Licence to Drive
The Importance of Reliability for the Validity of the Swedish Driving Licence Test

Susanne Alger

Department of Applied Educational Science
Educational Measurement
Umeå 2019
Responsible publisher under Swedish law: the Dean of the Social Faculty
This work is protected by the Swedish Copyright Legislation (Act 1960:729)
ISBN: 978-91-7855-115-6
ISSN: 1652-9650
Cover art and design: Björn Sigurdsson
Electronic version available at http://umu.diva-portal.org/
Printed by Cityprint i Norr AB
Umeå, Sweden 2019
“So reliability is about measuring things right and validity about measuring the right things?”

My husband after reading this licentiate thesis
# Table of Contents

Abstract ........................................................................................................................... iii

List of papers .................................................................................................................. v

Introduction .................................................................................................................... 1  
Aims and research questions ........................................................................... 3  
Structure ....................................................................................................................... 4

The Swedish driving licence system ............................................................................ 5  
The role of assessment ............................................................................................. 5  
Historical background ............................................................................................... 6  
Goals in the curriculum and their theoretical foundation ...................................... 8  
Operationalization – from goals to test format ....................................................... 10  
The test and test administration ........................................................................... 11  
Previous research about the quality of the Swedish driving licence test .............. 13

Quality in measurement ............................................................................................. 16  
Validity ....................................................................................................................... 16  
Reliability .................................................................................................................... 19

Data and methods for analysis .................................................................................... 22  
Data collection – Study I .......................................................................................... 22  
Data collection – Study II ....................................................................................... 23  
Analysis – Study I ..................................................................................................... 23  
Analysis – Study II ..................................................................................................... 23

Summary of the studies ............................................................................................... 25  
Study I ......................................................................................................................... 25  
Study II ........................................................................................................................ 26

Discussion .................................................................................................................... 28  
Empirical studies – background and results ........................................................... 28  
Validation ..................................................................................................................... 29  
Interpretation/use argument — What is the test supposed to do? ....................... 29  
Validity argument ...................................................................................................... 32  
Further studies .......................................................................................................... 37  
Roads not yet taken – possible knowledge gaps ................................................... 37  
A closer look at validity .............................................................................................. 38  
Conclusions .................................................................................................................. 39

Sammanfattning på svenska ......................................................................................... 41  
Inledning och syfte ..................................................................................................... 41  
Validitet ....................................................................................................................... 41  
Reliabilitet .................................................................................................................... 42  
Ger kunskapsproven samma resultat? ................................................................. 42  
Gör förarprövarna samma bedömning? ................................................................. 43  
Hur kan man tolka resultaten? .............................................................................. 44  
Vad betyder reliabiliteten för validiteten? .............................................................. 44  
Fortsatt forskning och utvecklingsarbete ............................................................... 45

Acknowledgements .................................................................................................... 46

References ...................................................................................................................... 47
Abstract

Background The Swedish driving licence test is a criterion-referenced test resulting in a pass or fail. It currently consists of two parts - a theory test with 65 multiple-choice items and a practical driving test where at least 25 minutes are spent driving in traffic. It is a high-stakes test in the sense that the results are used to determine whether the test-taker should be allowed to drive a car without supervision. As the only other requirements for obtaining a licence is a few hours of hazard education (and a short introduction if you intend to drive with a lay instructor) it is important that the test result, in terms of pass or fail, is reliable and valid. If this is not the case it could have detrimental effects on traffic safety. Examining all relevant aspects is beyond the scope of this licentiate thesis so I have focused on reliability.

Methods Reliability for both the theoretical and practical test results was examined. As these are very different types of tests the types of reliability examined also differed. In order to examine inter-rater reliability of the driving test 83 examiners were accompanied by one of five selected supervising examiners for a day of tests. All in all 535 tests were conducted with two examiners assessing the same performance. At the end of the day the examiners compared notes and tried to determine the reason for any inconsistencies. Both examiners and students also filled in questionnaires with questions about background and preparation. As for studying decision consistency and decision accuracy of the theory test, three test versions (a total of around 12,000 tests) were examined with the help of methods devised by Subkoviak (Subkoviak, 1976, 1988) and Hanson & Brennan (Brennan, 2004; Hanson & Brennan, 1990).

Results The results from two research studies concerning reliability were presented. Study I focused on inter-rater reliability in the driving test and in 93 per cent of cases the examiners made the same assessment. For the tests where their opinions differed there was no correlation to any of the background variables or other variables examined except for three, which had logical explanations and did not constitute a problem. Although there were cases where the differences were due to different stances on matters of interpretation the most common suggested cause was the placement in the car (back seat vs. front seat). Although the supervising examiners gave both praise and criticism as to how the test was carried out the study does not answer the question whether the tests were equal in terms of composition and difficulty.

In Study II the focus was on decision consistency and decision accuracy in the theory test. Three versions of the theory tests were examined and, on the
whole, found to be fairly similar in terms of item difficulty and score
distribution, but the mean was so close to the cut-score (i.e. the score required
to pass) that the pass rate differed somewhat between versions. Agreement
coefficients were around .80 for all test versions (between .79 and .82
depending on method). Classification accuracy indicated an .87 probability of
a correct classification.

**Conclusion** It is important to examine the reliability and validity of the
driving licence test since a misclassification can have serious consequences in
terms of traffic safety. In the studies included here the rate of agreement
between examiners is deemed as satisfactory. It would be preferable if the
classification consistency and classification accuracy, as estimated by the
methods used, were higher for the theory test, given its importance.

While reliability in terms of agreement between raters/examiners or
consistency and accuracy of classification are routinely examined in other
contexts, such as large-scale educational testing, this is not often done for the
driving licence tests. At the same time, the methods used here can be
transferred to contexts where such properties are generally not examined.
Collecting information about test-takers and examiners, like in Study I, can
provide evidence concerning possible bias.

Examining to what extent decisions are consistent is one important aspect
of collecting evidence that shows that test results can be used to draw
conclusions about driver competence. Still, regardless of outcome, validation
is a process that never ends. There is always reason to examine various aspects
and make further improvements. There are also many other relevant aspects
to examine. A prerequisite for the validity of the score interpretation of a
criterion-referenced test like this one is that the cut-score is appropriate and
the content relevant. This should therefore be the subject of further research
as the validation process continues.
List of papers

This licentiate thesis is based on the following studies, which are referred to in the text with the following enumeration:


---

1 Study I is co-authored. I performed the data analyses and wrote about previous studies. Together we prepared the text for printing and made changes after comments from the reviewers.
Introduction

This licentiate thesis examines quality aspects of the Swedish driving licence test, primarily in terms of reliability. This is an important issue, not only for people interested in measurement, but for everyone on the streets, since this is the test that entitles successful test-takers to drive a car in traffic. A large number of people take this test each year to obtain a licence, which will entitle them to drive not only in Sweden, but in several other countries around the world. Awarding a licence to a person without the necessary qualifications can have negative consequences for traffic safety and withholding a licence from a qualified driver can have serious consequences for the individual and others affected by that decision. Or, in other words, wrongly failing someone can significantly impact their life, but wrongly awarding a licence to someone can end lives. Whether the test manages to differentiate between those who fulfil the requirements and those who do not in a consistent manner is therefore a matter worth examining.

Not only in Sweden, but also in many other countries, prospective drivers have to pass one or more tests to obtain a licence. While the format of those tests varies from country to country, having a theory part and a practical driving part is common (Genschow, Sturzbecher, & Willmes-Lenz, 2014). Despite the widespread use of such tests there are surprisingly few scientific articles published about the quality of driving licence tests in terms of reliability and validity. In simplified terms, reliability concerns the accuracy of the measurement and validity to what degree one can justify the interpretation and use of the results (whether the “right thing” is being assessed). It follows that an unreliable test is a poor starting point for a valid conclusion from the test results (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014). The measurement or assessment of a test-taker’s performance has to provide a certain accuracy to make valid conclusions possible, and this was the starting point for this licentiate thesis.

The lack of published research on the quality of driving licence tests does not mean that there are no such studies, but these are not always documented in a systematic manner. Even those that are may not be easily accessible and/or only be available in the native language. For example, many of the reports mentioned in the section about previous research in this licentiate thesis are only available in Swedish.

When the validity of driving licence tests has been discussed, it has primarily been in terms of accident liability (Elvik, Vaa, Hoye, & Sorensen, 2009; Haire, Williams, Preusser, & Solomon, 2011; Maycock & Forsyth, 1997). The reliability of the test seems to have attracted even less research interest, but there are a few examples. There has been a study of test-retest reliability...
of driving tests in Britain (Baughan & Simpson, 1999) and in Germany inter-rater reliability has been examined in small studies using simulators, video recordings and driving tests in real traffic (Sturzbecher, Luniak, & Mörl, 2016; Sturzbecher, Mörl, & Kaltenbaek, 2015).

What can one examine to ascertain the quality of a test? In general, the point of a test is not to see whether the test-taker gets a score of 34 or 52 on a particular set of items but, for example, to draw conclusions about the knowledge or skills in a particular field. Whether such interpretations are defensible depends on the reliability and validity of scores.

What conclusions can be drawn based on a test result partly depends on what type of test it is. Norm-referenced tests are designed to be used to compare the test-takers to each other or to a norm group. Criterion-referenced tests are used to determine the achievement level of the test-taker in relation to a well-defined criterion or behaviour domain (Hambleton, Swaminathan, Algina, & Coulson, 1978).

Tests and assessments can also come in very different formats, which places different demands on design, administration and evaluation. The test formats discussed here are standardised multiple-choice tests, with fixed questions and fixed answers, and performance assessment, where knowledge and skills are demonstrated through more practical tasks. The term performance assessment has been used for a wide variety of test types, but authenticity in terms of being “true” or “real” is often a characteristic attributed to the process, conditions or context (Palm, 2008).

Furthermore, tests can have different stakes attached to them. Low-stakes tests are tests where the results usually have small consequences for test-takers or other stakeholders. Homework assignments or internet quizzes about your Star Wars knowledge could be characterized as low-stakes tests. High-stakes tests are tests where test results, and their interpretation and use, have serious consequences for the test-taker and/or other stakeholders. Admission tests for higher education or selection tests for air pilots could be viewed as high-stakes tests, and it follows that these tests need to be of high quality, providing reliable and valid results.

The test studied in this licentiate thesis is a criterion-referenced test. When such test are used to differentiate between those who have mastered an objective and those who have not they may be called mastery tests. Tests used to grant certifications or licences are sometimes referred to as certification tests or licensing tests, respectively. There are several definitions of these terms (Hambleton et al., 1978). This particular test results in a pass/fail distinction and those who pass (i.e. reach the minimum required level) are granted a licence, so in this licentiate thesis all those terms are used.

Another two terms I have used interchangeably are classification consistency and decision consistency. These are terms used for describing to
what degree a test-taker would be classified into the same category over multiple forms of a test or in repeated administrations of the same test. When the classification is a pass/fail decision *classification* and *decision* become synonymous.

