Developing and Validating Self-Report Instruments
Assessing Perceived Driver Competence

Anna Sundström

Academic Dissertations at the Department of Educational Measurement
Umeå University
No. 5 2009
Abstract

The overall aim of this thesis was to develop and validate a self-report instrument for perceived driver competence. The thesis includes six papers and a summary. All papers focus on perceived driver competence from a measurement perspective; that is, how to develop an instrument for perceived driver competence and how to use and interpret the scores from the instrument in a reliable and valid manner.

Study I reviews how perceived driver competence has been measured in other studies and discusses these methods from a measurement perspective. Most studies have examined perceived driver competence by asking drivers to compare their own skill to that of the average driver. That method is problematic, since it is not possible to determine if drivers are overconfident or not, when empirical information of their own skills is missing. In order to examine if drivers overestimate their skills or not, perceived driver competence should be compared with actual driving performance.

Study II reports on the development and psychometric evaluation of a self-report instrument for perceived driver competence - the Self-Efficacy Scale for Driver Competence (SSDC). The findings provide support for construct validity, as the SSDC demonstrated sound psychometric properties and as the internal structure of the SSDC corresponded to the theoretical model used as a basis for instrument development.

In study III, the psychometric properties of the SSDC were further examined using an item response theory (IRT) model. The findings confirmed the results indicated by the classical analyses in Study II. Additional information was provided by the IRT analyses, as it was indicated that the scale would benefit from fewer scale points or by putting labels on each scale point.

In study IV, Swedish and Finnish candidates’ self-assessment accuracy was examined by comparing candidates’ scores on the SSDC and a similar instrument for self-assessment of driving skill used in Finland, with driving test performance. Unlike previous studies, in which drivers compared their perceived skills to that of the average driver, a relatively large proportion made a realistic assessment of their own skills. In addition, in contrast to previous studies, no gender differences were found. These results were also confirmed in study V, where the results from the Finnish instrument for self-assessment of driving skill were compared with the results from a similar instrument used in the Netherlands.

Study VI further examined the construct validity of a revised version of the SSDC, combining qualitative and quantitative sources of evidence. There was a strong relationship between the SSDC and an instrument for self-assessment of driving skills, providing support for convergent validity. No relationship was found between the SSDC and driving test performance. Explanations of the lack
of relationship were provided from semi-structured interviews, as they indicated that confidence in performing different tasks in the test are different from being confident of passing the test, and that the candidates are familiar neither with assessing their own skills nor with the requirements for passing the test.

In conclusion, the results from this thesis indicated that the choice of methods for assessing perceived driver competence as well as the quality of these methods affect the validity. The results provided support for different aspects of construct validity of the SSDC. Moreover, the findings illustrated the benefits of combining different methods in test validation, as each method contributed information about the validity of the SSDC. The studies in this thesis mainly examined internal and external aspects of construct validity. Future studies should examine procedural validity of the SSDC.

*Keywords*: Test construction, test validity, measurement, drivers, self-confidence
Acknowledgments

I wish to thank all those who have, in various ways, contributed to my work with this thesis. I would especially like to express my gratitude to my supervisor Widar Henriksson, for his never-ending support and encouragement, his well-structured feedback, and for guiding me through the jungle of psychometrics. I am also most grateful to Marie Wiberg, my assistant supervisor, for encouraging me to become a doctoral student in the first place, and for supporting me and providing me with helpful comments and feedback on my work. During these years you have become a good friend with whom I have shared many great times.

I am also very grateful to the co-authors of papers IV and V in this thesis; Sami Mynttinen, Esko Keskinen, Marita Koivukoski, Jan Vissers and Kari Hakuli, for our fruitful collaboration. For reviewing the English in my papers, I would like to thank Susanne and Dave Alger, Gunnar Persson, as well as Nick Sanders. For applying his excellent drawing skills on Figure 1 in the summary, and for helping me with the cover-layout, I want to thank Björn Sigurdsson.

I also want to thank the entire staff at Vägverket [Swedish Road Administration], who have been involved in the development and use of the self-assessment instrument, for great cooperation. I especially, want to thank Inga-Lill Bogefors, for giving me the opportunity to take part in the project for a new driver testing model, and to administer the SSDC to driving-license candidates within the project. I also want to thank Gunnar Holgersson, for giving us the opportunity to conduct pilot studies at the test centre in Umeå, and for being so enthusiastic about these studies. I am also very grateful to all driving-license candidates that have answered the questionnaire and participated in interviews, without you this thesis would not have been possible!

Many thanks to my friends and colleagues at the Department of Educational Measurement at Umeå University for creating such a great atmosphere; it is a pleasure working with you! In particular, I would like to thank my fellow doctoral students Per-Erik Lyrén, Tova Stenlund, Anna Lind Pantzare, Peter Vestergren and Gunilla Näsström, for the great discussions about matters big and small at our doktorandfika, and for providing me with recipes, movie-tips, as well as for reading and providing helpful comments on my work. I would also like to give a special thanks to Lotta Jarl, Simon Wolming, Anna Lind Pantzare and Per-Erik Lyrén for contributing to my physical (and perhaps also mental?) health by regularly taking me to the sports centre IKSU.

Finally, to all my friends and family, Fredrik, my parents, my two younger brothers and my parents-in-law, for your love and constant support, and for always reminding me of what is important in life!

Anna Sundström

Umeå, October 2009
List of Papers

This thesis is based on the following papers:


All published papers are reprinted with the permission of the respective copyright holders.

---

# Contents

1. Introduction 1  
   Disposition of the thesis 1  
2. Self-assessment 2  
   Metacognition and self-regulation 2  
   Self-assessment of competence 4  
   Self-assessment accuracy 4  
   Self-assessment of driver competence 5  
3. Assessing perceived competence 8  
   Constructs of perceived competence 8  
   A construct for perceived driver competence 9  
   Instrument development 10  
4. Validity theory 12  
   The traditional view on validity 12  
   A modern perspective on validity 13  
5. Method 16  
   Procedure of administration, scoring and feedback 16  
   Methodological considerations 17  
6. Summary of studies 19  
   Study I 20  
   Study II 20  
   Study III 21  
   Study IV 21  
   Study V 22  
   Study VI 22  
7. General discussion 23  
   Identifying threats to a valid use of the SSDC 23  
   Empirical and procedural validity evidence 28  
   Applying self-assessment in driver education 29  
   Novice drivers' perceived driver competence 29  
   Limitations and generalisations 31  
   Suggestions for further studies 32  
8. Swedish summary 34  
References 45  
Appendix 52
1. Introduction

During the past two decades there has been increasing interest in self-assessment of competence in different educational settings, as the benefits of self-assessment for learning have been recognized. Lately, the importance of self-assessment for development of professional competence has been recognized within driver education as well. Some European countries have incorporated educational goals in their driver education systems that state that drivers should develop a realistic view of their own skills as a driver. In educational systems that are criterion-referenced, there should be a correspondence between the goals of the curriculum, the education and the test. That is, the goals in the curriculum should control the content of the education and the content of the examination. When goals for self-assessment of driver competence are introduced in the curricula for driver education, strategies for measuring these goals in reliable and valid manners need to be developed. Within the traffic psychological field there are some examples of self-report instruments that have been used to measure drivers’ perceptions of their driving skills. However, the majority of these instruments focus on the drivers’ perceived driving skill compared to that of the average driver, and not compared to their actual competence. In addition, there is little information about the reliability and validity of these instruments, as most of them have not been subject to psychometric evaluations.

In 2006 a new curriculum for driver education in Sweden, including goals for drivers’ self-assessment, was introduced. In order to measure these goals it was decided that a self-report instrument assessing perceived driver competence should be developed. The main aim of this thesis was to develop and evaluate the construct validity of this instrument. The theoretical foundation of the instrument is based on research about self-assessment of competence, both in general and with respect to driving skills. Although the field of application is perceived driver competence, the studies attached to this thesis have a strong psychometric focus, as the development and evaluation of the instrument is based on research on test construction, classical and modern test theory, as well as on a modern perspective on construct validity.

Disposition of the thesis
The thesis consists of a summary and six papers. After this introductory chapter, research on self-assessment of competence in general and self-assessment of driver competence in particular is presented in the second chapter. The third chapter presents the theoretical background to developing the instrument for perceived driver competence as well as the steps in the process of developing the instrument. In the fourth chapter the modern, unified validity perspective that was applied in the evaluation of the instru-
ment is outlined. In the fifth chapter the methodological choices made in the five empirical studies are described. In the sixth chapter the studies included in the thesis are summarized. In the seventh chapter the general findings of the six papers is discussed and suggestions for further research are provided. In the eight and final chapter a Swedish summary is presented. Then, the six papers follow in numerical order.

2. Self-assessment

Metacognition and self-regulation

Research in different fields has recognized that metacognition in terms of students’ knowledge and control of their own cognition play an important role in learning (Bransford, Brown, Cocking, 1999). One of the characteristics of effective learners is that they can monitor their current level of understanding and decide when it is not adequate. In other words, they have a realistic sense of their own strengths and weaknesses and they can use knowledge of their own achievements to direct their studying in productive directions (Boud & Falchikov, 1989). The ability to recognise the limitations in one’s current abilities and to identify what has to be learned in order to improve is extremely important for learners at all ages. The concept of metacognition has been used in many different ways, but an important general distinction concerns two aspects of metacognition: metacognitive knowledge, i.e. knowledge about cognition, and self-regulation, i.e. control, monitoring and regulation of cognitive processes (Bransford, Brown, & Cocking, 1999).

Metacognitive knowledge refers to students’ knowledge about their own cognition and control of their own cognition. Metacognitive knowledge comprises three different parts. Firstly, it comprises strategic knowledge, which is students' knowledge of general strategies for learning and thinking. Secondly, it includes knowledge about cognitive tasks, which also refers to knowledge of when and why to use different strategies. Thirdly, metacognitive knowledge comprises self-knowledge, which is defined as knowledge about the self in relation to both cognitive and motivational components of performance (Flavell, 1979). Self-regulation is an active, constructive process where learners set goals for their learning and then attempt to monitor, regulate and control their cognition, motivation and behaviour in order to reach their goals (Pintrich, 2000). Most models of self-regulation include three categories of strategies: planning, monitoring and regulating. Planning includes setting goals, whereas the monitoring includes checking one’s own understanding against these goals. Regulation is closely related to monitoring. As students monitor their learning and performance against the goal that has been set, this process suggests the need for regulation (Pintrich, 1999).
Self-knowledge, which is based on an individual’s own self-awareness and knowledge base, is an important aspect of metacognitive knowledge. It includes knowledge of one’s strengths and weaknesses in relation to cognition and learning. The accuracy of self-knowledge seems to be crucial for learning. It is much more important for students to have accurate perceptions and judgements of their knowledge base and expertise than to have inflated and inaccurate self-knowledge (Pintrich & Schunk, 1996). Figure 1 illustrates how the constructs presented above are related to one another.

Self-knowledge can be divided into two major parts: knowledge of one’s general cognition and beliefs about motivation. The first part concerns knowledge of one’s strengths and weaknesses in relation to cognition and learning. It also includes awareness of the different types of general strategies students are likely to rely on in different situations. In addition to knowledge of their general cognition, individuals have beliefs about their motivation. Motivation is a complicated and broad field with many available theories. A consensus has emerged, however, around social cognitive models of motivation that propose three sets of motivational beliefs (Pintrich & Schunk, 1996). The first set is self-efficacy beliefs, that is, students’ judgements of their capability to accomplish a specific task. The second set includes beliefs about goals or reasons students have for pursuing a specific task. The third set contains value and interest beliefs, which represent students’ perceptions of their personal interest in a task as well as judgements of how important and useful a task is to them.

