Deciding who is the best
Validity issues in selections and judgements in elite sport

Annika Johansson

Department of Education
Umeå 2010
DECIDING WHO IS THE BEST
Validity issues in selections and judgements in elite sport

ANNIKA JOHANSSON

Department of Education
UMEÅ UNIVERSITY
Nr 95 • 2010
Abstract

This thesis is about selection processes and processes of measuring and judging athletes in competitions in top-level sport. The purpose was to increase the knowledge of these processes and to analyse them from a validity perspective in order to contribute to the discussion of whether the “right” athletes are selected to participate in teams, competitions and games and whether the “right” athletes win. The rule and judging systems were investigated in the judged sports acroski and rhythmic gymnastics. Information was gathered through individual interviews with two judges, two coaches, and four elite athletes from each of the sports, and in addition to this the respective sport’s rule systems, judging manuals, meeting protocols and historical documents were studied. Selections to top-level sport teams were investigated by individual interviews with 14 top-level coaches (selectors) from the national league in soccer and a national team and from national teams in alpine skiing. The results from the judging study showed that both studied sports had undergone major changes in their rule and judging regulations, changes that had a considerable impact on the sports and the judgements. The level of definition of the rules and regulations was raised to increase the opportunities for clear and reliable judgements, but this became problematic for the overall validity of the judgements. The reason for this was that the new rules and regulations did not clearly correspond to the original idea of the sport, since the specified and detailed regulations lead to less originality and freedom in the performances. In the selection study, the results pointed to great differences in how precisely defined the selection criteria were among the teams. The selectors stated that well-defined selection criteria or grounds could be helpful in many ways, but they also emphasised how important it was for them that some parts of the selections were based on their subjective valuations of the athletes. Quite a few coaches from both sports argued that they would choose an athlete with good behaviour and favourable personality over an athlete with better sports skills, if they had an opportunity to do that depending on the selection system that was used. Overall, this research displays how validity issues connected to the selection and judging criteria and these processes might affect the outcome of the processes. It is notable that high reliability is in the main focus of the measuring and judging processes, while considerably vaguer and more subjective assessments are considered important in the selection process. The thesis points to the importance of discussing and understanding the consequences of rules, rule changes, selection and judging criteria as well as how these processes are performed, if the desired outcomes and consequences of the selection and judging processes are to be reached.

Key words: selections in sports, predictions of sport performance, assessment of sport performance, consequences of selections, judgement, judging, measurements, sport, validity, reliability, acroski, alpine skiing, rhythmic gymnastics, soccer.
ACKNOWLEDGEMENTS

I have been involved in sports for as long as I can remember and this has influenced my life in many ways. Except for being a joyful activity, sports have also been a place where I have learned a lot of things, met many wonderful people, experienced extreme happiness but also felt anger and despair. Sports have played and still play a big role for my development and my choices in life. My involvement in different levels and positions in sports lead to many questions and thoughts about sport and further to studies at Idrottspedagoglinjen (The study program for Sports Pedagogy) and to my current work with sport related educations at Umeå University. My experiences as a national team athlete in freestyle skiing triggered more questions and contributed to the writing of this thesis.

Writing a doctoral thesis can in many ways be compared to competing in elite sports, as it takes commitment, skills and knowledge and it is a process involving emotional ups and downs. Many of the abilities like goal commitment, hard work ethics and a strong internal motivation that are highly valued in elite sports are also important in the process of conducting research and writing a doctoral thesis. An important factor for success in elite sport has proved to be the influences and support from significant others and I have learned how important this support from others also is in this process. Therefore I would like to take the opportunity to let you who have supported me through this time know that you are greatly appreciated and that I would not have been able to finish this thesis without you. The reason that I at all started the work on this thesis is the now deceased Professor Ingemar Wedman who was the one that introduced me to the world of research and the field of measurements and judgements. He was my first supervisor at the time of writing my licentiate thesis and played a great role in my choice of an academic career. I am thankful for his help and advice and happy that I got to know him and could work together with him. To Professor Widar Henriksson, my second supervisor during the writing of my licentiate thesis, I also owe many thanks, as besides being a knowledgeable and patient supervisor, he also understood how important my sports career was for me and allowed me to combine my work with the licentiate thesis with my career as a skier in the Swedish freestyle ski team with all that that meant.

Eva Olofsson and Josef Fahlen have been my supervisors since 2007, when I resumed my research and continued to work on what finally became this doctoral thesis. To supervise me includes some challenges and can be a demanding task, but they have made an amazing effort to guide me through. They have
read through long manuscripts, contributed constructive feedback and believed in me more then I sometimes believed in myself and this has encouraged me to continue my work. Thanks you so much Eva and Josef! Gunnar Persson, who has revised my English, is worth a big thanks for his efforts. To Thierry Deschamps, who helped me with all technical issues related to the layout of the thesis and solving of many technical problems, Thanks! You are my saviour! To all of my colleagues at the Department of Education I like to express how happy I am to work with you and be cheered up by you. Special thanks to my doctoral seminar group and our seminar leader Britt-Marie Berge, and to Ann-Marie Smeds, Ulrika Sahlén, Ulrika Wikström, Inga-Britt Eriksson-Moström, Simon Wolming, Gerhard Nordlund, Jörgen From, Katrina Norberg, Kent Lövgren, Staffan Karp, Tor Söderström, Anna Frohm, Carina Holmgren, Katarina Weinheall and Seppo Salonen, who in different ways have helped me in connection to the work of my thesis. Thanks also to Professor Jan Wright for fast and valuable answers and reflections. To everyone who was involved in the work with my licentiate thesis: Thanks for help and support!!!

I am also very grateful to all of the informants who took their time to share their perceptions with me; without you this thesis would never have been written. To all my friends: Thank you for remembering me, although I have not been very social for some time. For financial support throughout this process I would like to thank the Department of Education, and the Department of Educational Measurements, Centrum för Idrottsforskning (the Swedish National Centre for Research in Sports) and Stiftelsen JCKempes Minnes Stipendiefond (the J C Kempe Memorial Foundation for Scholarships).

Finally I would like to round off these acknowledgements with some words to the people that have meant most for me during the writing of my thesis but also for my life in general. To my parents Barbro and Rolf and my siblings Magnus and Sophia, who have supported me throughout this period and during all my years of competing: You are my true heroes! I know I have been very busy and I am so grateful that you have put up with that and supported me the way you have. Thanks also Cyril and Sophia for always offering a place where I could get away from all work and to their little Chloé who makes me smile and forget about the hardships of writing a doctoral thesis. To my dear Rickard and Ronja and Jesper: Thanks for bringing me away from theories and work and back to life. Thanks for giving me inspiration, fun and love! Thanks for standing by my side and supporting me in this effort. You deserve a golden star for patience! Now I look forward to spending more time with you, skiing more with you, spending more time in the summer house in Öre and on the whole spending more time on fun, life and sports.

Umeå, April 29th, 2010

Annika
<table>
<thead>
<tr>
<th>Table of Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTRODUCTION</strong></td>
</tr>
<tr>
<td><strong>OUTLINE OF THE THESIS</strong></td>
</tr>
<tr>
<td><strong>DOING RESEARCH USING VALIDITY THEORY</strong></td>
</tr>
<tr>
<td><strong>The evolution of the concept of validity</strong></td>
</tr>
<tr>
<td>1. Criterion validity</td>
</tr>
<tr>
<td>2. Content validity</td>
</tr>
<tr>
<td>3. Construct validity</td>
</tr>
<tr>
<td>4. Face validity</td>
</tr>
<tr>
<td>5. From different types of validity to a unitary validity concept</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
</tr>
<tr>
<td><strong>Ongoing discussions about validity</strong></td>
</tr>
<tr>
<td>1. A recent proposal for deconstructing the unitary validity model</td>
</tr>
<tr>
<td>2. Fundamental aspects of modern validity theory</td>
</tr>
<tr>
<td><strong>Validity theory as the framework in my studies</strong></td>
</tr>
<tr>
<td>1. The validity concept in licentiate thesis</td>
</tr>
<tr>
<td>2. The validity concept in the study of selections</td>
</tr>
<tr>
<td><strong>RESEARCH ON MEASUREMENTS, JUDGEMENTS AND SELECTIONS</strong></td>
</tr>
<tr>
<td><strong>Personnel selection</strong></td>
</tr>
<tr>
<td><strong>Measurements and judgements in sport</strong></td>
</tr>
<tr>
<td>1. Reliability concerns in measurements and judgements</td>
</tr>
<tr>
<td>2. Validity concerns in measurements and judgements</td>
</tr>
<tr>
<td><strong>Selections in sport</strong></td>
</tr>
<tr>
<td>1. Validity concerns in talent identifications</td>
</tr>
<tr>
<td>2. Performance indicators and other factors influencing selections</td>
</tr>
<tr>
<td>3. Validity and reliability issues in models for assessing performance</td>
</tr>
<tr>
<td>4. Tipsters’ ability to predict performances</td>
</tr>
<tr>
<td>5. Fairness in selections – a validity issue?</td>
</tr>
<tr>
<td><strong>Introduction to my research focus</strong></td>
</tr>
</tbody>
</table>
AIM OF THE STUDY........................................................................................................59

METHODOLOGICAL REFLECTIONS..............................................................61
Selection of data collecting methods.............................................................62
Selection of the sports ..................................................................................63
Selection of the informants .........................................................................64
The research process ...................................................................................67
  The research questions .............................................................................67
  The interviews .........................................................................................69
  The transcriptions of the interviews .........................................................70
  The analysis of the empirical data .............................................................70
Validity in sport processes .........................................................................71
Validity in my research ..............................................................................72

SUMMARIES OF THE EMPIRICAL STUDIES ..............................................77
The licentiate thesis (empirical studies 1 & 2) ...........................................77
Empirical study 3 .........................................................................................81

DISCUSSION AND CONCLUSIONS.............................................................85
Benefits of and drawbacks to well-defined criteria ......................................85
The weight of different criteria and grounds ...............................................90
Interactions between judges and selectors and bias concerns .................92
Education of judges and selectors .............................................................94
Face validity and the practical value of criteria ..........................................96
Implications for practice ...........................................................................99
Concluding remarks ..................................................................................101

REFERENCES................................................................................................105

REPORT: SELECTIONS TO TOP-LEVEL SPORT TEAMS
INTRODUCTION

This thesis includes studies about factors related to the most fundamental ideas of competitive sport, namely the selection of athletes and the determining of winners. In competitive sports and especially in elite level sports, where athletes spend countless hours on training and preparation, the outcome of games and competitions play an enormous role. Even if the differences in time, length or points are marginal, the consequences of being the best or second best are big. To win a gold medal in a championship is about the honour of being the best, but is also connected to a number of other opportunities for athletes, coaches and sport organizations, because successful results are also connected to sponsorships, media value and respect, which in its turn render money and opportunities for further training and coaching, greater interest in the sport and perhaps even an altered life situation for the athlete.

For the results to be meaningful, the competitions must be performed in a fair setting with fair treatment for the ones involved. Unfortunately, events and occasions constantly remind us that sport is not always fair. The vision that all athletes compete fairly under the same conditions is partly obscured, as doping, economic, medical and social differences are threats to the fairness. In order to make the competitions as fair as possible and on the whole to be able to arrange competitions, rules and regulations decided by international governing sports federations define the content of the sports, and what performances and behaviours will be rewarded and what will lead to punishments. The significance of those rules, rule changes and measurement procedures, and how well they are defined and followed, affect the outcome of games and competitions as well as a sport’s development.

After winning the bronze medal in the Olympic Games in Beijing in 2008, the Swedish wrestler Ara Abrahamian took off his medal and placed it on the wrestling mat and left the arena in anger and despair. As a result of this behaviour at the prize giving ceremony, the International Olympic Committee (IOC) later disqualified Abrahamian and stripped him of his medal. The origin of this incident was the semi-final match in which Abrahamian had lost due to what he and the Swedish federation perceived as doubtful judging. After the match the Swedish coach had made a request to the International Wrestling Federation (FILA) for a video check of the situation, a request that was denied. This led the Swedish wrestling federation and Abrahamian to file a petition to the Court of Arbitration for Sport (CAS), which later stated that FILA had not acted in accordance with the ethical rules and that in the future they will have to establish a jury or body of appeal to determine the validity of complaints.
DE C I D I N G  WHO IS THE BEST

(Court of Arbitration for Sport, 2008). This loss bereaved Abrahamian of the opportunity to participate in the match about the gold medal and crushed all his hopes and dreams to conquer the Olympic gold medal. This incident does not only show an athlete’s frustration, but also how the credibility of the sport was questioned when this happened and when media revealed articles about unclear rules, corruption, politics, bribed judges, prearranged matches and threats within the sport.

Another situation where the judges’ decision played a big role and where the sport’s credibility was also questioned and debated was when the Irish soccer team was eliminated from the soccer World Cup play-offs in 2009. The incident that was discussed was the French player Henry Thierry’s assist that led to the victory for France and dismissal of the Irish team from the World Cup. The reasons for the discussions were that the assist was made by the use of a hand, something that is a not allowed in the game of soccer. In finding a responsible scapegoat, the roles of the referee missing this incident, Henry’s role as a cheater playing against the rules, the International Soccer Federation’s (FIFA) responsibility for not having established rules for video assessment and separate goal referees were debated. In the discussions following this incident, the importance of this result led Ireland to ask FIFA’s executive committee to give them an extra spot in the World Cup so they could participate despite this loss or to compensate them through better seeding in next year’s World Cup playoffs. To avoid situations like this, Ireland also proposed goal cameras or goal referees for future games (Macdonald, 2009; Missen kan ge måldomare till VM, 2009) [The miss may lead to goal referees for the World Championship].

In the striving for success, good recruitment and selection strategies are key factors in organizations aiming for successful results. In sport, selections of athletes to teams, competitions and games play an important role both for the outcome of games and competitions but also in regard to the consequences these selections have for the involved parties. A coach who does not select a winning team might lose her/his job; an athlete who is not selected into a national team might never be able to catch up with the training and opportunities of the athletes who were selected. Due to the interest in sport and sport results, the decisions taken in sports at the elite level are of public interest. Media put pressure on individuals and organizations through publicly discussing those measurements and judgements in a way that might not be done in the same way when it comes to selections to higher education, personnel selection and the like.

Selections to top-level teams should also be seen in the light of the total number of top teams and the opportunities to try out for another team if not being selected into the desired one. In the selections to national teams, this is especially interesting, as there is no other team (if not changing nationality) to try out for. As the top-level teams are the goal for many athletes, it is possible
that the actions, values and norms included in selections to top teams also influence the sport as a whole and not only the specific team the selection concerns.

When athletes are selected to teams, competitions and games, many factors have to be considered and numerous cases of selections reveal the complexity of the selection process. Athletes are usually selected by responsible functionaries in the respective federations and clubs, through different ways of assessing and predicting the athletes’ performances. How these selections are carried out, and what selection criteria the selections are based on, are decided within the federations and clubs, which leads to different selection procedures in different sports and also among different federations and clubs within the same sports.

Due to the impact that results and selections have on top-level sport and how the importance of results and selections seem to increase, and due to my personal interest in these questions, measurements, judgements and selection processes in sport are investigated in this thesis. The purpose is to increase the knowledge of these processes and to discuss them from a validity perspective. By using a validity perspective, the relevance, utility, fairness, reliability and usefulness of these procedures are elucidated.
OUTLINE OF THE THESIS

In this thesis, the processes of selections of athletes to top-level sport teams and the process of measuring and judging athletes in competitions are in focus. In two empirical studies, which are included in my licentiate thesis (Johansson, 2001) I studied the rule systems and judgements in the freestyle discipline acroski (Johansson, 1998) and in rhythmic gymnastics (Johansson, 2001). In 2007, I resumed the work on my doctoral studies and conducted a study of the selection processes to top-level sport teams in soccer and alpine skiing (Johansson, 2010). The model below is an illustration to clarify how the different empirical studies and the licentiate thesis are included in this doctoral thesis.

Figure 1. The content of the doctoral thesis clarified through a visual model of the empirical studies and the licentiate thesis included in the doctoral thesis.
Following this presentation there is a chapter about validity theory, which is the theoretical point of departure in my studies. Validity theory has undergone quite a few changes and the development of the concept and my way of understanding and using validity theory is a determinable factor for the structuring and understanding of my studies. In order to introduce the research questions into a bigger perspective and to increase the understanding of the field and my research approach, a presentation of previous research related to measurements, judgements and selections in society are subsequently included and followed by a presentation of the aim of my study.

Next there are some reflections on the methodological considerations and the validity in my research and after that the empirical studies are summarised. First there is a summary of the licentiate thesis including the two empirical reports about judged sports, and then the third report about selections is summarized. In the following and last chapter, the results are discussed and further conclusions about measurements, judgements and selections in sports are drawn and some implications for practice are suggested. Very last the full report on Selections to top-level sport teams is included.
A number of different angles could be of interest when measurements, judgements and selections in elite sports are studied, and depending on the aim and interest of the study, different theories of a sociological, educational, psychological as well as economic nature could be used to analyse and discuss the underlying factors of measurements and selections. Due to the importance of the results and selections in elite sport, the quality and usefulness of these processes are the main interest in my studies. In many studies where the quality and usefulness of a process, for example in studies of selections to higher educations, the usefulness of rehabilitation programs for athletes, selections of personnel to a company, the usefulness of tests in school, the theoretical concepts of validity and the integrated concept of reliability are used as the theoretical framework. Simplified validity is concerned with the relevance and utility in the measurement process and its outcomes, and reliability is concerned with the accuracy and consistency of the procedures and their outcomes. As the purpose of this thesis is to increase the understanding of measurement, judgement and selection processes in sport with a focus on the usefulness, the quality and the fairness of these processes, the validity concept is regarded as suitable and used as the theoretical point of departure.

In this chapter, I will provide a presentation of these concepts and the role they have played in my studies. The validity concept has undergone quite a few changes and there is still an ongoing debate about how to understand and use the validity concept. As this has implications for how I interpret the validity concept, a presentation of this development and the ongoing debate is included in this chapter. The aim of providing this information is to increase the possibilities of understanding my results and of examining my interpretation and incorporation of the theoretical concepts in my research.

The validity concept was introduced in the area of tests and testing practices foremost in the educational and psychological settings, and when the development of the concept is described in the validity literature, it is primarily described in connection with tests, testing practices or sometimes measurements. The validity concept is possible to apply to all kinds of measurement and assessment methods, but when I describe the development of the validity concept here, I have chosen to refer to ‘test’ or sometimes ‘measurements’ in conformity with much of the validity literature. When I further on in this chapter and the
rest of this thesis (including the empirical reports) describe my own research, I will instead refer to judgements or selections, as this is what I have investigated.

The evolution of the concept of validity

For decades it has been agreed that when evaluating the quality of a test, especially educational and psychological tests, validity is the primacy concept to study. In the present Standards for Educational and Psychological Testing from the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA; APA & NCME, 1999) it was stated that validity is “the most fundamental consideration in developing and evaluating tests” (p. 9). The concept of validity has evolved over time due to different ways of using the validity concept, and due to the different researchers’ views on validity. A summary of the concept’s development will show how it has gone from a narrow and evidence-based concept to a much broader concept.

Criterion validity

The validity concept was introduced around 1915, when researchers in psychological testing made attempts to correlate tests with an external criterion to demonstrate a test’s utility. Around the 1930s the validity concept was defined in the literature and then validity was foremost about criterion validity (Kane, 2006). Criterion-related validity referred to how well a test estimates a criterion. This means that in order for the test to have high criterion validity, the test (including one or a set of variables) estimates a certain criterion, which may be a certain performance or certain behaviour. The criterion that the test was supposed to correlate with (the behavioural or performance criterion) was decided and agreed upon by psychologists or other experts in the field. When tests were supposed to measure certain behaviours or performances, criterion validity was useful. Criterion validity was divided into concurrent validity and predictive validity depending on when the criterion was observed. Concurrent validity refers to when the criterion was observed during the time of the test and was primarily used to control if a new test was valid by comparing its outcome with an already validated test. Predictive validity was used when the criterion was to be observed in the future and the test was used to predict a future behaviour or performance. In tests included in selection processes, such as for example the selection of students to college or workers to specific jobs, predictive validity was a way to measure if the test was a valid selection device.

According to Kane (2006), criterion validity should be seen as the main standard for validity from the 1920s to the 1950s, and validity was then described as a correlation between the test score and a “true” criterion score. Criterion validity had its advantages, as it could be considered relevant, understandable
and objective, but it also had its limitations. One of the big limitations of the
criterion model was the difficulty in developing an adequate criterion, as it was
difficult to validate the criterion itself (Kane, 2006). In the 1930s Thorndike
argued that it was difficult to present the relevance of the criterion in relation to
the purpose of the testing, and in the 1940s Jenkins meant that there was a
serious problem with the validity of the criterion itself (Sireci, 1998). The diffi-
culties in developing a valid criterion were also questioned by Toops (1944),
who questioned the time it would take to construct and perfect a test.

Content validity
During this time period the focus was on the test instrument, and the correla-
tion between the test measure and the criterion showed the validity of the test
instrument. This purely statistic perspective of validity was seen by scientists
such as Kelly, Thorndike and Toops as a limitation in validations. They
thought that supplementary evaluations like professional judgement and the
like should be used as a supplement to correlational evidence (Sireci, 1998).
The limitations of and problems with the criterion model led to discussions and
alternative methods for understanding and studying validity. As a way to in-
crease the validity of the criterion that was used, Gulliksen (1950) introduced a
concept he called intrinsic content validity. To increase the intrinsic content
validity of the test, he argued that the criterion should be evaluated by using
many different kinds of systematic measures. If expert judges are used, their
judgements should be refined by assessing the expert judges against each other,
pre- and post-test results of tasks should be assessed and the test should also be
assessed in comparison with other kinds of achievement tests to reach greater
intrinsic content validity. This was a step towards more empirical procedures in
validations, and the concept of content validity started to gain more and more
ground.

Cureton and his contribution about validity and content validity in Educa-
tional Measurement in 1951 is described by Sireci (1998) as an early introduc-
tion of the term "content validity". Content validity had to do with how well the
content of a test correlated with the conclusions that were drawn, i.e. how well
the instrument actually measured what it was supposed to measure. Here it
should be noted that it was samples of performances that were the empirical
basis for the validations. Content validity had to do with the relevance of the
instrument as a measuring tool for a skill. In the content model, interpretation
of test scores based on a sample of performances estimated an overall skill in
that activity (Kane, 2006). When content validity was evaluated, experts in the
field gave their opinions of whether the things measured were the relevant and
essential things to measure. The content model was criticised for being subjec-
tive and having confirmatory bias, as the analyses tended to rely on judgements
about the relevance and representativeness of test tasks (Kane, 2006). There was
some dissatisfaction with content validity, as it was felt that it was insufficient for meeting the needs of all kinds of validations.

