issues. This is to a large extent related to the conflicting purposes of the grades. The increasing competition among students, teachers, schools and municipalities, using grades as quality indicators when competing for attractive positions has put strong pressure on the grading process. This negatively affects the usefulness and relevance of the criterion-referenced grades as instruments for educational evaluation and selection, which, in turn, leads to negative consequences for students, teachers, the educational system and society in general.

7.1. Suggestions for future research

A large number of research issues remain to be investigated. For instance, more information is needed concerning the GPA as an instrument for ranking students and what alternatives there may be to the present GPA model. It would also be interesting and useful to compare the benefits and problems of the Swedish assessment and selection model with similar issues in a diametrically opposed system, where teacher influence is low and external instruments are used for educational evaluation and selection, or perhaps systems where a middle course is practiced. There are most likely benefits and problems in all approaches, but an international overview would make the process of finding better solutions easier and more efficient, by learning from other experiences and not always having to "re-invent the wheel".

References

Anderson, L.W., & Krathwohl, D.R. (Eds.), (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York: Addison Wesley Longman.

Andersson, H. (1999). *Varför betyg? Historiskt och aktuellt om betygen* [Why marks? Historical and actual views on grading]. Lund: Studentlitteratur.

Andersson, H. (Ed.). (2002). *Att bedöma eller döma. Tio artiklar om betygssättning och bedömning* [To assess or to be assessed. Ten articles about grading and assessment]. Stockholm: The Swedish National Agency for Education.

Arnman, G. Järnek, M. & Lindskog, E. (2004). *Valfrihet - fiktion och verklighet* [Freedom of choice - fiction and reality]. (Research reports 4). Uppsala: Uppsala University, Department of Education.

Black, H.D., & Dockrell, W.B. (1987). *Criterion-referenced assessment in the classroom*. Glasgow: The Scottish Council for Research in Education.

Bloom, B.S. (1956). *Taxonomy of educational objectives: the classification of educational goals. Handbook 1, Cognitive domain*. New York: David McKay.

Brualdi, A. (1999). Traditional and modern concepts of validity. *ERIC Clearinghouse on Assessment and Evaluation.* Retrieved from www.ericdigests.org/2000-3/validity.htm

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

Cliffordson, C. (2004b). *De målrelaterade gymnasiebetygens prognosförmåga* [The predictive validity of goal-related grades from upper secondary school]. *Pedagogisk Forskning i Sverige, 9*(2), 129-140.

Ghiselli, E.E., Campbell, J.P. & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W.H. Freeman & Co.

Glaser, R., & Klaus, D. J. (1962). Proficiency measurement: Assessing human performance. In R. Gagne (Ed.). *Psychological principles in system development*. pp. 419-474. New York: Holt, Rinehart and Winston.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. Some questions, *American Psychologist, 18,* 519-523.

Gregory, R. J. (2004). *Psychological testing: history, principles and applications* (4th edition). Needham Hights: Allyn & Bacon.

Gustafsson, J-E. (2003). *The predictive validity of grades and aptitude tests in higher education.* Paper presented at the AEA-E conference, Lyon, France. Retrieved from www.aea-europe.net

Hambleton, R.K. (1994). The rise and fall of criterion-referenced measurement? *Educational Measurement: Issues and Practice, 13*(4), 21-27.

Hambleton, R.K., Swaminathan, H., Algina, J., & Coulson, D. (1978). Criterion referenced testing and measurement: a review of technical issues and developments. *Review of Educational Research, 48*(1), 1-47.

Henriksson, W., & Bränberg, K. (1994). The effects of practice on the Swedish Scholastic Aptitude test (SweSAT). *Scandinavian Journal of Educational Research, 28*(2), 129-148.

Henriksson, W., & Wolming, S. (1998). Academic performance in four study programmes: a comparison of students admitted on the basis of GPA and SweSAT-scores with and without credits for work experience. *Scandinavian Journal of Educational Research, 42*(2), 135-150.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23*(3), 6-15.

Lindström, J-O. (2003). *The Swedish national course tests in mathematics* (Em No. 43). Umeå: Umeå university, Sweden: Department of Educational Measurement.

Linn, R.L. (Ed.). (1989). *Educational measurement* (3rd edition). New York: Macmillan.

Linn, R.L. (1994). Criterion-referenced measurement: a valuable perspective clouded by surplus meaning, *Educational Measurement: Issues and Practice, 13* (4), 12-14.

Lyrén, P-E. (2005). *Approaches to accountability and modernisation in assessment practices in Sweden.* Paper presented at the NCME conference in Montreal, Canada.

Löfgren, K. (2004a). *Utbyteskompletteringar bland dem som avslutade gymnasiet 1997-2001* [Supplementary of grades among students who graduated from upper secondary school 1997-2001]. (BVM No. 6). Umeå: Umeå University, Department of Educational Measurement.

