TØI rapport
662/2003

Stryk eller stå

En undersøkelse av faktorer som påvirker resultatene av praktisk førerprøve

Torkel Bjørnskau

ISSN 0802-0175
ISBN 82-480-0361-2

Oslo, august 2003

Forord

Rapporten beskriver resultatene fra et forsøk der kandidater til førerprøven fra Oslo, Fredrikstad, Mysen og Drøbak ble samlet og kjørte opp på en felles oppkjøringsrute ved Moss Trafikkstasjon. Hver kandidat ble vurdert av to sensorer fra de samme områdene. Forsøket ble gjennomført for å undersøke om det er systematiske forskjeller på kandidater og på sensorvurderinger avhengig av hvilket trafikkmiljø de har erfaring fra.

Designet av eksperimentet og innsamlingen av data har Statens vegvesen stått for, mens Transportøkonomisk institutt har hatt ansvaret for analysen av data og skrivingen av rapporten.

Ved Statens vegvesen har Alf Glad og Dag Terje Langnes organisert forsøket og gjennomført den praktiske datainnsamlingen. Alf Glad har også innhentet data fra de forskjellige sensorene.

Ved TØI har Torkel Bjørnskau vært prosjektleder, punchet og analysert data samt skrevet sluttrapporten. Forskningsleder Fridulv Sagberg har gjennomført kvalitetssikringen, og avdelingssekretær Trude Rømming har tilrettelagt rapporten for trykking.

Oslo, august 2003
Transportøkonomisk institutt

Sønneve Ølnes
Fung. instituttsjef

Fridulv Sagberg
forskningsleder

Innhold

TØI rapport 662/2003
Forfatter: Torkel Bjørnskau
Oslo 2003, 26 sider

Sammendrag:

Stryk eller stå

En undersøkelse av faktorer som påvirker resultatene av praktisk førerprøve

Bakgrunn og problemstilling

Statens vegvesen har ønsket å få undersøkt om det er systematiske forskjeller i hvor vanskelig den praktiske førerprøven i kl. B er, avhengig av hva slags trafikkmiljø kandidatene har hatt opplæring i og hva slags trafikkmiljø de kjører opp i. Utgangspunktet er at man mistenker at det kan være vanskeligere å kjøre opp i komplisert trafikk i store byer enn på landsbygda eller i små byer og følgelig at kandidater som kjører opp i storbyene både har hatt en mer krevende opplæring og får en mer krevende førerprøve enn kandidater fra områder der trafikkmiljøet er enklere.

For å undersøke om dette er tilfellet ble det arrangert et forsøk høsten 2002 ved trafikkstasjonen i Moss. 30 kandidater fra Oslo, 30 fra Fredrikstad og 30 fra Mysen/Drøbak avla praktisk førerprøve på en og samme kjørerute og ble vurdert av 12 sensorer hentet fra de samme trafikkstasjonene (4 fra Oslo, 4 fra Fredrikstad, 2 fra Mysen og 2 fra Drøbak)

Hver kandidat ble vurdert av en sensor i forsetet (som avgjorde prøven) og en i baksetet som også vurderte kandidaten i henhold til vanlige prosedyrer. Alle kombinasjoner av sensorer og kandidater ble benyttet.

Gjennom et slikt undersøkelsesopplegg var det mulig å finne ut om kandidatene fra storbyområdene var flinkere ved oppkjøringen til den praktiske prøven, og det var mulig å undersøke om sensorene fra ulike områder vurderte kandidatene forskjellig.

Resultater

Resultatene viser at sensorene ikke alltid vurderer en og samme kandidat likt. I 28 prosent av prøvene vurderte den ene sensoren at kandidaten besto prøven, mens den andre sensoren vurderte kandidaten til stryk. Det viste seg også å være relativt store variasjoner mellom sensorenes registreringer av antall feil. Nå vil det alltid være en viss uoverensstemmelse mellom flere sensorers vurdering av samme kandidat, men resultatene her viser relativt svak reliabilitet ut fra vanlige kriterier.