Figure 1. Construct map of metacognition and related constructs.
Self-assessment of competence
When an individual makes judgements and evaluations based on self-knowledge, a self-assessment is made (Shrauger & Osberg, 1981). Self-assessment refers to the involvement of learners in making judgements about their own learning, particularly about their achievements and the outcomes of their learning (Boud & Falchikov, 1989). In line with this, self-assessment has been defined by Klenowski (1995) as

the evaluation or judgement of the worth of one's performance and the identification of one's strengths and weaknesses with a view to improving one's learning outcomes. (p. 146)

Self-assessment plays an important role in learning generally, and in the development of professional competence (Boud, 1995). The accuracy of self-knowledge seems to be crucial for learning. One of the characteristics of effective learners is that they have a realistic sense of their own strengths and weaknesses and that they can use knowledge of their own achievements to steer their studying in productive directions. If students are not aware of what they do not know, it is unlikely that they will make any effort to learn new material (Pintrich & Schunk, 1996). Research has indicated that engaging in self-assessment is positive for learning in that it contributes to higher student achievement and improved performance (Boud & Falchikov, 1989; Ross, 2006). In a study presented by McDonald and Boud (2003), high school students receiving self-assessment training were compared with a control group that did not receive training in self-assessment with respect to performance in different curricular areas. The result suggested that the students receiving self-assessment training performed significantly better than their peers that did not receive training.

Self-assessment accuracy
In many studies the accuracy of self-assessment of competence has been examined by comparing self-assessments of competence to measures of performance such as teacher marks, grades or test performance (see e.g. Ackerman, Beier, & Bowen, 2002; Boud & Falchikov, 1989; Falchikov & Boud, 1989). Some studies indicate that people can make fairly accurate assessments of their own competence (Dochy, Segers, & Sluijsmans, 1999; Mabe & West, 1982; Shrauger & Osberg, 1981), whereas other studies indicate that people are rather poor in assessing their own skills (Dunning, Heath, & Suls, 2004; Kirby & Downs, 2007; Relan, Guiton, Sopher, & Goldhaber, 2006; Ross, 2006).

The literature suggests that there are several factors that affect the accuracy of self-assessments. Some factors are related to the instrument used for self-assessment, whereas other factors are related to the characteristics of the respondents completing the self-assessment. Factors related to the
instrument include for example the difficulty of the tasks, the specificity or ambiguity of the domain as well as the scale used for self-assessment (Kruger, 1999). Ackerman et al. (2002) showed that when people assess their competence in broadly defined areas and compare their own competence to that of others, people tend to rate themselves as above average. The study also indicated that when domains for self-assessment are specific rather than general and when people assess their own skills against a criterion rather than comparing themselves with others, people make more accurate self-assessments.

As mentioned above, there are also factors related to the respondents that affect the accuracy of self-assessments. Such factors include for example their level of competence in the area. Studies indicate that proficient students tend to make more realistic self-assessments than their less proficient peers or even underestimate their competence. On the other hand, students who are less proficient tend to overestimate their competence more (Fox & Dinur, 1988; Hartman, 2001; Kruger & Dunning, 1999; Longhurst & Norton, 1997). Moreover, studies indicate that students need instructions in order to be able to accurately assess their own competence (Ross, 2006; Sullivan & Hall, 1997). Sullivan and Hall (1997) found that students that overestimated their competence were not familiar with the expected criteria and were unclear about how to evaluate their own work.

Inaccurate perceptions of one’s own competence that are due to factors associated with the respondents can be improved by different means. One important point is that self-assessment may be regarded as a skill and, as such, needs to be developed (Dochy & McDowell, 1997). In support of this, research has indicated that self-assessment practice improves the accuracy of self-assessments (Dragemark Oscarson, 2009; Jönsson, 2008). Self-assessments can also be improved by providing clear criteria for the assessment and making students familiar with these criteria. In addition, raising the competence of the students through education has shown to yield more accurate self-assessments (Kruger & Dunning, 1999). Other methods that have been effective in improving the self-assessment accuracy are observations of one’s own performance (Ward et al., 2003), as well as informing the students that the self-assessments might be cross-checked with external measures (Fox, Capsy, & Reisler, 1994).

**Self-assessment of driver competence**

Research on self-assessment in the field of driving is commonly categorised as assessments of one’s own driving style and driving skill. **Driving style** concerns the way people choose to drive, or individual driving habits that have become established over a period of years. It includes choice of speed, overtaking and traffic violations etc. Driving style is expected to be influenced by the driver’s attitudes and beliefs relating to driving as well as more general values. On the other hand, **driving skill** concerns the maximum level of
performance on different driving tasks, for example use of steering wheel and hazard detection. Driving skills are expected to improve with practice and training (Elander, West, & French, 1993). Driving skills have been classified into two main categories, one that focuses on technical skills and one that is safety oriented and focuses on defensive skills (Lajunen & Summala, 1995; Spolander, 1983). This thesis will focus on self-assessment of driving skill.

There are two main uses of self-assessments of driving skill. Firstly, they are frequently used as a means of measuring people’s driving skills in an effective, inexpensive way (Hatakka, Keskinen, Katila, & Laapotti, 1997). Some self-report measures have been developed for this purpose (see e.g. Hatakka, 1998; Lajunen & Summala, 1995; Spolander, 1983). Secondly, self-assessments of driving skill have been compared with observed driving performance in order to examine the accuracy of drivers’ self-assessments (Groeger, 2001). Although self-assessments have been less frequently used for this latter purpose, lately the importance of a realistic self-assessment has been stressed, and therefore self-assessment has been included in the driver education systems in some European countries (AKE, 2005; Hatakka, Keskinen, Gregersen, Glad, & Hernetkoski, 2002; Vissers, Mesken, Roelofs, & Claesen, 2008; VVFS 2004:110).

**Self-assessment in driver education**

Self-assessments of driving skill have important implications for driver education and traffic safety, as a realistic self-assessment can be expected to be important for safe driving. Many studies have examined novice drivers’ perceived driving skill (see Delhomme, 1991; Svenson, 1981; Williams, 2003). The results from these studies indicate that drivers overestimate their own driving skills, as the majority believe that they are more skilled than the average driver. Gregersen and Bjurulf (1996) presented a model for novice drivers’ behaviour and accident involvement, in which perception of driving skill is one important aspect. The model comprises two main processes: the process of learning how to drive and aspects of life that influence the driving. Three central aspects of the learning process are driving experience, perception of one’s own skills and perceived accident risk. Together with individual and social circumstances (e.g. lifestyle, personality, group norms, and values), these three aspects influence drivers’ motives, attitudes, and decision-making processes, which in turn influence the driving behaviour. Following this model, the overconfidence of young and novice drivers has been presented as one explanation of their overrepresentation in road accidents (Gregersen & Bjurulf, 1996; OECD, 2006).

The capability of making an accurate assessment of one’s own driving skill is of great importance in the driver’s regulation of his or her driving behaviour. To a large extent, driving is a self-paced task, which means that the driver can influence the difficulty of the driving task through his or her own driving
behaviour (Näätänen & Summala, 1974). In order to drive safely, the demands of the task should match the driver's competence. Adapting the behaviour to the demands of the task requires an accurate assessment of one's own driving skills and the complexity of the situation. If the self-assessment is inaccurate, the driver might engage in driving tasks that are too demanding and unsafe (De Craen, Twisk, Hagenzieker, Elffers, & Brookhuis, 2007). Inaccurate self-assessment refers to both underestimation and overestimation of one's own skills. However, from a traffic-safety perspective, overestimation is seen as the most problematic, as it is believed to be related to the high accident risk of young drivers (Gregersen, 1996).

In the Goals for Driver Education (GDE) model, the importance of self-assessment for safe driving is emphasized (Hatakka et al., 2002). The GDE model is a conceptual model of driver training and education comprising two dimensions. The first dimension consists of four hierarchical levels of driver behaviour. The first two levels, Vehicle manoeuvring and Mastery of traffic situations, are basic abilities. The two following levels comprise Goals and context of driving as well as Goals for life and skills for living. These levels are considered to be of great importance for how the driver decides to behave in traffic, and thus these goals are important for traffic safety (Hatakka et al., 2002). The second dimension in the GDE model is formed by three goals for training: Basic knowledge and skills, Knowledge and skills concerning risk-increasing factors and Skills for self-evaluation (see paper IV for an illustration of the GDE model). It has been emphasised that having a realistic perception of one's own driving skill and role as a driver is important for safe driving. From this perspective, self-assessment is an important tool not only in driver training, but in the development of driving skill after training as well (Engström, Gregersen, Hernetkoski, Keskinen, & Nyberg, 2003).

The GDE model constitutes the basis of the driver education systems in Finland, Norway, the Netherlands and Sweden. Thus, the curricula for driver education in these countries comprise goals of drivers' self-assessment (AKE, 2005; Statens vegvesen, 2002; Vissers et al., 2008; VVFS 2006:21). In order to evaluate if the candidates fulfill these goals two different strategies are used: by introducing these aspects in a compulsory driver education or by examining self-assessment of driving skills in the driving-license test. In Norway the self-assessment content is mainly included in the compulsory driver education, whereas in Finland the self-assessment is incorporated in the driving test. The Finnish candidates complete a self-assessment of their own driver competence before they take the driving test and the self-assessment is compared to their performance in the driving test when the test is finished. Similar procedures have also been tested in Sweden (Sundström, 2007) and the Netherlands (Vissers et al., 2008). By comparing candidates' perceived driver competence with their performance in the driving test it is possible to examine the goals of self-assessment, at least in the two lower levels of the
GDE model. The reason for choosing this approach for self-assessment of driver competence in Sweden instead of including self-assessment training in the compulsory education was that the amount of compulsory driver education was very small at the time of instrument development. The compulsory driver education consisted of a six-hour risk education on slippery surface.

3. Assessing perceived competence

Having made the decision of developing a self-report instrument for perceived driver competence, the next question to consider was the use of the instrument and the construct that was to be measured. The intended use of the instrument was to assess individuals' perceived driver competence for specific aspects of driving skill. Following the suggestion by Clark and Watson (1995), the literature of how subjective driving skill has been measured in previous studies was reviewed. This is reported on in study I. Another important issue was to decide which theoretical construct should form the basis of the instrument. Measurement of perceived competence is a fairly new area in the field of driver education and driver testing, but extensive research into perceived competence has been performed in other fields. Therefore, in order to examine how perceived driver competence should be measured within the Swedish driver education, the literature on constructs of perceived competence was reviewed.

Constructs of perceived competence

Four widely used theoretical constructs that focus on competence perceptions were compared with regard to definition, structure, methods of measurement and relationship to measures of observed competence. The constructs were expectancy beliefs and ability beliefs (Eccles & Wigfield, 2002), self-concept (Bracken, 1996; Marsh, 1990) and self-efficacy (Bandura, 1997). The construct of expectancy beliefs is defined as individuals' beliefs about how well they will do on upcoming tasks, either in the immediate or long-term future. The construct of ability beliefs, on the other hand, is defined as individual's evaluations of their competence in different areas (Eccles et al., 1983). Ability beliefs are distinguished conceptually from expectancy beliefs in at least two ways. First, ability beliefs are broad beliefs about one's competence and expectancy beliefs are domain-specific beliefs. Second, ability beliefs focus on present ability and expectancy beliefs focus on the future (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000).

Self-concept is defined as a person's self-perceptions formed through experience with and interpretations of his or her environment (Shavelson, Hubner, & Stanton, 1976). Self-concept is commonly divided into non-academic and academic self-concept. The non-academic self-concept comprises three
parts: physical, social and emotional self-concept. Each of these self-concepts is split into separate and more specific self-concepts. The academic self-concept is defined as self-perceptions that individuals have about their academic abilities, specifically their feelings and knowledge about these abilities and skills (Pajares, 1996).

Self-efficacy is an aspect of self-knowledge and refers to an individual's perceived capability of performing necessary tasks to achieve goals. Bandura (1986) defined the construct of self-efficacy as

people's judgements of their capabilities to organize and execute courses of actions required to attain designated types of performance. (p. 391)

Self-efficacy beliefs are future oriented, task- and situation-specific judgements that are made in reference to some type of goal (Bandura, 1986). Self-efficacy beliefs are multidimensional in the sense that individuals might have different perceptions of their own capabilities for different domains or tasks (Bong & Skaalvik, 2003).