**Construct validity**

As criterion and content validity did not cover validations of all kinds, another type of validity, *construct validity* developed as a third type of validity. Cronbach and Meehl, who were greatly involved in the innovation of construct validity, thought that construct validation should be used when there was no definite criterion measure of the quality that was measured and more indirect measures therefore had to be used in the validation (Chronbach & Meehl, 1995). Construct validity was mostly used to measure psychological traits like attitudes, intelligence and other personality characteristics that are not operationally defined. Construct validity refers to the degree that the actual “test” measures, the so-called theoretical construct (for example intelligence) it is supposed to measure. Morrow, Jackson, Allen, Disch, Moos and Dale (2005) describe construct validity as "if the construct is valid, then such and such would occur". According to Chronbach and Meehl (1955) it is required that there is a chain of inferences for a test to measure a construct. This meant that a nomological (lawful) net about the concept should exist, and if this nomological net was not accepted by several users, public validation would not be possible. In connection to their development of construct validity, Cronbach and Meehl (1955) also developed the so-called nomological network. This nomological network was a way to see validation as a logical process following certain guidelines or “laws”. The nomological network is to be seen as a system of laws constituting a theory. The laws included theoretical and empirical frames for validations together with interrelationships between those.

**Face validity**

Face validity may be explained as “perceived validity”. As this is a kind of validity that cannot be “proved” by any empirical data, it is discussed if it should be regarded as part of the validity concept. Still it would be hard to disregard face validity as it may be of importance for the test results, and test interpretations, especially if focusing on consequences of a test. One aspect of test fairness, which was long overlooked, was research about applicants’ reactions to selection procedures. During the last few decades, it seems that more focus has been placed on the more social side of the selection process (the social consequences). How applicants view the selection process, the face validity of the process, has become of more interest both from organizations and from researchers in the field of recruitment and organizational justice, resulting in the expansion of research in this field (Ryan & Ployhart, 2000). How the applicant perceives the selection procedure is studied both to understand possible perceptions of the selection procedures but also in connection with how the selection procedure
affects the applicant’s view of the organization. Gilliland developed a model of applicants’ reactions to selection systems (Gilliland, 1993), and this model has since then been widely cited in studies of applicants’ reactions. The model includes both psychological theory from organizational justice and prior research on reactions to selection procedures and is said by Gilliland to be an extension of organizational justice theory to the selection domain.

From different types of validity to a unitary validity concept

Along with the development of the different types of validity, the American Psychological Association (1952) developed a preliminary proposal for technical recommendations for psychological tests and diagnostic techniques. In this it says that there “are several types of validity depending on the type of inference for which the test is to be used” (p.467). These recommendations also state that the “unqualified term validity” (p. 467) should be avoided and instead the type of validity used should be reported. In this proposal four categories of validity were distinguished: predictive validity, status validity (concurrent validity), content validity and congruent validity (construct validity). In the 1954 Technical recommendations (APA, AERA & NCME, 1954), validity was referred to in terms of how to indicate to what degree a test was capable of achieving certain aims. In these recommendations four “aspects” (not types) of validity were defined and those were content validity, predictive validity, concurrent validity and construct validity.

It has been discussed that this change from referring to validity as different “types” to different “aspects” was not unproblematic, and Guion (1980) thought that many practitioners had difficulties with this change and continued referring to validity as three different things (different types). According to him one example of this problem was that validity in the “Uniform Guidelines on Employee Selection Procedures” was treated as “a holy trinity”, something he did not see as helpful in the process of seeing validity as different aspects. Predictive and concurrent validity were in those guidelines reduced to one category, criterion-related validity, and the other categories were described as content validity and construct validity. According to him the interpretation of treating validity as a trio of concepts led both to confusion and to three ways of studying validity. He thought this increased the risk of evidence being gathered from only one of the aspects instead of from as many sources as possible.

In the 1985 Standards for Educational and Psychological Testing (APA, AERA & NCME) a change is seen in the development of the validity concept. Here the concept of validity refers to the “appropriateness, meaningfulness and usefulness of the specific inferences made from the test scores” (p. 9). It was stated there that validity now was to be treated as a “unitary concept”. Instead of different validity types or aspects, validity was to be regarded as a unitary concept where the evidence for supporting inferences could be accumulated in different
ways. The ideal validation would include several types of evidence. In these Standards the categories of validity evidence were content-related, criterion-related and construct-related evidence. This broadened description of validity is also seen in Messick’s (1989) description of the concept:

*Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other models of assessment (p. 13).*

The development of the validity concept had gone from an early focus on specific criteria to different types and later different aspects of validity. These different types or aspects had then been included as evidence in a unitary concept of validity. There had also been a shift from a focus on predictions to a focus on explanations of the test scores (Messick, 1989). Earlier the validation had been aimed towards the test instrument but now it started to become a shift to validation of the interpretations of the test. Validity had developed to be something that referred to the accuracy of the inferences of the test scores (Tenopyr, 1977). The following statement from Cronbach (1971) can be seen as a milestone in validity literature and describes the shift of focus in the validity concept:

*One validates, not a test, but an interpretation of data arising from a specified procedure* (p. 447).

In this new view construct validity was seen as the “unitary validity concept” and was getting more and more importance over time, as it was argued that it embraced all forms of validity evidence. Here it should be noted that this new form of construct validity should not be confused with the previously described type of validity with the same name, as this new concept included much more than the original concept. From the early 1980s this so-called unitary construct model became the general approach to validity, including content and criterion evidence as well as reliability and other methods of theory testing (Kane, 2006). The nomological network, which had been a core in construct validation, was more and more questioned. It was developed to prove regularities through laws and was strongly influenced by logical positivism in the 1950s. To explain and understand human behaviour from the standpoint that strict laws and regularities govern it was no longer seen by social scientists as sufficient, although it was still seen to be of importance to use some form of “conceptual framework” or “validity framework” in validity studies (Shepard, 1993).

Using construct validity as his base, Messick (1989) developed a validity model called “facets of validity” that became very important in the discussion
about the validity concept. In addition to the traditional test score meaning, he added test relevance, values, and social consequences to be taken into consideration. The model gave possibilities of seeing validity in a different perspective than before, but it was also seen as a rather challenging way of investigating and understanding validity. It focused on two aspects of validity, the first one being the measure itself and the interpretation of the construct and the relevance and utility of the measurement, and the second one being the consequences studied through values about the measurements and the social consequences. The inclusion of the social consequences of a test was something that was new and something that became a topic of discussion (see further on in this text). Messick argued that the inclusion of social consequences was something that was necessary to include as a part of the validity concept:

For a fully unified view of validity, it must also be recognised that the appropriateness, meaningfulness and usefulness of score based inferences depend as well on the social consequences of the testing. Therefore, social values cannot be ignored in considerations of validity (Messick, 1989, p. 19).

As mentioned above, the new and accepted validity concept, the "unitary concept", is called construct validity but construct validity as a more comprehensive concept than the traditional concept. Modern construct validity integrates the traditional concepts of content and criteria together with consequences (Messick, 1994). In the present Standards for Educational and Psychological Testing (AERA et al. 1999), it says, "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p.9). There it is clearly stated that validity is a unitary concept and that the "sources of validity evidence may illustrate different aspects of validity" (p. 11) but that they do not represent different types of validity. The validity evidence mentioned there is evidence based on test content, evidence based on response process, evidence based on internal structure, and evidence based on relations to other variables.

Evidence based on test content has according to the Standards to do with the relationship between the test’s content and what it is intended to measure. This may "include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretations of the test scores" (p.11). According to these Standards this kind of evidence may also come from expert judgements of the relationships between the test and the construct. Evidence based on response processes are in these Standards concerned with the construct and the test taker’s responses to the test and its content. Analyses of test takers’ performance strategies or responses to particular parts of the test are there seen to give information about the test and the definition of the construct. The response processes of
Evidence based on internal structure "can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p.13). Evidence based on relations to other variables is expressed as evidence based on relationships to other tests that are expected to predict the same things as the test concerned; other external variables like performance criteria can also be used. It is stated in the Standards, "a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory supports the intended interpretation of test scores for specific uses" (p.17).

Constant questioning of the tests, procedures, results and interpretations became important and Cronbach, Kane and Messick all advocated argumentation as an important part of validity. Validation should always be seen in its context to be understandable and validation benefits from more than one single empirical study. Specification of what is going to be evaluated and the context it is involved in (the so-called interpretive argument) and evaluation of the proposed interpretations and the way the test scores are used (the so-called validity argument) should always be parts of validation (Kane, 2006). Kane (1992, 2006) argues for the use of interpretive arguments in validation. In his argumentative approach, he brings forward three kinds of criteria to be helpful in the interpretive argumentation. First, he talks about clarity of the argument, meaning that the interpretive argument should be clearly stated. It should be clearly known what is claimed by the argument. This is helpful and necessary to understand the conclusions. Secondly, he argues for coherence of the argument. This argument has to do with how the conclusions follow the specified assumptions. The network of inferences drawn from the observed performances to conclusions has to make sense and be plausible. No essential inferences or assumptions should be left out. The third criterion in this argumentative approach has to do with the plausibility of inferences and assumptions. The arguments used have to be discussed in terms of their plausibility and this can be done using different approaches. He explains how some assumptions can be taken for granted, some are the result of careful documentation and some are received from empirical evidence and analyses of procedures. Here several parallel types of evidence can be used. The most serious problem with the interpretive arguments is according to Kane the risk that there are "hidden assumptions". It is important to identify all the assumptions, as vague assumptions can be discussed but hidden assumptions will not be discussed.

Reliability
When the quality of a test is studied not only validity but also reliability is of importance. Reliability issues are closely connected to validity issues when it
comes to the quality of any kind of test or measurement. Reliability, which earlier was treated as a separate construct, is now seen as an integrated part of the evolved construct validity concept as reliability is to be treated as necessary condition for validity (see for example AERA et al. 1999; Stobart, 2001; Kane, 2006).

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses…. Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees (AERA et al. 1999, p. 17).

In the licentiate thesis I referred to reliability and validity as separate but connected constructs. Due to the point of departure in the evolved construct validity concept, I discuss reliability as one source of evidence for validity in the study of selections and in the final conclusions of this thesis. Hence I refer to the reliability of the measurements as its own concept although integrated in the theoretical validity concept.

The reliability of a test has to do with the accuracy, consistency, stability and precision of the measurement. High reliability of a measurement or judgement means that all the graders/judges come up with the same, or nearly the same scores, or values independent of each other (Morrow, Jackson, Disch, Mood, 2005). This means that the measuring procedures are so precisely developed that the results will be the same irrespective of who measures or judges the procedures (if it is assumed that what is being measured is not something that is supposed to be changing). It is stated that the more consistent the repeated results of a measuring procedure are, the higher is the reliability of the procedure (Carmines & Zeller, 1979). Consistency and repeatability are key words and indicate the importance of the test being reproducible under the same conditions.

When quantitative data is used, reliability is measured by different correlation methods. Examples of those methods are the test-retest method, the equivalence method, the split-half method and the Kuder-Richardson method. In the test-retest method a correlation coefficient is calculated when the same test is used on two occasions and the correlation coefficient measures the stability of the test. In the equivalence method two similar tests that are supposed to measure the same thing are given to the same group of people and the correlation coefficient shows the correlation between the tests, i.e. how well the test
measures the same thing. In the split-half method the test is divided into two parts and a correlation coefficient is calculated between the two halves and in the test. In the Kuder-Richardson method a correlation coefficient is calculated measuring the correlation among different questions in a test (Morrow et al. 2005; Johansson, 2001) se also Burns (2000); Crocker and Algina (1986), Ellis and Fouts (1996), Fhaner (1968) and Fraenkel and Wallen (2000).

When qualitative data is handled, reliability issues are sometimes discussed in terms of credibility, confirmability, consistency or dependability. I have chosen to use the term reliability, as this term is the one used in the Standards of Educational and Psychological Testing for Measurements and Assessments (APA, AERA, NCME, 1952, 1954, 1985, 1999). The term reliability is used there both when data is of a qualitative and a quantitative nature. In studies where the empirical data is gathered from interviews, observations or the like and where the data will not be categorized in measurable ways suited for correlations, other ways are used to discuss and determine reliability. Reliability is then analysed and discussed in terms of the characteristics of the measurement instrument (in my studies the rule and judging systems as used by the judges, the selection criteria as used by the selectors etc.) and knowledge about the measurement process (who is involved, how long is the process etc.). The stability and clarity of the rules, criteria or the like are of major importance, as they will affect the possibilities of making accurate measurements. The knowledge and experience of the ones involved in the process as well as the length of the measurement procedures are of interest. Possible sources of measurement errors should be recognised when the level of reliability is discussed. These may be categorised as errors rooted within the persons involved in the measurement, judgement or assessment (level of experience, personal feelings, the state of mind, etc.) or connected with external sources (external influences and external pressure) (APA, AERA, NCME, 1999).

Ongoing discussions about validity

Messick’s validity model (1989) has been widely accepted but there has also been (and still is) criticism of this way of looking at validity. As argued by Shepard (1993), the multifaceted way of evaluating validity by Messick’s model might never reach as far as to the consequences when it comes to reality. She is not alone in her critique, as both researchers and others were also doubtful about the inclusion of social consequences as a part of validity theory. Mostly the critique had to do with the difficulties of measuring consequences, which made the validity concept be even harder to use than before. Few researchers see the social consequences as something that has no bearing on a test. Consequences are by many researchers regarded as something that should be considered when developing and evaluating tests and as something that affects how
valid or invalid a test is. The discussion is about whether the consequences should be included or not as a part of the validity concept.

Although Shepard (1997) sees the social consequences as being of vital importance, she does not think they should be regarded as facets, aspects or dimensions of validity measurements. She argues that the inclusion of “social consequences” will make it complicated for most researchers to understand and work with the validity concept. A risk of this she believes is that the inclusion of the consequences as an inherent part of the validity concept in the long run might lead to less attention to the intended and unintended consequences of test use, as the theoretical framework will be too complicated to use. Popham (1997) also agrees with the importance of the consequences, but he too also thinks that they should not be included in the concept, as it would only lead to more confusion. Mehrens (1997) agrees with Popham and states that the use of the word validity should be narrowed rather than expanded. He states, “if validity is everything, then validity is nothing” (p. 18) and thinks that the consequences and the valuing of them are political judgements that would not say anything about the accuracy of the inferences concerning if the assessments were good measures of a construct. Reckase (1998) pointed at the difficulties in predicting consequences of a measure before the measure is made.

According to Cizek, Rosenbergh and Koon (2008), evidence related to test consequences is rarely included by those who develop and market tests. They argue that the so-called consequential validity is difficult to incorporate and is seen as a burden for the ones working with the test. They discuss this and go as far as to state that incorporating test consequences into test validity is a “flaw in modern validity theory” (p. 410) and that “consequential validity simply does not exist” (p. 410). They point to the contradictions and conflicting notions in Kane’s work (2006), which in their opinion states that “(a) validity is, by definition concerned with the inferences that are to be made on the basis of a test score and (b) that test consequences are a source of validity evidence” (Cizek et al. pp. 410-411) and discuss whether these statements are compatible. They argue that a test score can have the same validity regardless of the implications and consequences it might have for the practice where the test is implemented. They also discuss the problems of being able to take test consequences into consideration when producing a test, as it is not possible to know the consequences before evaluating them. They see this as a problem with the inclusion of consequences as a part of validity theory as “it contradicts the ethical idea that no test procedure should be used operationally until it has been thoroughly validated”.

There are also other kinds of criticism of Messick’s unitary validity concept. Most of this dissatisfaction seems to be connected with opinions concerning the use of the model as being a too complex model resulting in difficulties when implementing it in practical settings. To clarify this criticism some of these critical views are explained below. Shepard (1993) is not against Messick’s va-
lidity model but believes that the faceted presentation Messick uses in his model complicates the understanding of the concept. She thinks the presentation gives the impression that “values” are distinct from a scientific evaluation of a test score meaning and that the complexity of the model is not helpful when it comes to identifying which validity questions are the essential ones. She also regards the way of locating construct in one cell but also using it in the other cells as confusing. Instead of Messick’s model she proposes a simpler way to understand validity. She thinks the validity evaluations should be organised in response to the question: What does a testing practice claim to do? This would include reasoning for and against the intended aims of the test, and what the test does in the system other than what it claims to do. Her model builds upon Messick’s model but is modified to some degree and focuses more on the intended test use.

Borsboom, Mellenbergh and van Heerden (2004) also criticise the unitary validity concept as being a concept that is difficult to handle in practice. They argue that the concept, which now has "come to treat every important test-related issue as relevant to the validity concept", (p.1061) takes on a somewhat too big a task. By including all these concepts into one they believe the theory will fail both the theoretically oriented psychologist and the practically inclined tester. They do not see the need for a unified validity concept and argue for a realistic causation-based concept of validity as an alternative. Cizek, Rosenbergh and Koon (2008) also see problems with the unitary concept. They believe there is a general misconception that validity is connected to the test itself and that validity is represented as various kinds of validity. They regard this as partly a semantic problem. According to them few professionals use the terms of “sources of evidence” and many use instead “different kinds of validity” as their way to explain and discuss their findings. This results in the misconception that there are different kinds of validity. They believe this is mostly a semantic problem that does not have essential difference but reflects a diffusion problem in the current standards and validity theories. As mentioned earlier they also believe consequential validity to be problematic.

Wolming and Wikström (2010) also show difficulties in the practical use of the modern validity concept in a recent study where the aim was to investigate if practice had followed theory or if there is a gap between validity theory and validity in practice. They went through three Swedish research projects aiming to validate instruments used for selection to higher education. They studied the instruments’ designs, the research questions and outcome in comparison with how validity was described during the time of these projects. Their conclusion is that the practice had followed theory when it came to the research design but that there is a gap between validity theory and validation in practice. They believe the introduction of the unitary validity concept had increased the link between theory and practice and that the integration of validity evidence was a
weak link in the unitary concept. In order to decrease this present “gap” between theory and practice, they see a need for clarification and both theoretical and practical guidance for how validity evidence should be integrated and also a need for prioritisation of validity questions.

A recent proposal for deconstructing the unitary validity model

In an issue of *Educational Researcher* (number 36, 2007), Lissitz and Samuelson (2007a) propose a deconstruction of the unitary model and present a taxonomy of test evaluation procedures. This new taxonomy is divided into two dimensions, the investigative focus (internal and external sources of information) and the perspectives (theoretical or practical perspectives). They move partly away from the unitary concept with a focus on construct validity and on giving content validity a bigger focus. The reason for this deconstruction is according to them the difficulties involved in both using and describing the current construct.

They separate the analyses from test properties and the analysis from the construct measured. In their view internal characteristics of a test should be determined to be the content validity and should not be dependent on external factors. This means that the validity of the test has to do with the characteristics of the test itself rather than being dependent on other tests, nomothetic theory or on the purpose of the testing. In their concept they suggest that the focus should be on the test itself and the process that led to its development. This new approach by Lissitz and Samuelson is to a large extent being rejected by other validity theorists that in the same issue of *Educational Researcher* were asked to comment on this deconstructed validity concept. They believe it is good that the validity concept is brought under discussion, although none of them seem to agree that Lissitz and Samuelson’s deconstructed validity concept is better than the existing one.

Sireci (2007) says that because it has been some time since the unitary concept was agreed upon, it is the right time to reconsider the concept. He is therefore grateful to Lissitz and Samuelson for bringing life to the debate about the validity concept. Sireci is also partly critical of the unitary concept, although he still believes in the current state of the validity theory standards. His critique is about the unitary concept being so difficult to explain to lay audiences and practitioners who perform validations of different tests. He argues that it is especially hard to explain variables that are latent and “constructed” in comparison to explanations of the content domain. At the same time he argues that the current standards are doing a god job of describing validity, and as an example he brings up how the “sources of evidence” (APA, AERA & NCME, 1999) are helpful tools when gathering, analysing and documenting evidence to support the use of a test. Sireci (2007) argues that a theory for the measured construct is a necessity to be able to rule out possible rival hypotheses that may explain test
performance. Nor does he agree with Lissitz and Samuelson when it comes to their way of looking at validity as an inherent property of a test and is critical of their construct mostly seeming to be applicable in educational testing.

According to Kane (2008), the validity concept of Lissitz and Samuelson is a much narrower and more operational definition than the unitary validity concept. Kane agrees that their operational model might make the concept easier to use when making evaluations but disagrees with the concept, as he believes the interpretations would be too narrow and too operational. He also thinks there should be more attempts to justify the use of test scores than this model supports. He believes that abandoning a great extent of the assumptions and terminology that have been used during the last 50 years would also create a lot of confusion.

Embretson (2007) is critical of Lissitz and Samuelson’s opinion that only internal evidence should support test meaning. Although she agrees with their notion that “the role of external evidence in establishing test meaning should be minimised and that internal evidence should be strongly emphasised” (p. 454), she also argues that there is a need for multiple sources of evidence to establish test meaning. She argues that the construct validity concept including theoretical aspects is needed.

Gorin (2007) thinks that Lissitz and Samuelson’s new model is not as radical as they propose it to be and sees it more as an issue of semantics. Nor does she agree with the model concerning the importance of content validity. She argues that constructs exist across all contexts, which content validity does not. The use of content validity as the whole validity is something she believes stifles the recent advances in test design. She argues that validity is much more than score meaning and that the appropriateness of interpretations of uses and test scores has to be taken into consideration. The idea of Lissitz and Samuelson’s new concept was to provide a more useful and clear framework for validation but that purpose is according to her not reached, as she feels both the structure and the process to be unclear in their model. Mislevy (2007) argues that Lissitz and Samuelson’s concept does not emphasise the “constructive nature” as much as he would like it to and is not convinced by this new way of discussing validity. Moss (2007) is also sceptical about the new model. She thinks the model is a step back towards validity as a methodological issue instead of as a “generative understanding of validation as scientific inquiry reflected in the unitary approach” (p. 470).

In an article where Lissitz and Samuelson (2007b) gave responses to the above researchers’ comments, they stated that they wished it were clear they had not tried to propose a concept that was going to replace the unified theory. According to them their intention had been to show that there are quite different elements in what has been termed validity and that many techniques may be applied when validity is studied. Their proposal was an effort to change the
DOING RESEARCH USING VALIDITY THEORY

terminology to a new vocabulary that would make the validity concept “more accessible, understandable, usable and supportable” (p. 484). They felt they had not really succeeded, as their new terminology had been somewhat confusing to the other researchers. They clarified that they still think that a focus on construct definitions is important but that they see the content validity as a very important activity. They think that the test development phase is very critical in the completion of successful tests and argue that content and construct validity are not to be seen as separate from each other. According to them, they had also intended to show that the concept of reliability is not a totally different concept than validity but instead shares many of the same characteristics, especially in the area called utility. Their hope is their article and the comments from researchers in the field are going to serve a further discussion of the validity concept its development and usefulness.