Rapporten kan bestilles fra:
Transportøkonomisk institutt, Postboks 6110 Etterstad, 0602 Oslo
Telefon: 22 57 38 00 Telefax: 22 57 02 90

Om sensor kom fra en trafikkstasjon i storby eller en på landet, hadde ingen signifikant betydning for resultatene. Det var imidlertid en viss tendens til forskjeller i vurderingene mellom kvinnelige og mannlige sensorer, og en viss tendens til at de mest erfarne sensorene krysset av for færre feil på vurderingsskjemaet enn de mindre erfarne sensorene. Det var imidlertid ingen signifikant tendens til at det var lettere å bestå prøven med en erfaren sensor. Det betyr antakelig at de erfarne sensorene baserer vurderingene sine mer på skjønn og mindre på avkrysninger på skjemaet enn hva de mindre erfarne gjør.

Et interessant resultat var at mannlige kandidater i større grad besto den praktiske førerprøven med mannlig sensor, og at kvinnelige kandidater i større grad besto prøven med kvinnelig sensor. Det er uvisst hva dette skyldes, men det kan tenkes at mannlige sensorer, som gjennomgående er mer erfarne, aksepterer (og premierer) en kjørestil som er mindre korrekt i forhold til regelverket og mer "normal" enn det de kvinnelige sensorene gjør. Det kan føre til at mannlige kandidater favoriseres av mannlige sensorer og at kvinnelige kandidater favoriseres av kvinnelige sensorer.

Et annet interessant resultat var at uoverensstemmelsen mellom sensorvurderingene ser ut til å være større med utenlandske enn med norske kandidater. Hva det skyldes er uvisst, men det kan bety at det for mange sensorer er vanskeligere å tolke utenlandske kandidaters kjøreatferd.

Kandidatene fra Oslo var gjennomgående svakere enn kandidatene fra Mysen/Drøbak og Fredrikstad, til forskjell fra hva man på forhånd kanskje hadde ventet. Grunnen til dette var imidlertid først og fremst at kandidatene fra Oslo gjennomgående var eldre, og at flere var av utenlandsk opprinnelse. Generelt var det klare tendenser til at de som kjørte opp da de var 19 år eller eldre gjorde det dårligere enn de som kjørte opp som 18-åringer. Det var også helt klare tendenser til at utenlandske kandidater hadde høyere strykprosent enn norske kandidater.

Det viste seg at kandidatene fra Oslo hadde signifikant flere feil ved kjøring på landevei/motorvei enn kandidater fra Mysen/Drøbak og Fredrikstad, etter kontroll for alder og landbakgrunn. Det viste seg også at kandidatene fra Mysen/Drøbak hadde signifikant flere feil ved bykjøring enn kandidater fra Fredrikstad (og Oslo). Vi fant m.a.o. en viss tendens til at det trafikkmiljøet man har erfaring fra under øvelseskjøring påvirker resultatene til prøven. De som har erfaring fra bykjøring gjør det relativt bedre ved bykjøring enn ved landeveiskjøring, mens det omvendte er tilfellet for de som øvelseskjørt mest utenfor byområdene.

Den variabelen som var mest utslagsgivende for antall feil og for sjansen for å bestå prøven, var antall timer ved kjøreskole. Kandidater med mange timer på kjøreskole hadde flere feil og høyre andel stryk enn kandidater med få timer. Samtidig var det klare tendenser til at de eldste kandidatene og de utenlandske kandidatene hadde flere timer på kjøreskole enn andre. Grunnen til at de med mange kjøretimer på skole gjør det såpass dårlig til prøven er at disse har lite privat øvelseskjøring, og at de kjører opp før de har de nødvendige ferdighetene.

1 Bakgrunn

Statens vegvesen har ønsket å få undersøkt om det er systematiske forskjeller i hvor vanskelig den praktiske førerprøven i kl. B er, avhengig av hva slags trafikk-miljø kandidatene har hatt opplæring i og hva slags trafikkmiljø de kjører opp i. Utgangspunktet er at man mistenker at det kan være vanskeligere å kjøre opp i komplisert trafikk i store byer enn på landsbygda eller i små byer og følgelig at kandidater som kjører opp i storbyene både har hatt en mer krevende opplæring og får en mer krevende førerprøve enn kandidater fra områder der trafikkmiljøet er enklere.