A comparison of the four constructs indicated that the constructs of self-concept and ability beliefs are similar in some respects. Both constructs focus on fairly broad areas of competence and often measure a person's perceived competence in relation to that of other people (Sundström, 2006a). Self-efficacy and expectancy beliefs are also similar constructs as the assessment of one's own competence is made in relation to specific criteria rather than by comparison with others. Moreover, both these constructs focus on future-oriented measurements of perceived competence, and on more specific aspects of perceived competence than the constructs of ability beliefs and self-concept, even though self-efficacy is even more specific than expectancy-beliefs. A comparison of the predictive validity of the four constructs shows that they are related to achievement behaviour, including achievement, choice and persistence, to different degrees. Several studies have indicated that self-efficacy is a more efficient predictor of performance than self-concept because the self-efficacy construct is specific and tailored to the domain it is aimed to predict (Choi, 2005; Pajares & Miller, 1994; Pietsch, Walker, & Chapman, 2003).

A construct for perceived driver competence

In order to determine which construct is most suitable for measuring perceived driver competence within the Swedish driver education system, the four constructs presented above were compared with the requirements for self-assessment within the driver education system. The goals for self-assessment in the new curriculum for driver education state for example that the driver should judge his or her ability to manoeuvre a car, to identify risks of the function and manoeuvring of the car and judge his or her own driving
behaviour. In order to ensure that the goals concerning self-assessment are measured for all learner drivers, one possibility was to use a self-report instrument of perceived driver competence that is completed at the test centre shortly before the administration of the test, since not all examinees would assess their perceived competence if the self-assessment were conducted at a driving school. If the measurement of drivers’ perceived knowledge and abilities is intended to be conducted within the driving-license test, the theory test and the practical test would be suitable external measures of drivers’ observed knowledge and abilities.

As the idea was to compare the subjective driving skill with the subsequent performance on the driving-license test, it was important that the assessment of perceived competence met certain criteria. Firstly, the content of the self-report instrument should correspond to the contents of the driving test. Secondly, candidates should assess their driver competence with respect to different criteria and not in relation to the competence of other drivers. Thirdly, the candidates should judge their competence in relation to specific domains or certain tasks. Fourthly, the level of specificity in the measurement of perceived competence should correspond to the level of specificity in the content of the theory and practical driving-license test. Both the constructs of expectancy beliefs and self-efficacy seemed useful in the measurement of perceived competence, because they fulfil the criteria mentioned above. In the next section, the development of the self-report instrument for perceived driver competence is described.

**Instrument development**

The process of instrument development has been described by many psychometricians (see e.g. Clark & Watson, 1995; Downing & Haladyna, 2006; Schmeiser & Welch, 2006). Netemeyer, Bearden and Sharma (2003) identified four steps in the process of instrument development. In the first step, the construct is defined and the content domain is outlined. In the second step, items are generated to sample the construct domain and these items are assembled into an instrument. In the third step, the instrument is pilot tested and the quality of the items is examined via item analyses. Based on these analyses the instrument is revised if necessary. In the fourth step, the instrument is finalized and its psychometric properties and construct validity are further examined. These steps were followed in the development of the instrument for perceived driver competence and will be described in more detail below.

First, a theoretical model for the construct of perceived driver competence was developed, based on the literature reviewed. This model comprised two dimensions of perceived competence, theory and practical. In the model, perceived competence referred to the two constructs of expectancy beliefs and self-efficacy. The two dimensions of perceived driver competence
comprised five aspects of driver competence that are represented in the curriculum for driver education, driving-license test as well as in the GDE model: Vehicle knowledge and manoeuvring, Economic driving, Traffic regulations, Traffic safety and Personal circumstances and goals in life (see Table 1).

Table 1. Theoretical model for perceived driver competence.

<table>
<thead>
<tr>
<th></th>
<th>Self-efficacy</th>
<th>Expectancy beliefs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Perceived theory competence</td>
<td>Perceived practical competence</td>
</tr>
<tr>
<td>Vehicle knowledge and manoeuvring</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economic driving</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic regulations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic safety</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal circumstances and goals in life</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the second step, the theoretical model for perceived driver competence was used as the basis for item construction. Items were generated to measure perceived theory and practical driver competence with regard to the five different aspects of driver competence. Since both self-efficacy and expectancy beliefs were considered suitable for measuring perceived driver competence, items tapping both constructs were generated for a pilot version of the instrument. The self-efficacy items asked participants how confident they were of successfully performing different tasks in the theory and practical driving-license test, whereas the expectancy-beliefs items asked participants how well they think they would perform in the theory and practical driving-license test. Previous research has indicated that a ten-point scale is suitable for use when attempting to measure self-efficacy (Bandura, 1997). This format was used for all items. The self-efficacy items used a rating scale where the extremes were labelled “very unsure” and “very sure”, and for the expectancy-belief items, the extremes were labelled “very badly” and “very well”. The items generated were scrutinized by three researchers with expertise in educational measurement, and were also administered to four driving-school students, who were asked to review the items with regard to clarity and readability. These reviews resulted in some minor changes.

In the third step, the items were pre-tested on two small samples of respondents. The evaluation of the pilot version indicated that most items demonstrated sound psychometric properties, although the self-efficacy items seemed to function slightly better than the expectancy beliefs items.
Based on these findings it was decided that the self-efficacy items should be used in order to measure perceived driver competence. The self-efficacy items were slightly revised due to the findings in the pilot study. For example, the labels of the extremes of the rating scale were altered to “Not confident at all” and “Completely confident”. A label “fairly confident” was added to the middle of the scale. The revised items were then assembled into an instrument that was titled the Self-efficacy Scale for Driver Competence (SSDC). In order to cover all aspects of driver competence in the curriculum but still reduce the number of items, the 45 self-efficacy items about theory and practical perceived competence were split into two versions parallel in content. Each version included 28 items about perceived theory and practical driver competence. To examine the comparability of the groups taking the two different versions, 17 of the items were common to both versions. These items covered the five areas of driving skill in the theoretical model. Each version also included 11 unique items. The 28 items were distributed over the five aspects of driver competence specified in the theoretical model. The number of items in each domain was the same in versions A and B. As a fourth step in the instrument development process, the SSDC was administered to driving-license candidates and the construct validity of the SSDC was evaluated. Four of the six papers included in this thesis reports on the construct validation of the SSDC.

4. Validity theory

Validity is a fundamental concept in educational and psychological measurement. The importance of the concept of validity has been recognized for many years, but the scope of the concept has changed over time, from being a narrow concept focusing on correlations between test scores and criterion measures into being a broad concept embracing the use and interpretation of test scores as well as their value implications and social consequences.

The traditional view on validity

The traditional view on validity has been that a test is valid if it measures what it purports to measure. For a long time, validity was considered to be of different types: criterion-related validity, content validity, and construct validity (Cronbach, 1961). Criterion-related validity is evaluated by analyzing the correlations between the test scores and external criteria that are related to the construct being measured. These criteria are considered to provide a direct measure of the construct that is to be measured. Criterion-related validity evidence was further divided into two categories, concurrent or predictive validity, depending on when criterion data is available. Content validity is evaluated by examining how well the content of the test corresponds
to the domain that the test is constructed to measure. Construct validity is evaluated by gathering evidence that shows that the test scores are indicators of the construct being measured (Crocker & Algina, 1986).

During the first half of the twentieth century there was a strong focus on criterion-related validity, as validity was viewed as the correspondence between test scores and criterion score. The main limitation of the criterion-related validity model is the difficulty in obtaining an adequate criterion and validating this criterion (Kane, 2006). This problem contributed to the development of alternative strategies of finding validity evidence. By the early 1950s, more emphasis was put on content validity, and validation became concerned with the question of whether the behaviours measured by a test represent the behaviours that the test is designed to measure. The main criticism that has been directed towards the content-validity model is that evaluations of content representativeness are based on subjective judgments (Kane, 2006). The need for the construct validity model developed out of personality testing where there is neither a criterion to predict, nor a content domain to sample (Kane, 2006). Cronbach and Meehl (1955) suggested that the construct validity approach was a fundamental concern and that it could be used as a substitute for criterion and content validity, when a test is to be interpreted as a measure of a construct. The traditional view of validity, in terms of criterion-related, content and construct validity, was the predominating view of validity for a long time. During the last 20 years, however, the concept of validity has undergone important changes as the traditional view of validity largely has been replaced by a broad, unified validity concept.

A modern perspective on validity
One of those who have come to largely influence the modern view of validity is Messick. In his chapter on validity, Messick (1989) argued that the traditional conception of validity is incomplete since it fails to take into account the value implications of score interpretation as well as the social consequences of test score use. He also argued that validity is not to be regarded as a property of the instrument itself. Rather, it is the interpretation of the test scores as well as the implications for action this interpretation results in that is to be validated. Messick proposed a unified validity concept, in which various types of evidence, content-related, criterion-related and construct-related evidence, are gathered to support the interpretation of a score (Messick, 1989). Validity was defined as

an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment. (p. 13; emphasis in original)
This definition suggests that validation is an ongoing process in which several sources of validity evidence are integrated and where the interpretation of a test score is to be validated, not the test itself.

This unitary validity perspective, where construct validity is seen as an overarching aspect of validity embracing content and criterion-related validity, is today the most widely accepted view of validity, although there are those advocating narrower perspectives on validity (see e.g. Boorsboom, Mellenbergh, & van Heerden, 2004; Lissitz & Samuelsen, 2007). The current perspective on validation described in the latest edition of the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) is also influenced by Kane’s (1992) argument-based approach to validity. It is stated that the process of validation begins with stating the proposed interpretation and uses of the test scores. Further, according to AERA, APA, & NCME (1999)

*Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. (p. 9)*

Modern validity theorists emphasize that the articulation of a construct theory is a fundamental issue in developing a measure and evaluating a trait interpretation (Kane, 2006; Messick, 1989). This was however recognized more than 50 years ago by Cronbach and Meehl (1955). They argued that evaluation of construct validity involves at least three steps; firstly, defining a set of theoretical constructs and their interrelations, secondly, developing ways to measure these proposed constructs and thirdly, empirically testing the hypothesized relationships between the constructs. The correspondence between the construct domain and the measure is a central part of the validity theories presented by both Messick (1989) and Kane (2006). When it comes to the match between the construct domain and the content of the measure developed to tap this domain, there are two major threats to validity; *construct under-representation*, which refers to the degree to which an instrument fails to tap important aspects of the construct it is intended to measure, and *construct-irrelevant variance*, which refers to the extent to which scores from an instrument are affected by other things than those that the instrument is designed to measure (Kane, 2006).

Moreover, both Messick (1995) and Kane (2006) recognized that the social consequences of the measurement should be incorporated in the validity concept. The question of whether the consequences of a measurement should be a part of the validity concept has been subject to debate, however. Those who advocate the inclusion of social consequences in the validity concept claim that validity is called into question when the consequences of test score use are related to flaws in the measurement process, such as
construct under-representation or irrelevant variance (Nichols & Williams, 2009). On the other hand, the opponents argue that the inclusion of social consequences will make the validity concept vague and therefore lead to confusion instead of clarity (Boorsboom & Mellenbergh, 2004; Mehrens, 1997; Popham, 1997). It seems however that consensus has emerged lately, and that the current view is that consequences of test use are an important source of validity evidence (Nichols & Williams, 2009).

**Applying the modern validity perspective**

Even though Messick’s (1989; 1995) validity theory has been very influential on the modern validity perspective, it has been criticized for being too complex to use in practice (Brennan, 1998). In fact, two recent studies have noticed that there is a gap between the modern validity theory and current validation practices (Cizek, Rosenberg, & Koons, 2008; Wolming & Wikström, in press). From these studies it is concluded that in order to bridge this gap, a more practically oriented approach to validation is needed.

Frameworks that provide more detailed guidance for validation practices have been presented for example by Crooks and Kane (1996) and Kane (2006). Largely, these validity frameworks build on the theory presented by Messick (1989), but they are more practically oriented in that they provide more detailed guidance of the kinds of validity evidence that could be collected to support the interpretation of a test score.

In the papers attached to the thesis I discuss the construct validity of the SSDC from the modern validity perspective presented above. In this chapter, I use the validity framework presented by Crooks and Kane (1996) as a starting point for discussing the findings of this thesis. The rationale for this choice is that their framework focuses on the testing process and validity threats that are present at each step in this process.