What is described as the *driving licence test* in this licentiate thesis is the total test requirement for obtaining a licence category B (motor vehicles with a maximum authorized mass of 3,500 kilos). In Sweden this test consists of two parts — a *theory test* and a practical *driving test*. Since the theory test is a computerised multiple-choice test and the driving test a type of performance assessment judged by examiners, they are each accompanied by a particular set of issues when it comes to test quality.

**Aims and research questions**

The overall aim of this licentiate thesis is to examine the quality of the Swedish driving licence test, primarily in terms of reliability. The two empirical studies are based on the following research questions:

1) To what extent do the driving examiners agree in their assessment of the test-takers in the driving test?

2) How reliable is the classification of test-takers, in terms of pass/fail, based on the theory test?

In order to be able to assess the quality of a high-stakes criterion-referenced test issues concerning reliability and validity must be examined and discussed. What evidence is there that the decision based on the Swedish driving licence test is reliable and valid? The scope of this licentiate thesis does not allow for a complete study of all aspects of test quality so the enclosed studies are focused primarily on issues of reliability. Study I focuses on the driving test — a performance assessment — and examines to what extent driving examiners agree in their assessment of the test-taker (inter-rater reliability). Study II is aimed at the theory test — a standardised multiple-choice test — and its reliability in terms of pass/fail classification (decision consistency and decision accuracy).

Reliability in itself is not enough to establish the quality of a test — it is perfectly possible to measure irrelevant factors with great accuracy. On the other hand a reliable test result is a prerequisite for valid conclusions. In this licentiate thesis reliability is viewed as an integral part of the validity concept. Therefore, in this part of the thesis the answers to the research questions above are therefore placed in a validity context and a framework for further studies of the validity of the driving licence test is suggested, based on the argument-based approach to validation (described further in the section about Validity) to see if there is reason to believe that the results of this particular test — the Swedish driving licence test — can be used to award licences.
Structure

First the role of the Swedish driving licence test in the system is presented. The background to the test is then described in some detail, since the type of test, the definition of the goals, the process for development and administration are all factors that can affect the reliability and validity of scores. The concepts of reliability and validity are more fully described after a short presentation of previous research about the quality of the Swedish driving licence test. After the data collection and methods used in the two empirical studies in this licentiate thesis are presented, the results of these studies are then summarized and finally discussed within the framework of a model for validation and further studies are suggested.
The Swedish driving licence system

The role of assessment

The term assessment can refer to the process of gathering information about a certain construct. Construct is a term that can refer to a concept or characteristic intended to be assessed. Sometimes this is done with the help of test information. A test is a “device or procedure in which a sample of an examinee’s behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process” (AERA, APA & NCME, 2014, p. 2). Tests can be used in many contexts and for many purposes. A common use is within an educational system. One way of describing the components in an educational system is in the form of a triangle as in figure 1.

![Diagram](image)

Figure 1 Constructive alignment adapted from Biggs (1999)

The teaching (teaching activities) is guided by the curricular goals (intended learning outcome), which should also be reflected in the test (assessment). What is taught can affect what is assessed and what is tested can affect what is taught (Biggs, 1999). The reason for having a test can vary. Tests can be used to rank students in an admissions process, assess progress over time, find out what should be the topic of remedial teaching, and evaluate teaching strategies, among other things. The test discussed in this licentiate thesis is an example of a summative test used to assess to what degree test-takers have reached the stipulated goals.

Although there is a curriculum (TSFS 2011:20), which stipulates what driver education should contain and what students should be able to do when
goals are met, it is not compulsory to attend a driving school in the Swedish system. This means that the driving licence test is an essential tool for making sure that prospective drivers have achieved the intended learning outcome. There are only a few compulsory courses. There is a mandatory hazard perception course in two parts (TSFS 2012:40). Those who want to practice driving with a lay instructor also have to attend an introductory course (as does the instructor in question) (TSFS 2010:127, and others). This means in this case the test is not necessarily a good indicator of how well students have understood the teaching within an educational system – it is quite possible to come to the test with very little preparation (unless you are part of the small group that takes the test within a particular course in upper secondary school). This also means that the test plays a vital role in making sure licensed drivers possess the relevant skills and knowledge, which places high demands on its validity and reliability. Historically, this has not always been the case.

**Historical background**

At the start of the 20th century there were no particular requirements a driver had to fulfil to obtain a licence in Sweden (Franke, Larsson, & Mårdsjö, 1995). As cars became more common and traffic situations more complex the requirements for obtaining a licence and the measures for ascertaining these were fulfilled changed (see box below). More information about some of these changes can be found in Franke, Larsson & Mårdsjö (1995) and Alger & Eklöf (2016b).

Over time, not only the administration and format of the test has changed but also what is required. Driver training has moved from the initial focus on manoeuvring skills to include understanding of the traffic rules, hazard perception and insights into one’s own motivation and possible risk behaviours. The changes are, however, more a result of changing values and policies than driven by development of theoretical models for driver training.
Over the years it became common to attend driving schools, but before 1975 there was no compulsory education or training for car drivers in Sweden. However, that year a compulsory session at a skid pan facility was introduced. In 1999 a new curriculum for such training was introduced, in which the focus was moved from manoeuvring skills to risk awareness and insights. A graduated driver education system was suggested but rejected, but some of those ideas (Ekblad, Andersson, Gregersen, Jarneving, & Östbring-Carlsson, 1999), and perceived shortcomings in driving practice supervised by lay instructors (Gregersen & Nyberg, 2002; A. Nyberg & Gregersen, 2005) eventually led to a compulsory introductory course for students and lay instructors being introduced in 2006. The previous hazard perception training was expanded to two parts, one concerning risks with alcohol, drugs and tiredness, in 2009.

Some dates in the history of Swedish driving licence testing

- 1902 Certificate of competence
- 1907 Regulated that drivers must have a certificate
- 1916 Requirements to obtain a licence
- 1920 Both a theory test and a test drive
- 1927 The test drive carries the most weight in the test
- 1948 Tests more comparable across the country (eventually three parts – a written knowledge test of 25 items in 15 minutes, an oral examination more focused on understanding and a practical driving test of at least 10 minutes)
- 1958 Written theory tests generally used
- 1959 Driving test 15 minutes
- 1972 Oral theory test replaced by multiple-choice questions
- 1973 Theory test changed (60+20 items, scored in a non-compensatory way, with one correct answer)
- 1990 Theory test changed (40 items, several options can be correct)
- 1996 Driving test 25 minutes
- 1997 Driving test standardised
- 1999 Theory test computerised
- 1999 Theory test now 65 items + 5 try-out items, one correct option)
- 2000 Dual control cars required for the driving test
- 2008 No longer necessary to pass the theory test to be allowed to take the driving test. Both parts have to be passed within a 2-month period to result in a licence
- 2017 Theory test design changed to linear-on-the-fly testing where the test is assembled by a computer for each test taker in accordance with given parameters
One reason for introducing compulsory courses is to make sure that curricular content that is difficult to assess in the driving licence test is still covered. Although other training is not compulsory a majority of prospective Swedish car drivers still take lessons at a driving school (Alger, 2018). Currently more changes of the Swedish driving licence process are under consideration (Trafikverket & Transportstyrelsen, 2019). A summary of studies concerning the current Swedish driving licence system can be found in Alger (2018), but focus in this licentiate thesis is on the driving licence test, not the system as a whole.

**Goals in the curriculum and their theoretical foundation**

If the test results are to be interpreted as a demonstration that the test-taker has the necessary skills and knowledge to safely drive a car unsupervised in traffic there has to be a definition of what those are. There are some theories regarding driver behaviour, such as the GDE-matrix (see below), and driving skills, but not many that have been empirically tested (for examples see e.g. Backman, 2001; Bredow & Sturzbecher, 2016). Within the European union some specifications regarding the knowledge, skills and behaviour required to pass the driving licence tests can be found in the EU Driving Licence Directive (2006/126/EC), and within CIECA (Commission Internationale des Examens de Conduite Automobile, i.e. the International Commission for Driver Testing) some recommendations regarding minimum competence standards have been developed (CIECA, 2015). However, this does not automatically translate into a consensus as to the specific content, form and assessment of the driving licence tests.

Part of the issue is that there is no generally agreed upon definition of the construct “driver competence”, why the theoretical underpinnings leave much to be desired from a scientific point of view.

"Driver behaviour”, on the other hand, has been studied by many and is often described in terms of interaction between skills and motivational factors (MacDonald & Department of Transport, 1987).

The development of frameworks and curricular goals for driver education and training (Keskinen & Hernetkoski, 2011) has not been theory driven, but rather based on the ideas that (a) practice makes perfect, (b) solo driving is safer when novice drivers are older and (c) professionals are good at teaching complex skills.

Some countries have now used the The Goals for Driver Education framework (GDE-framework) as a starting point for adjusting driving courses and testing procedures (Grattenthaler & Krüger, 2011; Roelofs, van Onna, & Vissers, 2010). The Goals for Driver Education framework (GDE-framework) was presented in the GADGET project, a EU funded research project (Hatakka, Keskinen, Gregersen, & Glad, 1999), and published in 2002.
The researchers behind the GDE-framework found the above principles insufficient for fostering safe drivers. A safe driver is not only skilful at handling the vehicle but also motivated to drive in a safe way. The framework is often presented as four hierarchical levels, but has also been described in more detail in a matrix, see Table 1 (Peräaho, Keskinen, & Hatakka, 2003).

The GDE-matrix represents an effort to include not only the knowledge base for driving behaviour, but also motivational factors and their interaction. When the curriculum for driver education in Sweden underwent a major revision in 2006 the ideas in the GDE-matrix was one influence (Stenlund, Henriksson, & Sundström, 2006).

Table 1 GDE-model (Goals for Driver Education) (Peräaho, Keskinen, & Hatakka, 2003)

<table>
<thead>
<tr>
<th>Hierarchical level of behaviour (extent of generalisation)</th>
<th>Central content of driver education</th>
<th>Self-evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goals for life and skills for living (global)</td>
<td>Knowledge about/control over how general life goals and values, behavioural style, group norms etc. affect driving.</td>
<td>Knowledge about/control over risks connected with life goals and values, behavioural style, social pressure, substance abuse etc.</td>
</tr>
<tr>
<td>Goals and context of driving (specific trip)</td>
<td>Knowledge and skills re. trip-related considerations (effect of goals, environment choice, effects of social pressure, evaluation of necessity, etc.).</td>
<td>Knowledge and skills re. risks connected with trip goals, driving state, social pressure, purpose of driving etc.).</td>
</tr>
<tr>
<td>Mastery of traffic situations (specific situation)</td>
<td>General knowledge and skills re. rules, speed adjustment, safety margins, signalling etc.</td>
<td>Knowledge and skills re inappropriate speed, narrow safety margins, neglect of rules, difficult driving conditions, vulnerable road-users, etc.</td>
</tr>
<tr>
<td>Vehicle manoeuvring (specific task)</td>
<td>Basic knowledge and skills re. car control, vehicle properties, friction, etc.</td>
<td>Knowledge and skills re. risks connected with car control, vehicle properties, friction etc.</td>
</tr>
</tbody>
</table>
In 2010 a fifth level was added to include social aspects of driving.