Dette kan bety at kandidater som består førerprøven i storbyene er dyktigere enn kandidater som består førerprøven på landsbygda eller i mindre byer. Det kan imidlertid tenkes at sensorene også tilpasser sin bedømmelse til hvor vanskelig trafikkmiljøet er, og at de dermed er strengere overfor elever som kjører opp i enkle trafikkmiljøer enn overfor elever som kjører opp i kompliserte trafikkmiljøer. En slik praksis kan helt eller delvis kompensere eller overkompensere for eventuelle forskjeller i kjøreprøvens vanskelighetsgrad. Dermed kan det være lettere å kjøre opp i byen enn på bygda, det kan være like lett, eller det kan være vanskeligere.

2 Metode

2.1 Utvalg

For å undersøke om det er systematiske forskjeller i ferdigheter blant kandidater fra by og land ble det gjennomført et eksperiment høsten 2002. 30 kandidater fra Oslo, 30 fra Fredrikstad og 30 fra Drøbak/Mysen ble rekruttert til å kjøre opp på en felles oppkjøringsrute i Moss. Kandidatene ble vurdert av til sammen 12 sensorer fra trafikkstasjonene i hhv Oslo (4 stk), Fredrikstad (4 stk), Follo (Drøbak) (2 stk) og Mysen (2 stk).

Kandidatene ble rekruttert ved at et utvalg kjøreskoler ble kontaktet og forespurt om de ville være med på forsøket i hvert av de fire områdene. Kjøreskolene bestemte selv hvilke kandidater som skulle tilbys å være med på opplegget og kandidatene kunne også selv velge om de ville være med eller ikke. Utvalget av kandidater er følgelig ikke trukket tilfeldig. Det kan være grunn til å anta at det er de mest seriøse kjøreskolene som er med, og det kan ikke utelukkes at kjøreskolene har valgt ut kandidater som er flinkere enn gjennomsnittet. Det innebærer at kandidatene som er med på forsøket kan være bedre enn gjennomsnittet og at antall feil ved oppkjøring og andel stryk dermed ikke er representative for kandidater generelt. I og med at formålet med forsøket er å sammenligne kandidater fra ulike trafikkmiljøer, er det liten grunn til å tro at denne selvseleksjonen skal ha betydning for resultatet. Det som kan forstyrre resultatet er dersom kjøreskolene fra f. eks. Fredrikstad i større grad har valgt ut de beste kandidatene enn kjøreskolene fra f. eks. Oslo. Det er imidlertid ingen grunn til at det skulle være tilfellet.

2.2 Feilindeks

Hver kandidat ble vurdert av to sensorer, men kun vurderingen av sensoren i forsetet (sensor 1) var bestemmende for utfallet av prøven. Alle kombinasjoner av sensorer ble benyttet. Alle kandidatene kjørte opp på en og samme rute ved Moss. Dette var en meget sammensatt rute med kjøring både på landevei, motorvei, i by og tettsted og i boligområder.

Andelen av kandidatene som består eller stryker til førerprøven er et meget grovt mål på kandidatenes ferdigheter. Det kan være ørsmå forskjeller som skiller mellom en som består og en som stryker, og det kan være store forskjeller både blant de som består og blant de som stryker.

For å vurdere kandidatenes prestasjoner utover om de har bestått eller strøket til prøven har vi benyttet en indeks utviklet av Statens vegvesen. Denne indeksen er basert på evalueringsskjemaet som sensor fyller ut til prøven, der sensor angir både positive og negative aspekter ved kandidatens kjøring. Dersom kandidaten viser positiv atferd, bedre enn gjennomsnittet, belønnes dette med ”+” i skjemaet.

Negativ atferd, klassifiseres som feil av tre ulike typer; 1 mindre feil, 2 større feil og 3 alvorlige feil.

For å lage en samleindeks som tar hensyn til både positiv og negativ atferd, og som vekter dette har Statens vegvesen laget en indeks der hver + multipliseres med 1, og hver feil type 1 med -1 og hver feil type 2 multipliseres med -2. Feil av type tre multipliseres med -5. En kandidat som har 2 +, 10 feil av type 1, 2 feil av type 2 og en feil av type 3 får følgelig (2+ (-10) + (-4) + (-5)) = -17 på indeksen. Vi har valgt å benytte samme indeks, men snudd fortegnet slik at høye indeks-verdier innebærer mange feil.