According to Crooks and Kane (1996), validation of an assessment can be viewed as a chain of eight links: administration, scoring, aggregation, generalization, extrapolation, evaluation, decision and impact. Evaluating the validity of an assessment requires careful consideration of threats to validity that are associated with each link. As regards the administration of the instrument, threats to validity might be low motivation among examinees, test anxiety or inappropriate assessment conditions. Validity threats associated with the scoring of the students’ performance are that the scoring fails to capture important qualities of task performance and that the scorers are inconsistent in their scoring. For the aggregation of the scores, potential validity threats are that the aggregated tasks are too diverse, which will limit the reliability of the scores. Regarding the generalization from the particular tasks to the whole domain of similar tasks, potential threats are for example that the assessment conditions are too variable and that the scoring criteria are not consistent for different tasks. Threats that are associated with the
extrapolation from the assessed domain to the construct domain are for example that the conditions of assessment are too constrained, i.e. they are narrower than the range of conditions permitted in the construct domain. Another threat is that parts of the target domain are not assessed or are given little weight. There is a trade-off between generalization and extrapolation, as extrapolation can be strengthened at the expense of generalization by making the assessment tasks as representative of the construct domain as possible. Or, generalization can be improved at the expense of extrapolation by increasing the number of highly standardised items in the assessment. The goal is to compromise and find a way to support generalization and extrapolation adequately. For the evaluation of students' performance, threats to validity include that the person evaluating the assessment information does not properly understand the information and its limitations. As regards the decisions based on test scores, threats to validity are for example inappropriate standards. For the link that concerns the impact of the assessment on the test takers, potential validity threats are if positive consequences are not achieved or if negative consequences occur.

As the focus of this thesis is the development and construct validation of a self-report instrument, construct validity is an essential concept. Four of the six studies (studies II to IV and VI) presented in this thesis are empirical studies that directly examine the validity of the interpretations of test scores of the SSDC. In chapter 7 the findings of the studies included in the thesis will be discussed in relation to the eight links in the chain of validation discussed by Crooks and Kane (1996).

5. Method

Procedure of administration, scoring and feedback
Data used in studies II, III and IV was collected in a project conducted by the Swedish Road Administration (SRA) in 2006. The aim of the project was to test and evaluate a new model for driver testing. The candidates participating in the project received the SSDC at the driving test centre and were asked to complete it before they took the theory test. When they booked the test they received information about the project and about a new model of driving test being examined as well as the self-assessment. They were informed that responding to the SSDC was voluntary, and that if they did not want to answer the questions they should return a blank questionnaire. The SSDC was administered between July and October 2006. Of the 1559 candidates that received the questionnaire, 33 candidates did not answer any of the items.

After completing the SSDC they returned it to the staff at the test centre. The staff scored the questionnaires by summing the responses to items that
belonged to the same content area and dividing by the number of items, producing a mean score of the responses to each content area. To facilitate comparisons of the SSDC scores and driving test performance, the mean scores were categorised as low, medium or high perceived driver competence (1 - 3.4 = low, 3.5 - 7.4 = medium, 7.5 – 10 = high). The reason for using a larger interval in the middle of the scale was that the scores on the SSDC were assumed to be normally distributed. When the candidates had completed the theory test, the staff returned the questionnaire with the SSDC to them and asked them to bring it to the driving test. Copies of the questionnaires were sent to the Department of Educational Measurement at Umeå University for data analyses. Data for study II, III, and IV was obtained by optically scanning the questionnaires and thereby constructing a database.

During the driving test, the driver examiners judged the candidates’ performance as poor, mediocre or good in the four competence areas Vehicle manoeuvring, Economic driving, Traffic regulations, and Traffic safety. The assessment “poor” was given if the candidate failed the test due to errors in the competence area. The assessment “mediocre” was given if the candidate barely passed the standards for the area, and the assessment “good” was given if the candidate demonstrated a good performance. When the driving test was completed, the driver examiner asked the candidate to hand over the results from the SSDC. (The driver examiners did not look at the candidates’ score on the SSDC until the driving test was completed.) The driver examiner compared the results obtained from the SSDC (low, medium, high perceived competence) with the driving test performance (poor, mediocre, good) in each competence area. If the candidates had rated their perceived competence as low and had a good driving performance this was interpreted as an underestimation. If the candidates had high ratings of perceived competence and poor driving performance this was interpreted as an overestimation. If the perceived competence was in the same category as the driving performance, the perceived competence was interpreted as realistic. Based on this, the candidates received feedback on the correspondence between their perceived driver competence and their competence as demonstrated in the driving test.

**Methodological considerations**

The focus in this thesis was on gathering evidence for different aspects of construct validity of the scores from the SSDC. Both quantitative and qualitative sources of evidence were used to support the interpretations of the scores from the SSDC. With respect to quantitative evidence, one step in the validation process was to examine the psychometric properties of the items in the SSDC and the functioning of the rating scale. There are two theoretical traditions for psychometric evaluation: classical test theory (CTT) and item response theory (IRT). Although self-report instruments have commonly been evaluated
using classical item statistics, IRT applications are being more frequently used (see e.g. Beck & Gable, 2001; Fletcher & Hattie, 2004). There are benefits of applying both CTT and IRT as both methods provide information about the functioning of an instrument. Study II reports on the psychometric evaluation of the SSDC using CTT, while study III uses IRT to examine the psychometric properties and functioning of the SSDC.

Benefits of CTT analyses are for example that they do not require as large a sample as IRT analyses, and that CTT does not make assumptions of unidimensionality and local independence, which IRT does (Crocker & Algina, 1986). Furthermore, there are many user-friendly software packages available for CTT analyses. The main drawback of CTT is that the respondent characteristics cannot be separated from the item and test characteristics. This means that the psychometric properties are dependent on the items and persons examined, and that a person’s score can only be interpreted in the context of the specific items included in the instrument. The main advantage of IRT is that item statistics are independent of the subset of items and persons examined. This means that the psychometric properties are not sample-dependent and scores describing examinee trait level are not test-dependent (Hambleton, Swaminathan, & Rogers, 1991). Other advantages of IRT are that the item location and person trait level are indexed on the same metric and that IRT provides estimates of standard errors for individual trait estimates rather than a single estimate of error for all examinees. Due to this, IRT can be used for developing effective rating scales, as the efficiency of the scale can be adapted to a specific level of the construct (Hambleton, Robin, & Xing, 2000; Hambleton, Swaminathan, & Rogers, 1991).

Further validity evidence for the SSDC was gathered by examining the internal structure through the use of factor analysis. Factor analysis, which is the far most common method used for construct validation, can be divided into two categories: exploratory and confirmatory (Thompson, 2004). Exploratory factor analysis (EFA) is used to identify the underlying dimensions among items in an instrument. EFA is exploratory in the sense that it requires no theoretical expectations about the relationships between the items. In confirmatory factor analysis (CFA), on the other hand, hypothesised relationships between variables are tested and indices of model fit are provided to evaluate to what extent the data fits the theoretical expectations. In study II, EFA was used on the sample that completed version A of the SSDC to examine the number of factors underlying the data in each version of the SSDC. CFA was carried out on data from version B to confirm the factors determined in the exploratory factor analysis.

Criterion-related validity evidence was gathered in studies II and VI by examining the relationship between scores on the SSDC and performance on the driving-license test. Moreover, in studies IV and V, the patterns found in the SSDC were compared with patterns found in two similar instruments
used for perceived driving skills administered in a Finnish and Dutch sample. In addition, groups that could be expected to differ with regard to level of accuracy in their perceived driver competence were also compared.

As Kane (2006) pointed out, the strength of the validity evidence is increased by integrating different sources of validity evidence. Therefore, qualitative and quantitative evidence of the external aspect of construct validity was gathered in study VI. A sample of driving-license candidates completed a revised version of the SSDC and took the driving test. Seven of these candidates also participated in a semi-structured interview about perceived driver competence. The information obtained from the interviews provided explanations for the patterns observed in the SSDC scores.

6. Summary of studies

In this chapter the six studies included in this thesis are summarised. The studies are strongly related to one another in the sense that they focus on the assessment of perceived driver competence and development and construct validation of the SSDC. Study I presents a review of how subjective driving skill has been measured in previous studies and these methods are discussed from a measurement perspective. The conclusion was that there is a need to develop an instrument for perceived driver competence that could be compared to observed driver competence in order to examine the accuracy of novice drivers’ self-assessment of their driving skills. This was the point of departure of studies II and III, which reported on the development and evaluation of an instrument for assessing perceived driver competence – the SSDC. Similar self-assessment instruments as the SSDC have been used in other driver education systems. For this reason it was interesting to compare the results from the SSDC to those of other instruments. In study IV, Swedish candidates’ responses to the SSDC were compared with Finnish candidates’ responses to an instrument for perceived driving skill used in Finland. In study V, results from the instrument used in Finland were further compared with results from a similar instrument for driving skills used in the Netherlands. In study VI, the external aspect of construct validity was further examined by studying the relationship between the SSDC and the Finnish instrument for perceived driving skill, as well as the relationship between the SSDC and observed driving performance. A quantitative and a qualitative approach was combined as information obtained from a number of semi-structured interviews about perceived driver competence provided explanations of the patterns found in the SSDC responses.
Study I

The purpose of the literature review was to examine how subjective driving skill has been measured and to discuss these methods from a measurement perspective. The studies reviewed were divided into three different categories, depending on how they assessed subjective driving skill. The first category included studies that assessed subjective driving skill in relation to the ‘average driver’ or ‘peers’. The studies included in the second category assessed subjective driving skill by relating it to different internal criteria, e.g. assessing weak and strong competences of a person’s own driving skill out of a number of specific components of driving skill. In the third category, assessments of driving skill using internal criteria were compared to external measures of performance, e.g. driver examiner assessments. The findings indicated that the studies most frequently performed are those in the first category. These studies indicate that the majority of drivers believe that they are better than the average driver, and this tendency is most prevalent among young male drivers. The main conclusion from the review was that the studies where subjective driving skill is assessed by asking drivers to compare their own skill to that of the average driver are problematic, partly because they can result in biased assessments, and partly because it is difficult to determine whether drivers are overconfident with respect to their own skills or not, as empirical information of their skills is missing. Therefore, there is a need to compare subjective driving skill with external measures of performance in order to determine if drivers overestimate their skills or not.

Study II

This study reports on the development and validation of an instrument for measuring perceived driver competence. The aim of the study was to examine the construct validity of the SSDC. This paper adopts a modern validity perspective, as the construct validity of the scores from the SSDC is discussed in relation to Messick’s (1995) six aspects of construct validity. Two parallel versions of the SSDC were administered to two samples of driving-license examinees ($n = 721$, $n = 805$). The analysis indicated that the SSDC demonstrated sound psychometric properties, which provided support for the substantive aspect of construct validity. The internal structure of the SSDC was examined through an exploratory factor analysis of the first sample. The results indicated that the SSDC consists of two dimensions of perceived driver competence, theory and practical. The two-dimensional structure of the SSDC was confirmed in the second sample using a confirmatory factor analysis. These findings provided support for the structural aspect of construct validity, as the two empirical dimensions corresponded to the two dimensions of perceived theory and practical driver competence that the SSDC aimed to measure. In order to gather evidence for the external aspect of construct validity, the relationships between candidates’ scores on the SSDC
and their performance in the theory and practical driving-license test were examined. Although these relationships were weaker than what could be expected based on studies with other self-efficacy instruments, the magnitudes of the relationships were in line with other instruments of subjective driving skill. In conclusion, the examination of the psychometric properties and the internal structure indicates that the SSDC seems to function adequately as a measure of perceived theory and practical driver competence.

Study III
In this study the psychometric properties and functioning of the rating scale were further examined using IRT analysis. 795 and 714 candidates completed the two versions A and B, respectively. The 39 items in versions A and B of the SSDC were calibrated together using the polytomous rating scale model (RSM). Five items demonstrated poor fit to the RSM, and therefore these items were removed. The classical analyses and IRT analyses agreed on which items should be removed from the SSDC, although two additional misfitting items were detected by RSM analysis. In line with previous research, it was found that trait estimates and item location estimates obtained from RSM were comparable to the classical item statistics, as the item means and item locations provided similar item hierarchies. The RSM analysis indicated that the rating scale appears to function fairly well, but that it could be improved by decreasing the number of categories and hence making each category more clear to the respondents. Another solution that can be used to help respondents discriminate between categories is to provide labels on more categories, making each category more comprehensible. Moreover, the RSM analysis also indicated that the measurement precision of the SSDC can be improved by adding items at the upper part of the self-efficacy continuum.