The current curriculum for driving education for driving B-class vehicles in Sweden (TSFS 2011:20) is divided into four sections:

1) Manoeuvring, vehicle and environment
2) Driving in different traffic situations
3) Travelling by car in special contexts
4) Personal qualities and goals in life

Each section is then divided into two areas:

- Theory and skill
- Self-assessment

**Operationalization – from goals to test format**

Once the need for an assessment has been established and the construct to be measured has been defined one has to decide in what way this construct should be measured and create a suitable instrument (Wilson, 2005). The test design process involves many choices regarding content (structure and representativity), format, phrasing, scoring, standards and try-out.

A careful and systematic approach to item development is critical for test quality, and documentation of this can be used to support the idea that test results are valid. Downing and Haladyna (1997) details types of evidence concerning content specification, test specifications and various parts of the item development process. The construct and purpose can be defined in specific terms not only when it comes to content, but also different cognitive levels, assuming that it is a continuous construct that test-takers can have to an extremely low or extremely high degree or at any point in between. What behaviour or responses are typical for each of these levels and what differentiates test-takers at different levels? What level are the test-takers expected to be at?

As is evident from the section about the historical background driving licence testing in Sweden has consisted of a theory part and a practical part since the 20’s. The driving test can be regarded as a type of performance assessment and the theory test is a standardised multiple-choice test. (Standardisation here refers to unified administration conditions, not a norm-based reporting scale.) The quality of the test is not dependent on the test format – the same format can be used to produce both good and bad tests.

The theory test specification is the result of expert opinion, the item developers receive some in-house training in item-writing principles and there is an established process for item development and review. After an item is constructed by one of the item developers, it is fact-checked, the language is revised, it is formatted to be included in a test and finally tested as one of the five try-out items in the test. At each phase, a test developer collects and
evaluates input received and makes the necessary revisions. Once the item has been tested, it is evaluated in terms of psychometric properties and either rejected or deemed ready to be used. The development of an individual item can be tracked in the item database, but there are no other technical reports about test development.

Although a lot of effort has been made to develop and improve the test, documentation of the rationale behind the test specification seems to be lacking, or at least not easily found, and there is no explicit cognitive classification of items. Item selection is based on content regulations and psychometric qualities based on classical test theory measures. Five items are included in the regular test to be tried out, which means they are not awarded points. As test-takers are unaware of which five items are, it is assumed that they will do their best and provide an insight into how test-takers would respond if the items were to be approved and included in a later test. Information about content and difficulty (from the try-out) is used for item selection (M. Stenberg, personal communication, May 11, 2018).

Difficulty finding information about the rationale behind choices made is probably partly because test specifications have been refined over time. Test administration specifications have also changed. There are instructions about how the tests should be carried out and regulations as to their content (TSFS 2012:43). Test regulations specify what the questions in the theory test concerning the four sections in the curriculum listed above should cover. As for what tasks should be included in the driving tests only the first two sections are mentioned. Results on either of these tests should not depend on random circumstances such as who designed or administered the test, where it was carried out, and who assessed it, hence the need for reliability studies.

The process of test development should start with the specification of the purpose of the test. As I mentioned earlier tests can be norm-referenced or criterion-referenced, which has consequences for what the test ideally should do. Decision making in norm-referenced tests are based on the individual score, but in certification tests the passing score is the decision point (Berk, 2000). That the passing score invariably is set to 52 points for the theory test places great responsibility on the test designers to assemble parallel tests where this is a reasonable standard.

Since there are two parts to the driving licence test there has to be a rule as to how results should be interpreted together. As these tests are regarded as minimum-level test, test-takers have to pass both to get a licence.

The test and test administration
Test administration is standardised and carried out by trained examiners, but, given the nature of the tests, the driving test varies more between test occasions then the theory test.
The items in the theory test concern five areas of competence:

1) Vehicle knowledge/manoeuvring
2) Environment
3) Traffic safety
4) Traffic rules and
5) Personal qualities.

The computerised theory test consists of 65 multiple-choice items and five try-out items and each item has between two and six response options. The time allotted is 50 minutes. The test is available in Swedish and 14 other languages. The test versions in Study II were assembled by the test developers, but in 2017 the process was changed to a linear-on-the fly process. This means that the test is assembled for each test-taker with the help of a computer algorithm. The process for item development, however, remains the same.

Each correct response gives one point. The tests are scored automatically so there are no rater discrepancies. In order to pass the test the test-taker must obtain at least 52 points. The result is given on the screen at the end of the test and also sent via e-mail.

In the practical driving test tasks concern:

5) Vehicle knowledge/manoeuvring
6) Environment/eco-driving
7) Traffic rules and
8) Traffic safety/behaviour.

The test consists of a safety check and at least 25 minutes of driving in traffic. The examiner will direct the test-taker or give instructions to drive to a specific destination. The test will also include some manoeuvring tasks like reversing, parking or starting from an incline. There are no set routes. The examiner selects the content of the test and can choose to test the same ability again if the outcome is ambiguous. The test report form, which these days is electronic, lists possible content in terms of traffic situations like crossing, roundabout, turning left, slippery surface and so on. On the form the examiner marks what situations have been tested and, if there are tasks that the test-taker fails, which curricular goals he or she has failed to meet in that particular situation.

The examiner makes a holistic assessment with special attention to risk assessment in terms of attention, manoeuvring, placement, speed adjustment and behaviour in traffic. Once the test is finished the examiner presents the result to the test-taker, who later also receives a copy of the driving-test report form and a letter describing any curricular goals not reached.

Efforts to standardise the administration of the practical driving test have resulted in quite detailed descriptions of how the test should be carried out, what information should be presented to the test-taker and what elements of
the situation are assumed in order to test a specific task. For example, in order to test “crossing” it is not enough to just drive through one crossing, but several of varying types (in terms of size, visibility etc.) and there must also be other road-users to interact with. There are also recommendations as to how the test time should be divided between built-up areas and main roads. The test content should be varied and of a normal difficulty and, if possible, all content on the driving-test report form should be covered over the course of a day. Many of the directions are aimed at making sure the test-takers will feel welcome and be given the best possible chance to show their abilities. The document with instructions lists the regulations it builds on, as well as curricular goals and competence areas as they are the grounds for assessment.

**Previous research about the quality of the Swedish driving licence test**

Aspects relevant for the quality of the test, but not necessarily explicitly addressing reliability and validity, have been examined in the Swedish driving licence system over the years, many in projects conducted at Umeå university. However, very few studies have been made under the current curriculum.

Here follows a brief description of what types of studies have been carried out and summarized results from studies concerning reliability. Among the earlier studies there were many that focused on improving the test and the test administration, as a need for improvement in these areas was identified. For example, theoretical models for describing the test content and suggestions for additions to the test construction process were presented (Henriksson, Wikström, & Zolland, 1995; Wiberg & Henriksson, 2000; Zolland, 1999).

New ideas, such as delegating part of the testing to driving schools, adding an instrument for self-assessment or not requiring that test-takers have to pass the theory test before the driving test, have been examined and tested in pilot studies and changes in the curriculum have been followed up and reported. An English summary of some of the studies can be found in Henriksson, Sundström och Wiberg (2004).

There are studies where the theory test versions have been examined in terms of statistical parallelity (Sundström, 2003; Wolming, 2000; Zolland & Henriksson, 1998), with mixed results. Wolming (2000) found that test versions were parallel, whereas Sundström (2003) found that they were not (and attributed Wolming’s results to the fact that only test-takers who had passed had been included). To what extent test versions are parallel was therefore also examined in Study II.

The theory test and driving test are very different tests purporting to measure aspects of driver competence. That test-takers who do well on the theory test are more likely to pass the driving test has been shown even when only those who passed the theory test could take the driving test (Sundström,
2003; Wolming, 2000; Wolming & Wiberg, 2004). This has also been examined when those who failed could take it too (Alger, Henriksson, & Wänglund, 2010). In analyses of test results it is important that both tests provide reliable results.

Before the curriculum and test underwent major changes in 2006 the content and format for the tests were discussed and different driving licence systems were compared (Henriksson, Sundström, & Wiberg, 2002; Henriksson et al., 2004; Jonsson, Sundström, & Henriksson, 2003). One of the conclusions was that in countries without compulsory education more effort had been put into the development of the theory test than in countries with compulsory education. When it comes to the practical driving test, it was more regulated in countries with compulsory education (Jonsson et al., 2003). The idea of and methods for assessing self-assessment were also examined and tested (Sundström, 2004, 2009).

Part of the effort to develop the Swedish curriculum and tests is to follow up on changes made and the introduction of the 2006 curriculum was accompanied by training programmes and research studies (e.g. Berg & Thulin, 2009; J. Nyberg & Henriksson, 2009). The alignment between the test and the goals in the curriculum was examined for the old and the new curriculum in 2006 (Stenlund, 2006, 2007; Stenlund, Henriksson, & Sundström, 2006; Stenlund et al., 2006; Wiberg, 2007).

Most studies of the driving licence test after 2006 have primarily focused on the driving test. Results from the tests have been analysed in terms of passing rates, content, trends, correlation between test parts and differences between categories of test-takers (Alger & Eklöf, 2012, 2013, 2016a, 2016b, 2017; Alger & Sundström, 2011a, 2011b; Forward, Nyberg, & Henriksson, 2016).

When it comes to reliability in terms of internal consistency it is often described with the help of Cronbach’s alpha. It can have a value between 0 and 1, and the closer to 1 it is the better. As this is a coefficient developed for norm-referenced tests it is not always a good indicator for criterion referenced tests, but, with that in mind, it can still provide some information (Popham & Husek, 1969; Wiberg, 1999; Zolland & Henriksson, 1998). Back when the theory test had 40 items Cronbach’s alpha has varied between .82 and .85 for different test versions (Zolland & Henriksson, 1998) and in studies after it was lengthened to 65 items it has been between .78 and .84 (Sundström, 2003; Sundström & Wiberg, 2005; Wolming, 2000).

When it comes to the specific issues examined in this licentiate thesis – inter-rater reliability and classification consistency – there have not been any recent studies. The power of the test to classify applicants correctly was examined in 1976 after the test forms had been revised (Spolander, 1977). The tests then consisted of two parts – each with a cut-score (minimum required
points). The likelihood of erroneously passing a test-taker with a true score below one point below the cut-score for both tests was 6 per cent, and 1 percent if the true score was two points below the cut-score. The total number of misclassified test-takers was estimated to around 24 per cent, mainly for test-takers being failed despite sufficient knowledge.

The theory test has not been studied in terms of classification accuracy and classification consistency after the curriculum was changed in 2006. In previous studies decision consistency for six test versions was examined and found to be between .80 and .81 for the three test versions from 2004 and between .84 and .85 for the three test versions from 2003 (Sundström, 2003; Wiberg, 2004). It was noted that decision consistency had decreased but could still be regarded as high (Wiberg, 2004). In 2005 classification accuracy varied between .77 and .84 for three test versions (Sundström & Wiberg, 2005). The lack of studies after the revision of the curriculum indicated a need for a follow-up and the choice of method in Study II meant that comparisons could be made with these earlier studies.