2.3 Vektet antall feil

I tillegg til feilindeksen har vi også beregnet såkalt vektet antall feil både totalt og for ulike trafikkmiljøer. Dette målet skiller seg fra feilindeksen kun ved at eventuelle pluss-merknader ikke er med i beregningen. Grunnen til at vi ikke har med eventuelle pluss-merknader når vi fordeler sensors vurdering på ulike trafikk-miljøer er at det i mange tilfeller ikke har vært mulig å bestemme hvor sensor har plassert slike pluss-merknader. I en del tilfeller har dessuten sensor angitt pluss-merknader i samletabellen i skjemaet, men ikke i skjemaet som fylles ut underveis.

2.4 Analyse

I og med at hver kandidat vurderes av to sensorer, kan data analyseres på prinsipielt to ulike måter. En kan for det første analysere hvordan hver kandidat presterer i henhold til vurderingene til begge sensorene. Det innebærer at en har i alt 90 enheter, det samme som antall kandidater.

En annen mulighet er å betrakte hver sensors vurdering som en enhet. Det innebærer at en har i alt 180 enheter, men der hver enhet har to identiske verdier på en del av variablene. Kandidat nr. x vurderes av både sensor A og B. Det gir to enheter i datamatrisen, men egenskaper knyttet til kandidaten vil være de samme for begge enhetene. Fordelen ved å analysere data slik er at en får dobbelt så mange enheter, og det gjør det mulig å analysere alle sensorvurderingene samlet.

Vi har valgt begge framgangsmåter i analysen av data. Med hver kandidat som enhet og to sensorvurderinger for hver kandidat vil det for det første være mulig å undersøke reliabiliteten i sensorvurderingene, dvs i hvilken grad ulike sensorer vurderer en og samme kandidat på samme måte. I utgangspunktet er idealet at begge sensorene skal vurderer en og samme kandidat likt. Ved å sammenligne de to sensorenes vurdering av en og samme kandidat, kan man få testet i hvilken grad sensorene er samstemte i sine vurderinger.

For det andre vil det være mulig å undersøke om det er bestemte konstellasjoner av sensorer som gir større eller mindre forskjeller i vurderingene av den enkelte kandidat. Er det for eksempel slik at sensorer fra samme trafikkstasjon er mer samstemte i sine vurderinger av den enkelte kandidat enn sensorer fra ulike trafikkstasjoner?

I den andre fremgangsmåten, med hver sensorvurdering som enhet, får vi dobbelt så mange enheter, og dette datasettet vil være egnet til å undersøke om sensorvurderingene varierer ut fra kjennetegn ved kandidat og ved sensor. I og med at alle kandidater har kjørt opp på samme rute, vil vi kunne teste om kandidater fra Oslo er bedre (eller dårligere) enn kandidater fra andre distrikter, og vi vil kunne teste om for eksempel sensorer fra Oslo vurderer kandidatene systematisk annerledes enn sensorer fra de andre områdene.

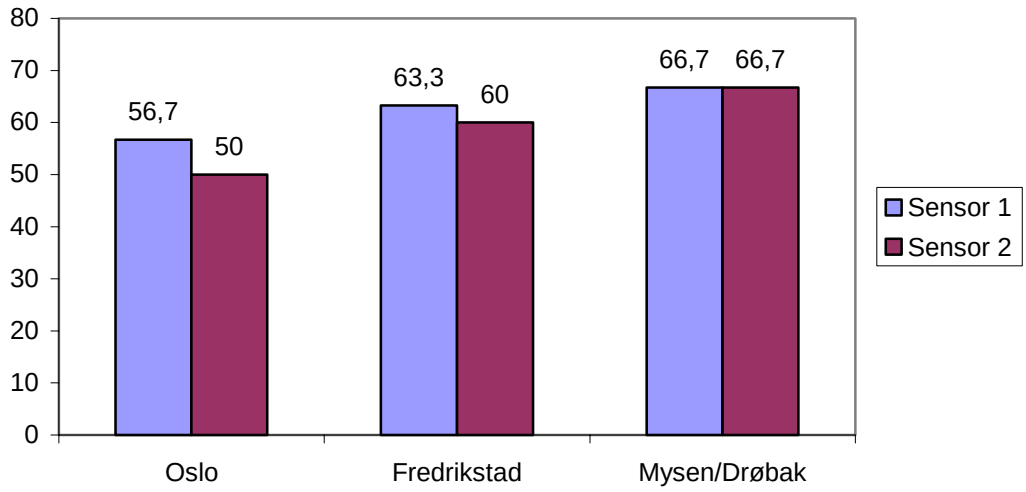
Vi vil både benytte tradisjonell tabellanalyse og multivariat regresjon. I regresjonsanalysen vil vi både benytte bestått/ikke bestått som avhengig variabel, og antall feil (vektet) i ulike trafikkmiljøer.