In the APA, AERA and NCME standards from 1999 it is written that the validity argument indicates a need for refining the definition of the construct. This, together with the above and similar kinds of discussions and arguments from researchers in the field of validity indicates that the validity concept will continue to develop and possibly include other aspects in the future.

Fundamental aspects of modern validity theory

Even if the debate about the validity concept still is going on, some main features are regarded as the core of modern validity theory. Before the presentation of the implementation of validity theory in my studies, some fundamental aspects of the modern validity concept are summarised and presented below. As mentioned earlier the unitary construct model has become the general approach to validity, including content and criterion evidence as well as reliability (Kane, 2006). The most explicit change is the broadened understanding of the concept where interpretations of the test use are a key. It is no longer only the test instrument that is evaluated but also the results of the test scores and their relevance in the specific setting. This is a bigger challenge for both the test developers and test evaluators (researchers) to understand the context where the test is used. Validity in its modern version is to be seen as something that changes over time and contexts and not as something that is stable. Unlike before validity is now also to be seen as “a matter of degree” instead of something that is all or nothing (Cronbach & Meehl, 1955; Messick, 1989). This seems to be the common conception, although there still are some researchers that are sceptical of this. Borsboom, Mellenbergh and van Heerden (2004) think for example that validity could be an either-or question and that “if one wants to measure something, then one’s test must be valid for that something – however suboptimal it may be with respect to properties like reliability” (p.1070).

Validity theory in its early form was more associated with educational and psychological testing, but according to Sireci (1998) it is now applied to all
measures in social science. The narrow evidence-based validity concept was to a great extent connected with correlations between the test and the criterion, and validity was proved by these correlations (i.e. the proof is logically valid). With the evolved validity concept this started to change and validity was no longer seen as an exact correlation. Guion (1977) argues that validity is an evaluation and not a fact. He thinks that validity may be week, poor, satisfactory or similar but that it can never be an exact number. He even believes it to be a "mortal, or at least a linguistic sin" (p. 408) to confuse an interpretation of validity with an obtained validity coefficient.

The broadened or should I say the modern way of understanding validity theory provides a lot of possibilities when trying to understand and evaluate measurements, but it also takes on a rather big role when so many aspects should be taken into consideration in the validation process. The integration of validity evidence has for example been seen as a weak link in validation theory, as there is little theoretical and practical guidance and few or no empirical examples of how to integrate validity evidence (Wikström & Wolming, 2010).

These discussions and the unanimity among researchers in the validity field may in one way be perceived as threatening, as there is no exact way to follow when making validity studies. At the same time I can see this as an advantage and a challenge. By studying the development of the validity concept and the argumentation from different validity researchers, I have been forced to think about and broaden my own understanding of the concept. This is helpful for me in the understanding of the totality of the concept. As the concept is used in so many settings, I believe it has to be a concept that is not too rigid but a concept that is possible to function in different settings. To finish this section about modern validity theory, I would like to use Kane’s words from 2006, which I now see as a guideline when working with and understanding validity issues:

*It is not the test that is validated and it is not the test scores that are validated. It is the claims and the decisions based on the test results that are validated* (Kane, 2006, pp. 59-60).

To put the focus on the claims and decisions of the outcome instead of focusing only on the inherent properties of the test gives a lot more useful information and is for me a much more satisfactory way of understanding and studying validity in a measurement. It demands knowledge of all the parts in the measurement process, which makes the validation complex and time-consuming, but it has the potential to give answers that really matter in the practical settings. To learn how to make validations in accord with this unitary concept demands both time and a deep knowledge of the concept, and hopefully clearer research tools can and will be developed.
Validity theory as the framework in my studies

Validity theory is, as discussed earlier, applicable when all kinds of measures are investigated scientifically in social science, and I found it suitable and applicable also for my studies of different kinds of measurements in the field of sport. In the empirical studies, the validity and reliability concepts were used as the theoretical framework for structuring the research process, the data collection and the understanding and analysis of the results. In the study included in my licentiate thesis (Johansson, 2001), content validity was the point of departure, while in the last empirical study of selections I used the unitary construct validity concept as the theoretical framework. As mentioned earlier, I treated validity and reliability as two separate concepts in the licentiate degree, while I in the selection study and in this thesis, treat reliability as integrated in the theoretical validity concept.

In the study included in the licentiate degree, the overarching aim was to analyse validity and reliability issues in the measurements and judgements in the judged sports across ski and rhythmic gymnastics (RG). Content validity was used to describe how well the instrument, in this case the judges performing their judgements in accord with the judging regulations, measure what is supposed to be measured, i.e. the idea of the sport. In that study a main focus was on the instrument and content validity was therefore useful, but I also included discussions about constructs like artistic impression, creativity, and personality and consequences of rule and judging systems and changes in them, although it may be discussed whether these are inherent parts of the content validity concept. Although content validity was helpful for the structure of the study, in hindsight I saw that I could have benefited from using the unitary validity concept in that study too, as I then would have been able, within the concept, to discuss evidence based on test content but also to include other evidence to be able to describe and discuss validity and reliability from a broader standpoint.

Many researchers in the field discuss how the complex and rather complicated meaning and development of the validity concept often results in different types of validity (as for example content or criterion related validity) instead of the unitary construct concept with different types of validity evidence being used, which was what I did in my first study. This is said to be fairly common both by students and researchers of validity and also by practitioners (Gorin, 2007; Lissitz & Samuelson, 2007; Sireci, 2007; Cizek et al. 2008; Wolming & Wikström, 2010). Even though many validity theorists had stopped using the term content validity at the time when my first study was introduced, this ‘abandoning’ of the concept was not something that was universally accepted by all validity researchers (Sireci, 1998).

Today I have a broader understanding of the validity concept, and for the selection study I therefore used the unitary construct validity concept as the
theoretical framework. In this broadened concept different sources of validity evidence serve as the basis for validations of measurement, judgement or assessment. The claims and decisions of the test results are of importance and in the selection study the ‘claims’ refer to what the selection procedure is supposed to produce, which presumably is a selection of the “right” athletes and how these decisions are made and discussed by the selectors. In the unitary validity concept the context should also be considered, and in studies of selections this is an important concern, as the different teams operate in different contexts. In the unitary concept validity is to be discussed in terms of degrees of validity, as there are no absolute measures of validity and that is also suitable when it comes to discussions of selections.

The validity concept in licentiate thesis

In the two reports about measurements and judgements included in my licentiate thesis, the validity and reliability concepts were used when the official rule and judging regulations in acroski and rhythmic gymnastics (RG) were analysed. My interest was to study how well the content of the rule and judging systems corresponded with what people involved in those sports perceived as the idea of the sport. In content validity, experts from a field are used to give their opinions concerning whether what is measured is relevant and essential to measure, and in my studies experts from acroski and RG were interviewed.

The first step was to get an understanding of what the experts from the field perceived as the ‘idea of the sport’ in order to understand what they felt was most important in the sports. Secondly, the aim was to get their view of whether the current rule and judging systems were optimal and valid. If the experts believed that high scores were given to the performances that they thought corresponded to ‘the idea of the sport’, the rule and judging system were discussed in terms of having a high degree of validity. If high scores on the other hand were given to performances they did not think corresponded with the idea of the sport, the validity of the rules and judging systems was discussed in terms of a low degree of validity.

The reliability in judged sports has to do with the judges’ ability to interpret and judge in a reliable way, and the reliability was studied through the experts’ perceptions of the rule and judging systems and the use of them. If different judges independently come up with scores for a performance that are very close or the same as the other judges’, the reliability is to be discussed in terms of high reliability. The level of definition of the rule and judging systems, the knowledge of the rule and judging systems together with factors like the number and distribution of judges were studied here. The differences allowed between the judges’ scores, in relation to the other judges’ scores, whether the scores were to be included in the final score, and how the scores were divided among different parts of the performances were of interest. Other issues that
might influence the reliability of the judges such as external factors (bribes etc.) and internal factors (biased judgements, personal feelings etc.) were also considered.

In addition to how experts from the field perceived the judgements and the rule and judging system rulebooks, judging manuals and meeting protocols were also studied. What changes the rule and judging system had gone through were investigated and the reasons for these changes were analysed and categorised depending on whether they were mostly made to improve the validity or the reliability, or if they were made for some other reasons. My analyses of these sources were the basis for the discussions about the validity and reliability of the measurements and judgements in those sports.

Acroski and RG were studied separately and the results were presented in two separate reports, which later on were included in the licentiate thesis. In one chapter in the licentiate thesis, measurements in sports measured in other ways than by judges’ evaluations were also discussed and problematised from a validity and reliability perspective.

The validity concept in the study of selections

In the report where the selections to top-level sports teams were studied, the purpose was to increase the knowledge of selections of individuals to top-level sport teams and to discuss the selections from a validity perspective. Only a few written documents explaining the selection criteria and selection procedures were available, so the empirical data in this study was mainly the selectors’ perceptions of the selections. Those selectors are coaches for the teams in the study and they are treated as experts in their field (see method chapter). Different kinds of validity evidence were gathered and the empirical source of this evidence is chiefly the selectors’ expertise.

Studying the selections involved looking into the different parts of the selection process. The starting point was to clarify and define what the involved selectors saw as the goals for the selections. The next step was to obtain knowledge of the criteria or guidelines and the experts’ perceptions of them. This was related to sampling of “evidence based on test content”, as this kind of evidence has to do with the relationship between the content of the selection criteria and what these are intended to measure. This was concerned with what abilities and factors the selectors felt were the crucial ones in the selections. There were some difficulties involved in this, as not all selections even at the top level are in accord with clearly defined and explicit criteria. This might be related to many factors, but difficulties in finding satisfying ways of measuring some of the achievements might play a role here.

How these criteria were measured or judged was studied and discussed, and how they were considered in comparison to physical tests, statistics or the like,
DEcIDING WHO IS THE BEST

and how they were weighed against each other was regarded as “validity evidence based on other variables and internal structure”. “Evidence based on response processes” was gathered through the selectors’ views of the athletes’ reactions in the selection process. Who are responsible for the selections and the assessments of the selection criteria were important parts to study in order to be able to discuss the selections from a validity perspective. In team selections, sources for reliability evidence are concerned with the possibilities for consistency among the different selectors’ choices, if there is more than one person responsible for the selections. Therefore, questions about the number of selectors and their relation to one another were included. How well the different selectors’ choices (if there was more than one selector) independently of each other corresponded with the other selectors’ choices was of interest due to reliability concerns. As the correspondence between the selectors’ decisions might be related to the design and knowledge of the selection criteria, this was also considered.

Taken into consideration were also factors like the stability and clarity of the selection criteria, the experiences of the selectors and the length of the selection process. The selectors’ perceptions of difficulties in the selections and factors of influence (like media, agents, feelings, bias etc.) and how they dealt with these factors were discussed. Another factor of interest was the length of the selection process, as longer selection processes possibly could increase the opportunities for avoiding errors depending on situational factors. Factors related to the outcome were also included as well as whether the selectors thought they had selected the “right athlete” and questions concerning the consequences of the selections.

As discussed earlier in this chapter, there are different views of the researchers in the field of whether consequences should or should not be an integral part of the validity concept. In a field where the consequences of many of the measurements and judgements are closely connected to the future of individuals and where these measurements and judgements are often publicly dissected, I consider that the consequences proposed by Messick (1989) should be seen as an integral part of the validity concept. Therefore I aimed also to elucidate the consequences of selections in the discussions about the validity of the selections. I also agree with Moss (1998) in her statement that it should be an obligation to consider how testing procedures of different kinds are incorporated in the particular contexts in which they are implemented. In my studies the applicants’ (athletes’) reactions are in accord with the unitary validity concept studied through the experts’ evaluation of the athletes’ reactions.
Below the research process for the third study is illustrated to give a visible picture of the different steps in the process of gathering empirical data about selections to top-level sport teams.

1. Identification of the idea of the selections. What are the purposes/goals of the selection?

2. Identification of the existing criteria or guidelines used in the selection. What abilities are seen as important in the selections? What information are these criteria supposed to give?

3. Identification of the selection process. How is the selection process performed? Who are involved? Who has knowledge about the criteria?

4. Understanding of the outcome and the social consequences of the outcome.

5. Discussion of the validity (including the reliability) of the selections based on knowledge of the above and values and context of importance for the selections.

Figure 2. The steps in the research process in the selection study
Measurements, judgements and selection processes occur in all areas in society and depending on the type, the level and the implications of those decisions the demands on high validity in these processes are more or less prominent. In Western society we are daily involved in judgements, assessments and selections of some kind. We decide what clothes to put on in the morning, what to eat for dinner and similar relatively small things and we also make more vital decisions regarding things like who we are going to share our lives with, where we are going to live, who we shall vote for in the government, what kind of job we should apply for and similar things. In society at large people are assessed and judged in many different ways and with different aims. Depending on the level of importance and how much people are involved in measurements or selections, validity and reliability issues get more or less attention. While some of these decision procedures take a lot of time and energy, others are probably only vaguely registered in our minds.

Opportunities to have many choices and take individual decisions have been treated as an advantage and a benefit in Western society, as it means that people can decide over their own lives. Freedom of choice has been seen as an opportunity, but when assessments and selections get more and more involved in every step in our everyday lives, this freedom of choice is no longer seen with only positive eyes (Schwartz, 2005). Opportunities of having choices may and in many cases also do put a lot of stress on people, as selecting one thing also means choosing not to do another thing or choosing something else and persuading ourselves why the one choice is better than the other. There are quite a few theories created to explain our selection choices. One of these is the prospect theory, a psychological theory developed by Kahneman and Tversky (1979), which is a so-called descriptive theory where real life choices are exemplified. The theory describes people’s decisions between different alternatives where the alternative decisions involve some risk. Economic theories like the game theory, decision theory, an expected utility theorem or expected utility hypothesis are other examples of describing and explaining choices. The economic theories predict uncertain outcomes through mathematical formulas. In ecology the r/K selection theory explains the selection of traits. The Selection Theory Bibliography (STB) or as the original theory "evolutionary epistemology" by Cziko, Gary & Campbell (1990) was called, is yet another selection theory.
When students are enrolled at a University or other higher educational institutions, they are usually subject to a selection process. In Sweden the selection criteria are developed by Högskoleverket (HSV) [‘the National Agency for Higher Education’] a state authority and to some degree by the universities themselves. The main goal of the selection process is here to select the students that have the best possibilities to pass the education. In order to do this the selection criteria are connected to the level of the students’ previous knowledge in certain fields. If the students’ previous knowledge varies too much, this might have effects on the quality of the education. Varying previous knowledge is a challenge for the university and the teachers and other students, as it might demand different kinds of teaching and examination methods, very well functioning educational and strategic leadership at the institutional level, better supporting functions for the students as well as well functioning quality work on both the national level and within the universities (Högskoleverket, 2009).

A number of studies have been performed in the area of selections to higher education. In the area of education there are also a number of studies concerned with assessment and measurement of students’ performances.

In the process of selecting persons to a job, a working group or similar body, different methods and criteria are often involved. Examples of selection tools in personnel assessment in the work-related domain are for example cognitive ability tests, both structured and unstructured interviews, personal inventories, biographical information, work samples, personal references, peer rating and even graphology analyses. As selections to sport teams have a lot in common with personnel selection, a little review of the literature in this area was made.

**Personnel selection**

In personnel selection, predictive validity is a term that is often used. Predictive validity there relates to the ability a method has to predict future job performance, and in a number of studies different methods for selecting personnel are investigated. A conclusion from this research is that when hiring personnel without previous work experience of the job, general mental abilities (GMA) are seen as the strongest predictor of future performance and learning (Hunter & Hunter, 1984; Ree & Earles 1992). GMA refers here to intelligence or general cognitive ability. Research studies estimating the abilities of assessment and selection methods in personnel selection have been conducted since the 1920s. In an overview of 85 years of research on personnel assessment methods Schmidt and Hunter (1998) examined this research. Their results show that different methods have very different validity when it comes to predicting future job performance. The already mentioned GMA factor as well as work samples (for workers who have experience of the job) has for example high validity, whereas interests and amount of education have low validity and graphology no
Some combinations of predictors also prove to be valuable. Here the GMA combined with integrity tests and the combination of GMA tests and a structured interview stand out as the most valid predictors. The last two procedures can be used both for applicants that have no previous experience of the job and experienced applicants. In their article, Schmidt and Hunter also discuss the validity of the methods in relation to their costs. One method they label as costly but well worth its price, especially for higher-level jobs, is connected to the behavioural consistency theory. When using this theory, previous training and experience are taken into consideration, as the theory is built upon the principle that past performances are the best predictors of future performances. According to Robertson and Smith (2001), a problematic area in personnel selections is job analysis. Jobs are getting more and more difficult to describe due to the complexity involved in them. This results in both researchers and employers having to be concerned with cross-functional skills of workers instead of looking at static aspects of jobs. They also describe the increased confidence researches have gained in personnel selection methods as a significant change in personnel selection research.

How applicants view the selection process has become of more interest both from organizations and from researchers in the field of recruitment and organizational justice resulting in the expansion of research in this field (Ryan & Ployhart, 2000). How the applicants perceive the selection procedure is studied both to understand possible perceptions of the selection procedures and in connection to how the selection procedures affect the applicants’ view of the organization. Gilliland (1993) states that it is likely that applicants’ experiences during the selection process have an impact on the hired individuals and that this might also affect the climate in the organization. To be able to understand applicants’ reactions to selection processes, Gilliland’s model from 1993 has been of importance to the field and is widely cited in further studies of applicants’ reactions. The model is created using organizational justice theory as its basis. The model should be seen as an extension of organizational justice theory to the selection domain according to Gilliland himself, and it is a model that includes both psychological theory from organizational justice and earlier research on reactions to selection procedures (Gilliland, 1993).

In organizational justice theory the two main components are distributive justice and procedural justice. Distributive justice refers to the probabilities of the result of the distribution of the outcomes regarding perspectives of fairness, equity, equality and needs. Procedural justice has to do with the perceived fairness of the selection and the selection outcome on the basis of the selection procedure. How the process is perceived has to do with factors like how job related the tests are, the individuals’ opportunity to influence the procedure, possibilities of bias, treatment of the individuals, etc. The model tries to provide a framework for how future research may be conducted through an organiza-
tion of previous research. Special focus is on the 10 procedural rules that are shown to be of importance in the applicants’ perceptions of the selection. These ten rules were according to Gilliland (1993) initiated by Leventhal in the 1980s but also influenced by researchers as Sheppard & Lewicki, Greenberg, Bies and Moag, Tyler and Bies. These ten procedural rules are grouped in three main categories: Formal characteristics (job relatedness, opportunity to perform, reconsideration opportunity and consistency) and explanation/feedback (selection information, honesty) and finally interpersonal treatment (interpersonal effectiveness, two-way communication, property of questions). All the areas mentioned above are of importance to the applicants’ reactions to a selection.

Although it seems as if more evidence needs to be gathered to be able to give a fair and true picture of factors of importance for the applicants’ perceptions of the selection process, Ryan and Ployhart (2000), after reviewing the research literature on applicants’ perceptions of selection procedures from 1985 to 1999, state what they call five “good ideas” to help both the organization and the individuals in the selection process. These “good ideas” are the following 1. Recognize that test taking attitudes influence performance in the selection process. 2. Provide explanations that give information and are delivered in an interpersonally sensitive manner. 3. Regularly monitor applicant perceptions. 4. Assess perception-behaviour links and 5. Recognise that selection involves evaluation.

Hausknecht, Day and Thomas (2004) mention five reasons why it is important to study and understand applicants’ reactions to selection systems. They think it is important because if applicants find part of the selections invasive, they may view the company as less attractive in the job search process, which might also influence how they advise others not to seek employment with the organization. They also discuss that the applicant might be less likely to accept an offer from a company with a selection process perceived to be unfavourable and that applicants who think a selection is inappropriate may also be more likely to bring suit than applicants who perceive the process as fair. Finally they discuss that applicants may be less likely to reapply to an organization or even buy the company’s products if they feel mistreated during the selection process. Much of these results from personnel selection and applicants’ reactions can be translated into the field of sport.

Measurements and judgements in sport
Rule systems and the measurements and judgements based on them are the core of competitive sport, as these rule systems define what the sports are about including how they should be measured and judged. When rule and judging systems have been studied scientifically, the main focus seem to have been directed towards reliability issues (see for example Steensaaen, 1969; Wirdheim, 1986;
Grimwall, 1987; Bassett, 1994; Garciano, Palacios-Huerta, Prendergast, 2005; Zitzewitz, 2006; Emerson, Seltzer & Lin, 2009 etc.). Articles related to reliability concerns in different kinds of measurement methods like the reliability of testing instruments and methods for rehabilitation training are also prevalent although of minor interest for my studies. More seldom the validity question seems to be in focus.

Lately more and more research articles focusing on rule changes in different sports have been presented (see for example Miah, 2000; O’Donoghue & Williams 2004; Wu & Yang, 2004; Williams, Hughes, O’Donoghue & Davies, 2005; Williams, Hughes, O’Donoghue & Davies, 2007; Harding, Toohey, Martin, Mackintosh, Lindh & James, 2007; Harding, Toohey, Martin, Hahn & James, 2008). Many of these changes are made with the aim of improving the reliability of the measurements and judgements.

Reliability concerns in measurements and judgements

Judging systems especially in judged sports have been discussed and researched and in much of the research the focus seems to be mostly on reliability issues within the judging and judging systems. Bassett and Perskey (1994) point to how figure skating, unlike many other judged sports, has adopted a judging system based on median ranks, which the authors show is a system that uniquely captures an important meaning of majority rule and at the same time provides for manipulation by a minority of judges. Still this system later on allowed for a judging scandal that involved rigged voting and so-called cronyism at the pairs competition in the 2002 Olympic Games, and this judging system was changed and a new system was introduced in the 2004 season (Park, 2006). This scandal addressed the vulnerability of the previous nine-judge placement rule and enforced a new judging system. According to Wu and Yang (2004), this new system is not better than the previous one when it comes to counteracting manipulation of scores, even if a stricter code of ethics for the judges and changes of the decision rule have been implemented. In their article Wu and Yang construct an alternative rule that they find to be more robust against manipulation than the now existing rules.