As for inter-rater reliability and application of the scoring guide for the driving test this had previously only been studied at the level of local offices, not at national level. That the quality assessment work differed between offices had been pointed out by Riksrevisionen (2007). A study such as Study I in this licentiate thesis was therefore a necessary next step.
Quality in measurement

The quality of a measurement is often discussed in terms of reliability and validity. These concepts will be explained further below, but in broad terms this reliability concerns the accuracy and consistency of test scores and validity to what degree one can justify the interpretation and use of the results.

The inferences from a test result are far more interesting than the result in itself. Knowing that a test-taker scored 52 points on a particular test version at a particular time under certain conditions is easily proved, but if that was the sole purpose of the test there would be no point in it. The test was constructed to measure certain skills, knowledge or traits. Test versions were constructed to be similar, at least in certain aspects, and test administration regulated to avoid any irrelevant differences to play too great a part for the performance of the test-taker. The reason for this is that we want to be able to say, for example, that this result means that a test-taker who passes has the theoretical knowledge required to safely drive a car unsupervised, as specified in the curriculum. This is not the only requirement for a driver’s licence in Sweden (and many other countries) but, nevertheless, there has to be evidence that this is a plausible interpretation. In order to actually obtain a licence in Sweden the test-taker also has to successfully take part in some compulsory education and pass a driving test, where similar inferences are made under other test conditions. What evidence is there that this process gives us reason to think that that decision is correct? There are numerous aspects of this issue, some of which will be discussed here.

Validity

When examining the quality of a test the most crucial aspect is validity, which can be said to be “the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use” (AERA, APA & NCME, 2014, p. 14).

The concept of validity has changed considerably over the years (for an overview see e.g. Brennan, 2006; Hathcoat, 2013; Kane, 2013). The initial focus on coefficients has been replaced by a validity concept that encompasses a wide range of possible issues. The process for validation has also received more attention. Since Cureton discussed relevance and reliability as two aspects of validity in the first edition of Educational Measurement (1951) the concept of validity has also changed from an attribute of the test (“is this test measuring what it is supposed to?”) to the interpretation of scores (“how can this test result be interpreted and used?”).

One of the intentions of this licentiate thesis was to present my findings in a validity context. There are a several models for validation, each with their
advantages and drawbacks. Traditional categorisation into content validity, predictive validity, concurrent validity and construct validity still live on. Messick’s matrix and categories of validity evidence caused debate, not the least concerning consequences (Messick, 1989). Some meant that the concept of validity was getting too complex (e.g. Borsboom, Mellenbergh, & van Heerden, 2004) and many designations of different types of validity were used (Newton & Shaw, 2013). New types of studies also places new demands on validation evidence (Mislevy, 2016; Moss, 1994). Even within the argument-based model there are variants as to how many and which steps should be considered (e.g. Crooks, Kane, & Cohen, 1996; Haertel & Lorie, 2004). I will mainly focus on the argument-based model as presented by Kane (Kane, 2006, 2013, 2016; Kane, Crooks, & Cohen, 1999). One reason for choosing Kane’s model for validation is that it clarifies what assumptions and claims are used and illustrates that validation is a process. I am also attracted by the use of logic to question and evaluate whether there is evidence for the assumptions made. At the same time, these are very challenging tasks, particularly when the claims are not clearly stated before the test is designed.

The argument-based approach to validation builds on the idea that all the claims that are to be based on test scores are specified, and then validity evidence for (or against) these claims is collected, and the plausibility of them is evaluated. Kane refers to the first part (claims) as interpretation/use argument and to the second part (evaluation) as validity argument. The process entails scrutinizing assumptions and then re-evaluating and amending the argument in a process that continues until the argument is considered sufficiently plausible or rejected (Brennan, 2006).

Say, for example, that a driving test only involved driving around a special track with no other traffic. How can a good score be interpreted? It is unlikely that this is a good indication of how well the test-taker can handle different traffic situations. It is possible that it is a good indication of manoeuvring skill, provided that certain assumptions were met concerning the tasks, the assessment criteria, the test situation, examiner and so on. If the tasks covered various relevant situations (like reversing around corners, shifting gears, driving at different speeds, turning left and right and so on) and these were assessed in a systematic, fair and relevant way by trained examiners in a car free of distracting insects there would be more reasons to think that results could be interpreted as a measure of manoeuvring skill than if it was a question of driving round an oval track at low speed for five minutes while being assessed by random strangers according to their personal preferences and with the car full of bees.

In the latter case alternative interpretations may have more to do with erroneous scoring and chance than competence. It might be that the test-taker
has the required skills, but without an opportunity to demonstrate this the test results might not reflect that.

In the initial scenario the main interpretation/use argument might be that the test is fair and suitable for assessing basic manoeuvring for grownups without certain medical conditions before moving on to the next phase in their training which builds on these skills. The next step is then to find evidence for the assumptions made. Validity evidence can come in many forms and relate to the content and format of the test, administration, scoring, and consequences of its use. If all tasks in the test focused on reversing or some examiners though the acceptable limit for stalling was once and some forty times that is evidence against the plausibility of the claim. If the result on the manoeuvring test was linked to how well they could process instruction during the next level of training such facts can be seen as support for the claim. All the evidence is then weighed and perhaps found to be enough to come to a conclusion about validity, or perhaps more evidence is necessary. If relevant conditions or assumptions change, new information should be collected and evaluated.

Kane (2006, 2013) describes the validity argument in four steps - scoring, generalization, extrapolation and decision. Since the results from the studies in this licentiate thesis will be presented in relation to these four steps I will include a short explanation of how the interpretation of a test-taker’s performance can be described as four steps (or links in a chain, or bridges if you prefer that analogy). Scoring is the first inference – the step from performance to an observed score. Assumptions made here include that the scoring guide is appropriate and has been used correctly under suitable conditions. The second inference is generalization from observed scores to universe score. Universe scores refers to the expected performance on similar tasks (across samples of tasks, occasions, examiners, etc.). Consistency can be improved by including more tasks or standardizing aspects of the test administration and tasks (Kane et al., 1999). If tasks are not representative of the domain the scores may be biased.

Moving beyond the universe score to the target score (extrapolation) requires support for the idea that the skills tested overlap those needed for the target domain or at least include the essential ones. If it cannot be proven that this is the case perhaps it is necessary to examine the importance of identified differences. Kane, Crooks & Cohen (1999) argues that focus should be on the weakest link in the argument, which for performance assessment (such as the driving test) is generalization and for standardised tests (such as the theory test) is extrapolation. One reason for this is that practical tasks usually takes longer so it is more difficult to have a sample of tasks large enough to cover enough of the content domain within the time allotted for the test. As for standardised tests, knowledge of the content domain tested is usually
demonstrated in ways very different from the test situation. In the context of the driving licence test this means finding evidence that the situations in an individual test are typical for that type of test and for the test situation. Scrutinizing the rules for decisions and their consequences is important for all assessments.

In this licentiate thesis no proper validation studies have been carried out, but since reliability studies can be used as evidence in this process it seems relevant to put the results from the two empirical studies into this context. Another reason for discussing validity to this extent is that I think it is important to consider what interpretations one intends to make from the test score, and whether there is enough justification for such interpretations, when examining evidence for validity.

**Reliability**

Reliability is necessary for validity, since, as Kane puts it, “almost all test-score interpretations involve generalizations over some conditions of observation (e.g., over tasks, occasions, raters, and/or contexts) and our estimates of precision characterize the dependability of such generalizations” (Kane, 2013, p. 3). If the test result is unreliable, it is impossible to draw valid conclusion about the construct one attempts to measure. Reliability can therefore be viewed as essential to the validation effort.

Reliability is, like validity, not a characteristic of tests or test versions, but of the test scores (Brennan, 2001). It concerns precision and replicability. The term has, over time, not only been used to refer to correlations between equivalent forms of the test, but also to other forms of replication of the test procedure. Efforts to estimate reliability have been presented in the form of reliability coefficients, but also in terms of standard errors, generalizability coefficients, item response theory information functions and agreement indices (AERA, APA & NCME, 2014). What is relevant depends on context.

According to classical test theory a test score consists of a “true score” and random error. The trustworthiness of any score would depend on the size of that error. Sources of error can be connected to human performance, environment, the process of evaluating or the selection of questions (Kane, 2011). Errors can be categorized into two types – systematic errors and random error. Systematic errors consistently bias test results due to something that is irrelevant to the construct measured (e.g. a faulty scoring guide, or items that require knowledge of specialist terms not relevant to the domain in question) whereas random error refers to unpredictable fluctuations due to chance (Crocker & Algina, 1986). Reliability is primarily concerned with random error – the more random error, the lower the reliability. Systematic errors can introduce construct irrelevant variance and therefore affect validity,
Test score interpretations can be described as norm-referenced (or relative), where test-takers are compared to each other or a reference population, or criterion-referenced (or absolute), where the score is related to a specified performance standard. Errors that affect all test-takers equally will not impact the error of a relative interpretation, but may contribute to the absolute error.

Traditional reliability coefficients were developed with the norm-referenced type of interpretation in mind (AERA, APA & NCME, 2014). In a norm-referenced test, where the main aim is to rank test-takers, the reliability increases when the observed score variance is large. For a criterion-referenced tests large variance is not as essential to its reliability.

For tests with many test-takers the scores are often normally distributed. If it is a mastery test where the outcome is pass or fail this is impossible, which is why a binomial or a beta-binomial distribution is sometimes assumed. Methods developed for examining reliability in norm-referenced tests are therefore not always suitable for criterion-referenced tests (Hambleton et al., 1978; Popham & Husek, 1969). The methods used in this licentiate thesis have been chosen with this in mind.

There are a number of different ways of estimating reliability. Having calculated a coefficient or index for the reliability it still has to be interpreted. Whether this number indicates a serious problem depends on the situation. To what degree error can be tolerated when it comes to test results depends on what the scores are used for and what potential problems might occur. Perhaps the errors vary between test-takers. If the errors are such that they will not affect the outcome this is probably not a serious problem, but if they are systematic, rather than random, and bias the score unfairly it is more of an issue.

As was stated in the introduction about high-stakes tests – the more importance attached to the outcome, the lower the tolerance for error. If the interpretation of the scores is very narrow (e.g. performance on these particular items in this particular context rather than the examined ability in general) some of the potential sources of error have less of an impact (Kane, 2013).

When assessment is carried out in order to come to a decision in terms of classification (e.g. mastery tests), certain errors are more critical than others. Errors that do not affect the outcome (in terms of pass/fail or whatever the decision is) can be tolerated to a greater degree. Errors for test-takers near the cut-score are more likely to result in misclassification. The consistency or accuracy of the classification/decision can be expressed with indices, examples of which are given in Study II.