3 Resultater

3.1 Kandidater

3.1.1 By/land

Som nevnt har noe av bakgrunnen for prosjektet vært at man har hatt en mistanke om at kandidater fra store byer lærer å kjøre under mer krevende forhold enn kandidater fra mindre steder, og at kandidater fra store byer derfor gjennomgående er flinkere når de kjører opp enn det kandidater fra mindre steder er. En enkel test på om det er slik er å undersøke om strykprosenten varierer mellom kandidater fra ulike steder når de kjører opp på en og samme rute. Figur 3.1 viser andelene av kandidatene fra de forskjellige områdene som har bestått førerprøven i henhold til vurderingene fra sensor 1 og sensor 2.



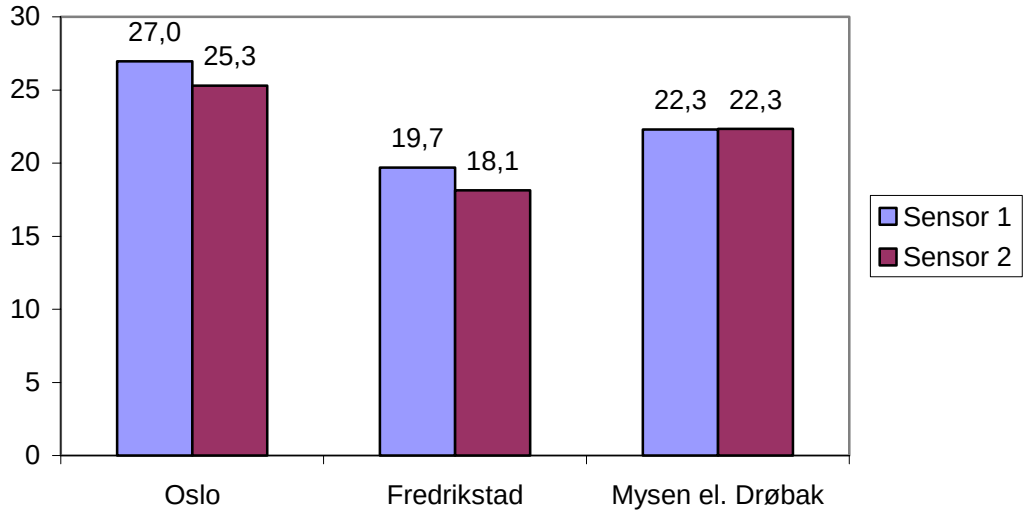
Kilde: TØI rapport 662/2003

Figur 3.1 Andel av kandidatene fra ulike områder som har bestått førerprøven i henhold til vurderingene til sensor 1 og sensor 2. Prosent.

Kandidatene fra Oslo har høyere strykprosent enn kandidatene fra Fredrikstad og fra Drøbak/Mysen. Forskjellene er imidlertid ikke statistisk signifikante, men uansett tyder dette i hvert fall på at kandidatene fra Oslo ikke har bedre ferdigheter når det kjører opp enn kandidater fra mindre steder. [1]

[1] Statistikk for 2002 for hver at trafikkstasjonene som inngår i forsøket viser at strykprosentene var 40 % i Mysen, 25 % i Drøbak, 37 % i Fredrikstad og 43 prosent i Oslo. Strykprosenten for kandidatene som er med i forsøket stemmer godt overens med disse tallene. Det tyder på at kandidatene som er med i forsøket er relativt representative.

Det samme mønsteret gjenfinnes delvis dersom vi sammenligner indeksverdiene til kandidater fra Oslo, Fredrikstad og Mysen/Drøbak.



Kilde: TØI rapport 662/2003

Figur 3.2 Gjennomsnittlig skåre på feilindeks for kandidater fra ulike områder i henhold til vurderingene til sensor 1 og sensor 2.