Study IV
This study aimed to examine the accuracy of novice drivers’ perceived driver competence. In previous studies this was mainly done by asking drivers to compare their own skill to that of the average driver. In order to examine whether novice drivers are overconfident with respect to their actual skills, specific aspects of perceived driver competence were compared with assessments made by a driver examiner. A Finnish (n=2739) and a Swedish (n=805) sample of candidates completed two different instruments of perceived driver competence. Unlike previous studies, which indicate that the majority of drivers are overconfident, the result from this study indicated that a fairly large proportion of drivers made an accurate assessment of their skills. About 50 percent of the Finnish and between 25 and 35 percent of the Swedish candidates made realistic assessments of their competence in the areas Vehicle manoeuvring, Economical driving and Traffic safety. The proportion of those who overestimated their competence was greater among
the Swedish candidates than among the Finnish ones. One possible explanation of this is that the Finnish candidates have had greater opportunities of practicing self-assessment in the driver education. Results from previous studies, where perceived driver competence is examined by comparing a person’s own skills to that of the average driver, indicate that men are more overconfident than women. However, the results from the present study indicate that men are not overconfident to a greater extent than women. Thus, it seems that when perceived competence is related to actual competence instead of the skills of the average driver, the majority of drivers are no longer found to overestimate their skills and no gender difference is found.

Study V
This study aimed to further examine the accuracy of novice drivers’ perceived competence, by comparing candidates’ perceived driver competence with driver examiner assessments of their performance in the driving test. A Finnish (n = 2739) and a Dutch sample (n = 239) of candidates assessed their driver competence in six areas and took the driving test. The results found in the present study, where candidates’ assessments of perceived driver competence were compared with examiner assessments, differed from those found in previous studies where drivers have assessed their skill in comparison to the average driver. Compared to previous studies, a smaller proportion overestimated their competence and a larger proportion made realistic self-assessments of their driver competence. Between 40 and 50 percent of the candidates in both samples made realistic assessments and 30 to 40 percent overestimated their competence. In contrast to previous studies, men did not overestimate their skills more than women, and younger candidates were not more prone to overestimation than older drivers. In the field of self-assessment research it has been found that there is a relationship between level of competence and self-assessment accuracy, in the sense that more competent persons tend to have a more realistic view of their competence and the other way around. This phenomenon seemed also to be present among the candidates, as a larger proportion of candidates that failed the test overestimated their skill compared to those who passed.

Study VI
The aim of this study was to further examine the external aspect of validity of the SSDC by integrating qualitative and quantitative sources of evidence. More specifically, the relationships between perceived driver competence, in terms of scores on the SSDC, and scores from a similar instrument for self-assessment of driving skills were examined. In addition, the relationship between SSDC scores and performance on the driving test was examined. 357 candidates completed a questionnaire including a revised version of the SSDC, a scale for self-assessment of driving skill used in Finland and items
about confidence in passing the test, test anxiety etc. Seven candidates also participated in semi-structured interviews where their own driver competence was further explored. The findings provided support for convergent validity, as there were strong relationships between the SSDC scores and the scores from the instrument for self-assessment of driving skill. There was a weak relationship between the SSDC scores and the performance on the driving test. Possible explanations for the weak relationship were obtained partly from semi-structured interviews about the respondents’ perceptions of their skills as a driver and partly from questionnaire data about confidence in passing the test etc. Firstly, it was indicated that confidence in performing different tasks in the test (measured by the SSDC) is different from confidence in passing the test, which might explain the weak relationship. Secondly, the interviews indicated that candidates were not familiar with the requirements for passing the driving test, and that they were not used to assessing their own driver competence with respect to those skills included in the SSDC. The unfamiliarity with assessing one’s own driver competence in relation to what it takes to pass the driving test is a possible explanation for the weak relationship between the SSDC and the driving test performance.

Errata for the papers are presented in Appendix.

7. General discussion

The aim of this chapter is to discuss the results from the studies included in this thesis from a modern validity perspective, as well as to provide suggestions for further studies. The point of departure for the validity discussion is the framework provided by Crooks and Kane (1996), where the process of validation is viewed as a chain of eight links. In each link threats to the validity should be identified. In order to discuss and draw conclusions of the validity of the score interpretations, the use of the SSDC as well as the results from the studies presented in this thesis are put in relation to these links in the following section.

Identifying threats to a valid use of the SSDC

Administration
The first link in the chain of validation is the administration link. Although validity threats to the administration of the SSDC were not explicitly examined in any of the studies in this thesis, possible threats to validity need to be identified and discussed. The candidates completed the SSDC at the driving test centre shortly before they took their theory test. This procedure could result in lack of time for filling out the SSDC. Another factor is that they came to the test centre for the purpose of taking the theory test, and therefore
another possible threat to the validity of the administration is that they were unable to concentrate and focus on answering the questions in the SSDC. Moreover, most candidates completed the SSDC in the waiting room at the test centre. One cannot be sure that the testing environment fulfils the requirements of reasonable comfort and minimal distractions.

**Scoring**
The second link in the chain of validation is that of scoring. When the SSDC was administrated to candidates at the test centres, assistants at the test centres scored the questionnaires by summing the responses to items that belonged to the same content area. The reason why the scoring was made at the test centre was that the driver examiners were to use the scores from the SSDC after the driving test in order to provide feedback to the candidates on the correspondence of self-assessment and driving test performance. With regard to this procedure, possible threats to the validity are scoring errors. Copies of the questionnaires were sent to the department of Educational Measurement, Umeå University, and were optically scanned. The data file resulting from this scanning was used in studies II, III, and IV. Although optical scanning can result in errors too, the potential risk of errors is probably smaller compared to manual data recording.

**Aggregation**
When the SSDC was used in the project conducted by the SRA, the scores from the items were aggregated to produce subscores for the content areas Vehicle manoeuvring, Economic driving, Traffic regulations and Traffic safety. These subscores were compared to the performance in the driving test and feedback about self-assessment accuracy was provided on the basis of this comparison. However, the psychometric evaluation indicated that the SSDC consists of two factors: theory and practical perceived driver competence. Therefore, in future use of the SSDC, one should consider whether it is useful to report scores for the subscales corresponding to the areas of driver competence, or if subscores are to be provided only for theory and practical perceived competence.

**Generalization**
Evidence to support generalizations over replications of the measurement procedure was gathered through an examination of the internal consistency. The internal consistency for the subscales of theory and practical perceived competence was high indicating that items in the SSDC were homogenous.

Moreover, the item-scale correlations indicated that most items discriminate well between participants who rate their perceived competence as being high and low on the SSDC. The psychometric properties of the SSDC and the functioning of the rating scale were further examined with the RSM, which
provided additional support for construct validity. The RSM analysis confirmed the findings from the classical analysis that the psychometric properties were sound. However, the analysis also provided additional information about the functioning of the rating scale as it was indicated that the categories were located too close to one another. The conclusion was that the rating scale can be improved by decreasing the number of categories and hence making each category more clear to the respondents. Another solution that could be used to help respondents discriminate between categories is to provide labels on more categories, making each category more comprehensible. Based on these findings, the SSDC was revised to consist of a five-point scale, as research has indicated that such scales give sufficient discrimination for most purposes and are easily understood by respondents (Brace, 2004). By using a five-point scale it was also possible to provide each category with a label. This is beneficial from a validity perspective, since it makes the scale clearer to the respondents, and thereby making their responses easier to interpret (Berk, 2006). The revised SSDC was used in paper VI and classical item analyses indicated that the revised scale demonstrated good psychometric properties.

Extrapolation

The extrapolation inference can be evaluated using analytic and empirical types of evidence. According to Kane (2006) much of the analytic evidence tends to be collected during the test development process. As regards the SSDC, the process of development provided some support for adequate content-representation, as the item generation was based on the theoretical model of the construct domain. In addition, the item formulations were scrutinized by experienced item writers.

Empirical evidence of extrapolation was gathered by examining relationships between observed scores on the SSDC and other scores that are associated with the construct domain. The results indicated that the two-dimensional structure of the SSDC corresponded to the hypothesised structure of the construct domain, comprising perceived theory and practical driver competence. This provided support for the validity of the extrapolation link.

Furthermore, convergent validity evidence for the SSDC was gathered by examining the relationship between SSDC scores and scores obtained from a similar self-assessment instrument used in Finland. Because both measures were developed to assess perceived driver competence, the fact that there was a strong positive relationship between the two measures provides support for the extrapolation to the target domain.

Moreover, the relationship between scores on the SSDC and actual driving performance was examined and was found to be weak. From a self-efficacy research perspective, the relationship between perceived driver competence and performance on the driving-license test could be expected to be stronger,
as previous self-efficacy studies have reported moderate to strong relationships between perceived competence and observed performance in other areas of competence (see e.g. Choi, 2005). However, self-assessment research has indicated that people have difficulties in making accurate assessments of their own competence (Dunning, Griffin, Milojkovic, & Ross, 1990). Moreover, based on previous findings from self-assessment studies of driver competence, the relationship could not be expected to be much stronger, as these studies have reported rather weak relationships between perceived and observed driver competence (Eby, Molnar, Shope, Vivoda, & Fordyce, 2003; Hall & West, 1996).

Possible explanations of the weak relationship between perceived and observed driver competence were obtained by combining qualitative and quantitative methods in terms of semi-structured interviews and questionnaire responses, where candidates’ perceived driver competence and confidence in passing the driving test were further explored. It was found that confidence in successfully completing the tasks in the driving test was not the same as being confident of passing the test. This might explain the low correlation between SSDC scores and test performance. The findings also indicated that the candidates were not familiar with assessing their own driving skill and that they did not use the same competence areas as those in the SSDC to express their own strengths and weaknesses as a driver. One of the validity threats in relation to the extrapolation-link is that of construct underrepresentation, i.e. that the instrument fails to tap the whole domain. The fact that the interviewees described their competence using other terms than those included in the SSDC, indicates that the instrument does not tap the whole domain of perceived driver competence. However, when validating the interpretation of scores from an instrument, it is important to consider the intended purpose and use of the scores, as an interpretation of a test score can be valid for one purpose but not for another. The purpose of the SSDC was to assess specific aspects of perceived driver competence and to relate these to observed driving skill, in terms of the performance on the Swedish driving-license test. For this reason, the contents of the SSDC corresponded to the competences assessed in the driving-license test. The studies presented in this thesis suggest that the score interpretations from the SSDC can be considered valid for the intended use of the scores. However, it cannot be assumed that the scores from the SSDC can used as a general measure of perceived driver competence.

Further evidence for the extrapolation link was obtained when the results from the SSDC were compared to the results obtained from the instrument for self-assessment of driving skills used in Finland, and when the results from the Finnish instrument were compared to results from a similar instrument used in the Netherlands. The results from these studies indicated that the method used when assessing perceived driver competence affects
the results obtained from the assessment. When drivers are asked to assess their driving skill with respect to different aspects of skill rather than by comparing themselves to the average driver, a fairly large proportion makes a realistic assessment. Furthermore, in contrast to results from previous studies, in which drivers assessed their skills as compared to others, men are not found to be more overconfident than women. The fact that similar findings were obtained in three independent samples using three different instruments for perceived competence provides support for these findings as well as validity evidence for the scores from the SSDC.

Validity evidence can also be obtained by comparing groups of respondents that are believed to differ on the construct of interest (AERA et al., 1999). Research has indicated that practicing self-assessment improves the accuracy of the perceived competence (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Jönsson, 2008). In Finland, self-assessment has been incorporated in driver education since the year 2000, while self-assessment constitutes new content in the curriculum for driver education in Sweden and the Netherlands. Therefore, Finnish candidates can be expected to have a more accurate perception of their own skills than Swedish and Dutch candidates. This expectation was confirmed, as studies IV and V indicated that a greater proportion of Swedish and Dutch candidates overestimated their competence compared to the Finnish candidates.