Löfgren, K. (2004b). *Validation of the Swedish university entrance system. selected results from the VALUTA-project 2001-2004* (EM No. 53). Umeå: Umeå University, Department of Educational Measurement.

McLeod, W.T. (Ed.). (1987). *The new Collins Dictionary and Thesaurus in one volume.* London & Glasgow: William Collins Sons & Co Ltd.

McMillan J.H. (1997). *Classroom assessment. principles and practice for effective instruction.* Boston: Ally and Bacon.

Mehrens, W.A. (1997). The consequences of consequential validity*, Educational measurement: Issues and Practice, 16*(2), 16-18.

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

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

Messick, S. (1998). Test validity: a matter of consequence. *Social Indicators Research, 45,* 35-44.

Millman, J. (1994). Criterion referenced-testing 30 years later: promise broken, promise kept. *Educational Measurement: Issues and Practice, 13*(4), 19-39.

National Agency for Higher Education (2004). *TIMSS 2003. Svenska elevers kunskaper i matematik och naturvetenskap i skolår 8 i ett nationellt och internationellt perspektiv* [TIMSS 2003. Swedish students' skills in mathematics and natural science in 8th form, from a national and international perspective]. Stockholm: Fritzes.

Nitko, A.J. (1996). *Educational Assessment of Students* (2nd edition). Englewood Cliffs: Merril.

Nyström, P. (2004). *Rätt mätt på prov: Om validering av bedömningar i skolan* [Validation of educational assessments]. (Doctoral thesis). Umeå: Umeå University, Department of Education.

Parkes, J. (2000). The relationship between reliability and cost of performance assessments. *Education Policy Analysis Archives, 8*(16), retrieved from http://epaa.asu.edu/epaa/v8n16

Popham, W.J. (1994). The instructional consequences of criterion-referenced clarity. *Educational Measurement: Issues and Practice, 13*(4), 15-18.

Popham, W.J. (1997). Consequential validity: right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Popham, W.J., & Husek, T.R. (1969). Implications of Criterion-Referenced Measurement. *Journal of Educational Measurement, 6*(1), 1-9.

Roos, B. (2005). *ICT and formative assessment in the learning society.* Doctoral thesis. Umeå: Umeå University, Department of Education.

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Shavelson, R.J., Baxter, G.P. & Gao, X. (1993). Sampling variability of performance assessment, *Journal of Educational Measurement 30*(3), 215-232.

Shepard, L.A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.). *Review of research in education, 19.* Washington: American Educational Research Association.

Skolöverstyrelsen (1983). *Den eviga betygsfrågan. Historiskt och aktuellt om betygssättningen i skolan* [The eternal grading question. Historical and

current issues about grading in the schools]. Stockholm: Skolöverstyrelsen & LiberLäromedel/Utbildningsförlaget.

SOU 2004:29. *Tre vägar till den öppna högskolan* [Three routes to the open university]. Stockholm: National Agency for Higher Education.

Stage, C. (1994). *Use of Assessment Outcomes in Selecting Candidates for Secondary and Tertiary Education* (EM No. 11). Umeå: Umeå University, Department of Educational Measurement.

Stage, C., & Ögren, G. (2004). *The Swedish Scholastic Assessment Test (SweSAT). Development, Results and Experiences* (EM No. 49). Umeå: Umeå University, Department of Educational Measurement.

Törnkvist, B., & Henriksson, W. (2004). *Repeated test taking. Differences between social groups* (EM No. 47). Umeå: Umeå University, Department of Educational Measurement.

Wedman, I. (1973). *Kriterierelaterade prov: bakgrund, egenskaper och begränsningar* [Criterion-referenced tests: background, characteristics and limitations]. (ER No. 33). Umeå: Umeå University, Department of Education.

Wedman, I. (2000). *Behörighet, rekrytering och urval. Om övergången från gymnasieskola till högskola* [Eligibility, recruitment and selection. About the transition from upper secondary school to higher education]. (No. 2000:6 AR). Stockholm: National Agency for Higher Education.

Wiberg, M. (2004). *Classical test theory vs. item response theory. An evaluation of the theory test in the Swedish driving license test.* (EM No. 50). Umeå: Umeå University, Department of Educational Measurement.

Wolming, S. (1998). Validitet. Ett traditionellt begrepp i modern tillämpning [Validity. A traditional concept in modern application]. *Pedagogisk Forskning i Sverige, 3*(2), 81-103.

Wolming, S. (1999). Validity issues in higher education selection: a Swedish example. *Studies in Educational Evaluation, 25*(4), 335-351.

Wolming, S. (2001). Att värdera urvalsinstrument. Några reflektioner over begränsningar och möjligheter [To validate selection instruments. Reflections over limitations and possibilities]. *Pedagogisk forskning i Sverige*, *6*(2), 122-130.

Yen, W.M. (1993). Scaling Performance Assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.