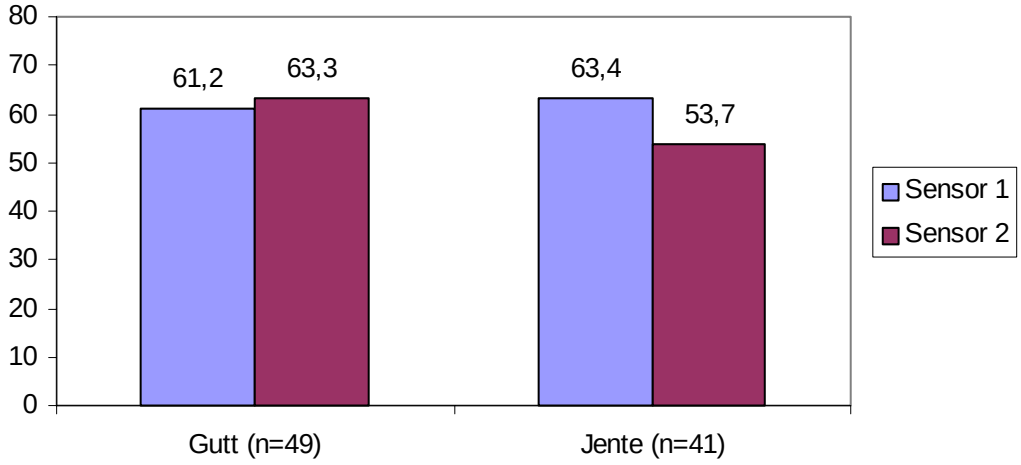
Kandidatene fra Oslo har flest feil i følge vurderingene til både sensor 1 og sensor 2, noe som overensstemmer bra med at de også har høyest andel stryk. Resultatene viser at kandidatene fra Fredrikstad skårer best, både i henhold til vurderingene til sensor 1 og sensor 2. Forskjellen mellom gjennomsnittlig indeksskåre til kandidatene fra Oslo og Fredrikstad er klart signifikant (t-verdi = 2,624 $p < 0,01$). Forskjellen mellom Oslo og Mysen/Drøbak er derimot ikke signifikant.

Det er litt overraskende at kandidatene fra Mysen/Drøbak har høyere skåre på feilindeksen enn kandidatene fra Fredrikstad i og med at strykprosenten er høyere blant kandidatene fra Fredrikstad. Det viser seg at kandidatene fra Fredrikstad har lavere skåre både blant de som har strøket og blant de som har stått enn kandidatene fra de andre områdene. Dette skyldes trolig tilfeldigheter. Det er større spredning i indeksskåren blant kandidatene fra Fredrikstad enn blant kandidatene fra de andre områdene.

Uansett om man benytter andelen bestått eller feilindeksen til å studere forskjeller mellom kandidatene etter område, kommer Oslo-kandidatene dårligere ut enn kandidatene fra Mysen/Drøbak og Fredrikstad. Antakelsen om at kandidatene fra Oslo blir flinkere før det kjører opp fordi de øvelseskjører i storbyen, får dermed ingen støtte i dette materialet.

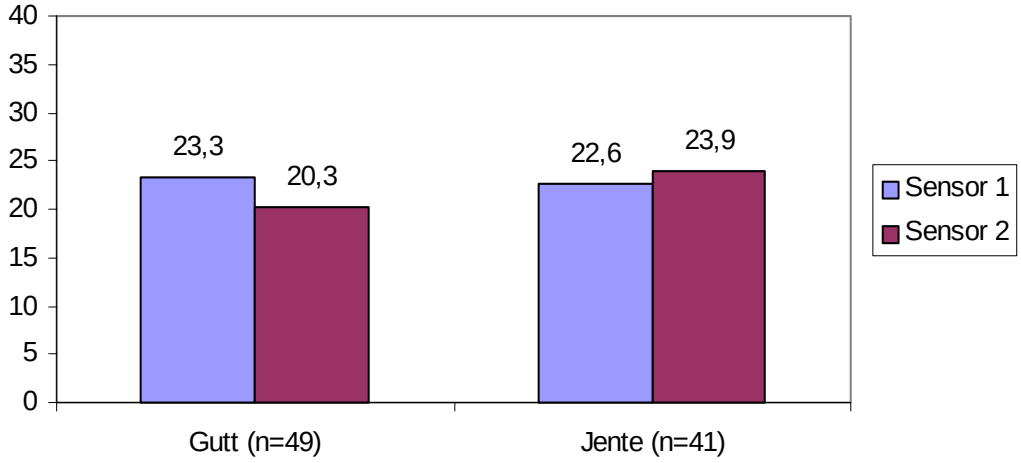
3.1.2 Kjønn

Figur 3.3. og 3.4 viser andelen som har bestått førerprøven og indeksskåre blant gutter og jenter. Forskjellene er små og ikke signifikante.