In other sports judging issues have also been studied with the aim of improving the judging. In recent studies the subjective performance assessment in the snowboarding discipline half-pipe was discussed and tested against an assessment system for automated and objective information of athletes’ performances (Harding, Toohey, Martin, Mackintosh, Lindh & James, 2007; Harding, Toohey, Martin, Hahn & James, 2008). The judging in the snowboarding discipline half-pipe is currently assessed with a subjective measure termed ‘overall impression’, and the automated system tested was capable of using technology and signal processing to calculate objective information on air-time and degree of rotation during snowboarding. This new system was initially devel-
oped as a tool to allow objective assessment of training, but was in these articles discussed as an assessment device in competitions. With the understanding of how new technology can effect change beyond the original purpose and possibly also generate unintended consequences, this automated assessment system was tried and evaluated through the perceptions of key members of the elite half-pipe snowboard community. The results indicated that the informants had a negative perception of the automated judging system, unless the system is integrated with the current subjective judging. If it were integrated with the subjective system and if the integration of technology were made without removing opportunities to allow for athletic freedom of expression, it was perceived more positively. Elite level judges expressed they were in favour of trying the system as a judging aid, but were in strong opposition to using it to determine competition scores or judging criteria (Harding et al. 2008). These results were said to have to do with freedom of expression and athletic individuality being key features in snowboarding and something that can only be judged subjectively. Another conclusion from these studies was that there were strong opinions that further development and integration of new assessment systems should be made in close cooperation with people in the snowboarding community and that they should be controlled from within the sport.

Another reliability issue in judged sports is whether the judges’ scores are biased towards athletes of the same nationality. Zitzewitz (2006) studied nationalism in winter sports and his results showed that the ski jumping and figure skating judges score their compatriots about 0.13 standard deviations higher than other judges (mogul skiing, aerials and snowboarding half-pipe). At the Olympic Games the ski jumping judges were shown to compensate for the bias of a particularly nationalistic judge, which caused the results to be fairer than they would have been without this compensation. In figure skating the judges instead showed that they reinforced each other’s biases and appeared to participate in bloc judging or vote trading. For a ski jumper this means that it is a disadvantage to have a judge of the same nationality judging the Olympic event. For a skater not to have a judge from her/his home country judging the competition this is a serious disadvantage. Reliability issues in how the judges for the Olympics were selected were also discussed in this article. In ski jumping the selections are performed by a subcommittee from the International Ski Federation with the aim to select the judges who were less biased than average in pre-Olympic competitions. In figure skating the national federations select the judges they want to represent them in the Olympic Games, and Zitzewitz gives examples of how the most nationally biased judges are chosen. As a result of this, strategic judging occurs prior the game, as judging the Olympics is a valued performance for judges. The article also contains a discussion of the truncation of extreme scores and how this can make bloc judging and vote trading easier to implement.
In a study of judges’ scores in rhythmic gymnastics during the 2000 Summer Olympics Popovic (2000) found that judges had a tendency to favour gymnasts from their home country, and in a study examining the judges’ scores in the diving competitions in the Olympic Games in 2000, there was also strong evidence of nationalistic favouritism being discovered (Emerson, Seltzer & Lin, 2009). In that article the authors’ even argue that the medal standings could have been changed if more unbiased judging had occurred.

The interest in the questions about nationalism and biased judging points to an interest in making the judgements more reliable and fairer for the athletes. It also reveals difficulties in assessing whether the scores really are biased or not, as it is hard to evaluate this because there is no objective measure of performance quality to compare with (Zitzewitz, 2006). The judges might come from countries with better athletes, which means that those athletes often deserve higher scores, and that the judges’ style of preference (which very well may be the style of their home nations) rather than an intended higher score could explain the “bias” (Zitzewitz, 2006; Emerson, Seltzer & Lin, 2009). The average of the judges’ scores for all athletes in relation to the other judges’ averages scores also plays a role, as they might be consistently lower or higher than the other judges’ scores, which then points to a consistent bias, which should not be confused with intended high or low scores for an athlete from the same nation as the judge (Emerson, Seltzer & Lin, 2009).

In soccer too, where the result is a combination of the number of goals made and the referee’s (judge’s) judgements, the question of reliability of the judgements was highlighted in a study by Garicano, Palacios-Huerta & Prendergast (2005), who investigated whether soccer referees internalise the crowd’s preferences in their judging by systematically favouring the home team. The results showed that if the home team was ahead, the referee favoured the home team by shortening the game and that s/he lengthened the game when the home team was behind. This could be done, as the referee decides how long the extra time added to the game as a compensation for irregular breaks will be. They also pointed out that the referees change their bias when the rewards for winning games increase. Crowd pressure also showed to affect the behaviours of referees. The authors also argue that this is unlikely to be the only form of bias in the judging, as subjective interpretation favouring one team may also occur by favouring a team with fouls, offsides, penalties and the like but that this form is the only form they can verify.

In an article by Plessner and Haar (2005), judgements of sport performance are discussed from a social cognitive perspective. From this perspective it is central how social information is perceived, encoded, transferred to or recalled from memory, and what processes are involved when people make judgements, attributions and decisions. They present an overview of empirical research that is focused on biases in judgements of sporting performances. The study con-
contains numerous examples of how the perception, categorisation, memory processes and information integration may be affected in different ways leading to biased decisions. Plessner and Haar also give examples of how identification of different cognitive processes at different stages can help to improve the decision making in sports.

Validity concerns in measurements and judgements

Validity issues related to sport measurements and judging do not seem to have been studied to the same extent, at least not with a focus on judging issues, measurements and judgements. In the above studies of the new automated system for snowboard judging (Harding et al. 2007; Harding, et al. 2008), validity issues were discussed, as the snowboarding community were afraid that this new system could interfere with the important values of athletic freedom of expression. Miah (2000) also discusses validity-related issues concerning the trials of different ball types in tennis and whether the changes in the game that this will lead to are in the interest of all concerned and how these changes have to do with the future of the game and the credibility in future tennis.

In a study by Lee (2008), the judging system in figure skating is investigated with reliability questions in mind but with a focus on validity issues. In the article the incentive structures for judges in figure skating are discussed and how these structures influence the judges’ scoring. It was found that the judges manipulate their scores to achieve a targeted level of agreement with the other judges due to the system that provides external and unintended incentives for the judges. These results were then discussed in relation to the validity in the judging system.

Selections in sport

In the search for literature about how to predict future performance and how to select athletes to teams, much of the literature found was related to talent identification and talent programs. There are some similarities between talent identification and selections to top-level teams, but there are also some differences that should be recognised. In the search for talented athletes the main purpose is to identify young athletes that have a future potential to develop and reach successful performances. In the selections of athletes to top-level sport teams and competitions, the athletes who are going to be selected might be older but also more ready to perform top results right away. This is of course not the whole truth. Some athletes (that might very well be young) are selected to a top-level team to develop within the team and to be ready to perform well in some years. When talent identification programs are made at earlier stages in an athlete’s career, questions about the interest of the child in comparison to the interest of the parent and the coach, financial opportunities, and how mature
the child is are questions to consider. In selections to national teams the athletes have often been in the sport for quite some time and have an understanding of what the sport is all about and what efforts and sacrifices have to be made to reach the top. It is of importance to point out that the discussions about previous performances as a tool for measuring future success at the elite level should not be confused with the performances that athletes made at a young age. Performing well at younger ages is not such a good indicator of future success as developmental capacity and developmental opportunities seem more important than the actual results at younger ages (Abbott & Collins, 2004). Even if there are quite a lot of differences between the selection process in a top-level team and talent identification and talent development for younger athletes, the latter should not be ignored when studying the selection process within sports. Similarities between talent identification and team selection might play a role in the selection process and the determination of the selection criteria.

Validity concerns in talent identifications

Talent identification programs play a big role in the search for young promising athletes. Still, a great deal of the research literature concerning talent identification is concerned with discussions about difficulties in being able to predict performances. In the 1960s and 1970s the research was trying to find stable characteristics that could differentiate athletes. No innate or genetic profile could be identified and the current research approach is orientated more towards studying individual interactive processes (Abbott & Collins, 2004; Serpa, 2008). Many different factors have to be considered when talent identification is made, and it seems hard to find any unique characteristic that determines future success (Williams & Reilly, 2000; Abbott & Collins 2004). Still, factors related to anthropometric (refers to the study of the history of height), physiological and skill attributes for example seem often to be the factors of major importance in different talent identification programs (Hoare & Warr, 2000; Reilly, Bangsbo & Franks, 2000; Philippaerts, Vaeyens, Janssens, Van Renterghem, Matthys, Craen, Bourgois, Vrijens, Beunen & Malina, 2006; Gabbett, Georgieff, Domrow, 2007; Sherar, Baxter-Jones, Faulkner & Russell, 2007). Because of the importance these factors are given and differences in young athletes’ maturation levels, this often favours athletes born early in the respective sport’s selection year. Birth date therefore seems to have a significant role in talent identifications (Sherar et al. 2007). This has become so evident that in research it is called the “relative age effect” (season of birth bias).

The research on talent identification has for a long time been unidimensional instead of multidimensional, even though most practitioners and researchers believe sporting feats are the results of a combination of many factors such as physiological, physical, psychological, behavioural and social factors. Multidimensional and dynamic talent identification with a focus on develop-
mental aspects such as mental behaviours and transferable skills should therefore be seen as more important than just the early identification of talents. Later research also seems to focus more on multidimensional studies and talent is also seen as a more dynamic concept than before (Abbott & Collins, 2004). According to Morris (2000) and Abbott and Collins (2004), many talent identifications and development models have little emphasis on psychological factors.

The time period of tryouts for talent identification programs is to be considered. If the selections are made only on one or a few occasions, there is a risk that the athlete does not have enough opportunities for showing his or her true talents well enough. In some research a time period for being able to identify talent and make good assessments is suggested to be around two to three months (Hoare & Warr, 2000), while in some talent programs a two-day tryout is judged as fair assessment time (Mutch, 2001). To be able to develop and reach full potential, it is important to have opportunities for development. Abbott and Collins (2004) think that it might be as important to identify over time what factors limit talent development as it is to identify the talents.

There is an interesting discussion in the research on talent versus deliberate practice. The 10-year rule (Ericsson & Lehmann, 1996) indicates that someone who is trying to perform at world-class level in any field must endure and take part in deliberate practice in the field for a period of at least 10 years. For practice to be defined as deliberate practice some conditions must be met. It has to be practice of a well-defined task on an appropriate level for the individual. Informative feedback should be available and there should also be opportunities for repetition and correction of errors. Howe (1999) thinks that the person that we perceive as a genius is the result of “doggedness, persistence, the capacity for fierce and sustained concentration, as well as intense curiosity” (p. 205). He has studied a number of geniuses and believes that many geniuses disclaim that they have superior intelligence but maintain that they possess abilities to work hard and be curious. According to this kind of research, everyone who is in a good environment and gets sufficient training for a long time should then be able to be successful in what they have been practicing.

In the Eastern European countries, systematic databases of personal and performance variables together with formal monitoring of progress and development were used in the searches for talent (Kluka, 2008). These systems were used mainly for individual sports. According to Kluka (2008), the most systematic models for talent identification in sports were probably the models in the former East Germany. In these programs, the young athletes who showed talent in their sport also needed to be healthy and have no medical anomalies, have psychological and physiological capabilities for hard training but also maintain good academic achievement. For the 2000, Olympic Games in Sydney, Australia adopted some elements from that talent identification approach when they implemented a talent search program. This talent identification and the talent
development programs were mostly targeted at individual sports such as rowing, cycling, swimming and track running. Some of those models for individual talent identification programs were experimentally applied to work for talent identification and development in the team sport soccer but the result of this did not manage to prove this talent identification to be the perfect way of selecting soccer players (Kluka, 2008). In a study by Hoare and Warr (2000), a model to identify and develop potentially talented female soccer players in Australia was tested. The program included a selection of 17 girls aged 15-19, who took part in a 12-month talent development programme. The programme was successful in that it demonstrated that anthropometric, physical and skill attributes made it possible to select potentially successful female soccer players. Some of the goals of this program were reached but the final goal that some of the athletes should make the national team was never reached. According to the researchers the program was seen as useful but should not be seen as the only way to select players. According to them it should rather be seen as an additional way to support existing programs and procedures (Hoare & Warr, 2000).

Performance indicators and other factors influencing selections

When athletes’ performances are valued in a selection process, different factors play a role for the outcome of these evaluations. What these critical factors are, and how they will be identified, is as already discussed in relation to talent identification, partly problematic and there are validity issues concerned with these processes. Performance indicators are action variables that aim to define some or all aspects of performance and can according to Hughes and Bartlett (2002) be categorized in either scoring indicators or indicators of the quality of a performance. These kinds of indicators can for example be of technical, tactical and biomechanical form. Hughes and Bartlett argue that for the possibilities of efficient interpretation of data of sampled performance indicators there has to be comparative data from previous performances or aggregated data of a peer groups or individuals competing at a similar standard. Without these comparisons the performance indicators are of little value as every sport performance is relative both in conjunction to previous performances and to the opposition. The use of non-dimensional ratios is also discussed as they are expressed as being independent although they are only implicitly independent of other variables. As sport performances are complex, especially in team sport where the interaction between players plays a role for the performance, the use of non-dimensional indicators might be very misleading. Hughes and Bartlett think a combined research approach is needed to develop the understanding and knowledge of performance indicators.

When individuals are selected in sports one common way to do this seems to be through use of current performances and current results to predict future performances. In both Boulier and Stekler (1999) and in Trewin, Hopkins and
Pyne (2004), the results show that rankings are good predictors of outcome in performance. In Trewin et al. the authors investigated the common belief of leading swimming coaches “that swimmers need to be ranked in the top ten in the world to be able to have a realistic chance of an Olympic or World Championship medal”. When investigating the relationship between world ranking and Olympic performance in swimming their results supported the coaches’ beliefs showing that 87% of the Olympic medallists in the 2002 Olympics had a top 10 world ranking prior to the Olympic Games.

In individual sports the results are visible and speak for themselves even if the results are not always a fair predictor of who is the best athlete (Johansson, 2001). In team sports it might be complicated to find good measurable variables to use if looking at previous results as a factor to predict future performance. In team sports the results show which team is the best but not which players in the team made the best performance. If previous results are to be a useful tool for predicting performance, it is of utmost importance that the right things are measured and looked at in the selection process. In team sports there are often a lot of statistics gathered during a game, statistics that could be used as predictive tools if they are valid for the performance expected in the future, in the sense that they measure what is important behaviour for different categories of players. As an example, it is important to score goals for a player in an offensive position in a team and statistics concerning the number of scored goals and assists etc. could therefore be useful statistics for that position. At the same time this kind of statistics would be a poor predictor of success for a goalkeeper. For some categories of players there might be more useful measurable statistics recorded than for other players where more subjectively measured variables might be a better predictor of future success. As a team player’s performance is dependent on the other players’ performance, the validity of statistical measurements has to be considered.

In a study of 740 players in the National Hockey League, Voyer and Wright (1998) attempted to determine what factors predicted performance in the National Hockey League. The studied variables mostly focused on previous playing experiences such as goals scored per game in a regular season junior career; assists per regular season game in junior hockey; points scored per regular season game in junior hockey; goals scored per game in junior playoff games, etc. but also variables such as entry draft rank; current height; and current weight. Two dependent variables were then used to estimate the NHL performance; points scored per regular season game in career, and points per playoff game in career. According to their research the most important variable for predicting performance both in NHL regular seasons and in playoffs is the points scored per game by players in the regular season junior hockey and in junior playoff games. Weight did not appear to be exceptionally important and entry draft rank did not at all seem to be an important factor in predicting performance in
the NHL. These writers also point out the difficulties in measuring mostly factors of offensive ability as a predictor for performance, as defensive-oriented players like defensive forwards, defence men and goalkeepers will not fairly benefit from looking only at offensive statistics. They suggest that variables like "finished checks", "hits", "take-aways", and "plus/minus" should maybe be looked at to create a fairer and more valid way to predict performance for all kinds of players in a team. Besides these problems numerous statistics need to be accumulated over time to give the needed information (Nadeau et al. 2007).

From the 1950s to the 1970s studies to identify personality characteristics of successful athletes were carried through but no personality profiles could be established. Rather it seems that the use of certain psychological skills like for instance goal setting were successful determinants of performance. While physical characteristics have been able to explain differences between athletes, only psychological factors seem to be able to explain the maintained success (Abbott & Collins, 2004). Mental and psycho-behavioural skills have in a number of studies proved to play an important role in high level performance and dealing with stressors related to elite sports (see for example Orlick & Partington, 1988; Gould, Guinan, Greenleaf, Medbury & Peterson, 1999; Abbott & Collins, 2002; Durand-Bush & Salmela, 2002; Gould, Dieffenbach & Moffett, 2002; Pensgaard & Duda, 2002; Abbott & Collins, 2004; Baker & Horton, 2004; Van Yperen, 2009).

In addition to mental and psycho-behavioural skills influencing the athletes’ performances, physical, environmental and social factors are also argued to be of importance for the performances (see for example Gould et al. 1999; Durand-Bush et al. 2002). In a study where the purpose was to investigate the perceptions athletes and coaches had of physical, psychological, environmental and social factors influencing peak performance at the Olympics, Gould et al. (1999) interviewed athletes and coaches from eight Atlanta US Olympic teams. Four of the teams had met or exceeded the expectations and four of the teams failed to perform according to the predictions. It showed that there were differences between the ‘successful’ and the ‘unsuccessful’ teams. The teams that reached or exceeded the expectations had participated in resident training programs, experienced crowd and family or friend support, utilized mental preparation, and were also highly focused and committed. This in contrast to the teams who had not reached the expectations and showed team cohesion problems, lacked experience, faced travel problems, experienced coaching problems and problems related to focus and commitment. The conclusion of this study pointed to a number of factors influencing peak performances, factors including mental skills but also factors related to practical, social, tactical and physical aspects.

When the psychological characteristics and their development in Olympic champions from USA were studied (Gould et al. 2002), the results showed that
the athletes were characterised by ability to cope with and control anxiety; confidence; mental toughness/resiliency; sport intelligence, ability to focus and block out distractions; competitiveness; hard work-ethics; ability to set and achieve goals; coachability; high levels of dispositional hope; optimism; and adaptive perfectionism. It was also pointed out that a number of individuals and institutions played a big role for the athletes’ psychological development and the influences from the coach and the family seemed to be particularly important.

In a study aimed to identify psychological factors that predict success in professional adult soccer, initial performance level, goal importance and goal commitment, potential stressors, coping, seeking social support, number of siblings, ethnic origin and parental divorce were investigated (Van Yperen, 2009). The predictor variables were measured in the initial phase of the soccer players and career success was assessed 15 years later. Success was defined as “actually playing for a premier soccer team in a European competition for the last ten years in the 15 year period following data collection” (p. 317). The results showed that the psychological factors that predicted career success were goal commitment, engagement in problem-focused coping behaviour and social support seeking. The successful players also had more siblings and were more often of non-white (non-Dutch) ethnic origin. They also had divorced parents more often.

One psychological factor that in the last decade seems to have attracted increasing interest in studies concerning performance is the research on perfectionism and its effects on sport performances. There is a discussion among researchers of whether striving for perfection will enhance or undermine performance (Bieling, Israeli & Antony, 2004; Stoll, Lau & Stoebber 2008). Perfectionism may have different forms and affect the athletes very differently. Perfectionism may be very positive for the athlete who is able to use it to take the performances a step further. If the perfectionism is instead connected to stress and fear of not succeeding, it may instead be very negative for the athlete. How perfectionism is dealt with is also discussed in relation to burnout in sports (Lemyre, Hall, & Roberts 2008). Serpa and Barrerios (2008) studied perfectionism in young Portuguese elite soccer players with young talents and non-talents in two age groups. They concluded that the talented players did have a higher adjusted and lower negative perfectionism than the non-talented players.

According to Morris (2000), psychological factors are valued in the practical field of soccer, but in talent identifications in soccer this part is still in its infancy. Coaches and administrators consider psychological issues but they do not use any objective or systematic subjective ways to measure or assess those factors. Humara (2000) discusses the coaches as experts on identifying the physical characteristics needed for success, but thinks that they do not have sufficient psychological skills to validate the psychological factors that are also of importance in the selections.
There are also other factors than the above kinds of performance indicators that are of importance in the selection process and thus will affect the selections. In top-level sports it is important to achieve the best results at specific events such as playoffs, finals, World Championships, Olympic Games and other major competitions. To have the best team ready for those main events might affect the way teams are selected. There is a need for balance between reaching the short-term and long-term goals, between winning games and competitions and development of athletes. To sometimes sacrifice short-term victories for the benefit of long-term goals is important for all levels of athletic competition (Naylor, 2006). This might therefore influence the selections. Letting other athletes than the best ones play and compete may in the long run lead to more victories and more success for the team, even if it might cost a victory in the short run. Even at the elite level the learning and development process is a key to future success. The balance between “coaching to win” and “coaching for learning” is a key factor for a coach to consider (Naylor, 2006).

The number of athletes that are available in a selection process in comparison to the available number of spots in the teams, the so-called selection ratio, is also of importance in the selection process and a determinant of the practical value of the selection method (Schmidt and Hunter, 1998). The number of available athletes might play a role in the final selection outcome as well as for how the selections are carried out. Gelade and Dobson (2007) show that 70 percent of the variance in international team ratings in soccer is dependent on factors such as the number of men who regularly play soccer, the length of the country’s soccer tradition, the wealth of its population, the percentage of expatriate players in the national team and climate conditions. Besides the selection ratio, financial opportunities and limitations are of importance for selection processes and should be regarded when selection processes are discussed. In my study the economic frameworks are not in focus, but the importance they have is still of importance for the selections and the outcomes in competitions. When it comes to results in Olympic Games, it has been discussed whether bigger countries with more athletes to choose from have greater opportunities for winning medals, but it has been shown that the population rate is not sufficient to explain the number of Olympic medals. GDP (gross domestic product, a measure of national income and output for a given country’s economy) per capita is instead the best single predictor for the number of Olympic medals (Bernard & Busse, 2004). The host nation also has a notable increase in medals estimated at 1.8% of the medals beyond the ones predicted in the GDP quota.