When discussing reliability in the context of performance assessment relevant issues are standardisation and rater agreement. Standardisation can
be seen as an aspect of replicability, a way to diminish random error, but is also closely linked to issues of fairness and validity. If students are given tasks that vary too much it is not a fair comparison. A certain standardisation is necessary to obtain comparable and non-biased results but not to the extent that the task at hand no longer reflects the reality it is supposed to mimic (Kane, Crooks, & Cohen, 1999). Just because a person can manoeuvre a car on a special track, it does not mean that they can handle the car well in a more complicated traffic situation. Any tasks given should be assessed in the same way by examiners, which places certain demands on the quality of criteria and training. To what degree raters agree about how to assess a certain performance is examined in Study I.
Data and methods for analysis

In order to gather information about the reliability of the Swedish driving licence test two empirical studies were carried out. The Swedish driving test consists of two very different parts and the methods chosen to analyse to what extent the pass/fail decision can be trusted are therefore also different. As the outcome of the practical driving test is based on the examiner’s judgement, the critical issue is whether this is reliable. The judgement should be independent of which examiner carries out the test. In Study I the inter-rater reliability was examined by having two persons assessing each driving test. The theory test, on the other hand, is a computerised multiple-choice test, where the test administrator does not have such a pivotal role, which is why the focus in Study II was placed on the statistical analysis of the test in terms of decision consistency and decision accuracy.

Data collection – Study I

Examiners in study I were selected to be representative of the test situation for a majority of the test-takers. Therefore only examiners who had carried out more than 700 tests in the previous two years were included. 92 fulfilled the requirements, but 7 were not currently active and another two were unable to participate. Data for Study I was collected from 535 driving tests conducted by a sample of 83 driver examiners who were accompanied by a supervising examiner over the course of a day (usually 7 tests).

There were five supervising examiners who had been vetted by the Swedish Road Administration and had the opportunity to discuss and partly formulate the criteria to be assessed.

During the tests, the ordinary examiners filled in the driving-test report form and the examiner filled in a special form based on the criteria for assessing the examiner used for quality control, and also the test-taker’s achievement. The forms and questionnaires were developed by Henriksson, Sundström and me at Umeå university, in collaboration with staff from the Swedish Road Administration, and tested during a pilot study in Umeå (Alger, Henriksson, & Sundström, 2008).

The tests were carried out over a three and a half month period in the towns/cities where the main offices were located. (The Swedish Transport Administration has just over 30 main offices, but also carry out driving tests in just over a hundred other designated places.) Some of the test-takers declined having an extra person in the car, but 93 per cent agreed to participate. Information from the test-takers, the ordinary examiners and the supervising examiners was also collected via paper questionnaires. 372 of the 535 test-takers filled in the questionnaire (70%).
Data collection – Study II

Data for Study II consisted of test data from three versions of the theory test administered over a seven-week period in 2012. The test versions were among those with the largest number of completed tests that year. To avoid multiple responses from the same test-takers only their first attempt at the theory test was included. A total of 12,072 test-takers were included in the sample. All the test items were multiple-choice questions where the correct answer was awarded one point.

Analysis – Study I

An analysis of inter-rater reliability was a suitable choice of method for the driving test. More specifically the driving test was assessed by two examiners simultaneously and the degree to which they agreed on the outcome was stated as a percentage. The possible correlations between disagreement and background variables were examined in order to see whether there were systematic differences (i.e. whether the disagreement was not merely a result of chance but due to specific characteristic of the test-taker or examiner). Analyses of differences based on variables from the questionnaires were done with the help of statistical tests suitable for the type of variable concerned (i.e. \( \chi^2 \) for nominal level, Kolmogorov-Smirnov two-sample test for ordinal variables and t-test for variables at interval level).

Analysis – Study II

If a test-taker re-took the test (assuming no new knowledge was acquired), what is the probability that they would obtain the same classification both times? This is often estimated as a proportion or percentage. Classification consistency can be based on one or several test administrations. It is recommended in the standards (AERA, APA & NCME, 2014) to use more than one, but often this is not possible or at least not done. The study of the theory test has been carried out from available data, which means that there is only information about one test administration, which limits the methods available. Methods for estimating classification consistency from a single administration can be based on distributional assumptions (e.g. Huynh or Hanson & Brennan) or individual results (e.g. Subkoviak or Lee) (Lee, 2010).

As this is a single-format theory test, where there is no complex scoring or weighting of subtests, decision consistency was examined with methods from Subkoviak (Peng & Subkoviak, 1980; Subkoviak, 1976, 1988) and Hanson and Brennan (Hanson & Brennan, 1990). Classification consistency of the latter type and classification accuracy was calculated with the help of the software BB-Class (Brennan, 2004). For the calculations according to Subkoviak a syntax file for SPSS was used, and the results were also compared with a
lookup table. (There are other options, such as the software Lertap 5 or R-package rcrtan.) The focus of the study was not on comparing methods, but on obtaining information about classification consistency and classification accuracy these methods were chosen as being suitable for this type of data. The test developers who produce the theory test are currently using classical test theory methods for analysis. Subkoviak’s method was used in a previous study by Wiberg & Sundström which means results can be compared.
Summary of the studies

Whether a prospective driver receives a licence mainly depends on the results of the two parts of the Swedish driving licence test. It is therefore important that the results are reliable.

Study I

Agreement of driving examiners’ assessments – Evaluating the reliability of the Swedish driving test

If results on the driving test are to reflect the ability of the test-taker it is important that tests are assessed in a reliable manner. The reliability of Swedish driving examiners’ assessments was therefore studied in terms of examiner agreement (inter-rater reliability). 83 driving examiners were accompanied by one of five supervising examiners for a day. All in all, 535 driving tests were included in the study. Both examiners in the car assessed the test-takers performance on a two-grade rating scale (pass/fail) as well as on a six-point scale. At the end of the day they compared notes and tried to determine the reason for any discrepancies. In order to determine or rule out whether any disagreement could be linked to specifics related to the test-taker or the examiner, information was collected via questionnaires. Test-takers, examiners and supervising examiners all had to fill in questionnaires, providing information about background variables, preparation for and attitudes to the test. Only three of the many variables had any connection to disagreement in the study and these could be expected. The variables in question concerned when the examiner made the decision, how difficult he/she found it and to what extent the supervising examiner considered it to be an overall assessment.

In 93 percent of cases both examiners agreed on whether the test-taker should pass or fail the driving test. For 14 tests the supervising examiner would have passed the test-takers although the ordinary examiner failed them, whereas the roles were reversed for 23 cases.

When they disagreed they sometimes attributed the cause to their position in the car (front seat or back seat) and occasionally different views on where to draw the line between admissible advice and a reprimand. There were also instances where the examiners disagreed about the severity of a perceived error. In a couple of cases previous tests with the test-taker were thought to play a part.

The assessment was also carried out according to a six-point scale and the agreement was then lower (63%). The scale was, however, new to the examiners and their level of preparation for using it varied.
The results do indeed seem to be a reflection of the test-taker's performance rather than other qualities in the test-taker or examiner. All in all, the inter-rater reliability is deemed satisfactory.

**Study II**

*Is This Reliable Enough? Examining Classification Consistency and Accuracy in a Criterion-Referenced Test*

This study focused on the reliability of the theory test. As it is a criterion-referenced licensing test classic reliability measures developed for norm-referenced tests may not always be appropriate. The important issue for the theory test is the reliability of the classification (i.e. the decision of pass or fail) more than reliability in terms of general stability.

Results from three test versions distributed over a seven-week period in 2012 were studied to estimate to what extent they can be considered to be consistent in terms of classification (pass/fail) and to what extent these classifications are accurate. Only tests distributed in Swedish were included. For test-takers with more than one attempt only the first was included. The sample then included 12,072 tests — around 4,000 for each test version. There were no statistically significant differences between test-takers taking the three versions when it comes to age, gender and method of registration for the test. The test versions had similar means and variance. Since the mean was close to the cut-score the percentage of students who passed the test differed somewhat. The p-value varied between items, but in a similar manner for all three versions, with the exception of one item that was considerably more difficult than the rest.

In order to examine classification consistency, and to some degree classification accuracy two methods were used – Subkoviak and Hanson and Brennan respectively (Hanson & Brennan, 1990; Subkoviak, 1976, 1988).

Both methods showed similar results – a consistency agreement coefficient of around .80. Results for Subkoviak’s calculations varied between .78 and .81 if different estimators of probability of a correct item responses were used (KR20, KR21 and maximum likelihood). Hanson-Brennan’s 4-parameter method resulted in .82 for all test versions. This can be interpreted to mean that if a student took the test again (with no memory of previous attempts), eight times out of ten the result would be the same. A simplified version of the Subkoviak method was also used with similar results. As for accuracy, around 6-7 per cent of tests were positively misclassified and an equal amount negatively misclassified in terms of pass/fail. The probability of accurate classification was 87 per cent.
Considering the importance and length of the test the classification consistency falls slightly short of the recommended level. In order to improve classification reliability lengthening the test should be considered and any measure to improve the quality of the items taken. Further studies of classification consistency should include reliability measures from item response theory and other aspects important for the validity of test scores should also be examined, such as the placement of the cut-score.
Discussion

The purpose of this licentiate thesis was to examine some aspects of the quality of the Swedish driving licence test, primarily in terms of reliability. As there are only a few studies about the reliability of this test, and none at all since the curriculum revision in 2006, the studies presented here are an important contribution to this field. As reliability here is viewed as an integral part of validity I also aim to place my reliability studies in a larger validity framework. With this in mind, I will first present and discuss the main findings from the empirical studies and then continue to clarify how this may serve as part of a larger validation effort with the help of the argument-based approach to validation.

Empirical studies – background and results

The Swedish driving licence test is a criterion-related test where the results are expressed in terms of pass or fail, which means that methods for analysis have to be suitable for such tests. The two test parts are of fundamentally different types — a performance assessment and a standardised multiple-choice test — and the types of data from them differ. One is a judgement from an examiner, who also designed the testing to a certain degree. The other is an automatically corrected score from a number of multiple-choice items. The studies included in this licentiate thesis focused on inter-rater reliability (Study I) and decision consistency (Study II).

Two questions shaped the studies in this licentiate thesis:

1) To what extent do the driving examiners agree in their assessment of the test-takers in the driving test?

2) How reliable is the classification of test-takers, in terms of pass/fail, based on the theory test?

The simple answer to the first question is 93 per cent. That is the degree to which the examiners in the study agreed on whether the test-taker they had observed should pass or fail the driving test. There are other coefficients that could be used to account for agreement by chance (see e.g. Gwet, 2001; Hallgren, 2012). Given that more than half of the test-takers failed a 93 per cent agreement rate is unlikely to be solely the result of chance. The examiners in the sample may not be entirely typical for the group of examiners as a whole, but the differences we know about do not appear to be influencing results. The outcome indicates that the test result on the driving test is not dependant on who the examiner is. However, evidence that examiners agree does not necessarily mean the individual examiners’ assessments are consistent or that the driving tests are equivalent in terms of content and difficulty, but it does imply that examiners use the same criteria when assessing the test-taker’s
performance and that the assessment is based on those criteria rather than irrelevant variables.