Kilde: TØI rapport 662/2003

Figur 3.3 Andel bestått til førerprøven fordelt etter kjønn og sensor. Prosent.

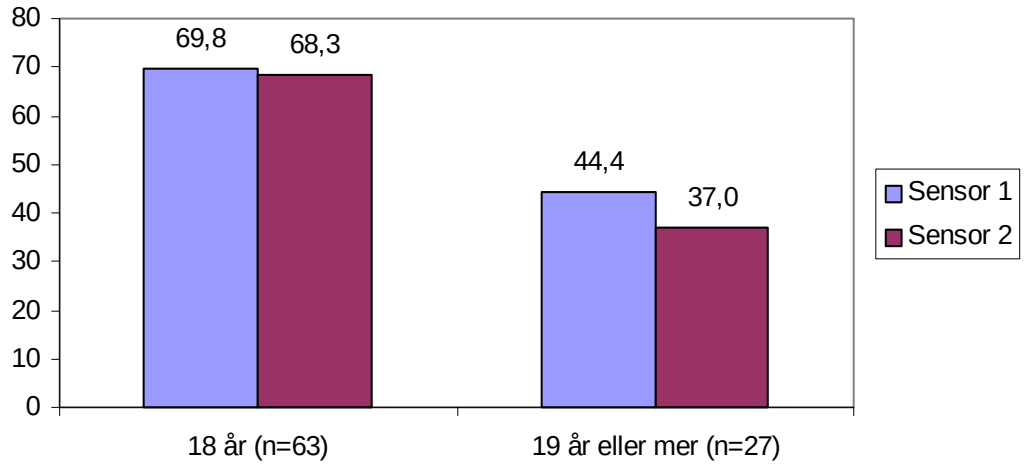


Kilde: TØI rapport 662/2003

Figur 3.4 Indeksskåre etter kjønn og sensor

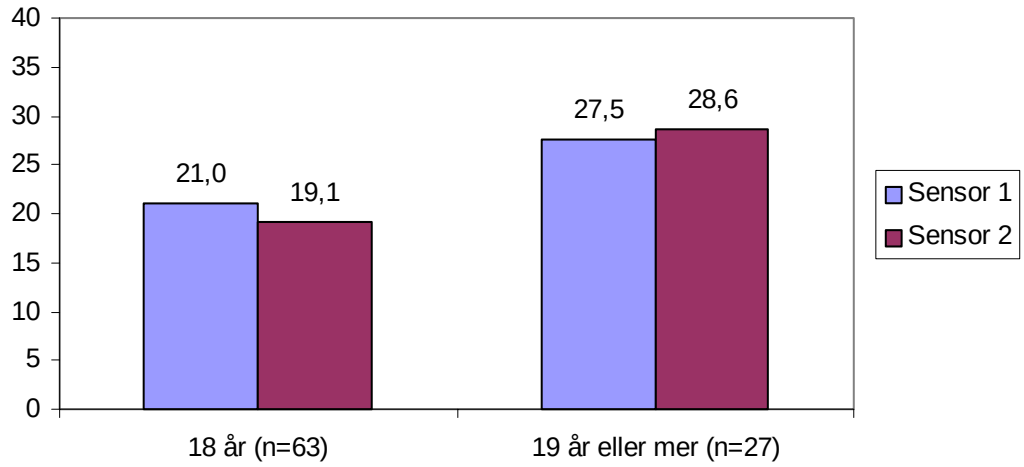
3.1.3 Alder

Figur 3.5 og 3.6 viser tilsvarende fordelinger etter kandidatenes alder, dvs. om de er 18 år eller 19 år og elder. Her er det klare forskjeller både i andelen som består prøven og indeksskåre. De som kjører opp når de er 18 år er gjennomgående flinkere og langt flere består prøven enn de som kjører opp når de er over 19 år. Disse forskjellene er klart signifikante.