**Evaluation**
Although validity threats regarding the evaluation of the scores were not explicitly studied in this thesis, the procedure of using the SSDC initially involved driver examiners comparing the scores on the SSDC to the performance in the driving test in order to provide feedback to the candidates about the accuracy of their self-assessment. One problem with this procedure, however, was that neither driver examiners nor candidates were familiar with self-assessment procedures. Therefore, if the SSDC is going to be used in the driver testing in the future, and feedback regarding the accuracy of self-assessment is to be provided to the candidates, there is a need to develop procedures for score interpretation and feedback that are easy to use for driver examiners. In addition, driver examiners would also need training to be able to make accurate score interpretations and to deliver feedback in an adequate manner.

**Decision and impact**
In the studies included in this thesis, the SSDC was used in a research project, and therefore, no decisions were made based on the SSDC scores. Consequently, the use of the instrument did not result in any consequences for the respondents. When the SSDC was developed, the idea was however that the scores of the SSDC could be compared with the performance in the theory and driving test.
in order to examine the accuracy of candidates’ subjective driving skill. The idea was also that the comparison of SSDC scores and test performance could be made by the driver examiner after the driving test, and that he or she could provide feedback on the accuracy of self-assessment to the candidate. If the SSDC is to be implemented in the driver education system, and if the scores from the instrument are to be used for examining the accuracy of drivers’ perceived competence, there is a need to evaluate the potential negative and unintended consequences of the use of the instrument. Moreover, if decisions regarding self-assessment accuracy should be made based on the scores from the SSDC and the performance in the driving test, it is important that standards for the different outcomes (e.g. underestimation, accurate assessment, and overestimation) are developed and validated. Furthermore, if the purpose of the self-assessment is that it should contribute to more realistic self-assessments and increased reflection on one’s own competence, evaluations need to be made of whether the use of the SSDC had such an impact or not.

**Empirical and procedural validity evidence**
The studies included in this thesis have evaluated and provided support for different aspects of construct validity of the SSDC. From a modern validity perspective, validation is an ongoing process and the collection of different kinds of validity evidence strengthens the interpretations made from the test scores. According to Downing & Haladyna (1997) validity evidence comes in empirical and procedural forms. So far the validity evidence gathered for the SSDC has mainly been empirical, focusing on internal structure, psychometric properties as well as relationships to other similar constructs. Although the findings from these studies have provided support for construct validity of the scores from the SSDC, the procedural validity also needs to be considered. Factors related to the testing procedure, such as administration and scoring, are important for making valid interpretations of the scores from an instrument (AERA et al., 1999). The identification of possible threats to validity associated with the eight links in the testing process indicates that the procedures for using and reporting scores from the SSDC have to be further developed and evaluated. For example, even though the procedure used when evaluating the SSDC was similar for the respondents, it was not completely standardised. If the SSDC is to be used in the driving testing in the future, a standardised test procedure should be developed. Moreover, it is crucial that those who administer the SSDC be familiar with this procedure and that the driver examiners that interpret the scores have received adequate training in using the SSDC and providing feedback to the candidates.
Applying self-assessment in driver education

Commonly, the driver education in many countries has tended to focus on knowledge and skills related to vehicle manoeuvring and mastery of traffic situations. With the formulation of the GDE model, the importance of risk-increasing factors, individual and social circumstances as well as self-assessment have been recognised (Hatakka et al., 2002). In Sweden, the introduction of the new curriculum for driver education based on the GDE model brought attention to the importance of self-assessment. In order to measure the goals for self-assessment included in the curriculum, a self-report instrument, the SSDC, was developed and evaluated in this thesis. The design of comparing perceived with observed driver competence allowed us to examine how accurately the drivers assessed their own competence.

When the SSDC was used in this thesis, most participants did not have any previous experience of assessing their own driver competence. Research on self-assessment has indicated that practice in self-assessment and feedback is essential in order to be capable of assessing one’s own skill accurately (Boud & Falchikov, 1989; Dochy et al., 1999). It has also been found that self-assessment is a contextual rather than general skill and therefore it is important that self-assessment practice is included in the particular subject (Jönsson, 2008). Thus, in order to improve candidates' self-assessments of their own driver competence, practice of self-assessment and feedback on the accuracy of self-assessment need to be included in driver education. This is however somewhat problematic, since the amount of compulsory driver education is very small in Sweden. However, compared to the situation when the SSDC was first developed and used, the compulsory education has now been extended. Firstly, the risk education has been extended and consists of two parts; the first part concerns alcohol, drugs, fatigue, and other risk behaviours, and the second part concerns speed, safety as well as driving under special circumstances (VVFS 2008:251). Secondly, an introductory education for lay instructors and their students has been introduced (VVFS 2005:76). The aim of this education is to provide guidance for those engaging in private driver training. Due to the fact that the compulsory driver education has been extended, there is now a greater opportunity for including self-assessment training in the compulsory education.

Novice drivers’ perceived driver competence

In the research on young novice drivers, overconfidence in one's own driving skill has been presented as one explanation of why this group is over-represented in accidents (Gregersen & Bjurulf, 1996). Previous research on perceived driver competence has indicated that the majority believe that they are better than the average driver. This has been interpreted as overconfidence. In this thesis, another approach to assessing perceived driver competence was used. Driver candidates assessed specific aspects of their own driver
competence, and their assessments were compared with their performance in the driving test in order to examine the accuracy of their self-assessments. In contrast to previous research, the results from this thesis indicated that a fairly large proportion of the candidates made a realistic assessment of their driver competence, especially in the Finnish sample. Around 50 percent of the candidates in the Finnish sample, and between 25 and 35 percent in the Swedish sample made a realistic assessment. Worth to note, however, is that only assessments that were in the same categories were considered realistic (examiner assessment poor/candidate assessment low, examiner assessment mediocre/candidate assessment medium, examiner assessment good/candidate assessment high). If ratings in adjacent categories would have been considered realistic assessments as well (e.g. examiner assessment mediocre and candidate assessment high) the proportion of candidates making realistic assessments would have been much larger (between 78 and 91 percent in the Swedish sample).

Moreover, in contrast to previous studies (DeJoy, 1992; Finn & Bragg, 1986), the findings from this thesis did not show any gender differences in the accuracy of self-assessment. This indicates that the method used for measuring perceived driver competence has great implications for the results obtained. Previous self-assessment research also indicates that the comparison with the average person is problematic and that it might result in biased assessments. Firstly, when people make self-assessments based on traits that are ambiguous, people tend to define the traits to their own advantage (Dunning, Meyerowitz, & Holzberg, 1989). This results in people tending to believe they are above average. Both the terms average driver and driving skill can be considered ambiguous terms that can be interpreted in different ways, which might explain why drivers believe they are better than average. Secondly, it has been shown that when assessing their own skill as compared to that of others, people fail to take into account the skill of the comparison group. This results in people tending to rate themselves as more skilled than the average when the task is easy and less skilled than the average person when the task is difficult (Kruger, 1999). These findings might provide an explanation for why drivers’ seem overconfident when asked to compare their own skill to that of the average driver. In conclusion, these problems can be avoided by choosing another strategy for assessing perceived driver competence. In order to examine if drivers have an accurate perception of their own skills, perceived driver competence should be assessed with respect to specific aspects of driver competence, and should be compared to the observed driving performance. Developing an instrument for perceived driver competence that provides valid assessments is one important step in the understanding of novice drivers’ driving behaviour and accident involvement.

The content of the SSDC was based on the goals of the Swedish curriculum, which in turn is based on the GDE model. The fact that there are other driving
license systems that have also based their driver education and testing on the GDE model provides a possibility to adapt the SSDC to other driver education systems. The findings from studies IV and V indicated that similar results are provided from the SSDC and two instruments for self-assessment for driving skills used in Finland and the Netherlands. Because these countries share many similarities with regard to contents in the driving test, it would probably be possible to have one instrument for self-assessment of driving skills common to all three countries.

**Limitations and generalisations**

The studies presented in this thesis have some limitations that need to be mentioned. One limitation is related to the examiner assessments. When the SSDC was used in the project for testing and evaluating a new model for driver testing, examiners assessed candidates’ driving performance in the four competence areas as poor, mediocre or good, instead of only in terms of pass/fail. Even though an effort was made to inform examiners about the scale, how to use the scale, and how to record the performance on the driving test form, we have no information regarding the inter-rater reliability of examiner assessments. A similar problem was probably also present in the Dutch data used in study V. Another problem in the Dutch sample related to examiner assessments was the large proportion of missing values. Measures were taken to reduce the problem as missing values were estimated, but still the results need to be interpreted with caution because the proportion of missing was fairly large. Because no information was available about the reliability of the examiner assessments, the comparisons of examiner and candidate assessment should be interpreted with some caution. These comparisons should be replicated on a sample where the reliability of examiner assessment can be examined.

Another limitation is related to the comparison of examiner and candidate assessments. As the assessment of perceived driver competence was a novel feature in the Swedish driver education system, there was no established procedure for comparing candidate and examiner assessments. There are different ways in which a realistic assessment can be conceptualised. In study IV in this thesis the assessments were considered realistic only if they were in the same category as examiner assessments. However, it can be argued that adjacent assessments can be viewed as realistic. For example, if the candidate rated his/her perceived driver competence ‘high’, this means that the candidate is very confident in passing the test. If the performance on the driving test is assessed as ‘mediocre’, this means that the candidate passed the test, and thus the self-assessment was fairly realistic. Depending on how a realistic assessment is defined, the proportions of candidates underestimating, overestimating and making a realistic assessment of their own driver competence will differ. If the SSDC is going to be used with in
driver testing, a procedure for comparing candidate and examiner assessment, as well as standards for what should be regarded as a realistic assessment have to be developed.

The comparison of candidates’ self-assessments in Sweden, Finland and the Netherlands are also somewhat problematic due to the different instruments for perceived driver competence. The instruments used in Sweden and Finland had different number of scale points and the extremes on the scales had different labels. Due to these differences, the comparisons of Finnish and Swedish candidates with respect to self-assessment accuracy should be viewed only as preliminary results. In order to make more reliable comparisons of candidates’ self-assessment accuracy in different countries, there is a need to use the same instrument for perceived competence in all samples included in the study.

The results from studies II to IV is generalisable to all Swedish driving-license candidates, as the samples used can be regarded representative for the population. However, we cannot draw any conclusions about the results we would obtain if the SSDC would be administered to another population, e.g. experienced drivers.

**Suggestions for further studies**

The methods used for assessing subjective driving skill seem to affect the results from the assessment. Therefore, it is of great importance that psychometric evaluations are made in order to examine the reliability and validity of the instruments used for assessing subjective driving skills. So far, the validity evidence collected for the SSDC has mainly focused on internal structure and relations to other variables. In addition to these kinds of validity evidence, there are other types of evidence that can be collected. This would be a good suggestion for further studies. For example, it would be desirable to examine the procedural validity in terms of administration and scoring. If the SSDC is to be used within driver testing, the consequences of the use of the instrument should be examined. In order to obtain further support for the fact that the strategy for measurement of perceived driver competence affects the results, it would be interesting to let candidates rate their own driving skill as compared to the average driver in addition to completing the SSDC. Then the results obtained from the SSDC could be contrasted to the self-assessments made in relation to the average driver.

There are also topics for future studies that would be of interest for the applied field. Previous self-assessment research indicates that experts make more accurate self-assessments than novices (Kruger & Dunning, 1999). The participants in the studies included in this thesis were driver candidates. Therefore, there was no large variation in driving experience. Because of this it was not possible to fully examine the relationship between driving experience and accuracy of self-assessment. It would therefore be interesting to compare a
group of novice drivers and experienced drivers with respect to self-assessment accuracy. Moreover, it would be interesting to examine how the accuracy of perceived driver competence could be improved. Boud (1995) emphasised that familiarity with criteria is important for the accuracy of self-assessment. Firstly, it is important that students are familiar with the standards and criteria used for assessing their work, and secondly that they can assess to what extent these criteria have been reached. The results from study VI indicated that candidates are not familiar with the requirements for passing the driving test. For this reason it would be interesting to examine if the accuracy of self-assessment would improve if the candidates received additional information of the standards for the driving test.