In answer to the second research question, about the classification consistency in the theory test, two types of agreement indexes were calculated and the estimates indicated that a test-taker would have the same result on a comparable version of the theory test around eight times out of ten, which is acceptable for a low-stakes test (Subkoviak, 1988). However, given the high stakes of the theory test it would be preferable if the agreement was higher.

Not only classification consistency, but also classification accuracy was examined and the probability for a correct classification was 87 per cent. The false positive rate and the false negative rate varied between six and seven per cent for all test versions. For certification tests where the aim is to protect the public false positives constitute more of a problem than false negatives (Millman, 1989).

The consequences of misclassification depend on whether the cut-score really separates the competent from the incompetent. Even for those tests where cut-scores are carefully deliberated, they are still somewhat arbitrary and the difference between 51 and 52 may not be a matter of life and death, but even when the underlying competence is a continuous variable there has to be a cut-score somewhere (McCoubrie, 2004). The total probability of a correct classification on a previous, shorter test was lower (76 per cent) (Spolander, 1977).

Validation
How do these reliability studies contribute to the chain of validity evidence? In the section about validity I mentioned that the argument-based approach to validation has two parts according to Kane (Brennan, 2006). First all the claims that are to be based on test scores are specified (interpretation/use argument), and then validity evidence for (or against) these claims is collected and their plausibility is evaluated (validity argument).

Therefore I will first present a table of possible claims to validate (interpretation/use argument) specifying the inferences and assumptions involved in the interpretation and discuss some alternative interpretations. Then I will go on to discuss the studies included in this licentiate thesis, as well as some previous studies, in terms of evidence for inferences regarding scoring, generalization, extrapolation and decision.

**Interpretation/use argument — What is the test supposed to do?**
As was mentioned above - the first step in the argument-based approach to validity is to specify the proposed interpretations and uses of the test result and then examine what evidence can be found in support of this (AERA, APA
The advantage of specifying claims is that it becomes clear what the expectations are, and it might turn out that some are problematic or that some are not shared by all stakeholders. Preferably, the proposed interpretations and uses of test scores should be something stakeholders can agree on. There might be several different interpretations or uses of a test result and validation should be done for each separately.

Here I will use fairly general interpretations based on Kane’s examples (Kane, 2006), see Table 2. It would be possible to have a number of more detailed and specific claims, but that exceeds the scope of this licentiate thesis.

Table 2 Possible claims, or interpretation/use argument, about the driving licence test scores

<table>
<thead>
<tr>
<th>A</th>
<th>Scoring – From observed performance to an observed score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The assessment guide is appropriate</td>
</tr>
<tr>
<td>2</td>
<td>The assessment guide is applied as specified</td>
</tr>
<tr>
<td>3</td>
<td>The assessment is free of bias</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>Generalization – From observed score to universe score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The observations are representative of the universe of observations defining the testing procedure</td>
</tr>
<tr>
<td>2</td>
<td>The sample of observations is large enough to control sampling error</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th>Extrapolation – From universe score to the level of skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The test tasks require the competences required to independently drive a vehicle in traffic in a safe way</td>
</tr>
<tr>
<td>2</td>
<td>There are no skill-irrelevant sources of variability that would seriously bias the interpretation of results as measures of level of skill in the competencies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D</th>
<th>Decision – From conclusion about level of skill to awarding a licence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Students who pass the test have the skills necessary to independently drive a car in a safe way in most situations</td>
</tr>
</tbody>
</table>

What is the purpose of the driving license test? As a mastery test its purpose is to differentiate between those who have reached a specified level of a construct or domain and those who have not. In order to do an in-depth validation it is necessary to specify whether it is claimed that the test adequately covers the domain as specified in the curricular goals or if there are other, perhaps more ambitious, claims.
In Table 2 I have specified an interpretation/use argument based on what having a licence allows one to do. There are, however, other ways to look at it. If the test is supposed to measure driver competence – is there evidence supporting that the curricular goals are a good specification of this construct? Is it there to make test-takers prepare and train in a certain way? Or is the purpose of the test to differentiate between safe and unsafe drivers?

Admittedly, I have not examined all documentation from the Swedish authorities involved in driver licensing, but only regulations and official investigations, but I have not found a clear statement of what the overall aim of the driving licence test is. There are, however many specifications of what types of knowledge and skills are required from a driver and also what the test should cover.

In the driving regulations (1998:980) different areas of knowledge and skills that should be included in the driving licence test are listed, but also that the test should concern the prospective driver’s judgement and presence of mind. It is difficult to measure such things in a theory test since answers might not be truthful. (If it is reflected in unsafe driving that is a different matter.) In the Swedish system it is hoped that the compulsory hazard perception training will make students aware of possible risks and change their behaviour accordingly. The most recent evaluation indicated that the training is more successful at the former than the latter (Forward, Warner, & Berg, 2010).

There are several complicating factors if the test score is to be interpreted as a means of differentiating between safe and unsafe drivers. An important aspect of safe driving is how the driver estimates and handles risk but accidents can be the result of a lack of skills or of violations (Keskinen & Hermetkoski, 2011). Even if the test differentiates between test-takers who have and who lack the necessary skills and knowledge that is not necessarily differentiating between safe and unsafe drivers.

It is possible to do your best on a test and drive in a way that does not reflect your later driving style. On the other hand it could be claimed that it is unreasonable to make claims regarding other than the minimum level. A test can only be required to measure a test-takers current skills, knowledge and abilities, not how he or she chooses to use those in the future and how they develop (Kane, 2013).

Different claims require different types of validation, but if the proposed interpretations and uses are not clearly stated it is difficult to know what evidence is necessary. It is probably easier to find support that the curricular aims have been met than that this can be interpreted as a sign of being sufficiently qualified to drive unsupervised. If there is not enough evidence to support a certain interpretation it might be necessary to amend it.

In a proper validation study the process of scrutinizing assumptions and then re-evaluating and amending the interpretation/use argument continues
until it is deemed plausible or is rejected. The next step is to set up a validity argument in which the chain of inferences when interpreting the test results are evaluated.

**Validity argument**

Even though they could be refined further, say we accept the claims in Table 2. where do the results from the studies presented here fit into the model for validation? When using the argument-based approach the inferences might be described as four links in a chain – scoring, generalization, extrapolation and decision. As can be seen from Table 2, issues include, for example, the appropriateness of the assessment guide and its application, the representativeness of the observations, whether the task-demands correspond to the target domain, the influence of skill-irrelevant sources and whether the test-takers who pass have the necessary skills. In order to provide evidence for (or against) the plausibility of the claims in Table 2 I will not only present findings from Study I and II, but also from some previous studies.

**Scoring**

Validity inferences from observations to scores are based on the quality of the test development process and the skills of the test developers and examiners. Ideally, validity evidence should be documented from the entire process — from overall plan to item banking (Downing & Haladyna, 2006) Test development for the theory test was described in an earlier section and instructions for examiners and the driving-test report form have been developed over time by experienced examiners (sometimes with some help from measurement experts). That there are regulations ensuring that examiners have certain qualifications and are trained (not only initially, but regularly) is a factor that should enhance the quality of the assessment. This supports, or at least does not falsify, the assumptions but is no guarantee for appropriate and correctly-applied scoring guides (Schuwirth & van der Vleuten, 2011).

As for correct application and possible bias (see Table 2) the main focus for Study I was whether both examiners made the same assessment of students, and in 93 per cent of tests they did. This result supports the reliability and to some extent also the validity of the scoring/assessment: as examiners to a high degree agreed on whether the student should be awarded a pass grade, this indicates that they not only have assessed “in the same way” (a reliability issue), but also “the same thing” (a validity issue).

When they did disagree there was no connection to any student variables, which can be interpreted as an indication that perceived student background variables do not affect the examiners’ assessments (i.e. that the assessment is free of bias, also a validity issue). Of course, in order to cover alternative
interpretations the possibility that both examiners are equally biased should also be considered.

Although there was no association between differing assessments between examiners in Study I and the background variables of students, there are differences in pass rates between different test-taker groups (Alger & Eklöf, 2017). However, there are no indications that these are the result of a biased assessment.

As the theory test is scored automatically, the scoring guide is consistently applied, but the other assumptions regarding the scoring aspect of the validation (appropriateness of the scoring guide and freedom from bias) still have to be supported (see Table 2). There are differences in results between test-taker groups, which could suggest bias, but more likely they could be interpreted to mean that there are actual differences between, for instance, those who register for the test via a driving school and those who do not (Alger & Eklöf, 2016a). A previous study of the theory test showed that some items seemed to be biased as they displayed differential item functioning (DIF, a measure of item bias) with respect to educational background. However, no DIF was found with respect to gender, even though females outperformed males (Wiberg, 2006).

**Generalization**

The reliability/precision of scores affects the generalizability and the consistency of classifications of test-takers based on the scores. If the scores are not consistent across replications of the testing procedure their usefulness as the basis for wise decisions is limited (AERA, APA & NCME, 2014). The three versions of the theory tests examined in Study II contained items that were reasonably similar in terms of statistical properties. The test specification gives different weights to different content areas by having different number of questions and gives boundaries for level of difficulty within each content area, so theory tests should be representative of the test domain in these respects. However, the most recent study of alignment between curriculum and the theory test, just after the curriculum change in 2006, indicated that alignment between the previous curriculum and theory test was better than between the new curriculum and theory test (Stenlund, Henriksson, & Sundström, 2007).

One way of improving representativity and reducing bias is to change the testing procedure to reduce sampling variability, i.e. include more items or observations in the test (Kane, 2013). Generalizability across items increases as the number of observations for each student increases (Kane et al., 1999). A test of 65 questions the theory test gives a lot of information, but a longer test would reduce sampling variability even further. Given that it is a high-stakes test in terms of possible consequences for traffic safety it would be
preferable if the decision consistency and decision accuracy for the test was higher than was the case in Study II. On the other hand, decision consistency is still on a similar level as in the study from 2004. Although Wiberg (2004) then considered this to be high, Subkoviak (1988) was of the opinion that high-stakes tests should have coefficients exceeding .85. Results in 2005 varied between versions, but none exceeded .85 (Sundström & Wiberg, 2005).

The assessment should be based on curricular goals. In the driving test the examiners not only need to know what the curricular goals are, but they also have to interpret what they see in terms of evidence for knowledge and skills and make sure they have enough information to make a holistic assessment. As there are no set routes this includes choosing tasks to determine whether minor lapses that are not deemed hazardous should be regarded as a one-off due to nerves or a habitual behaviour. Which shortcomings can be deemed of minor importance to traffic safety can be an issue of contention. Training of examiners and discussing assessment of potential situations where disagreement might occur can therefore be important for improving reliability.

Other studies of performance assessments have shown substantial variability over tasks (Kane et al., 1999). This could mean that results are not generalizable beyond what has been assessed. As examiners cannot cover all possible content in an individual driving test they have to make a selection. This introduces variance in the content of tests. Even when exactly the same tasks are given, they vary in difficulty depending on the situation (location, other road users, weather conditions etc.). If the selection of tasks vary in a way that gives unfair advantages to some students over others that is a problem. The supervising examiners in Study I did not assess to what extent the tests were similar but they did assess other aspects of the test administration such as the difficulty of different parts of the test, the extent of various tasks, information provided, and so forth. There was no significant connection between assessment of the content and the disagreement between examiners.