Validity and reliability issues in models for assessing performance

When it comes to assessing motor performance in team sports, Godbaut designed a two-dimensional model identifying four categories of information that are of interest when team sports are assessed (Nadeau, Richard & Godbout,
The categories are technical and tactical aspects of the players’ performance and the different foci are on the result of the players’ actions or on the way they are conducted (the process). All of these four categories may be considered when team sport performance is assessed. Godbaut has further summarized various measurement strategies for collecting information of team sports performance into another two-dimensional model (Nadeau et al. 2007). He identifies how the assessments are made in standardized set-ups or derived during game play but also recognizes the nature of the measurement procedure. The measurement procedure is described as quantitative, relying on physical units of measurement or qualitative relying on the use of rating instruments. Through combining these strategies four general strategies for collecting information are identified. Nadeau et al. (2007) are summarizing the theoretical validity of each of those four strategies. He states that the standardised motor skill tests, although they have been the most common way to test efforts related to team sport performance, are hard to validate, as they do not take into account any tactical components of the game but only the technical performance skills. The inter-observer reliability and stability of these kinds of tests are referred to as high. During the two last decades there has according to them been a growing interest in using rating instruments in standardized settings. They discuss how (most often) the elements rated are related to process but may also be related to the result of the motor skills. Even if both these parts are considered in the evaluation, the tactical components are not taken into account resulting in a limited validity of the measurement strategy. When a rating instrument is used in a natural setting (for example during a game), this approach should be more valid, but when all facets of performance are to be evaluated by a single observer in live conditions, this has its difficulties. Theoretically, they see this strategy as the most valid way of assessing the game performance. When it comes to statistics derived from game play, they argue that this kind of statistics focuses only on the end result of various aspects of the sporting performance and not on the technical or tactical aspects of the players’ performance. The inter-observer reliability is also said to be problematic.

In the field of physical education, research concerning assessment of sporting performances has lately (in the last few decades) gained ground (Memmert & Havey, 2008). French sport pedagogy researchers tried out various procedures to assess game play in context, and in that field of experimentation an assessment procedure developed in 1989 is in focus in an article by Gréhaigne, Bouthier and Godbout (1997). They discuss this assessment procedure with the aim to integrate assessment with the teaching and learning process and to reach a high level of ecological validity in the procedure (the relationship of measurement with what is taught and that the evaluation does not disturb the “ecology” of the classroom) and assess students’ performance in various team sports. The assessment procedure is based on observations of players’ actions in
matches where the volume of play and the efficiency index are evaluated. Problems related to the assessments are discussed and how there are numerous interacting elements to consider when a player’s performance is going to be assessed. How the report of strength may vary depends on the opposing team’s condition and other situations, how the members are interdependent and how the player is assessed within a system that has its own coherence. According to the authors the procedure appears to produce objective, reliable and valid information of players’ overall offensive performance in team sport. Integrating this assessment method into the teaching and learning process was said to be something that should occur without stopping the learning activity to conduct the assessments. By the use of this assessment method tactical and not only technical efforts could be included in the evaluation. This assessment procedure was developed by Gréhaigne and various colleagues and modified for the Team Sport Assessment Procedure (TSAP) in 1992.

The TSAP was designed to measure individual performances of players in game situations by looking at players’ involvement in game play and a player’s efficiency. How a player gains possession of the projectile (ball, puck, etc.) and how a player disposes of the projectile were the main factors evaluated. Although this assessment procedure was not developed for ice hockey, Nadeau et al. (2007) wanted to verify the use of the model in the setting of ice hockey and conducted a study to establish the reliability and validity of the TSAP in the context of ice hockey. Two simulated games (training games) of two minutes were assessed and two student observers were assigned for each of the players. Their results pointed to the TSAP as a valid and reliable instrument for assessments of ice hockey performances. The validations were made through two correlational studies where concurrent validity (Spearman rank correlation) and expert judges were used. Reliability was tested through inter-rater reliability. In another study Nadeau, Godbout and Richard (2008) aimed to develop and validate an ice hockey assessment in real-game conditions by using an adaptation of the TSAP. This was according to them to be seen as the first study when real-game conditions were used. Performance analysis of 103 players at the age of 11-12 playing in the region championships in the Quebec area was included. The TSAP was adapted and more categories were added so that a total of 10 categories were assessed. Video recordings were analysed by three trained observers and ‘volume of play per minute’ and an ‘efficiency index’ were calculated and combined to a composite score that was the TSAP performance score. The measurement procedure was analysed and found to be both valid and objective. The performance stability was inconclusive but explained by the nature and duration of the study) and it was stated that more studies are needed. If the TSAP is going to be used for summative purposes, it will require observations over a series of matches. The variables were still said to provide more extensive information on performance than traditional measures.
In 1994, Mitchell, Griffin, and Oslin proposed the Game Performance Assessment Instrument (GPAI) as a way to enhance game play performance evaluations. The GPAI is a multidimensional system designed to measure game performance behaviours. The focus is on tactical understanding of the game and on players’ abilities to solve practical problems and apply appropriate skills. In the GPAI system, analyses are made both of individual game performance components such as, for example, decisions made, skills execution, and support and/or overall performances such as, e.g., game involvement and game performance. The GPAI was initially developed to be a teaching tool for Physical Education teachers. It was developed as a means for teachers wishing to teach games from a tactical standpoint. As the tactical aspects of a game and tactical understanding are an essential part of game play, this system was thought to provide a possibility to make the learning environment more developmentally appropriate. Seven components of main importance for game performance were developed through consultation with expertise from the field (teachers and coaches). Descriptions of each of those components were formulated and reformulated due to the experts’ inputs. Discrepancies among the experts were taken into consideration and led to discussions and modifications of the formulation of the component. Independently of each other, the experts watched videotapes of segments of the games and measured the game play based on these components, and the inter-observer agreements were calculated. Only one discrepancy was observed (concerning the definition of support). When individual performances were measured, the use of (number of) appropriate/efficient and inappropriate/inefficient responses could be included to give a more comprehensive picture of the athletes’ performances.

In 1998, Oslin, Mitchell, and Griffith reported on the development and validity aspects of GPAI. They tested four components of the GPAI to determine the degree of validity and reliability of three games (soccer, basketball, and volleyball). Their results showed that the face content and ecological validity were favourable but that construct validity was only partly favourable due to differences in the skill execution component. Skill execution was analysed as valid across all three games but the decision-making component was not significant in the study of soccer, and there were also some validity concerns when it came to the support component in soccer. The results indicate that the GPAI could differentiate between high and low ability performers in the different components that were studied.

Memmert and Harvey (2008) discuss concerns with the GPAI and suggest a re-examination of the scoring and coding systems in order to make the assessment tool more efficient. They identified five different problems with the GPAI and also offered solutions to these problems. The problems concerned the calculation of the component indices (a player must register in the appropriate category for the calculation to be mathematically possible, as zero is used as a
numerator in the equation and therefore not a divisible number to use in the calculation), how the relation and use of the index game performance (GP) and game involvement (GI) can conceal the true level of performance/involvement of the player (as the calculation formula does not make the number of actions visible and it is necessary to go back to the individual total scores of the components to understand which player is involved more and which has the better performance). The other problems they highlighted were concerned with observer reliability (the GPAI does not take into consideration the number of observers, non-linearity (good performances are treated differently than bad decisions) and the usefulness of action (it being quite hard to define what actions are appropriate or inappropriate when the behaviours are coded). In the article solutions to these problems are proposed and discussed. Although Memmert and Harvey see benefits of using the GPAI and regard the GPAI as a useful educational tool, they still argue for a re-examination of the GPAI if it is to be used as a valid tool for assessing game performances.

In an article about sport league drafts, Fry, Lundberg and Ohlmann (2007) model a decision-making process for player selection drafts. The model suggests that team selection be based on a combination of the player’s estimated value, the value of the other players currently available, and the team’s need at each position. As they felt that this was not enough, they also implemented a model with a spreadsheet-based decision support system to adjust it better to varied conditions. With perfect information about opposing teams’ selection strategies, they believe their method is better than competing strategies. If they lack good information about other teams, they do not believe their strategy will dominate but still be better than average strategies.

**Tipsters’ ability to predict performances**

So-called tipsters’ possibilities to predict and foresee performances have also been studied. In an investigation of the 1994 English soccer league games, the result showed that only one of the three tipsters seemed to use the information that was relevant for the games’ outcomes in a successful way (Forrest & Simmons, 2000). The tipsters’ success rates are higher than random forecasting methods like for example predicting wins for a home team or the like. Still, the expertise they claim to provide is very limited (Forrest & Simmons, 2000). Andersson, Edman and Ekman (2005) discussed the prediction of performance through letting experts and non-experts forecast the results of the World Cup in soccer 2002. Their study points to the same results as the one above and shows that the expert tipsters have limited forecasting ability. This study supported the hypothesis that sport experts and non-experts forecast the World Cup in soccer equally well. They predicted the results better than chance, but none of them outperformed the prediction rule based on former results. This prediction rule means that the team in a better position in the world ranking would win over a
Deciding Who is the Best

When the difference between the home and visiting teams’ scores is modelled as a Brownian motion process, it can predict the probability of victory for the home team (Stern, 1994). Factors such as the time into the game, the result at that time and the difference between the teams are considered in the model and suit the predictions of basketball games best due to the nearly continuous nature of the game and the score.

Fairness in selections – a validity issue?

In an article by Stevenson (1989) the athletes’ perceptions of the fairness of the selection to national sport teams was examined. Distributive and procedural justice was researched by means of two main questions, one of which was about whether the athletes perceived the selections to be fair and the second whether their perceptions of the fairness of the selection outcomes were related to their perceptions of the fairness of the selection procedures.

Three selection procedures were identified: selections by a board of selectors, selection by the national coach and selections as a mix of these two methods. The results clearly showed that selections by a board was perceived as unfair when it came to the selection procedures, while the outcomes and selections by the national coach were perceived as fair when it came to both the procedure and the outcome. The athletes’ perception of the mixed selection procedure was ambivalent with regard to both the procedure and the outcome and was placed somewhere in-between the above two categories. These results clearly showed that the athletes’ perceptions of the fairness in the selection outcome were related to the selection procedures. The author’s message to sport governing bodies is “this is best accomplished by establishing the competency of the selectors and the perceived objectivity of the selection criteria as well as reducing the diverse influences of bias and favouritism” (p. 379). The empirical data in this study consisted of interviews with athletes that were selected to national teams.

Introduction to my research focus

Much of the literature about judging, measurements, rule systems and new rules in sports is related to reliability concerns. Biased judging and the reliability of new rule systems are some of the topics investigated. There is some but little literature focusing on validity issues in measurements and selections. In the literature related to selections in sport, works on talent identification and devel-
opment together with literature about assessments and predictions of performance are the most prevalent publications.

As it is acknowledged that validity is the most fundamental and important characteristic when a test or assessments are developed, my interest in validity issues within sport was raised but not satisfied through this literature review. The result of this review was therefore a contributing factor for the direction of my research. The literature review together with my personal interest in these questions led to a focus on validity issues in judging, rule and judging systems and in selections to top-level sport teams.

To date there are few direct empirical studies of the perceptions of experts in measurements, judgements and selection in the sport field and the purpose of my studies therefore became to elucidate these processes through my interpretations of the experts’ views.
AIM OF THE STUDY

Two core processes in competitive sport are the selection of athletes to teams, competitions and games and the determining of winners. The purpose of this thesis is to increase the knowledge of these processes and to analyse them from a validity perspective.

This is studied from the following points of departure:

− To describe rule changes in the judged sports acroski and rhythmic gymnastics and the implications these changes have and to find out what perceptions experts from these sports have of the measurements, judgements and the rule changes in their sports.

− To gather information about selections to top-level sports teams in soccer and alpine skiing through selectors’ perceptions of the goals and criteria for the selections, the selection processes and the outcome and consequences of the selections in their sports.

Through conducting this thesis, I aim to contribute to the discussion of whether the “right” athletes are selected to participate in teams, competitions and games and if the “right” athletes are winning.
METHODOLOGICAL REFLECTIONS

I see the research process as a process where aspects of craftsmanship have to be handled with concern and care but where it is not necessary to position oneself in one certain ontological standpoint but still be aware of ontological and epistemological questions. The awareness of different ontological and epistemological perspectives has widened my perspectives of what knowledge and research are about. It has made me consider how I look at research and its truth claims as well as methods for data collections. I see the research question as a point of departure in the research process and that the existence or non-existence of suitable theories has to be considered. My research approach was therefore dependent on the aim of my studies and my research questions. Validity theory had proven to be successful and is widely used when questions about quality, usefulness and fairness in measurements in other fields have been studied and was therefore used as the theoretical framework.

I am a firm believer that all of us see the world and the events in our world with references to our context, our knowledge, our experiences and our social relations. This means that we acquire knowledge from different sources around us but also that we interpret events differently related to our different knowledge, different experiences but also because of different surroundings and our temporary state of mind. According to this reasoning, interpretation is always made when something is described. Writing or telling someone else about a phenomenon will always involve a certain amount of interpretation on the part of both the one presenting the data and the one trying to understand it.

The data acquired in the interviews was mostly of a qualitative kind, as they focused mostly on the participants’ thoughts, feelings, perceptions and experiences of judgements and selections rather than on numbers and statistics in the processes. Interviews were the main data collection method and this means that the ‘knowledge’ presented in my studies is my interpretations of the informants’ perceptions of the investigated area rather than an all-encompassing truth. Some of their perceptions of the area are based on official documents, rules and the like and some are perceptions based on experience and gathered knowledge. ‘Knowledge’ of this kind is still meaningful, as it may reveal both patterns and differences that play a role in the processes that are studied and should be seen as a contribution to understanding the studied area.

I have continuously through the process worked thoroughly with critical re-consideration of my interpretation and understanding. I have worked both to
reflect systematically on the inherent meaning of the empiric material but also reflected on myself as the researcher and on my context and former experiences. Describing the research process is a way to increase the transparency and the credibility of the research process. Complete transparency probably does not exist, as it would mean that I should be able to explain every thought process I had during my research, and this would be impossible and probably more confusing than clarifying. Secondly, if my standpoint were that I could explain every step in my research process and my interpretations of my data, this would imply that I believe in total objectivity, which I do not.

**Selection of data collecting methods**

As stated above, I regard the aim and research questions as the point of departure of the research process and the deciding factor for the data collecting methods. With the aim of understanding the measurement, judging and selection processes and being able to analyse them from a validity perspective, extensive information about these processes was needed. Some information was found in official documents, but to obtain more extensive information of these processes and information of feelings and perceptions of these processes, interviews, together with gathering of official documents were selected as the most suitable methods to collect data. In the interviews I reached beyond the knowledge that was possible to obtain through official documents, as the informants related to and discussed the answers based on their own experiences and feelings. For my research question this was of great importance and provided me with a more comprehensive picture of the processes. In the communications with the informants I was also able to ask them to ask follow-up questions and to develop their thoughts about certain areas that seemed of importance to them. When in some interviews I was uncertain about whether I had understood what the informants tried to express, I asked for clarifications and opened up for a more extensive discussion, which in some cases led me into new questions and resulted in a better understanding of these processes. The informants also asked me for clarifications when they were not sure about the meaning of a question. By using questionnaires I would have had an opportunity to get information from more informants, but as I needed extensive information that in some cases was very sport specific and quite complicated to specify in a questionnaire, I believed the choice of interviews as the main data collecting method was a valid choice, even if this resulted in information from a smaller number of informants than questionnaires would have allowed. Observation was also considered especially for the study of selections, but as the process of selection in many cases is a process extending over quite a long time, and as it was in my interest to study different selection processes, I decided that observations would be too time-consuming and would have limited me to studying and obtaining knowledge of fewer selection situations.
Selection of the sports

The selections of what sports were to be included in the studies were dependent on a combination of factors. In the studies of measurements and judgements, so-called judged sports were of special interest as there the rules and the judges' interpretation of the rules are so closely connected to the athletes' results. Judged sports are also often the target when measurements in sports are discussed and many times they are discussed in terms of being unfair. Since I have been involved in judged sports my whole life, I had many times got questions about the fairness and subjectivity in judged sports and had thought a lot about what is really measured and judged and how these processes function in judged sports. The choice of the judged sports acroski and RG was partly related to my personal experience and knowledge of these sports and the network I had within the sports. I had been involved in both sports for a number of years and my knowledge of the rule and judging systems was thought to be helpful in the studies. Other factors that played a role were that the sports had official and organised rules and regulations for the content and judging of them. Both of these sports were organised through national and international governing bodies, performed worldwide and were interesting as they were young sports that had gone through several rule changes in the last 20 years. My previous knowledge of the sports was very helpful in the analysis and understanding of the rather complicated judging criteria and rules for the sports. Without the previous knowledge I had as an athlete, a coach and a judge, it would have been very difficult to perform the analysis I made of the rules and regulations. The network I had with persons highly involved in these sports was also helpful in the process of reaching the informants.

In the study of selections to top-level sports teams, the purpose was to increase the knowledge of selections to sport teams from a wide perspective. Therefore, one team sport, soccer, and one individual sport, alpine skiing, were targeted. In a team sport the players are interdependent, which means that the performance of the team and the outcome of a game are a result of the cooperation between the players. In an individual sport, the athletes are not directly dependent on one another’s performance, but still individual athletes also interact with one another to some degree. For this reason, selecting athletes to team sports and individual sports was of interest, as it could reveal different ideas.

1 In rhythmic gymnastics I have been a competitor and organizer at the national level and a coach and judge at the district level for many years, and in acroski I competed at the time of the licentiate thesis in the Swedish National Freestyle Ski team and was one of the top skiers in the world. In acroski, I had been a member of the freestyle skiing committee in the Swedish Federation and was during the time of the study the chairperson of the international freestyle skiers’ athletes committee and a member of the FIS Athletes Commission. I was also working as a co-producer for ski shows in USA, Argentina and Italy, and performed in ski shows in a number of countries.
DECIDING WHO IS THE BEST

about the selections. This choice of including both a team and an individual sport proved to be interesting during the interviews, as there were both different and similar ways of working and thinking about selections in these different sports. To be able to talk about selections in a general way, the inclusion of a team and an individual sport helped to increase the validity in that a wider perspective on selections was portrayed.

The sports alpine skiing and soccer were selected as they were sports that are performed all over the world, have well-developed competition programs at all levels, are Olympic sports and have many contenders and public interest in many countries. These factors were thought to increase the demands on the selections and making the sports interesting to study. Although both sports were familiar to me I did not have the same personal experience and knowledge about them as I had about acroski and RG. This was something I firstly saw as a challenge but it turned out to work well and my connection to and involvement in sports and especially elite sports were helpful assets in these interviews too.

Selection of the informants

In the process of selecting informants my motives were to find informants that had extensive knowledge of the processes I was interested in. This was important for two reasons, one being that with extensive knowledge they would be able to provide a lot of information of the processes and the second that if they had this extensive knowledge of these processes, they could be regarded from a validation perspective as experts in their fields and their voices carried an impact not only as an information source but also as a validation source. Persons involved at the elite level in sports have most commonly been involved in sports for a number of years, which means that they should have extensive knowledge and experiences of the sports and processes within the sport. The importance of the results and the consequences that the measurement and judging process and the selection process have at the elite level also means that the demands on the processes on this level are very high. Therefore it was important for me to find informants that were and had been involved in the elite level of sports for quite some time. How things are valued and what kinds of decisions are taken at the elite level in sport might also have a significant impact on sport at lower levels and were another reason for selecting informants from the elite level.

In the studies of measurements and judgements I decided to gather information from judges, coaches and athletes, as persons from these categories are very much involved in measurements and judgements and would possibly also view the judgements and judging process differently due to their different roles, as the outcomes of these measurements and judgements could affect them differently. This resulted in a selection of totally sixteen informants, eight from each of the sports acroski and RG. In both of the sports two judges, two coaches and
four athletes were selected. The reason for selecting more athletes than judges and officials had foremost to do with their being so immediately affected by the judgements, rule and judging systems.

Since there were very few official documents about selections, the main source of information about selection processes was thought to be the ones taking the selection decisions, the selectors. In both soccer and alpine skiing this showed to be the coaches. As they are involved in the whole selection process and the thoughts and feelings during the whole process, and as selectors at the elite level have made many selections, they were expected to have a lot to share about the selections. This proved to be true, as they had many thoughts and feelings about the selections and decisions during the process. I also considered interviewing athletes, but realised that it was problematic to know which athletes to select, the ones that had been selected or the ones that were not selected. A representation of both groups probably would have been the best alternative, but how to reach those that had not been selected and to know who they were would have been partly problematic. In this study, which is my first study of selections, I decided that the most important thing was to gather information from the ones performing the selections, the ones that were supposed to have “the total picture” of the selections, i.e. the selectors. As the selectors are responsible for the selections and also are able to affect the criteria and the process as well as the outcome and consequences, this study focused on them.

The number of selectors I interviewed was fourteen, eight soccer coaches and six ski coaches. The total number of informants that this thesis is based on is 30 informants.

Table 1. Distribution of the informants

<table>
<thead>
<tr>
<th>Informants</th>
<th>Coaches</th>
<th>Judges</th>
<th>Athletes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acroski study</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Rhythmical gymnastics study</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Selection study</td>
<td>14</td>
<td></td>
<td></td>
<td>14</td>
</tr>
</tbody>
</table>

When I defined “elite level”, I chose to consider persons involved in the top level of the sports on the national level and persons who in addition to this were involved in international competitions or games. Being involved at this level was thought to be a guarantee that they had long experience and extensive knowledge of these processes. All of the informants selected for the interviews also had this desirable long experience and extensive knowledge of these processes at the elite level.
As the sports are organised somewhat differently, this resulted in informants both from outside and within Sweden. In RG it was possible to find experts on the elite level that lived in Sweden and belonged to the Swedish Gymnastic Federation, but in acroski there were not enough experts on this level within Sweden. In the search for informants that suited the above criteria in acroski, informants were found in Sweden, Norway, France and USA.

In the selection study alpine skiing and soccer proved to be a bit different in their organization, and in alpine skiing the skiers at the top level perform most of their competitions for the national team and therefore the selectors for the national teams came to be the experts of interest in alpine skiing. In soccer, players at the elite level play most of the games for their club teams and some in the national team, so both the national team and the club selectors were therefore of interest. The selection of the informants was made by looking at the current rankings of the athletes/teams in the sports, in skiing global ranking and in soccer national ranking. Top teams on the ranking lists were targeted. In skiing both big and small nations were involved in the selections. Besides playing in the highest soccer league in Sweden, the club teams in soccer were also playing in international games and tournaments. I got an opportunity to interview six alpine ski coaches during the World Cup opening competition in Sölden, Austria 2008, and in soccer I ended up making eight interviews. The selected coaches were all highly involved in the selection process and some of them had the total responsibility for the selections, while others shared the responsibility with other coaches and/or team managers or the board of the club or federation.

My aim was to interview persons that were involved in measurements, judging and selections of both male and female athletes and also to include both men and women in the interviews. This aim was reached, but as I also wanted to have a valid selection of informants due to male and female representation within these sports, the actual number of males and females varied depending on the sports. In RG all the informants were females, as the sport does not allow male participation at any level (athlete, coach or judging level). In the acroski study a small majority of men (n=5) in comparison to women (n=3) were interviewed and this also mirrors what this sport is like, as there are more males than females in judging and coaching positions. In the selection study two ski coaches of men’s teams and two ski coaches of women’s teams and two overall ski coaches (responsible for both male and female athletes) were interviewed, this being due to different structures in different countries. In soccer five coaches of female teams and three coaches of male teams were interviewed. Except one all the interviewed coaches in the selection study were men.