Furthermore, research indicates that feedback is an important part of self-assessment practice and that it can be used to improve the accuracy of self-assessment. According to Taras (2001) feedback provides an opportunity for the student to understand his or her own strengths and weaknesses in the sense that the feedback provides information about the qualities of the work and what needs to be improved. The self-assessment combined with feedback helps the student learn how to assess his or her own performance realistically. Based on this it would be interesting to examine if feedback on self-assessment and performance could lead to more accurate self-assessments.
8. Swedish summary

Inledning och syfte

Validitetsbegreppet

I denna avhandling görs valideringen av instrumentet för upplevd förarkompetens med utgångspunkt i det moderna validitetsbegreppet. I de fem empiriska studierna i avhandlingen samlas olika typer av validitetsbevis för att ge stöd för att de tolkningar som görs av resultatet från instrumentet är rimliga och relevanta.
Utveckling och validering av ett självskattningsinstrument för upplevd förarkompetens

Hur mäter man upplevd förarkompetens?


För att undersöka hur realistisk syn blivande bilförare har på sin egen kompetens utvecklades ett självskattningsinstrument för specifik upplevd förarkompetens, för att användas inom ramen för det svenska förarutbildnings sistems. Instrumentet baserades på det teoretiska begreppet self-efficacy, vilket är ett begrepp som fokuserar på den egna upplevda kompetensen i en specifik, avgränsad domän. Instrumentet benämndes Self-Efficacy Scale for Driver Competence (SSDC), i den resterande delen av denna sammanfattning kommer förkortningen SSDC att användas. För att det skulle vara möjligt att jämföra den upplevda förarkompetensen med den observerade förarkompetensen, i termen av prestation på förarprovet, baserades instrumentet på målen i kursplanen för förarutbildning, behörighet B samt på innehållet i den så kallade Goals for Driver Education (GDE) modellen, som ligger till grund för innehållet i kursplanen. För att tydliggöra vad instrumentet skulle mäta utvecklades en teoretisk modell för begreppet upplevd förarkompetens. Denna modell bestod av två dimensioner av specifik upplevd kompetens; teoretisk och praktisk. Dessa dimensioner innefattade i sin tur fem aspekter av förarkompetens som återfinns i kursplanen för förarutbildning och i förarprovet. Nästa steg i instrumentutvecklingen var att generera frågor som mäter aspekterna i den teoretiska modellen. Dessa frågor granskades av

Utveckling och psykometrisk utvärdering av SSDC

En utvärdering av instrumentets inre struktur och mätsäkerhet redovisas i studie II och III i avhandlingen. I studie II administrerades de två versionerna av SSDC till två stickprov med körkortstagare (n = 805, n = 721). Resultatet visade att mätsäkerheten är god samt att frågorna i instrumentet är homogena, dvs. att de mäter samma underliggande begrepp. Resultatet från en explorativ faktoranalys visade att frågorna i instrumentet mäter två dimensioner av upplevd förarkompetens: teoretisk och praktisk. Frågorna som handlar om hur säker man är på att klara uppgifter i kunskapsprovet mäter en underlig-gande faktor (upplevd teoretisk förarkompetens) och frågorna som handlar om hur säker man är på att klara uppgifter i körprovet mäter en annan underliggande faktor (upplevd praktisk förarkompetens). Denna resultatbild bekräftades när en konfirmatorisk faktoranalys gjordes på det andra stickprovet. Dessa dimensioner överensstämmar med strukturen i den teore-tiska modellen som låg till grund för utvecklingen av instrumentet, vilket ger stöd för validiteten.

I studie III undersöktes kvaliteten i instrumentet vidare genom att använda en annan analysmetod; en item response theory (IRT) modell. Denna analysmetod användes också för att undersöka hur väl skattningsskalan fungerade. De två versionerna av SSDC besvarades av 795 respektive 714 körkortstagare. Resultaten bekräftade de resultat som erhölls i Studie II, dvs. att mätsäkerheten är god. Dessutom gav analysen ytterligare information
som visade att skattningsskalen kan förbättras ytterligare och göras tydligare för respondenterna, genom att minska antalet skalsteg, alternativt genom att förse varje skalsteg med en benämning.


**Körkortstagares upplevda förarkompetens**

Syftet med studie IV var att undersöka hur realistisk syn körkortstagare har på sin egen förarkompetens. I tidigare studier har detta i huvudsak gjorts genom att be förare skatta sin skicklighet i relation till genomsnittsföraren. Dessa studier har indikerat att en stor andel överskattar sig, eftersom majoriteten anser att de är bättre än genomsnittsföraren. Några av dessa studier har också visat att tendensen till överskattning verkar vara störst bland unga män. Till skillnad från dessa tidigare studier, användes en annan metod i studie IV; körkortstagare skattade sin upplevda förarkompetens inom ett antal specifika områden och deras upplevda kompetens jämfördes sedan med deras prestation på körprovet för att undersöka om de hade en realistisk syn på sin kompetens, eller om de överskattade eller underskattade sin kompetens. Några av dessa studier har också visat att trendens till överskattning verkar vara störst bland unga män. Till skillnad från dessa tidigare studier, användes en annan metod i studie IV; körkortstagare skattade sin upplevda förarkompetens inom ett antal specifika områden och deras upplevda kompetens jämfördes sedan med deras prestation på körprovet för att undersöka om de hade en realistisk syn på sin kompetens, eller om de överskattade eller underskattade sin kompetens. Ett stickprov med svenska körkortstagare (n = 805) besvarade SSDC och gjorde det svenska körprovet, och ett stickprov med finska körkortstagare (n = 2739) besvarade ett självskattningsinstrument för upplevd förarkompetens som liknar SSDC, och gjorde det finska körprovet. Till skillnad från tidigare studier, visade resultaten från denna studie att en relativt stor andel av körkortstagarna gjorde en realistisk skattning av sin förarkompetens. Ungefär 50 procent av de finska och 25 till 35 procent av de svenska körkortstagarna gjorde en realistisk skattning av sin förarkompetens inom de olika kompetensområdena. Beroende på vad som definieras som realistisk skattning får man dock olika resultatbilder. I studie IV krävdes att körkortstagarna skattade sin upplevda kompetens i kategorin som motsvarade förarprövarens bedömning för att skattningen skulle definieras som realistisk. Om skattningar som bara skiller ett steg också ses som realistiska (t.ex. hög upplevd förarkompetens och en prestation på körprovet som ligger runt god-
kändgränsen) är det en betydligt större andel som gör en realistisk skattning (78-91 % av de svenska körkortstagarna). Ett annat intressant resultat var att inga statistiskt säkerställda könsskillnader fanns med avseende på andelen som överskattar, gör en realistisk skattning eller underskattar sin förarkompetens. Vidare visade en jämförelse av de finska och svenska körkortstagarna att andelen som gjorde en realistisk skattning var större i det finska stickprovet och att andelen som överskattade sin kompetens var större i det svenska stickprovet. En möjlig förklaring till detta är att de finska körkortstagarna har mer erfarenhet av självvärdering. Forskning om självvärdering pekar på att övning i självvärdering är av stor vikt för att göra en realistisk bedömning av sin egen kompetens. I Finland är självvärdering i förarprovet en etablerad procedur sedan år 2000. I Sverige är självvärdering något nytt för provtagarna eftersom instrumentet för upplevd kompetens endast har testats i en försöksverksamhet under sju månader.

I studie V jämfördes resultaten från det självskattningsinstrument för upplevd förarkompetens som används i Finland (n = 2739) med resultaten från ett liknande instrument som testats i en försöksverksamhet i Nederländerna (n = 239). I likhet med studie IV visade resultaten att en relativt stor andel av körkortstagarna gjorde en realistisk skattning av sin förarkompetens. Mellan 40 och 50 procent av körkortstagarna i Finland och Nederländerna gjorde realistiska skattningar av sin förarkompetens. Vidare fanns inga betydande skillnader mellan män och kvinnor eller mellan körkortstagare i olika åldrar beträffande andelen som överskattade, gjorde en realistisk skattning eller underskattade sin förarkompetens. Detta indikerar att om man mäter upplevd förarkompetens, genom att jämföra specifika aspekter av upplevd förarkompetens med observerad förarkompetens, får man en annan resultatbild jämfört när man ber personer skatta sin förarkompetens i relation till genomsnittsföraren. Vidare visade resultaten från studien att det fanns skillnader mellan godkända och underkända provtagare med avseende på andelen som överskattade, gjorde en realistisk skattning eller underskattade sin förarkompetens. Andelen som gjorde en realistisk skattning var större bland dem som blev godkända än underkända och andelen som överskattade sig var större bland de underkända än bland de godkända. Detta resultat stöder tidigare studier om självvärdering som visar att mer kompetenta provtagare tenderar att göra en mer realistisk skattning av sin kompetens, medan mindre kompetenta provtagare tenderar att överskatta sin kompetens mer (Kruger & Dunning, 1999).

Relationen mellan upplevd och observerad förarkompetens
I studie VI undersökte validiteten i den reviderade versionen av SSDC vidare, dels genom att undersöka sambandet mellan SSDC och det instrument för självvärdering av förarkompetens som används i Finland, dels genom att undersöka sambandet mellan SSDC och prestationen på körprovet. Resultaten
nyanserades också med information från intervjuer med körkortstagare om deras upplevda förarkompetens. 357 körkortstagare som bokat tid för körprov besvarade en enkät som innehöll SSDC, en svensk översättning av det finska instrumentet för upplevd förarkompetens samt några frågor som handlade om hur säker man var på att bli godkänd. Sju av dessa körkortstagare medverkade också i intervjuer, där deras syn på sin förarkompetens utforskades vidare. Resultatet visade att det fanns ett starkt samband mellan resultatet på SSDC och det finska självvärderingsinstrumentet. Detta ger stöd för att båda instrumenten mäter samma underliggande begrepp, dvs. upplevd förarkompetens. Vidare visade resultaten också att det inte fanns något signifikant samband mellan svaren på SSDC (hur säker man är på att klara olika delar i körprovet) och prestation på körprovet. Möjliga förklaringar till detta erhölls via intervju- och enkätsvaren. Intervjuerna indikerade att körkortstagarna inte var vana att värdera sin egen förarkompetens, och att de inte heller var vana vid att beskriva sin förarkompetens i termer av de aspekter som finns med i SSDC. Vidare visade intervjuvaren att körkortstagarna inte har så stor kunskap om vad som krävs för att bli godkänd på körprovet. Detta kan vara möjliga förklaringar till varför det inte fanns något samband mellan SSDC och prestationen på körprovet. En annan möjlig förklaring till avsaknaden av samband erhölls från enkätsvaren. De visade att provtagarna tycker att säkerheten att bli godkänd på provet och säkerheten att genomföra olika delar i körprovet är skilda saker. Körkortstagarna kan vara ganska säkra på att klara de olika delarna som ingår i provet, men ändå vara tvetskamma till att de ska bli godkända på provet, på grund av faktorer som nervositet, okunskap om vad som krävs för att bli godkänd, och att något oväntat kan inträffa under provet. Detta kan också vara en möjlig förklaring till varför det inte fanns något samband mellan SSDC och provprestation. Sammantaget, visade denna studie på vikten av att kombinera kvalitativa och kvantitativa data i validering av instrument, då de olika metoderna bidrog med unik information om hur väl instrumentet fungerande. Denna information hade inte varit tillgänglig om endast en metod hade använts.

**Ger SSDC valida mätningar av upplevd förarkompetens?**

Studierna som presenteras inom ramen för denna avhandling har beskrivit utvecklingen och valideringen av ett självskattningsinstrument för specifik upplevd förarkompetens. Sammantaget ger resultaten från dessa studier stöd för att instrumentet fungerar bra sett i ett mätperspektiv och att de tolkningar som görs av resultatet från instrumentet är rimliga och relevanta.

I detta avsnitt förs en vidare diskussion om validiteten i resultaten från instrumentet och möjliga hot mot validiteten identifieras och diskuteras. Enligt Crooks och Kane (1996) kan validering av en mätning beskrivas som en kedja bestående av åtta länkar, där varje länk motsvarar ett steg i testpro-
cessen (Crooks och Kane, 1996). Valideringen innebär att man analyserar möjliga hot mot validiteten som förknippas med varje enskild länk.