Not all the data collected in the study are presented in Study I, but around round 12 per cent of the tests examined contained segments that were too hard or too easy according to the supervising examiner. The design of the test was criticized in 11 per cent of the tests, but was praised for 21 per cent of the tests in the study (Alger, Henriksson, & Sundström, 2009). As tests are carried out in various places all over the country the environment will differ somewhat and therefore the test conditions. It is difficult to determine the relative importance of the examiner and the place/traffic situations for the test result (Alger & Eklöf, 2016a).
Extrapolation

In order to move to extrapolation from the assessed domain to the target domain there has to be some evidence that test items are sufficiently representative of the universe of generalization for such inferences to be made and that efforts have been made to avoid bias, for instance by changing the test procedure or by changing the definition of the universe of generalization by standardisation (Kane, 2013). However there is a limit as to how standardised a performance assessment test can be and still give the desired information in terms of extrapolation. Although standardisation can improve generalization it can be problematic in terms of extrapolation (Kane, 2013). For example, if the tasks in the driving test are limited to a set route, or even a special test track, it means fixing aspects of the test domain that are not fixed in the target domain, which makes the test less authentic. If only certain aspects are tested it is difficult to draw conclusions about the areas in the target domain that are not included.

Support for extrapolation inferences (i.e. the leap from test situation to level of skill) can either be a judgement that the skills measured and those required for driving unsupervised in traffic overlap to a great degree, or correlation with some external measure. There is little evidence to support that the curricular goals and the content of the test reflect all competences required to drive a car independently in a safe way, but the similarity between the test situation and the real-world situation lends some credibility to this claim. This is most apparent in the driving test, but the theory test also reflects possible real-life scenarios. However, selecting the best option in a multiple-choice test is very different from making the right choice in the traffic situation in question.

The disagreement between examiners has hardly any connection to other collected variables in Study I, but this does not necessarily mean that there are no irrelevant sources of variability that could seriously bias the interpretation of results in terms of level of skill. However, measures have been taken to avoid bias due to skill-irrelevant sources of variance, like providing information in several languages and offering extended time or oral examination to test-takers with dyslexia. The most difficult aspect when assessing skill-irrelevant sources of variability is to know which possible sources to collect information about.

One reason for including the description of the test, its history and administration in this licentiate thesis is to illustrate that many choices are made during the process of developing a test. So what do the choices made entail for the driving licence test – a standardised criterion-referenced test consisting of one part performance assessment and one part multiple-choice test? Including both a high-fidelity task, which, by necessity, only can include as small sample of performances, and a lower fidelity test with a larger number
of tasks is one way of handling the balance between strengthening the generalization inference (by standardisation) or the extrapolation inference (by having tasks similar to the target domain).

**Decision**

The two test parts are of different types and focus on different aspects of the curriculum, but both are aimed at measuring driver competence. Those who do well on the theory test are more likely to do well on the practical driving test (Alger & Eklöf, 2014; Henriksson et al., 2004; Wolming & Wiberg, 2004). When it comes to the driving licence test the rules state that a test-taker has to pass both test parts to obtain a licence.

How is a pass on both parts of the driving licence test used and interpreted? That examiners agree in 93 per cent of cases in Study I does not necessarily mean that this decision is the right one in all but 7 per cent of cases. They could both be wrong or the ordinary examiner, whose decision determines whether the test-taker passed, could have made the wrong decision. However, when more than one examiner comes to the same conclusion at the end of a well-designed and well-executed assessment process it lends more credibility to the assumption that it is a correct assessment.

Since the theory test is administered on an almost daily basis and the incentive to cheat is strong it is necessary to have multiple test versions. If decisions based on those versions are to be fair and reliable all versions have to be parallel and interchangeable. Even when that is the case, we cannot be sure that the decision is right for all test-takers and this uncertainty is often expressed in terms of misclassification and accuracy, as in Study II.

As always, the validity of the score is dependent on its use and the appropriateness of its use is dependent on to what extent there is support for the validity of all steps of the chain.

Given the results in Study II it is apparent that some test-takers pass the theory test despite not having the required knowledge and some are erroneously failed. As it is impossible to demand error free measurement the question is rather a matter of what level of tolerance should be accepted. It might also be a question of what is achievable given the time and resources available to the authorities responsible for the testing and other stakeholders. Given those constraints I think this result is acceptable.

However, the main issue when it comes to decision rules is defining the cut-score (Kane, 2006). Study II provides some measures of to what extent the pass/fail decision is consistent given the current cut-score. Where the boundary is between those who can and those who can’t is always a matter of judgement, but more evidence that the current cut-score (a fixed cut-score of
So per cent correct) is suitable in all test versions would strengthen the argument for validity.

In conclusion, the outcome of the study of the driving test (Study I) indicates that the examiner agreement is good enough to support the idea of a correctly applied scoring guide. Examination of background variables did not indicate bias in the assessment.

The outcome of the study of the theory test shows that the pass/fail decision clearly is not a result of chance, but that there are some inconsistent decisions. Both methods used show similar results. Although the results give reason for examining possible ways of increasing classification consistency they do not give cause for alarm.

Other aspects described here also give support for the scoring inference of the theory score, some aspects of generalization and extrapolation, but still we do not know if driving tests differ too much in terms of design and difficulty, and whether all theory tests measure the same competences. Bias is still a possibility for both test parts, but it is more difficult to definitely prove an absence of something than a presence. The design of and consequences of the test have only been touched upon briefly, but both test parts have their pros and cons. These studies cannot show if the test differentiates between those who have the necessary qualifications and not, but they can offer information as to the reliability of the decision, which is important.

Further studies
Designing a study means making many choices regarding sample, method and assumptions. If some of those choices had been different in the empirical studies presented here that would have enabled other conclusions to be drawn. The studies presented in this licentiate thesis provide some support for the consistency of decisions, but are not designed to support claims regarding the appropriateness of content, standards or effects.

Roads not yet taken – possible knowledge gaps
The sample in Study I was selected to be representative of the tests, rather than the examiners, so although the selected examiners carry out a large proportion of the tests they are not representing all the main offices and all types of backgrounds. It is therefore not possible to say whether there are any differences typical for particular offices. That said, there is no indication that examiner disagreement is connected to educational background. If it is considered important to draw conclusions about differences between offices or types of examiners a study with a different sample should be carried out.

Although the methods and measures used for Study II were chosen because they are suitable for the type of test they also have limitations. The
reliability coefficient KR-20 can vary substantially over replications and may not be the best coefficient to use in Subkoviak’s calculations (Algina & Noe, 1978), but in this case it is probably more fruitful to estimate classification consistency with other methods than to recalculate Subkoviak’s index with a different reliability coefficient than the three used in Study II. As is recommended in Study II using several different measures will help to determine classification accuracy and classification consistency. As the study only used two methods it would be interesting to make additional studies using other methods, based on item response theory or generalizability theory, to see if they would result in other conclusions. Given the current linear-on-the-fly testing system for assembling theory tests, item response theory methods of analysis seem more appropriate given that they focus on items, whereas classical test theory analysis is based on tests as a whole.

Indices for classification consistency depend on a well-placed cut-score. Although standard-setting is a judgemental process and the point for separating the competent from the incompetent on the underlying continuum of competence is arbitrary there are many methods that can make standard-setting more systematic (McCoubrie, 2004). When the theory test was changed in 1999 possible procedures for establishing a suitable cut-score were discussed. Iterative Angoff or Informed judgement were suggested methods (Wiberg & Henriksson, 2000). However, none of these procedures have been used for the theory test. If the cut-score does not reflect a suitable boundary between those who have and do not have the necessary qualifications then the consistency and accuracy of that decision loses meaning. As the cut-score is set at 52 points out of 65 for all theory test versions and no scores are scaled it becomes very important that all versions are reasonably parallel and that results reflect the intended performance standard. If the claims only referred to a range of scores on the test then reliability evidence would be enough, but, as stated in the introduction, in practice there is always a broader interpretation, referring to non-test situations. Since the cut-score is the operationalization of the performance standard more evidence for the appropriateness of this cut-score is required.

**A closer look at validity**

Hopefully the presentation of the argument-based model of validation can inspire others to use it, but the model is not implemented to its full potential here. One issue that should be discussed further is what claims are made. A more developed definition of the inferences made from the test score may lead to collecting other types of validity evidence. It may also promote a discussion regarding what skills, knowledge and attitudes should be required. In a proper validation study the warrants for making certain interpretations are scrutinized and possible other explanations examined. This has not been done...
to a great extent here. A validation study of the driving licence test would address a wider range of aspects, for example whether the theory test specification and the position of the cut-score are reasonable.

As validation is a never-ending process and new situations require new evaluations there is always a need for collecting evidence regarding validity. In 2017 the Swedish Transport Administration changed to a system for linear-on-the-fly testing where the test is assembled for each test taker in accordance with given parameters. Providing unique tests for each candidate will make cheating more difficult, and (as opposed to completely random item selection) the system is supposed to ensure adequate content coverage and test difficulty in order to make tests comparable. It would be interesting to examine a couple of the traditionally assembled test versions with the help of a panel of experts, from the field of driver training and driving assessment as well as educational measurement, and examine the new tests to compare them in terms of curricular goals (and subgoals) covered, cognitive level, suitability of cut-score, and so forth.

Naturally, evaluating the validity of test inferences does not end there. The entire chain of arguments for both test parts must be evaluated, preferably in several manners. Identifying assumptions made is a difficult task, but, if done well, is a good starting point for the validation process. What support is there and what is missing? Does the argument hold? Changes to the driving licence system are currently being discussed, but proposed changes to the test concern administrative procedures (such as whether you have to pass the theory test in order to take the driving test) rather than content and format as the former are questioned more than the latter. This could be interpreted as widespread acceptance of the test among consulted stakeholders or a lack of critical examination of the assumptions made.

Conclusions

Swedish driver testing has come a long way in the last century and many of the changes have been made to increase reliability and validity of test results, although not always expressed in those terms. As a basis for future studies of reliability of the Swedish driving test Study I in this licentiate thesis provides results concerning inter-rater reliability (93 per cent agreement). As a continuation of the trend for previous studies of classification consistency of the results from the Swedish theory test Study II shows values around .80 for Subkoviak’s coefficient.

When it comes to classification consistency in the context of driving licence tests there is not a lot of international research available, despite it being a high-stakes test. There are many studies of classification consistency and accuracy of tests in an educational (i.e. school) context (e.g. He & Opposs,
Good psychometrical qualities are equally important in the context of the driving licence test. The studies presented here form an important piece in the puzzle of validity evidence for the Swedish driving licence test.