Before performing the interviews, it was difficult to know how many interviews would be performed. Kvale (1996) has an interesting answer to the question of the number of interviews that should be made and states, “Interview as
many subjects as necessary to find out what you need to know” (p.101). As it is not so easy to predict what is necessary to know, this is not as simple as it sounds, but it was nevertheless an objective in my work. Saturation in the answers (Gratton & Jones, 2004) is another concept used to express the same thing and a way to express that no or hardly any new knowledge will arise by conducting more interviews. In my study, I believe I received answers to be able to describe and discuss judgements and selection methods in the sports that were in focus. If more interviews would have rendered other kinds of answers is hard to say, but my belief is that the interviews gave me “the answers I need” to be able to understand and discuss judgements and selection processes in top-level sport. Still, there is always the possibility that other answers and explanations would have been found if more interviews had been made.

The research process
For the study of selections to top-level sport teams I developed a model to clearly define the steps in my research process (Figure 2, p. 35). This model proved to be useful for me as I built it upon concerns important in validity theory and it provided me with a simplified way of handling the complicated research process. By working through the steps in the model thoroughly I was able to implement the research process in a structured way.

The research questions
This model was very important in the development and formulation of the research questions in the selection study, as it was developed with validity theory as the basis and as each of the steps came to represent vital parts of what should be researched. These vital parts, or areas, were the point of departure for the formulation of the research questions. When I worked with the studies of measurements and judgements I had not developed this model but worked in a similar way as I used validity theory as the point of departure for the process and the research questions in that study too, but I had not structured my work in a model. The areas important for the formulation of the research questions and later on in the process also serving as the framework for the development of the interview guide were in the study about measurements and judgement: background information; ‘the idea of the sport’; the judgements’; ‘judges and influences’; ‘rules and regulations’; ‘protests’ and; ‘other thoughts’. In the selection study these areas were background information: the idea of the selections; the selections divided into (a) criteria, (b) values and (c) the process; difficulties in the selections; reactions from the athletes (perceived by the selectors); outcome and consequences and; other thoughts and ideas.

In the initial stage of the development of the interview guide, the formulations belonging to the theoretical concepts were used, but the questions were
then reformulated to contain a more “everyday” language to facilitate the understanding of the questions, which I believe contributed to a good flow and good understanding when the interviews were performed.

Some of the questions were rather tricky to formulate. An example of this is the formulation of the question aimed at obtaining information about the idea of the sport in the acroski and RG interview guides. What the informants perceived as ‘the idea of the sport’ is a very central question in these interviews. In one of my earlier studies (Andersson, Bergkvist, Johansson & Svahn, 1988), it had proved to be somewhat difficult for the informants to answer the direct question “What do you believe is the idea of the sport?” without them being too caught up in the current rule and measurement systems. When this question was asked in the former study to expertise in ice hockey, the answer was often “to make goals”, an answer directly linked to the current rule and measurement system in the sport. To open up for a more extensive discussion about ‘the idea of the sport’ and to release the informants from only answering in accord with the current measurement system and instead being able to value the sport and performances from what they really thought was the core of the sport, other kinds of questions were formulated. Questions like: What do you believe is the most important quality for an RG-gymnast? How would you describe an ideal performance in RG? What do you want to see when you are watching an RG competition? and similar questions were used instead of the direct question about the idea of the sport. These indirect questions gave more extensive answers and also opened up for a discussion of whether the current rule and judging systems were optimal or not for attaining valid judgements. Using these kinds of surrounding questions was a strategy I felt gave more valid answers to what the informants perceived as the idea of the sport.

The interview guide was then translated into English, as some of the interviews were performed in English. The work with the translations of the interview questions was quite time-consuming, as it was important that the formulation of the questions captured the same content irrespective of what language they were asked in. Although I think I managed to formulate questions with fairly equal meanings in both languages, the validity of the interview guide might have been increased if I had taken help from someone specialised in translations. Still, interpretation problems may appear both if the questions are formulated in a first or in a second language, and as I was very aware of these interpretation challenges, I worked during the interviews to make sure that the informants understood the questions in the way I had meant. For this reason I believe the validity of the questions was increased.

Before making the first interviews in both of the studies, I performed and recorded a couple of pilot interviews. The main purpose of these was to test the understanding of the questions, their content, the order of the questions and the length of the interviews. This was tested and except for some minor adjust-
ments of some of the questions and their order, most of the questions were left the same after the pilot interviews, as they seemed to work well. Going through and listening to the recordings of the pilot interviews were helpful not only regarding the content of the questions and answers and the order of the questions but foremost as a help to increase my awareness of the influence I might have on the interviews in my capacity of interviewer. I became particularly aware of the risk of prolonging and explaining the questions too much. Listening to myself was an important part, as I realised that I really had to be aware of my influence on the interview and its outcome.

The interviews

Reaching the informants worked well. I contacted them through personal communication or e-mail letters, and except for a couple of soccer coaches the informants that were contacted also came to participate in the studies. Two of the selected soccer coaches were not possible to reach or get any replies from, and a couple of the soccer coaches who were going to participate never participated in the end, this due to busy schedules and transfers to a foreign club. With the same selection method as mentioned above, other soccer coaches were contacted and they participated in the study. A reflection concerning this is that, except for the soccer coaches that did not participate in the study, all seemed very interested in discussing these kinds of questions. The selectors that never answered or with whom no interviews were made later on in the season had some challenges with the results in the teams or as mentioned, switched clubs and this could possibly have influenced them not to want to or not to prioritise participation in my studies.

For me it was important to perform all of the interviews myself, as this would help me understand these processes better and I would then be able to ask the follow-up questions I felt were of importance, ask for clarifications and help clarify questions that the informants perhaps did not seem to interpret the way I intended. I preferred meeting the informants in person when I made the interviews, as I felt that this made it easier to create contact with them. Meeting them in person was also helpful, as both the verbal and the non-verbal part of the communication could then be considered, which is of importance to be able to understand the nuances of the words (Gratton & Jones, 2004; Kvale, 1996). Personal meetings could be arranged for 21 of the 30 interviews. The other eight interviews were telephone interviews (all performed in the studies about judged sports) and they worked well too, and I believe a contributing factor was that I was familiar with these informants and they knew who I was, even if I did not know all of them well. The interviews were recorded, which helped me to actively focus on the informants instead of being caught up with writing down the informants’ answer.
Sixteen of the interviews were performed in English and fourteen in Swedish. I regarded it as a challenge to perform interviews in a second language and to interview informants some of whom had English as their first and some as their second language, as the words in the interviews are used as result-carrying items and misunderstandings would affect the outcome of the research. In the situations I felt I had doubts about the informants’ understanding of the questions I tried to express myself by using different words and/or asking for clarifications, and sometimes the informants also asked me for clarifications. By being aware of the challenges involved in performing the interviews in a second language, I was observant of the risk of misunderstandings. As misunderstandings and misinterpretations might also occur in the interviews performed in Swedish, the awareness of the importance of the understanding of the questions was something I also worked with in the interviews performed in Swedish. My interpretation was that the interviews worked well and that clarifications had to be made in interviews performed in both languages and that they had more to do with my understanding of the sports than with the language in the interviews. As I was interested in the perceptions of the persons involved in the very top-level of sports, I felt it benefited my study to incorporate informants from other parts of the world, even if there was a slight risk that some misunderstandings would appear.

The transcriptions of the interviews
All the interviews were transcribed into texts where all of the words, hesitations etc. were rendered as precisely as possible. I transcribed most of the interviews, but for a couple transcriptions I got help. Regardless of who made the transcriptions, the same structure for how to deal with pauses, hesitations, emphasis on certain words etc. was used for all transcriptions. For me the transcription of all words in the interviews was important, as I wanted to work with transcriptions that were as close as possible to the interviews and not interpretations of what the informants had said. For me this was done in the analysis process and not in the transcription process, even if I of course became familiar with the material and started to analyse the interviews already in the transcription process.

The analysis of the empirical data
As it always is a challenge in research to analyse the empirical data and as this is very important from a quality perspective (Kvale, 1996), I worked very thoroughly with this part of the research process. In the RG and acroski study many hours were spent reading up on and understanding the rules and regulations of the sports to be able to discuss them. I also contacted a member of the European Union of Gymnastics as a source of knowledge and clarifications of these rules and regulations. In the selection study there was not so much official ma-
As I had performed all the interviews and transcribed most of them myself, I had started the process of analysing the data before I sat down to perform the analysis. This was good as it made me feel that I was quite familiar with the material, but I also felt that it was a risk, as it had made me think of some things as more important than others and I was unsure if they really were or if it was just the way they were expressed and how they connected to my previous experiences that made them interesting. For this reason I decided that my first step was to become familiar with all the material again and read through all the transcriptions to get an overview of the material. In the analysis of the interviews I used different analysing methods such as copying and pasting all the informants’ answers to a certain question into one document and trying to go through all words in the document through meaning condensations and meaning categorisation (Kvale, 1996), but I also went back and listened to the sound files of the interviews in order not to get too stuck with the transcribed words, as I felt the way the informants talked about the questions was as important as what they said. The use of different methods helped me see the material from different angles, which was useful as it helped me to avoid getting stuck with the first thoughts I had about the answers but also to look at them from different angles. As the statements in the interviews are statements the informants had at the time for the interview due to the situation of the interview and their previous thoughts and experiences the aim of the interviews was to contribute knowledge due to these perceptions. To be as close as possible to the informants’ statements, I tried from the beginning to use their wording in my texts, but the more I learned about my material, the more I realised that this made my texts more confusing than informative, as the informants used different words to describe the same thing and the same words to describe different things. Instead I had to interpret the material and try to understand the more underlying meanings of their words, if I was to present a text that would represent what they thought. To be as fair as possible to the material, I went through it a number of times and analysed it by means of different methods.

Validity in sport processes

The model (Figure 2, p. 35) I developed to clarify the steps in how to study validity and reliability issues in sport was very helpful for me in the research process. It gave me a framework for my study and guided me in all the steps from the planning of my work process to the development of the research and interview questions and the analysis process. The model helped me to avoid getting too lost in details and to follow the planned course of action in my research process. The last boxes in the model show the steps where validity and
reliability are to be discussed in relation to outcomes, social consequences, values and contexts, which are crucial parts in the modern validity theory. In the measurement and judging study this was discussed but not with such a big focus, as the point of departure there was content validity.

Validity in my research
As this thesis has shown, I see validity issues as being of great importance. When it comes to research, I believe validity issues have to be treated with great respect and understanding, otherwise the outcomes will be mere speculations. As there are many possible pitfalls in the research process, great awareness of validity issues and a structured research process are necessary for the quality and usefulness of the research. The model described above was helpful in my studies of validity in sports and below I have applied the same model (Figure 3) to analyse and discuss the validity of my thesis.

Figure 3. Steps to analyse and discuss the validity in my thesis
By using the steps in the model I will discuss the validity and reliability of my thesis. The first step in identifying the ideas of the thesis means identifying and deciding an overarching aim of the thesis. The overarching aim was strongly related to my interest in the sports field and these kinds of questions but also connected to some earlier reports and articles I had written on the subject. By having an overarching aim it was possible to start the research process and go further to the next step, identification of suitable theories for the study and identification and studies of literature related to the field. Studies of previous literature revealed that quite a few studies had been made in this field but none with the same aim as mine, and that in other fields some studies with similar aims had been performed. This strengthened my idea for the research and my initial aim of the study. Although the formulation of the aim changed some times during the research process, the main aim stayed the same, as did my goal for what I wanted to study. As validity and reliability concerns were often used to study processes of tests, assessments, judgements and similar processes of valuations, validity theory was regarded as a valid theory to use as the point of departure. The third measure in determining the steps in the research process, involving data collecting methods and how to best reach this information was discussed earlier under the headings of selecting data collecting methods, selecting sports and selecting informants. All these decisions were carefully considered and are decisions valued and determined with validity issues in mind.

The fourth step “Analysis of the outcomes and discussions of the consequences of the outcomes” was a complex and partly difficult step. The discussions and problems concerning the inclusion or non-inclusion of consequences as a part in the validity concept were exemplified earlier in this thesis. The reason why I wanted to make an attempt to incorporate a discussion of the outcomes and consequences of the selection process is my belief that the consequences of selections are of major importance in the sports world and should not be ignored. The consequences of the measurements, judgements and selections are also very visible and obvious in the sports field in comparison to many other fields, as the results and selections are public and officially displayed. In the conclusions and discussions of the selection study, the outcomes and consequences were discussed due to the informants’ perceptions of this but also analysed in a broader perspective due to information and perceptions that the informants contributed concerning the whole selection process. Although the consequences of certain selection procedures could have been discussed from many more angles and more information could have been gathered from the informants about this, I have still worked to include this discussion of the outcomes and consequences and not only the instruments. This was done in a more structured way in the selection study, as it was an aim to do this in that study, but consequences are discussed although more briefly in the measurement and judgement study. Even if this could have been done more extensively, I believe I have captured
many issues of importance for the outcome and consequences of the studied processes.

In connection to the fifth and final step in the model I would like to summarise my thoughts about the validity and reliability of my research and express some more validity and reliability concerns. The structured research approach was used to give a clear idea both to myself and to others of how my research was conducted and of the importance the validity theory had in all steps from locating the main areas to the research and interview questions within the framework of my aim of the study. In the above reflections on the research process, I explain the steps in my research process but also how I worked consciously to provide the research process with as high a degree of validity as possible. This shows that I am aware of the importance of all my decisions in the research process but also how I handled these concerns. By explaining the steps and considerations the transparency of my process will increase making it possible for all readers to evaluate my decisions and considerations from their own point of departure. Including not only my thoughts about my research process but also the interview guides (in the separate reports) has the purpose of increasing the possibilities for others to assess and analyse the validity of my research based on the content of these questions. Inclusion of the interview guide will also increase the possibilities of replicating the studies, or at least of making similar studies.

When it comes to reflections concerning the contexts, I have some validity concerns that I have tried to handle. The first one is related to the sport context and the second to the scientific context with me as a doctoral student. The practical sport context is discussed here in relation to my background and involvement in sports. As long as I can remember I have been involved in sport and sport organizations and I had a lot of thoughts and feelings concerning both rules and judging systems as well as judgements and selections prior to the studies. I had also actively worked in committees to improve and develop the rule and judging system in acroski, and as a judge and organiser been very much involved in these kinds of discussions in RG. As a member of a national team (in freestyle skiing – acroski) I had also been involved in and surrounded by numerous selections at the elite level in sports. My initial interest in both rule and judging systems and selections was mostly connected to the consequences of these judgements and selections for the athletes involved, which was probably related to my experiences of athletes’ reactions to these kinds of issues.

When I worked with trying to make my thoughts and feelings clear to myself, I realised that they were very much concerned with fairness and especially fairness for the athlete. The identification of this helped me to be consciously concerned with involving and treating all the informants’ perspectives but also with continuously challenging my preconceptions through incorporation of the knowledge gained from the literature and empirical data. Still, I will always
carry my experiences and feelings from sports within me and they will influence my understanding of new information. Identifying as much as possible of these experiences and feelings and challenging them against new knowledge were important for the outcome of the studies but also for my personal development. In total I believe my background in sport was essential for my research, as it was helpful in the understanding of the somewhat complicated rule and judging systems, and because my knowledge of the field together with the theoretical knowledge helped in the creation of interview questions and analyses. My background was also felt as essential when it came to gaining entrance and acceptance in the field.

The scientific context is during this thesis my role as a doctoral student with all that this involves. This has implications for my work and my research process in different ways and is also connected to the total validity of this research. Being a doctoral student and writing a doctoral thesis is for me and many others a major step into the world of scientific writing and a process of learning about the craftsmanship of research. Being a doctoral student involves having supervisors to discuss the research process and the analyses and outcome with. This means that the outcomes of the study are discussed by different persons but also that the final text has undergone changes after feedback from the supervisors. These are steps increasing the validity and reliability of the thesis, as it is done for the purpose of improving the research process but also in order to increase the understanding of the thesis.

Doing research means having limited time and many times as in my case a limited budget for the research factors that may affect the validity in the research. Thanks to some scholarships and research grants I was able to reach most of the selected informants in person, which I value as positive for the total validity of the research. As being a doctoral student means learning how to perform the research process and work with the research material, many of the steps were made by myself with little involvement of anyone else. This is part of the learning process but could be questioned from a validity standpoint. For example, I made the interpretations of the interviews as the single interpreter and I translated the interview guide myself. To increase the evidence for reliability these processes could have been done and controlled by help of others. With more training in research and use of analysis methods, the whole research process could probably be more valid, as knowledge is a key factor when quality is to be achieved. Still I believe the structured process and the help of my supervisors have contributed in the striving to produce a thesis with a high level of validity.
SUMMARIES OF THE EMPIRICAL STUDIES

In this chapter summaries of the licentiate thesis and the report on selections in sports are presented. These summaries mainly focus on the findings of the studies, as the theoretical framework, related literature and methodological reflections are presented in earlier parts of this thesis.

The licentiate thesis (empirical studies 1 & 2)


The overarching aim of the licentiate thesis was to study measurements and judgements in sports from a reliability and validity perspective. In two separate reports the judging systems in acroski and rhythmic gymnastics (RG) were studied to describe the rule changes in the sports, the implications these changes had for the sports and to find out what perceptions persons highly involved in these sports had of the measurements, judgements and rule changes in their respective sports. Information was gathered from individual one-hour interviews with two judges, two coaches, and four elite athletes from each of the sports acroski and RG. All interviewees were very much involved in their sport and had very long experience of their sports on the elite level. The respective sport’s rule systems, judging manuals, meeting protocols and historical documents were also studied.

The results show that in the period of study both sports underwent major changes in their rule and judging regulations, changes that had a considerable impact on the sports. The quick evolution of these sports seemed to be the main reason for these changes but other factors were also of vital importance. The reasons for these rule changes were divided into changes that were made to improve the reliability, changes made to improve the validity and changes due to other reasons.

Rule changes that had to do with improvements of the reliability were for example that the judgements had been divided so that the judges assessed separate parts of the athletes’ performances instead of the overall performances as earlier on. This was done so that the judges could concentrate and focus on fewer factors in the judgements. As regards changes in the number of judges,
there is a difference between acroski and RG. In acroski the number of judges stayed constant while it increased in RG. In RG more judges and juries to control the judges were introduced to make the judgements more reliable and fair. To make the judges’ scores as reliable as possible and avoid too divergent scores being included in the final scores, the highest and lowest judges’ scores were eliminated in some parts of the judging in both sports. In RG there were rules stating how much the judges’ scores were allowed to differ and these differences were decreased during the studied time period. In acroski an acro plan layout (APL) was introduced to make the judgements more reliable. The APL was a map of the athletes’ planned performance that was provided to the judges before the competition. The aim of the APL was to help the judges know what tricks would be performed and when the athletes would perform them.

The level of definition of the rules and regulations was raised to increase the opportunities for clear and reliable judgements. In acroski the detail level of the rules and judging regulations was changed from not being so detailed to being very detailed and then back to not being so detailed again. The change towards more detailed regulations took place around the time before the Olympic Games in Albertville, 1992, when the sport tried to adapt its regulations to the International Olympic Committee’s (IOC) ideas of necessary objectivity in the judgements, which was thought to be reached through more detailed judgements. In RG the already detailed regulations gradually became even more detailed resulting in a very detailed system with highly specified rules for how the athletes’ manoeuvres were to be performed and judged.

The informants’ perceptions of these changes were positive as regards the changes concerning the division of the judges so that they judge fewer factors and the elimination of high and low scores. The larger number of judges in RG was also perceived as positive. Concerning the inclusion of the APL, it was considered to have both advantages and disadvantages, and for this reason different ways of using the APL were discussed by the informants. The results of the changes made to improve reliability through a higher detail level in the regulations are something the informants had different opinions about. Many of these views had to do with these detailed rule and judging regulations resulting in stereotype performances, as the athletes would have to follow highly specified rules for the design of their programs. They saw this as negative, as the sport would then lose its basic idea when the originality and freedom of the performances were no longer present.

Rule changes to improve the validity in the sport are changes related to the regulations stating what factors are to be measured and judged and the content in the sports. In both acroski and RG major changes have been made according to the content of the sports, this being dependent both on the developments of the sports and on other demands for changes. How many difficulties, what kinds of difficulties and how they are to be valued against other parts of the
performance were discussed and changed in both acroski and RG. In both sports the regulations have moved in the direction of more equal valuation of technical and artistic abilities. In RG bonus points have been included both as a way to distinguish between the athletes’ performances and as a way to award extra points to certain parts. What factors to be included in these bonus points was discussed and changed in the different regulations.

The changes towards more equal valuation of technical and artistic abilities were regarded by the informants as positive for the development of the sports, as the athletes would then have to be versatile and good at all parts of the sport to win. The bonus points (RG) were up for much discussion, as they were thought to lead to monotony in the performances when all athletes tried to perform them. A discussion about excluding them and going back to using risk factors to make the sport more interesting and in accordance with the idea of the sport was going on during the time of the interviews.

When the more detailed regulations were introduced to improve reliability, this also affected the validity of the performances. The rule changes made to help acroski become an Olympic sport were seen as negative by the informants, because they thought that the sport would lose its basic idea, since the specified and detailed regulations would lead to less originality and freedom in the performances. Since it was perceived within the sport that the new rules were not in accordance with the idea of the sport, the rules were after a while changed again, this time back towards being less specified and with a lower number of obligatory tricks in the programs. In RG the detailed rules were also seen to limit the performances. This shows how the reliability and validity of the judgements interact with each other and how the efforts of raising the reliability had negative effects on the validity of the judgements. The importance of the validity issues was also highlighted, as these rule changes showed that if the wrong things are measured or judged, it does not matter if the reliability is perfect, the quality of the measurement is not optimal and the “right” athlete might not be the winner.

There were also some rule changes that were made to improve neither reliability nor validity issues and an example of a rule change like that was the change of the number of athletes who were to qualify for the final competitions. Changes like this could be the result of external forces such as media and sponsors, and here it was noted that that even if the purpose of the changes was related to neither reliability nor validity concerns, the results of them might be.