Den första länken handlar om *administreringen* av ett instrument. Inom ramen för Vägverkets försöksverksamhet besvarade körkortstagarna SSDC i vänturummet på Vägverket just innan de genomförde sitt kunskapsprov. Denna procedur kan ha inneburit att vissa körkortstagare inte hade tillräckligt med tid att besvara instrumentet, och att de hade svårt att fokusera och koncentrera sig därför att de snart skulle göra kunskapsprovet. Om SSDC ska användas i skarpt läge inom förarprovssystemet är det därför viktigt att utforma en standardiserad procedur för administrering av instrumentet.

Den andra länken handlar om *poängsättningen*. När SSDC användes i försöksverksamheten räknade personal på förarprovskontoret ut en genomsnittlig poäng för varje provtagare inom de fyra olika kompetensområdena för den upplevda praktiska förarkompetensen. I detta sammanhang kan möjliga hot mot validiteten i denna länk vara räknelån. Poängen från SSDC användes sedan för att avgöra om provtagarna hade en realistisk syn på sina kompetens inom de olika områdena, genom att jämföra poängen med prestationen i körprovet inom motsvarande områden.

Den tredje länken handlar om att *aggregera* poängen på de frågor som tillhör samma kompetensområde i SSDC. Utvärderingen av den inre strukturen i SSDC visade att frågorna mäter två underliggande dimensioner av upplevd förarkompetens: teoretisk och praktisk. Således finns det inget stöd för att rapportera poäng för varje kompetensområde, utan poäng bör rapporteras för frågorna om praktisk och teoretisk förarkompetens.

Den fjärde länken handlar om att det ska vara möjligt att *generalisera* från de specifika frågorna tillhor samma kompetensområde som ställdes i instrumentet, till begreppsdomen för upplevd förarkompetens. Stöd för detta erhölls då analysen visade att frågorna i instrumentet mäter de två dimensioner av upplevd förarkompetens som instrumentet var utformat för att mäta. Det starka sambandet som fanns mellan SSDC och instrumentet för självvärdering av förarkompetens som används i Finland gav också stöd för att frågorna i SSDC mäter upplevd förarkompetens. Även det faktum att SSDC, självvärderinginstrumenten i Finland och Nederländerna gav liknade resultat när de används i tre olika stickprov med körkortstagare, gav stöd för att instrumenten mäter samma underliggande begrepp.

Vidare visade resultaten från studie IV och V att metoden för hur man mäter upplevd förarkompetens är avgörande för resultatet. Tidigare studier där man mätt upplevd förarkompetens i relation till genomsnittsföreningen indikerade att majoriteten överkattar sig, medan dessa två studier, där specifika aspekter av upplevd förarkompetens jämförts med prestation på körprovet, visar att
en relativt stor andel gör en realistisk skattning av sin kompetens. Vidare noterades inte heller några könsskillnader beträffande hur realistisk bedömning man gör av sin förarkompetens, något man funnit i tidigare studier. Att resultat från tre olika oberoende stickprov visar att en relativt stor andel av körkortstagarna kan göra en realistisk bedömning av sin kompetens, och att det inte finns några könsskillnader, ger stöd för giltigheten i dessa slutsatser. Detta talar för att strategin för att mäta upplevd förarkompetens är avgörande för resultaten och validiteten i de tolkningar som görs av resultaten.

Jämförelsen mellan körkortstagares självvärdering i Finland och Sverige samt Finland och Nederländerna indikerade också att övning i självvärdering är viktigt för att kunna göra en realistisk bedömning av sin förmåga. En större andel av de finska körkortstagarna, jämfört med de svenska och nederländska, gjorde en realistisk bedömning av sin förarkompetens, medan andelen överskattning var större i Sverige och Nederländerna jämfört med i Finland. I Finland har självvärderingsproceduren varit etablerad sedan år 2000, medan man i Sverige och Nederländerna bara använt självvärdering inom ramen för försöksverksamheter. Att det krävs träning för att göra en realistisk bedömning av sin förmåga, stöds av tidigare forskning om självvärdering (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Jönsson, 2008).

Sambandet mellan SSDC och prestationen på körprovet kan också vara en indikator på validiteten i mätningen. Ett starkt samband indikerar att det finns en överensstämmelse mellan den upplevda och observerade förarkompetensen. Resultaten visade att det inte fanns något statistiskt säkerställt samband mellan resultatet på SSDC och prestationen på körprovet. Å ena sidan skulle man kunna förvänta sig ett relativt starkt samband då tidigare studier visat att instrument som mäter specifik upplevd kompetens (self-efficacy) inom andra områden än förarkompetens visar relativt starka samband med prestation i termen av resultat på prov. Å andra sidan, är detta resultat i linje med andra studier om självvärdering av förarkompetens som visat på svaga eller obeintliga samband mellan upplevd förarkompetens och observerad förarkompetens. Möjliga förklaringar till varför det inte fanns något samband erhålls via intervjuer om upplevd förarkompetens. Intervjuerna indikerade dels att körkortstagarna inte är vana att värdera sin egen förarkompetens, och dels att de inte är vana att beskriva sin kompetens i termen av de kompetensområden som beskrivs i SSDC.

Ett av validitetshoten när det gäller att dra slutsatser från frågorna i instrumentet till begreppsdomänen för upplevd förarkompetens, är att instrumentet inte lyckas mäta hela begreppsdomänen. Att så var fallet för SSDC indikerades av det faktum att intervjuerna beskrev sin förarkompetens med andra termer än de som presenterades i SSDC. Enligt det moderna synsättet på validitet är det dock viktigt att utvärdera validiteten med avseende på syftet med ett instrument och vad resultaten är tänkta att användas till. SSDC är utformat för att mäta upplevd förarkompetens i
förhållande till den observerade förarkompetens, i termer av prestation på det svenska körkortsprovet. Därför är det naturligt att instrumentet mäter upplevd kompetens med avseende på de aspekter av förarkompetens som mäts i förarprovet. De studier som redovisats i denna avhandling ger stöd för att resultaten från SSDC ger en valid mätning av körkortstagarnas upplevda kompetens för denna användning. Det är däremot inte säkert att SSDC ger en valid mätning av upplevd förarkompetens generellt.

Den sjätte länken i valideringskedjan är tolkningen av resultatet. Även om detta inte var föremål för någon av studierna i denna avhandling är tolkningen av resultatet en oerhört viktig aspekt av validiteten, särskilt om instrumentet ska användas i skarpt läge inom ramen för förarprovssystemet. När SSDC användes inom ramen för försöksverksamheten för utvecklat förarprov jämförde förarprövarna provtagarnas upplevda förarkompetens med deras prestation på körprovet. Utifrån denna jämförelse fick provtagarna feedback på hur realistiskt de hade bedömt sin egen förarkompetens. Under försöksverksamheten hade varken provtagarna eller förarprövarna erfarenheter av den här typen av procedur. Provtagarna var inte vana vid att värdera sin förarkompetens och förarprövarna hade inte någon tidigare erfarenhet av att ge feedback om självvärdering. När SSDC utvecklades var en av tankarna att instrumentet skulle kunna bidra till att körkortstagarna får insikter om sina styrkor och svagheter och att det kan bidra till en mer realistisk syn på sin kompetens. För att användningen av instrumentet ska få sådana effekter krävs dock att en relevant återkoppling görs av hur realistisk syn körkortstagaren har på sin kompetens. Om SSDC ska användas inom ramen för förarprovet är det därför viktigt att förarprövarna får utbildning i att tolka och förmedla resultatet på SSDC.

De två sista länkarna i valideringskedjan handlar om vilka beslut som fattas utifrån resultatet och de konsekvenser som dessa beslut får. Eftersom SSDC endast har använts inom ramen för försöksverksamhet så här långt har inga beslut fattats utifrån resultatet. Därmed är det inte möjligt att identifiera några validitetshot som har med beslut och konsekvenser att göra. Om SSDC ska implementeras i förarprovssystemet och användas för att undersöka hur realistisk syn blivande förare har på sin förarkompetens, är det dock av stor vikt att undersöka dessa aspekter av validiteten. Exempelvis är det viktigt att försätta sig om att användningen av instrumentet inte resulterar i några oönskade, negativa konsekvenser för körkortstagarna. På sikt är det också angeläget att undersöka om användningen av SSDC får de konsekvenser man tänkt sig, dvs. en ökad medvetenhet och en mer realistisk syn på sin förarkompetens.
Fortsatt forskning och utvecklingsarbete

Validitetsbevis kan ha två olika karaktärer, dels i form av resultat från empiriska undersökningar, dels i form av beskrivningar och utvärderingar av de procedurer som använts vid utvecklingen och administreringen av instrumentet (Downing & Haladyna, 1997). I denna avhandling har tonvikten legat på att samla in empiriska bevis för validitet. Sammantaget har de studier som gjorts inom ramen för denna avhandling gett stöd för att de tolkningar som görs av resultaten från SSDC är rimliga och relevanta. För att gå vidare i utvecklings- och valideringsarbetet är det angeläget att vidareutveckla procedurerna för användningen av instrumentet och utvärdera dessa, eftersom även de påverkar hur giltiga de tolkningarna är som görs av resultatet. En sådan utvärdering är framför allt angelägen om instrumentet ska användas inom ramen för förarprövningen för att undersöka körkortstagarnas syn på sin egen förarkompetens.


Vad gäller träning för körkortstagarna i att värdera sin förarkompetens vore det mest optimala att införa sådana inslag i den obligatoriska förarutbildningen. Vid tidpunkten för utvecklingen av SSDC var mängden obligatorisk förarutbildning begränsad till en riskutbildning på ca sex timmar, vilket medförde att det inte var möjligt att införa nya inslag i utbildningen. Idag är dock situationen delvis annorlunda. Dels har en obligatorisk introduktionsutbildning för handledare och elev vid privat övningskörning införts, och dels har riskutbildningen utökats till att bestå av två delar. En förhoppning
är därför att det skulle vara möjligt att införa självvärderingsövningar inom ramen för dessa utbildningsinslag.

Eftersom tidigare forskning om självvärdering har visat att övning i självvärdering bidrar till att värderingen av den egna kompetensen blir mer realistisk, vore det intressant att undersöka om utbildning i självvärdering inom förarutbildningen skulle ge liknande effekter. För att undersöka detta vore det intressant att genomföra en studie där man jämför två grupper av körkortstagare; en som fått självvärderingsträning och en som inte fått det, med avseende på hur realistiskt de bedömer sin egen förarkompetens. I detta sammanhang vore det också intressant att undersöka andra faktorer som kan påverka hur realistisk syn man har på sin egen förarkompetens. Exempelvis har tidigare studier visat att mer kompetenta personer tenderar att ha en mer realistisk uppfattning om sin egen kompetens än mindre kompetenta personer (Kruger & Dunning, 1999). Det vore intressant att undersöka om så också är fallet för upplevd förarkompetens. Därför vore det intressant att jämföra en grupp blivande förare och en grupp erfarna förare med avseende på hur realistisk syn de har på sin egen förarkompetens. Denna typ av studier är viktiga för att öka kunskapen om hur synen på den egna kompetensen påverkas, och kan förhoppningsvis på sikt användas för att hjälpa körkortstagare att utveckla en mer realistisk syn på sin förarkompetens.
References


Appendix

Errata

Study I
Page 5, paragraph 2, and page 7, paragraph 3, the references to Lajunen et al., 1997 and Lajunen, Corry, Summala, & Hartley, 1998 should be removed.

Page 6, last paragraph, and page 7, second paragraph, page 8, line 2; AERA, APA, & NCME (2004) should be AERA, APA, & NCME (1999).

Study IV
In the abstract, n= 2847 should be n = 2739.

In Table 6, the heading for the columns; Swedish candidate assessments should be Swedish examiner assessments, and the columns should be entitled Poor, Mediocre and Good. The rows should be entitled Swedish candidate assessments, and the categories should be entitled Low, Medium and High.

Study V
Page 3, Statistical analysis section, last paragraph; “and average number of driving lessons” should be deleted.

Table 4, column 5, row 10; 11. should be 11.5.