A chain is only as strong as its weakest link. In the chain of evidence for the validity of the interpretation of test scores all links are important. It is important to make sure that the test is covering relevant content, the standards are set at the right level and the test versions are parallel. Reliability studies are an important starting point in this work since, even if you have a test that covers all the right content and seems very impressive but where the decisions based on test scores cannot be relied upon, the measurement does not fulfil its purpose. This is why reliability studies are important for the validity of any test, not the least when it comes to the driving licence test – a high-stakes test that can result in a licence to drive.
Sammanfattning på svenska

Inledning och syfte
Varje år genomförs oerhört många förarprov för B-behörighet (personbil) i Sverige. Resultatet från provet avgör om provtagaren ska få körkort och därmed tillåtas framföra fordon på vägarna. Visserligen måste dessa förare ha genomgått en riskutbildning i två delar och i många fall även en introduktionsutbildning, men den främsta garantin för att de har de kunskaper och färdigheter som krävs är att de genomfört ett godkänt förarprov bestående av ett teoretiskt kunskapsprov och ett praktiskt körprov. Med tanke på att det är väldigt lite obligatorisk utbildning är det av största vikt, såväl för individen som för samhället, att förarprovet håller tillräckligt hög kvalitet. Syftet med denna licentiatavhandling är att undersöka kvaliteten i förarprovet utifrån några aspekter.

Kvalitet i provsammaehang brukar ofta beskrivas i termer av reliabilitet och validitet. Ytterligare beskrivning av dessa begrepp följer nedan, men i mycket övergripande termer kan man säga att det handlar om med vilken precision man mäter och hur resultatet kan tolkas utifrån vad man avsåg att mäta. Studierna som ingår i denna licentiatavhandling gäller huvudsakligt reliabilitet, men då reliabilitet ses som en del av validitetsbegreppet är de båda begreppen sammanflätade. Då god reliabilitet är en förutsättning för god validitet är det logiskt att börja med att fråga sig om resultatet från förarprovet är reliabelt. Provet består av två delar och studierna fokuserar på dessa delar. Den ena behandlar tillförlitligheten (reliabiliteten) hos förarprövarnas bedömning av körprovet, dvs. i vilken utsträckning olika förarprövare gör samma bedömning, och den andra i vilken mån olika versioner av kunskapsprovet skulle ge samma resultat, i termer av godkänt eller underkänt, om provet upprepades.

Validitet
Validitetsbegreppet har förändrats en hel del över tid (för en översikt se t.ex. (Brennan, 2006; Hathcoat, 2013; Kane, 2013). Från att vara en term för ”vad provet mäter” utvecklades olika validitetsbegrepp utifrån olika aspekter (innehåll, begrepp, prediktiv förmåga m m). På senare år har ett enhetligt validitetsbegrepp, som omfattar samtliga dessa aspekter och dessutom följderna av mätningen, fått stort genomslag. Samtidigt har fokus flyttats från egenskaper hos provet till tolkningen av resultat. Synen på validering har skiftat från uträkning av ett numeriskt resultat till en ständig pågående process med samlande av argument.
När det gäller praktiska prov som körprovet ska valideringen omfatta samtliga steg i processen från observation av provtagarens beteende till provresultat till generalisering till liknande uppgifter och extrapolering till i vilken mån målen i kursplanen uppnåtts (Kane, 2013). Förarprövarna spelar en stor roll då de både styr provinnehållet, utfärån givna riktlinjer och trafiksituationen på provorten, och bedömer provtagarnas prestation. Samtidigt är alla prov olika eftersom trafiksituationen aldrig är densamma. En standardisering av provet kan göra det mer reliabelt men om provsitenationen blir alltför restriktiv påverkas validiteten då provprestationen inte längre kan generaliseras till den verklighet provet ska spegla.

För det teoretiska kunskapsprovet, och andra standardiserade prov med flervalsfrågor som rättas automatiskt, är inte provsitenationen och bedömaren lika avgörande. Där ligger fokus mer på i vilken mån provinnehållet speglar målen, om provtagarnas respons kan tolkas som att de har de kunskaper som efterfrågas och om mätningen har de avsedda konsekvenserna.

**Reliabilitet**

Reliabilitet ses numera som en egenskap hos provresultatet snarare än hos provet och vilka aspekter som är mest intressanta att undersöka beror på vilken tolkning av provresultaten man avser göra. För kriterierelaterade prov där resultaten tolkas i termer av godkänt eller underkänt är det mer intressant att veta hur exakt den klassificeringen är snarare än hur väl provpoängen avspeglar förmågan hos provtagare långt under eller över kravgränsen för godkänt.

**Ger kunskapsproven samma resultat?**
En del av reliabilitetsaspekten utgörs av antagandet att provtagarna i teorin skulle få samma resultat om de gjorde provet två gånger. Ett sådant antagande förutsätter dock att de inte lärt sig mer och inte påverkats av att de besvarat samma frågor tidigare. En annan variant är att provtagarna gör två parallella prov, dvs. prov som mäter samma sak och ger samma medelvärde och varians. Många reliabilitetsmått förutsätter att man kan arrangera två provtillfällen, vilket inte alltid låter sig göras.
I avsaknad av sådana data kan man skatta utfallet av upprepade prov för att se om provtagarna konsekvent skulle få samma klassificering (godkänt/underkänt) för båda proven. Det finns flera olika metoder för att göra sådana skattningar, men i studien som ingår i denna licentiatuppsats används de som utarbetats av Subkoviak (1976, 1988) och Hanson & Brennan (Brennan, 2004; Hanson & Brennan, 1990).

Tre provversioner som distribuerats i hela landet under sommaren 2012 studerades. De fanns ha liknande statistiska egenskaper och provtagargrupperna skilde sig inte åt med avseende på ålder, kön och hur stor andel som anmält sig till provet via trafikskola eller på egen hand. För att bara ha ett resultat för varje provtagare valdes de som genomförde provet för första gången ut (ca 4 000 per version). Enligt de uträkningar som gjorts skulle omkring 80 procent av proven klassificeras på samma sätt om två prov genomförts. Uträkningarna av Subkoviak’s coefficient jämfördes med en förenklad version där man kan utläsa koefficienten i tabeller (Subkoviak, 1988) och utfallet blev detsamma för två provversioner (0,80) och liknande för det tredje (0,79 respektive 0,81), vilket gör att det vore en möjlig genväg. Hanson-Brennans modell med fyra parametrar gav 0,82 för samtliga provversioner.

Ett annat mått är i vilken mån klassificeringen utifrån observerad poäng stämmer överens med det hypotetiska genomsnittet av oändligt många upprepade prov av sammaprovtagare ("true score” – ”sant värde”). (Crocker & Algina, 1986). Utfallet beskrivs ofta i termer av andelen som skulle få godkänt respektive underkänt trots att deras "true score" ger en annan indikation. När det gäller de provversioner som undersökt i studie II skulle omkring 6-7 procent av provtagarna felaktigt få godkänt på kunskapsprovet enligt Hanson & Brennans beräkningar för beslutsprecision, vilket motsvarar ca 250 provtagare för respektive version.

Gör förarprövarna samma bedömning?

Vägverket (senare Trafikverket) har arbetat med att standardisera körprovet och bedömningen, genom att bland annat utarbeta en provnyckel där kriterier för bedömningen klargörs och genomföra samåkning följt av diskussioner av kvalitet i prov och bedömning. Som ett led i detta arbete genomfördes en studie av likvärdigheten i bedömningen. 83 förarprövare fick samäka med en av fem särskilt utvalda förarprövare under en dag. Alla förarprövare fyllde i protokoll under provet och svarade på enkäter i samband med studien. Även provtagarna fick fylla i enkäter där de redogjorde för sin bakgrund, sina förberedelser, hur de skattade sin kompetens samt hur de såg på provet. Totalt ingick 535 prov i studien som pågick från mitten av oktober 2008 till slutet av januari 2009.
Resultaten visade att de ordinarie förarprövarna och de särskilt utvalda prövarna som samäkte med dem gjorde samma bedömning av provtagarnas prestation i termer av godkänt/underkänt vid 93 procent av proven. I de fall bedömningarna skilde sig åt var det i några fall kopplat till olika åsikter om hur ett visst beteende skulle tolkas, men oftare handlade det om skillnaderna i hur situationen uppfattades från fram- respektive baksåte. Inte i något fall kunde skillnaderna i bedömning kopplas till de mer än hundratals variabler som kom från enkätsvar och medåkarnas bedömning. Endast tre av variablerna i enkätsvaren kunde kopplas till skillnaderna och dessa gällde hur svårt den ordinarie förarprövaren ansåg det var att bedöma provet, när under provet de fattade sitt beslut och medåkande förarprövares bedömning av ordinarie förarprövares helhetsbedömning.

Hur kan man tolka resultaten?


Vad betyder reliabiliteten för validiteten?

**Fortsatt forskning och utvecklingsarbete**


För att få tillräckligt underlag för slutsatser om man mäter det man tror sig mäta behövs valideringsstudier. Kanske kan en tydligare specificering av vilka slutsatser man tror sig kunna dra av resultatet på förarprov ge grund för en mer kritisk granskning av vilket stöd det finns för detta och ytterligare studier kring provets innehåll, svårighetsgrad, kravgränssättning och konsekvenser. Reliabilitetsstudier är en oumbärlig förutsättning för en systematisk validering av provresultaten och de inkluderade studierna utgör ett inledande steg i det arbetet.
Acknowledgements

First I would like to thank the staff at the Swedish Road Administration for making these studies possible and for continuing to provide information and help after they moved to the Transport Agency and the Transport Administration.

My heartfelt thanks go to Anna Sundström and Widar Henriksson who got the ball rolling and were my first supervisors. Anna also generously invited me to write a paper with her (study I in this thesis).

As I have been studying part time this has been a long process and when Widar retired, Hanna Eklöf stepped in as my supervisor (together with Anna). Both in her role as a supervisor and as a project leader in the projects I have been part of she has helped me improve and do better.

Over the years I have also benefitted from comments and advice from my fellow postgraduate students and others involved in seminars and discussion groups. Thanks to you all!

Many other current and former colleagues have also encouraged me to become a postgraduate student and helped me in various ways along the way. All of my colleagues have contributed to making working at the university enjoyable, both during work sessions and coffee breaks. It is a long list since I have worked at the department for a long time so won't name you all, but I hope you still realize how much I appreciate you.

Naturally, I am, as always, grateful for my wonderful family. You help me focus on the important things in life and are invaluable resources of inspiration, joy, inner strength and inside jokes. I love you more than I can express! You help me towards putting the right perspective on things.
References


Alger, S., & Eklöf, H. (2017). *Samma prov - andra provtagare: En studie av förändringar i körprovsresultat och provtagargruppens sammansättning* [The same test – different test-takers; a study of changes regarding the percentage of passed driving tests and the composition of test-taker population] (BVM 66). Retrieved from Umeå university:
http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-142165


http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-39256

http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-50504

http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-50506


Henriksson, W., Sundström, A., & Wiberg, M. (2004). *The Swedish driving-license test: A summary of studies from the department of educational measurement, Umeå University (EM 45).* Retrieved from Umeå university:


52