The overall conclusion from these studies indicates how important it is to discuss and understand the consequences of rule changes from many different perspectives, as rules intended to improve some part of the judging might have undesired effects on other parts of the performances. In these two judged sports the main challenge seemed to be finding the balance between making the judgements as objective as possible, without losing the idea of the sport, to-
Deciding Who is the Best

gether with the quick evolution of the sports. Many of the changes were made
to improve the reliability of the judgements but resulted in undesired effects on
the validity of the judgements.
Empirical study 3


The purpose of this study was to increase the knowledge of selections of individuals to top-level sport teams. The focus was on the selections in one individual sport, alpine skiing, and in one team sport, soccer. Discussions about the judgements and measurements were to be related to validity and reliability perspectives. The study was based on 14 semi-structured face-to-face interviews with top-level coaches (selectors) highly involved in the selection processes in soccer (mostly club teams in Sweden) and in alpine skiing (national ski teams).

The results showed that there were some disparities between the club and federation boards and the coaches as regards both the level and the types of goals and ideas for the selections. From a validity standpoint this is partly problematic, as there might then be different kinds of goals that the selection method has to be adapted to. This will be especially problematic if some of those “goals” are visible and explicit and some are neither visible nor communicated but still have an impact. The clearer and more common the goals are and the fewer underlying individual goals there are, the better the chances are of making the selection methods valid and reliable.

In this study there were great differences in the definitions of the selection criteria among the teams. How precise the selection criteria used by the coaches were ranged from very well defined to very vague. The stability of these criteria can be discussed based on how well defined they are. If they are written down and officially communicated, it is harder to change them during the season. If on the other hand they are not so clearly defined and not communicated, they can be changed during a season. They could also be differently applied in relation to different athletes. This is from a validity and reliability perspective an issue concerned with the total quality of the selections. Flexibility in a system is often good, but depending on the level of flexibility of the criteria the chances of stability in the selections are affected.

Some of the coaches seem to be convinced that clearly defined criteria are a necessity for good selections, while other coaches do not see well-defined criteria as a necessity. Having the selection criteria visible and communicated to the athletes has both advantages and disadvantages according to the coaches.

A mix of past performances and predictions of future performances determined the coaches’ selections, although past performances seemed to have a very big impact. Former performances were seen to be an important factor for the selections in both soccer and skiing. In skiing former performances were often measured in terms of the athletes’ positions on different ranking lists and
in soccer former performances were mostly based more on a subjective picture that the coach had of the athlete. The ski coaches brought up some concerns regarding the validity of the ranking lists, which are based on fairly advanced calculation systems. They stated that the validity of these lists varied depending on what context they were used in and that because of this some of the coaches barely paid any attention to them, while other coaches based most on their selections on these ranking lists.

In both sports the importance of getting the team to function well together was mentioned and this was also a factor that was said to influence the selections in both sports. In soccer it was said that it was not always the best players regarding the sport skills that were to be selected but the players that made the best team together. The soccer coaches pointed out how important the players’ personality and behaviour are and that they would not select players they know will behave badly. Quite a few coaches from both sports said that they would choose an athlete with good behaviour and favourable personality over an athlete with better sports skills, if they had a possibility to do that depending on the selection system that was used. All coaches emphasised the importance of the selection criteria being formulated in such a way that they gave the coaches opportunities to influence the judgements themselves.

Regarding the selection process all skiing teams had formal meetings where they discussed which athletes were going to be selected. Those selection meetings were said to contain thorough discussions of the athletes. In soccer some of the coaches have more informal and some have more formal meetings about the selections. The opportunity for a lot of experts to have their say about the skiers might possibly enhance the quality of the selections. It is difficult to know to what degree the coaches in this study come up with the same names in the selections independently of each other, as this is never really visible in the selection processes. It is also hard to say something about how much they influence each other during the selection decisions or to what degree the selections are made with the selection criteria as the basis. In the skiing team the head coaches or similar functionaries are the ones who make the final selection decisions (in some teams the decisions also have to be approved by higher bodies in the federation).

As the head coaches of the national skiing teams spend a lot of their time on tour with their teams, their opportunities for seeing other skiers than the ones that are already in the team are limited. Information about those athletes is instead provided by the coaches of these athletes and from result/ranking lists. When the soccer coaches recruit new players they trust to some extent other coaches’ and agents’ statements. In soccer it seemed that most coaches tried to get a lot of information about new players by travelling to see them, talking to them and by looking at video recordings of games they were in. They seemed to have a fairly good knowledge of the new players they selected. One soccer coach
differs from the others as he says that he makes all the selection decisions himself, while the other coaches say they decide the selections together with their assistant coach.

All coaches stated that their eyes and their feelings were their most valuable selection tool. These are “subjective tools” in comparison with test results, statistics, result lists and other more objective tools. The favouring of the “subjective tools” is interesting from a validity and reliability perspective. It raises questions of whether human performances are too complex to measure by means of objective and measurable factors, whether statistics and tests cannot measure human performance well enough, or whether this has to do with the coaches’ own wishes to control the selections.

External pressure in the selections was not seen as a big issue by the skiing coaches, although it was mentioned that parents sometimes tried to influence the coaches. In soccer, on the other hand, the external pressure on the coaches was perceived as big. The soccer coaches who continually made selections to games every week talked about the media, sponsors, agents, friends and the general public as having opinions about and wanting to influence the selections. All coaches seemed to have fairly formalised ways of informing the athletes about the selections. They also welcomed athletes to discuss the selections with them, but were clear that they would not change the selection decision due to these discussions. As regards the process of selections, the skiing coaches did not have many reflections concerning possible gender issues, while all soccer coaches had some thoughts about gender and its effects on the selections. It seemed fairly clear that some of the soccer coaches were of the opinion that both their own gender and the players’ gender had some influence on the way the selections were carried through.

When it comes to athletes’ concerns and reactions in the selection process, the coaches have different opinions about how much the athletes need to know about the selection criteria and the reasons for the selections. The different opinions seem to be related to how they are currently working with the criteria. It is interesting that all coaches said that their athletes did know about their selection criteria, but that only a few of them could describe what the athletes knew or how they would have obtained this knowledge. If the athletes do not know very much about the selection criteria and if for this reason they value other abilities then the coaches, this may be a validity issue in the selections. Concerning the fairness of the selections, quite a few of the coaches stated that they thought the athletes who were selected felt that the selection decisions were fair, while the athletes who were not selected felt that the selection decisions were unfair.

When it comes to the outcome and the consequences of the selections, the importance of the selections was seen as very big. To be selected to a national team in skiing was by some coaches described as a career-changing event,
whereas the soccer coaches described the selections in soccer to be very important although not crucial for the athletes’ career. For the coaches themselves it was also said to be of utmost importance to select a successful team, as they otherwise could lose both their job and their respect. Most of the skiing coaches thought they selected the “right” skier, even if the athletes did not always achieve the expected results, while all the soccer coaches thought that they sometimes had selected the “wrong” player.

Overall the results indicate that there are validity issues to be considered in the selection processes. It was also shown in this study that many of the coaches regard the selection process as rather difficult and the evaluation of the process as challenging. The fact the coaches do not have any education concerning selections and how to deal with factors affecting the selections means that they have to figure out strategies for dealing with these complex situations.
DISCUSSION AND CONCLUSIONS

In elite sport athletes and teams compete against one another with the goal to be the best. In this thesis I have studied rule and judging systems, discussed judgements and measurements that are the basis of the athletes’ results, and searched for knowledge about how selections to top-level sport teams are performed. I have analysed these processes from a validity perspective and thereby discussed issues that play a role for the fairness, utility and usefulness of the outcome of these processes. By doing this I have aimed to contribute to the discussion of whether the “right” athletes are selected to participate in teams, competitions and games and whether the “right” athletes are winning.

In practice the selection process and the measurements and judging of athletes’ in competitions and games are dealt with as separate, although sometimes connected, processes and this is also the way I have worked and described them in my studies. Although these processes are dealt with as separate, they also have many similarities, as both of them are based on evaluations of athletes and their performances. In this final chapter I aim to compare some of the results from the judging and selection processes that I see as being of high importance for the validity within these processes. By pointing at how similar concerns are dealt with in these different processes, and in the different sports, my aim is to contribute to an enhanced understanding of these kinds of processes but also to an enhanced understanding of the role validity and validity concerns have in measurements, judgements and selections in elite sport.

Benefits of and drawbacks to well-defined criteria

What athlete and performance are regarded as the “right” to select and give high scores to is connected not only to the athlete and the performance per se, but also to “the idea of the sports or the selections”, the current rules and criteria, i.e. the grounds for these decisions, as well as the experts’ opinions and use of the above. Because of this, the content as well as changes in these grounds play an important role for the outcome of these processes and are crucial in the discussion of whether the “right” athlete is selected and whether the “right” athlete wins. Hence, the “right” athlete should here be understood as the athlete that are the best athlete according to the selecting and judging criteria and other eventual grounds influencing the decisions.
If the criteria that are the ground for the judging or selections are pre-set, well thought out, well-defined and publicly displayed, it could be argued that this will raise the validity of the judgements, as it will then be clear to everyone what is going to be valued and how different abilities are weighed against one another. The risk of arbitrary or biased judgements will also be diminished when the decisions are based on well-defined criteria. Due to this kind of reasoning, the opposite, undefined or implicit selection criteria or grounds will then be treated as a validity issue in the sense that it is unclear what really serves as the basis of the judgements, which would increase the risk of arbitrary or biased judgements. It has been argued that the development phase of a test or the like is crucial for whether it will be successful (Lissitz and Samuelson, 2007b), but this development phase is not possible to value in the cases where the criteria are implicit and or undefined. Based on my studies and other research it has nevertheless been shown that the question concerning the level of validity in these processes is far more complex than the reasoning above, especially as the modern validity concept includes the outcome and the consequences of the judging processes and not only the content of the criteria.

The rules and judging criteria in acroski and RG are decided by the international governing bodies, and the same rules and criteria are hence used in elite level competitions worldwide. Even though the detail level of the rules and judging criteria were rather different in the studied sports, the rules and criteria were defined, official and stated in documents in both sports. Both acroski and RG developed considerably during the studied time period and this led to a number of rule changes in these sports. One of these changes was that the criteria decisive for the judgements developed towards more and more well-defined criteria. The reason for this was the demands for objectivity that were raised both within the sports and externally, as for example by the Olympic Committee (acroski). More detailed criteria specifying e.g. how much every spin, leg lift etc. should be rewarded, and thus elucidating how much every part of the performance was worth, were aimed to increase the objectivity of the judgements and the sport’s credibility. The focus on reliability issues as a way to improve the judging in sports has been prevalent in other studies too (Bassett & Perskey, 1994; Wu & Yang, 2004; Harding, Toohey, Martin, Mackintosh, Lindh & James, 2007; Harding, Toohey, Martin, Hahn & James, 2008).

The interviewed experts from acroski and RG pointed to some problematic issues with these well-defined criteria, when stating that this led to stereotype performances when the athletes applied these rules and criteria to their performances. To receive high scores the athletes were forced to include the exercises that were a part of the rules and criteria and perform them due to what gave the highest scores, which meant that the athletes’ performances became more and more alike. This was far from the idea of these sports, where creativity, personal expression and freedom in the performances were the core of the
sports. Nor did the experts agree exactly with what was valued in these well-defined criteria. The validity evidence based on test content was hence quite varied and should be treated as low in the cases where the interviewed experts thought that what was judged was not in accord with the idea of the sport. In these cases the well-defined criteria and rules did not lead to greater validity but rather to the opposite, and the consequences of this could be that other athletes than the one they thought should win were given the highest scores.

These rule changes, which were made with the aim to improve the judgements, showed that although the reliability of the judgements was perceived as improved, the validity as a whole was negatively influenced. These kinds of unplanned results of rule changes are problematic if they affect the validity of the process. In other studies it has also been discussed that the intentions of rule changes and the consequences of them are not necessarily the expected ones (Miah, 2000; Wu & Yang, 2004; Harding, Toohey, Martin, Mackintosh, Lindh & James, 2007; Harding, Toohey, Martin, Hahn & James, 2008).

These findings concerning the definition of the judging criteria can be compared to the discussion of judged sports in the X-Games. The X-Games is a competition for athletes in extreme sports. Many of these sports are judged sports, and have had rather vague judging criteria, and it has publicly been expressed that it has been as a risk that using well-defined criteria might stifle the development and creativity that are the core of these sports. The more recognition, more prize money and more sponsor contracts are involved in the X-Games, the higher will be the demands for fair competitions. These demands come from the athletes but also from the sponsors, and the judges’ criteria are therefore up for discussions more and more in these sports too (O’Brien, 2010). To follow the development of the rule and judging criteria in these kinds of sports, where many have shown resistance to the organization and regulation of the activities, but which now might be forced to be more regulated due to the demands for fair competitions, is something that would be of interest to study further and compare more thoroughly with the results from my studies of judged sports.

In the studied teams in soccer and alpine skiing the coaches were the ones that selected the athletes (although the boards or similar functions were also involved in finalising the decisions when the teams/squads were selected), and they are therefore referred to here as the selectors. In the selection studies the criteria relate to certain skills, abilities, characteristics or the like the selectors look for in the athletes. I have chosen to use the term ‘criteria’ as an analytic concept, even if some selectors used the word criteria, some talked about guidelines and yet others about bases or grounds for the selections.

Unlike the criteria for the judgements in acroski and RG, the criteria or grounds for the selections were not stated by any international governing organization but by the federations and clubs themselves, resulting in a variety of
different criteria and systems. Some of the selectors worked with criteria or grounds that were preset, official, written down and publicly displayed, whereas in other teams the selectors worked with guidelines that were written down but not official for anyone else than the coaches. There were also selectors who stated that they did not work with pre-defined official criteria or guidelines. Here it should be noted that this does not mean that these selectors make random selections based on nothing, but that their decision grounds are not clearly enunciated. The majority of the selectors in the alpine teams and a few of the selectors in the soccer teams said that they used well-defined criteria in the selection process. It was mentioned that some coaches would select other athletes than the ones that were now selected, if they had a little bit more freedom in the selections then they now had due to pre-set criteria. This means that the pre-set criteria sometimes restrict them to selecting the athlete they have the most belief in. Whether the selectors worked in teams where pre-set selection criteria or guidelines were used, or whether they worked in teams where more subjective grounds for the selections were used, they all pointed out how important it was that they had some kind of acting space where they could influence the selections according to their feelings and their intuition about the athletes. Instead of fixed criteria the selectors here talked about the importance of using their intuition and their “eyes” in the valuing of the athletes.

It might be questioned why the selectors do not proceed from the same criteria that are decisive for who will win competitions and games as the criteria when athletes are selected to teams, but in analysing what the selectors see as important in the selection process, it becomes evident that other skills and abilities than the ones directly linked to the sporting performance have to be regarded in the selection. The selectors in my study argued that the sport skills were crucial for the selections, but quite a few of the selectors also stated that they would choose an athlete with good behaviour and favourable personality over an athlete with better sports skills, if they had an opportunity to do that in terms of the selection system. How well the athletes work together plays, according to the selectors, a big role both in the team sport of soccer and in the individual sport of alpine skiing, where the athletes travel and live together for long periods of time. It was brought up that selecting athletes who would create homogeneous teams was important and that this sometimes became a decisive factor in the selections.

The ability to act and behave according to team norms and values and to work within the team was seen to be of major importance in both sports. The value of behavioural skills and personality is here to be compared with general mental abilities (GMA) in personnel selection. These abilities, which are a combination of intelligence and cognitive abilities, are together with work samples shown to be the most valid predictors for personnel hiring and selection when the workers have previous experience of the job (Hunter & Hunter,
and in job advertisements skills such as flexibility, leadership skills, ability to work in a team and similar abilities are often asked for.

The mental skills of an athlete have also in a number of sport-related studies shown to play an important role for the athletes’ performances (see for example Gould, Guinan, Greenleef, Medbury & Peterson, 1999; Gould, Dieffenbach & Moffett, 2002). Further support for the selectors’ ideas about other skills and abilities than the sport-specific ones as decisive for the athletes performances is provided by the research of Abbott & Collins (2004), where it is stated that while athletes’ physical characteristics have been able to explain differences among athletes, only psychological factors seem to be able to explain the maintained success. My research here also confirms previous findings that psychological factors are highly valued in the practical field of sports (Morris, 2000; Abbott and Collins, 2004). It should be noted that the behaviour and personality of the athlete were said to play this role in the selections to the teams, but when the selection to a championship competition is the issue, it seems to be the most recent results that are the most decisive for the selections. There the athletes’ most recent performances and sometimes their experience of former championships are what seemed to be of the most importance.

The development towards more and more defined criteria in the judged sports and the above discussed need for subjective valuations in the selections can be compared to how qualitative assessments of students’ performances have been dealt with in higher education. Since the mid-1990s the assessments have there become more structured with an aim to decrease the arbitrariness of the assessments and also to make the assessment grounds and procedures more explicit to students (Sadler, 2009). Sadler discusses how analytical grading (qualitative judgements based on preset criteria) has increased because of this, but argues for some benefits of holistic grading (an assessor’s more holistic complex mental response to a student’s work). Some of the problems he highlights in the analytical grading are how this system fails to capture special characteristics and the problem of weighing these criteria against one another and why just these criteria are the prevalent ones. The perceptions of the experts in my studies regarding how the more specified criteria lead to stereotype performances are in line with his arguments about how the analytical grading fails to capture the special characteristics in the performances. In RG, scores for bonus points and “risk manoeuvres” were introduced because of this as a way to keep the creativity and the originality intact. In acroski, changes in the content of the artistic score were made to solve this problem.
The weight of different criteria and grounds

The relation among the different judging or selection criteria (validity evidence based on internal structure) is of interest, as the outcome of the judging or selection is also dependent on how much value the different criteria have in relation to one another. The way in which some skills are treated as critical in the judgements and selections and others are not refers to which values are the dominating ones in the sport and/or in the teams. These values might be very clear and well thought out, but they might also be less well planned and thought out. How the different parts of the performance were valued changed over time in acroski and RG. In both sports the relation between the technical difficulties and the artistic impression/choreography changed so that they became more equally valued. These changes were perceived as positive by the interviewed experts in the sports. In acroski it was mentioned that despite this, the technical parts were found to play a bigger role in the outcomes of the competitions. When I studied the scales for the scoring, I discovered some problems related to them, as the scales had different maximum scores. The consequences of this were that the judges spread the points on the technical scale much more than they did on the artistic scale, resulting in the technical difficulties being the ones decisive for the outcome. This was a clear validity issue, as the rules then did not measure what they were supposed to measure.

In the selection study it is harder to discuss how the different criteria or grounds for the selections weigh against one another, as only a few of the teams had made their selection criteria and the relation between them explicit and clear (sources of validity evidence based on test content, evidence based on internal structure and evidence based on relations to other variables). The relation between statistics, rankings and the like and the more subjective valuations of an athlete’s sport skills, form, behaviour etc. is interesting here, but this is also hard to discuss for the same reason. The selectors agreed that the selections should include some element of subjective valuation and that statistics, test results and similar factors were not enough grounds for their selection decisions. In comparison to the former discussion about holistic assessments, these results also point to a similar need where subjective assessments based on many different factors are seen as important in the selections. These opinions concerning the insufficiency of only using rating instruments also correspond to earlier findings where rating instruments were criticised for not being valid, at least not when they are used in standardised settings (Nadeau, Richard & Godbout, 2007).

In the ski teams that used explicit criteria, it was evident that the criteria to a great extent focused on former results as being more valuable than other skills and abilities, and in the other ski teams former results were also said to play a big role. In skiing former performances were often measured in terms of the athletes’ places on ranking lists, and the use of these ranking lists were for a few
teams a clear exception to the above mentioned resistance to statistics as a
ground for the selections. Here it should be mentioned that especially the FIS-
ranking list that was often used was said to have different validity depending on
what context it was used in, and that for this reason the selectors used the list
rather differently in their selections. This ranking list is based on a quite com-
plex calculating system where the points given to the athletes are dependent on
how many top racers there are in the race and the time differences among the
racers. The consequence of this is that it is very critical what athletes participate
in the competitions, and as the number of highly ranked athletes is different in
different nations, the validity of the list was perceived to be very context de-
pendent. In soccer former performances were also said to be of importance, but
what was included in the judgements of these performances was based on the
coach’s subjective picture of the athlete. Previous performance and previous
results are factors that in other studies too have proved to be crucial for the
prediction of future results (McDaniel, Schmidt & Hunter 1988; Boulier &
The above focus on former results should here be related to the role that the
behaviour and the personality of the athlete was also said to play. In relation to
this it should also be noted that some of the selectors had stated that the behav-
iour and personality were of the highest importance. These rather contradictory
results point to the complexity of the variables to be taken into consideration,
but they also indicate that the weight among them did not seem to have been
totally elucidated, at least not in the minds of the selectors.

Some of the selectors said that they had reformulated some of the main goals
that were set for the teams and the selection, as they had some difficulties in
applying the pre-set goals in their practical work with their teams. Although the
new goals might be more valid and functional, this reformulation also influ-
ences what criteria the selections are based on. The use of different goals and
hence different criteria is a validity concern, as this means that the selectors
might have partly different grounds and reasons for their selections than the
clubs or federation boards. It is also of interest here which of these goals (if
any?) the athletes might be aware of. These reformulations might result in dif-
ferent selections than if the original goals had been used. In other areas there is
substantial research on how goals and incentives are determinants for the out-
come in performance ratings and how possible conflicts between the official
purpose of appraisal systems and the way raters use them affect the outcomes.
In a study by Murphy, Cleveland, Skattebo & Kinney (2004) where students
rated their teachers, the results show that raters who pursue different goals when
rating performances tend to give different ratings even when the have observed
the same performances. In connection to this it should be noted that criteria
that are vaguely enunciated most likely increase the risk that different incentives
influence the selection decisions.
In my results it was evident that many factors played a role in the selections, and teams selections can without exaggeration be regarded as very multifaceted processes. The coaches and especially the soccer coaches talked about the complexity in soccer and how the players have to be versatile and have skills in many areas. This is very similar to research on personnel selection, where job analysis is identified as problematic when more and more aspects have to be taken into consideration, as jobs are becoming more and more difficult to describe due to their complexity (Robertson & Smith, 2001).

Interactions between judges and selectors and bias concerns

The athletes’ performance is the main focus in both the judging and selection process. In competitions and games it is the current performances that are measured and judged, although the subjective part of judging can never really disregard that former performances also play a role. In the selections the former and current performances are used to predict the future performances. The number of experts involved in these judgements and their respective interactions and roles are somewhat different in these processes. The number and the roles of the judges in the competitions are clearly defined, but when it comes to the selectors, there did not seem to be any rules stating a set number of selectors, and their mutual relations were seen to be partly flexible. Consequences of this are for example that in the selections it is possible that one person selects the team, and this was also done in one of the teams in the selection study. From a validity perspective this is questionable, as it makes it difficult to evaluate the reliability of the judgement. From a practical perspective this does not mean that those selections might be better or worse than if more selectors had been involved, but they are very dependent on the knowledge and experiences of the single selector.

To decrease the risk of the judges in acroski and RG influencing each other’s scores, the judges have to give their scores independently of each other and will be punished if they talk with each other in the judging process. This is done to provide for high reliability, which will be reached if the judges’ independently of each other give scores that are very close or the same as each other’s. These scores are also official after the competitions.

In order to diminish the risk that biased judging or judging mistakes will have crucial impact on the athletes’ results, different systems for preventing this were incorporated in the judging systems in the studied sports. An example of this is how the rules were changed so that, instead of judging all parts of the athlete’s performance, the judges could concentrate on judging only smaller parts to improve their opportunities to judge according to the rules. Rules stating how the different parts of the performances are to be valued in relation to
each other, how big differences there are allowed to be among the judges’ scores, rules concerning the elimination of highest and lowest judging scores, and how a tie between athletes should be solved are other examples of this. The interviews revealed that the judges in RG many times were assembled, because their scores were too far away from each other in comparison to what the rule system allowed. The judges then had to either agree on scores that fitted within the allowed limits or a so-called “base point”, a mathematical formula used to calculate the judges’ scores, was used. This and truncation of the highest and lowest scores were used as methods to prevent mistakes or biased judging. In the licentiate thesis I questioned if these systems could also lead to strategic judging and how well these methods really increase the reliability of the judgements. Nationalistic bias has been studied in some articles, and Zitzewitz (2006) brings up this problem when discussing whether the truncation of extreme scores can make bloc voting and vote trading an issue. Concerning biased judging in RG, Popovic (2000) considered this to be an issue, as he found that the judges had a tendency to favour gymnasts from their home country. Nationalistic bias was also found in studies of diving competitions at the 2000 Olympic Games (Emerson, Seltzer & Lin, 2009).

In the selections the decisions are mostly a result of discussions between the involved coaches. In practice this means that the selectors can hide behind “common agreements” in comparison to the judges who are solely responsible for their scores. Whether the coaches’ opinions of the athletes are very distant from or very close to each other is hence not visible in these processes. A safety net preventing biased decisions does not seem to exist in the same way in the selections, but here the different coaches’ arguments and discussions with each other could be interpreted as a system for preventing biased selections from occurring. If there are discussions, the head coaches seem to be the ones that have the last word, but in the alpine teams the head coaches decisions had to be finalised by the federation, the board or some similar organization. In the soccer teams the boards or similar organizations seemed to be involved mostly in the selection to the squad, while the selections to the games mostly were the selectors’ (coaches’) responsibility. These discussions were perceived by the selectors as working well in most cases.

To follow these discussions would be a way to study and understand more about how the final selection decisions are taken. It has for example been discussed in the literature how different incentives for persons involved in judging and other decision processes influence their way of taking decisions (Prendergast, 1993; Murphy, Cleveland, Skattebo & Kinney, 2004). Prendergast argues that subordinates have an incentive to conform to the opinion of their superiors. Since the consequences for the selectors are said to be very serious if they do not succeed in their job and their selections, the incentives to follow the wishes of their subordinates should theoretically be serious too.
The relations among the head selectors (head coaches) and the other selectors (coaches) would therefore be of interest to investigate more. Most of the selectors stated that there were good relations among the boards of the federations or clubs, but some also said that there had been some controversial requests and decisions from the federations/clubs that they found hard to handle. When it came to possible pressure from external forces such as for example media, agents, sponsor and parents, there did not seem to be any prepared way to handle this, and it was instead dealt with by individual solutions and coping strategies by the selectors.

**Education of judges and selectors**

I interpret the conditions of the judges and the selectors to be quite different and something to consider in the discussion of these processes. To be allowed to judge a competition in acroski and RG, judging educations are obligatory. The purpose of these is to secure that the judges have acquired knowledge and ability of evaluating the athletes’ performances in accordance with the current rules and criteria. The selectors in my studies do not have any educations in how to perform selections, and according to them there are no educations or education materials concerning selections. This means that the selectors instead learn how to make selections and what skills to value through their own and possibly other’s experiences. If there are no available educations, one consequence is that problematic issues might be treated differently among different selectors depending on their personal ambitions and skills, their interest in selections and what experiences they have from former selections.

According to my studies, being a judge in acroski means that judging is the judges’ only and therefore also prioritised task, whereas being a judge in RG, or a selector in soccer or alpine skiing also means being a coach. These different arrangements have consequences involving both advantages and disadvantages. If a judge or a selector is too much involved with the athletes, there is a risk that other factors than the ones in the criteria will be decisive, as there is then a risk that emotional ties and other feelings will affect the judgements. A disadvantage mentioned in the acroski study was that besides judging acroski, the judges there also had to be judges in the two other freestyle disciplines, something that was perceived as taking away the focus from the acroski judging. It was also mentioned to be a disadvantage that the judges were not very much involved except when judging competitions (as they were not allowed to) and that they did not spend so much time watching the sport live. In RG, soccer, and alpine skiing the coaches (which as mentioned also were the judges and selectors) spent time with the athletes on a daily basis and therefore had a great deal of knowledge of the sport.
The selection processes were said to involve quite a few situations that were difficult and challenging for the selectors to handle. Often this seemed to be connected to the impact that bad or wrong selection outcomes could have on those involved. According to the selectors, the selections were of major importance for the athletes in soccer and alpine skiing. In alpine skiing, some of the coaches saw selection to a national team a “career changing event”, as this involved many opportunities for the athlete. The selectors also described the consequences of the selections as something that could also be serious for themselves, and they said that they could lose their jobs because of this. To handle this pressure, some of the coaches said that not thinking about the possible consequences of bad selections was a way to cope with this pressure. To study further what role the consequences of the selections has for the selectors could be a way to understand the selectors’ choices and incentives in a more complex way.

As reported earlier, technical sport skills, physical skills, tactical skills, mental skills, social skills and behaviour and psychological abilities are all skills taken into account in the selections. To manage the selections, the selectors should have a good knowledge of and experience in working with developing selection criteria involving all these skills, assessing all these skills, handling external pressure but also have knowledge of validity issues to be able to secure high quality selections. With these fairly high demands on the selectors, I argue that it is of great importance to create opportunities for them to handle these situations well. Many coaching educations have traditionally had a focus on sport skills rather than on valuing these kinds of skills and abilities. From a validity perspective the selectors’ possible lack of education and experience in valuing behavioural and psychological skills is hence a concern for the validity of the selections. This is also identified and discussed by Humara (2000), who sees the coaches as experts when athletes’ physical characteristics are identified but does not believe that they have the sufficient psychological skills to validate athletes’ psychological factors. The knowledge the selectors have of the different skills they are valuing in the selection process might here influence the weight that is given to them. In a validity perspective this is an issue if the selection outcomes are influenced by the selectors’ knowledge of valuing the different skills and abilities.

How the coaches consider the different skills and what opportunities they see for developing these skills may also lead to different selection outcomes. It is therefore interesting that some coaches state that they would hesitate or not at all choose to select an athlete with a personality or behaviour that does not fit in with the team values and norms. Does this mean that personality and behaviour in the view of the coaches cannot be developed and changed? Looking into research about this, more and more studies state that much of the personality is flexible and dynamic (Mischel & Shoda, 1995) and something that changes
over a life span and is shaped by experience (Roberts, Walton & Viechtbauer, 2006). This means that these factors should also be possible to practise and develop through proper training. What skills the selectors believe are evolvable and what kind of knowledge they have about the different areas they consider in the selections are aspects that would be interesting to research further. Some teams in the study worked closely with a behavioural scientist or psychologist as support in these issues.

Face validity and the practical value of criteria

Even if face validity is not an empirically supported form of validity, I argue that face validity has a considerable impact on the measurements and judgements in elite sport. In the results from the studies of judged sports it was clear that the reliability issues are in the main focus and that objective judgements were a principal goal, because it is of great importance for all involved, but also for the credibility of the sport, that the measurements and judgements are perceived as fair. In research on personnel selection the value given to the applicants’ reactions to selection processes (see for example Gilliland, 1993) also points to face validity as a considerable factor in selections. How the selection processes are performed and what knowledge the athletes have of the criteria and the processes could probably affect how the selections are perceived by the athletes, as it has proved to do in personnel selections (Ryan & Ployhart, 2000), something that may have consequences for the selections.

In acroski and RG it was stated that knowledge of the rules and criteria was a key factor for successful performances, but the opinions about how much the athletes need to know about the selection process varied among the selectors. The lack of information that some athletes seemed to have about the selection grounds due to vague or unofficial criteria could probably affect their feelings and reactions concerning the selections. It might also make them focus on the wrong things in the selection process. In the often-cited 10-year rule, according to which it takes 10 years of deliberate practice to be able to perform at world-class level in any field (Ericsson & Lehmann, 1996), informative feedback is argued to be crucial for the athletes’ development. In sports athletes daily get internal and external feedback on their sport skills, which means that their knowledge of these factors is probably fairly extensive, but in the light of the above reasoning it is interesting to highlight what kind of and how much feedback the athletes get concerning behaviour and mental abilities. If these factors are as important as mentioned, the feedback should from a validity perspective be evenly correlated to these skills in order for the athletes to know what to develop and work on. If the selection criteria and selection process are not clear to athletes that are trying to get a spot in a team, they might focus on the
“wrong” factors, or at least this might limit their possibilities to be tactical and focus on the factors that will contribute to their chances of being selected.

How the athletes are informed about the selections and what possibilities they have to argue against the selection decisions may also play a role for the face validity and the athletes’ feelings about and reactions to the processes. In all the teams the procedure of informing the athletes was fairly formalised. For the soccer selections it was for example almost always organised so that the athletes were informed about the selection decisions a day before the game. This was done to give the athletes opportunities to react and adapt to the decisions. All athletes were welcome to discuss the selection decision with the coaches, but none of the coaches said they would change their decision due to these kinds of discussions. Stevenson (1989), who studied how athletes perceived the fairness of selections to national sports teams, stated that athletes’ perceptions of the fairness of the selection outcome were related to the selection procedures. He pointed to the importance of establishing competency in the selectors and to a perceived objectivity in the selection criteria and ways to reduce bias and favouritism. In research about personnel selections, Gilliland (1993) stated that it is likely that the experiences the applicants have during a selection process also affect the attitudes and behaviour of the hired individuals and that this may even shape the organization’s climate. The athletes in the judging study felt that it was hard to accept that the judging was not always felt to be fair, which meant that the results in their view were not always right. They said that they worked to accept that these are human judgements and that they may not be totally fair. It would be interesting to study further how the athletes perceive selections and the rankings in competitions and what implications this might have for them and their behaviour.

In the understanding and developing of criteria, the practical value of the criteria also plays a role and this is also connected to the face validity. In competitions and games on the elite level, the criteria are decided to be the same worldwide, but when it comes to the selection criteria the practical value of the selections might affect how the criteria are developed and defined. When it comes to the practical value of valid selections, the context is, according to Kane (2006), of main importance. Because of the anonymity in my research, I have not discussed the context further, as this could have revealed what federations and teams I have studied, but the influence of content is critical for understanding when selections are discussed. Context-dependent factors such as the economic frameworks, the kind of sport, the society in which the selections take place and the selection processes might influence the selection criteria. This means that selection criteria that work extremely well in one context cannot be transferred to another context and be expected to reach the same validity there.

Another important determinant when it comes to the practical value of valid selections is the variability of the athletes’ skills. The variability of skills in my
study refers to the level and skills the different athletes have in comparison to
each other. The selection ratio, referring to the number of athletes that can be
chosen in comparison to the available number of spots in the teams, is also to
be regarded as a determinant of the practical value of the selection method
(Schmidt & Hunter, 1998). In soccer there are rules for the maximum number
of players in a team, and in alpine skiing the number of athletes that are allowed
to compete is regulated due to the athletes’ rankings. The number of spots in
the team is not correlated to the number of available athletes the selectors can
choose from, and this makes the practical value of valid selection methods extra
important when there are many athletes competing for few spots.

The selections to the Olympic Games can be seen as an example of when the
need for a valid selection method is extra important. In alpine skiing only four
athletes per discipline and nation can be selected to the Olympics. In teams
with many good skiers the consequences of this could be that a skier who is
ranked high on the World Cup standings (the current ranking of the athletes
participating in the FIS Alpine World Cup) and the WCSL-list (a ranking list
based on top 30 finishes in competitions such as the World Cup, the World
Championships and the Olympics) will not be selected if there are other ath-
letes from the same country that are predicted to perform better at the Olympics,
or have a better placing on these lists. For the teams that have many good
athletes, the process of selecting athletes to an Olympic Game is therefore extra
critical. If a selector has a number of skiers that have proved they can make
good results, and s/he has to choose only four of these skiers, the demands on
the selection criteria and process are very high. These selections matter a lot to
the athletes, the coaches, the sponsors and others and are of public interest and
very probably will be discussed in the media, which is a factor that makes them
even more critical. In my studies former results were said to play a big role for
these selections, and this would also be clearly understandable from a perspec-
tive of face validity. Still, some selectors said that these selections were much
more complicated than this, as having the results as the only or a very strong
indicator had it problems. They discussed what competitions these selections
should be based on and in what time frame the results were to be considered
but also how sickness, temporary problems with for example mistakes in the
runs, equipment issues, waxing or effects of the social environment were to be
dealt with. The pressure on athletes that are “on the edge” to make it might also
be higher then on the ones that are the absolutely best. To study further these
kinds of critical selections to championship competitions and the factors affect-
ing them would be interesting, as in my studies I did not do this thoroughly.
The consequences of a system where every nation has a restricted number of
spots not correlated to the numbers of athletes competing at the highest level,
might lead to the absence of some of the top ranked athletes in the Olympic
Games. The incentives and the outcome and consequences of this system would therefore also be of interest to further investigate further.

In my study the differences in the selection criteria, selection grounds and the selection processes could partly be seen as dependent on the selection ratio. In the ski teams where the applicant pool (the available number of athletes to choose from) is big in relation to the total amount of spots in the team, the selection criteria and the selection process were much more clearly defined than in the teams with smaller applicant pools. This could be explained by the opportunities that the coaches in the teams with a smaller applicant pool have of gathering information about their skiers. In soccer, where the club teams in theory have soccer players from the whole world as their applicant pool, the above reasoning would then lead to very structured selection criteria and selection processes, as the practical value of a valid selection method then is high. This is the case for some teams but the variations among the teams are big. An explanation of this might be that the theoretical possibility to select athletes from all over the world does not correlate with the practical possibilities the selectors perceive that they have. Economic conditions and how well foreign players fit into the teams to some degree limit their practical possibilities of having the world as the selection field. This in its turn means that the practical value of a valid selection can be seen as lower in practice than in theory, but it can still be high if there are many athletes to choose among.

**Implications for practice**

Due to the findings in my study and other studies discussed above, some issues are especially worth considering and relating to in the practical work with judging and selection processes. Below I have summarised some key issues to take into consideration.

- **Criteria.** Criteria can be of importance if the judges and selectors are to assess the same skills and abilities. For valid judgements well-defined criteria are also an asset, as with clearly defined criteria the judges and selectors should judge the same skills and abilities. Well-defined criteria also make the processes more transparent and clear to everyone involved, which is positive for the face validity of the processes. Shortcomings of very clearly defined criteria for performance appraisal are that they are time-consuming and complicated to develop, as they have to cover complex skills and abilities. When well-defined criteria are used, there is always a risk that some special characteristics of some athletes do not fit into these criteria. The risk is that these special characteristics will then be overlooked. In sports these special characteristics might be what develops the sport and attracts the audience and to overlook these performances and/or athletes because of too narrow criteria would not be beneficial for the sport and those involved. The selectors’ strong statements that their own subjective judgements should be a part
of their decisions point to the importance of holistic decisions and opportunities to also value the athletes and their performances a bit more freely than would be possible with the use of (only) well-defined criteria. For valid selections to take part, this freedom of the experts to use their expertise should therefore not be overlooked although reliability concerns are to be considered.

- **The weight among the different skills.** How the different skills and abilities are weighed against each other and in what ways they are judged are crucial, as the outcome of the processes might change because of this. The consequences for those involved and for the credibility of the sports if some but not other skills are valued are crucial for the validity of the total process. How these criteria are defined affects both the possibilities to interpret them and how the different skills are valued in relation to each other. In my results, where the non-sport-specific skills like behaviour and psychological characteristics were given high importance in selections, the weight of these kinds of skills vis-à-vis the more sport specific skills is of importance to consider here.

- **Changes in the criteria.** Due to the development in sports, changes in rules and regulations are to be treated as a continuous process, if valid measurements and selections are to be made. With the existing technology where video recordings and different video programs are used in almost all sports at the elite level, many changes could also be possible to test and evaluate on already performed competitions and games before they are implemented. The need for updated rules and criteria is a key issue if the “right” athletes are to be selected as the winners.

- **The processes.** What kinds of actions are taken to provide for high reliability within the processes and how does this correlate with the overall validity in the process? As bias and internal and external influences always occur when judgements are performed, ways to handle this should be incorporated into the judging and selection process. Awareness and possible actions concerning the risks and benefits of judges’ and selectors’ involvement and personal connections with the athletes, and their conditions for dealing with external pressure should be recognised here.

- **The outcome and consequences.** Crucial issues for all involved are the outcomes and the consequences of these outcomes, which are related to all the different steps in the judging and selection process. Understanding and obtaining more knowledge of these consequences is from a validity perspective both needed and meaningful, if the aim is to reach a high level of validity. Here evaluations of judgements and selections play an important role. Without the selectors’ acceptance of the goals for the selections, reformulation of goals and other incentives for the selections than the original ones might lead to other outcomes and consequences than if the goals had been
accepted. The influence that the goals have on the total process of selections is therefore of importance to consider.

Concluding remarks
The results in this thesis have shown that the judging criteria in acroski and RG were changed towards more well-defined and clear criteria to be more objective, but that this became problematic for the overall validity of the judgements, even if the evidence for reliability was increased. The reason for this is that the new rules and regulations did not clearly correspond to the original idea of the sport. It also has been pointed out that well-defined criteria or grounds for selections could be helpful in many ways, but the expert selectors in both soccer and alpine skiing emphasised how important it was for them that some parts of the selections were based on their subjective valuations of the athletes and that they had some freedom in how they interpreted the athletes and their performances. It is notable here that reliability issues are given a great deal of focus in the judgements, but that they do not seem to be considered as much in many of the selection processes. By discussing and comparing these processes I have aimed to increase the understanding and awareness of how these different methods and strategies played a role for the validity in these processes. Analyses, evaluations and discussions of measurements, judging and selection criteria and methods are crucial from a validity perspective but also a way to bring different values, norms and working methods up for discussion. Discussions that might lead to an enhanced understanding of the consequences that the outcomes have for the involved persons. These discussions and understanding can be decisive for the development of the sport, as the measurements, judgements and selections play a major role for how the sport will develop.

Future considerations connected to validity issues are also how new technology, new rating systems and more and more available statistics from different measured parts of the sporting performances will be implemented into the sports word. Although incorporation of new technology in sports may possibly develop the sports and improve the quality of measurements and judgements, consideration and understanding of how this will affect the idea of the sport/selections are crucial from a validity perspective. How different skills and abilities are valued in relation to each other is also connected to the incorporation of technical tools. Some skills and abilities will be easier to measure with these kinds of tools and the relation they get to other skills and abilities may change the outcome of a process.

With many external forces trying to influence sports, an interesting and significant question for the development in sports is who “owns” the issue of defining the idea of the sports? Is it the sports federations? The athletes who embody the sport through living and performing it daily? Or is it the Olympic
Committee or perhaps the media or sponsors? Besides the considerations that are linked to the idea of the sport, a lot of other considerations can also be discussed from a validity perspective. Economic considerations are for example something that can affect validity in a larger perspective. The incorporation of new technical systems for judging could possibly increase the validity of the outcome for those that can afford to use them, but it might also decrease the inclusion of some federations and clubs to stage the highest level competitions and lead to other consequences for the sport then those wished for.

Measurements and judgements of athletes’ performances are the core of elite sport, but in comparison to research on assessments in higher education, personnel selections and similar related fields where measurements and judgements also are key features, the research on measurements, judgements and selections in the sports area is relatively limited. Through writing this thesis and with the use of validity theory as the theoretical tool for analysing the judging and selection processes, I have aimed to contribute to an enhanced interest in issues about measurements, judging and selections in the sports field. With this thesis I have shown that validity issues are of major concern in the discussions and understanding of whether the “right” athletes are selected and the “right” athletes win. The experts in the thesis revealed that it happened that the “right” athletes were not always selected or won, and some of the reasons and issues related to this have been discussed above.

In the literature about the validity concept it was described that the validity concept developed due to the demands of being useful in different areas and with all kinds of measures. In my studies it may also be noted that the developments of the theoretical validity concept and the practical judging and selection processes are somewhat similar. In the all-embracing validity concept a variety of validity evidence is needed to validate a test or the like. In the sport judgements and selections different kinds of evidence are sampled to give a “true” picture of an athlete. Earlier on a test could theoretically be valid if it correlated with a criterion or if the content was perceived to measure what it was supposed to measure, but with this developed validity concept a validation also includes the outcome and the consequences of the ‘test’. The same kind of reasoning can be seen in the judging and selection processes, where specific criteria are supposed to measure the athletes’ skills and abilities but where it is questioned whether these specific criteria are a valid ground for the assessments, or if they limit the athletes’ acting space and hence their performances. How the theoretical concept has developed into being more complex and focusing also on the outcome and the consequences may be compared here to the selectors’ statements about the importance of including their subjective valuations to achieve a more complex assessment of the athletes. By including subjective valuations, the consequences of the selections could be considered and handled in a way that the selectors felt was beneficial due to the current situations for
the team and the involved athletes. Using only the pre-set criteria gave them less flexibility to adjust the selections to the consequences of their outcome. Whether more defined criteria or more subjective valuations will be used in the future in performance appraisals in the sport world is an open question and something to study further. Will the earlier mentioned demands for reliable judgements, concerns for face validity and the possible implementation of new technology lead to more defined and exact criteria or will the importance of also incorporating special skills and characteristics into the judgements lead to more holistic judgements? Or is the combination of these two the solution?

I also see the results in my studies as being of interest for ‘measurements’ in other fields with similar problems and where the outcome and the consequences of the methods play a big role. Research on personnel selection, educational measures and selection to higher education has inspired me in my work and my aim is that the results of this thesis will be beneficial also to other areas than sport. Especially in highly competitive environments like for example the use of bonus systems, results from this study could contribute extended knowledge. As a final statement, I would like to claim that the modern validity concept with its focus on a number of different sources of evidence for validity including the focus on the outcomes and consequences of the process has been valuable to use as the theoretical framework of my studies.
REFERENCES


DECIDING WHO IS THE BEST