Kilde: TØI rapport 662/2003

Figur 3.5 Andel bestått til førerprøven fordelt etter alder og sensor. Prosent.



Kilde: TØI rapport 662/2003

Figur 3.6 Indeksskåre fordelt etter alder og sensor

Det viser seg at denne klare aldersforskjellen er meget viktig også når det gjelder forskjellene mellom kandidater fra ulike steder. Det er langt flere eldre kandidater fra Oslo enn fra Fredrikstad og Mysen/Drøbak, og det er en viktig grunn til at kandidatene fra Oslo gjør det så dårlig på prøven. Dersom en kontrollerer for alder og sammenligner kandidater fra ulike steder som er like gamle, blir forskjellene visket ut.

Tabell 3.1 Andel av kandidatene som har bestått førerprøven fordelt etter alder og hjemsted. Prosent.

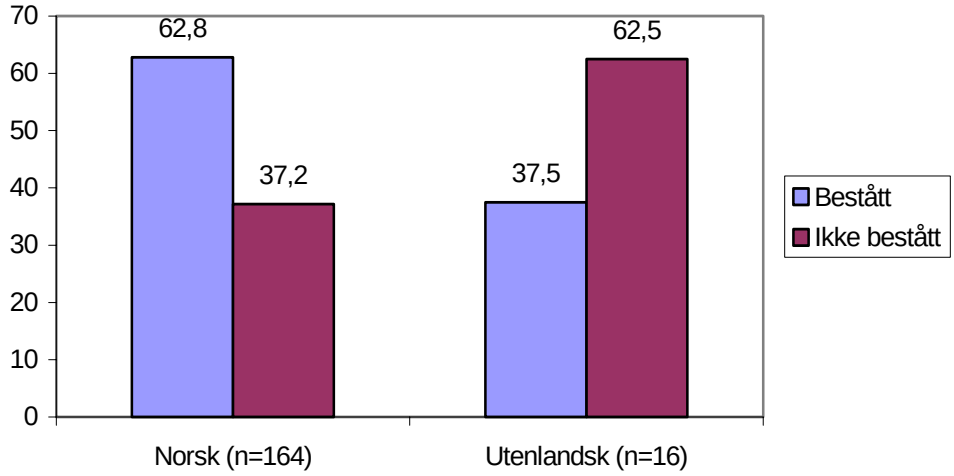
	Sensor 1		Sensor 2		Antall	
Kandidatsted	**18 år**	**19 år +**	**18 år**	**19 år +**	**N 18 år**	**N 19 år**
Oslo	78,6	37,5	71,4	31,3	14	16
Fredrikstad	73,9	28,6	69,6	28,6	23	7
Mysen/Drøbak	61,5	100,0	65,4	75,0	26	4
Total	69,8	44,4	68,3	37,0	63	27

Kilde: TØI rapport 662/2003

Tabell 3.1 viser at andelen som har bestått førerprøven blant 18-åringene er høyere blant kandidatene fra Oslo enn blant kandidatene fra Fredrikstad og Mysen/Drøbak både i følge vurderingene til sensor 1 og sensor 2. Forskjellene er imidlertid ikke statistisk signifikante. Blant de eldre kandidatene er det også flere av kandidatene fra Oslo som har bestått enn av kandidatene fra Fredrikstad, men færre enn kandidatene fra Mysen/Drøbak. Selv om det er veldig små tall her, er forskjellen i andelen som har bestått i følge bedømmelsen til sensor 1 signifikant både mellom Oslo og Mysen/Drøbak (p=0,025) og mellom Mysen/Drøbak og Fredrikstad (p=0,022).

3.1.4 Landbakgrunn

Ut fra navnelister på kandidatene er det laget en variabel for landbakgrunn. Her er kandidater med ikke-vestlige navn både som fornavn og etternavn kodet som ”utenlandske” og de resterende kodet som ”norske”. Det viser seg at det er en signifikant tendens til at strykprosenten er høyere blant de utenlandske kandidatene (kjikvadrat = 3,9, p = 0,048), noe man også har funnet indikasjoner på i andre undersøkelser (Rismark m.fl. 2002).



Kilde: TØI rapport 662/2003

Figur 3.7 Andel bestått til førerprøven fordelt etter landbakgrunn. Prosent.

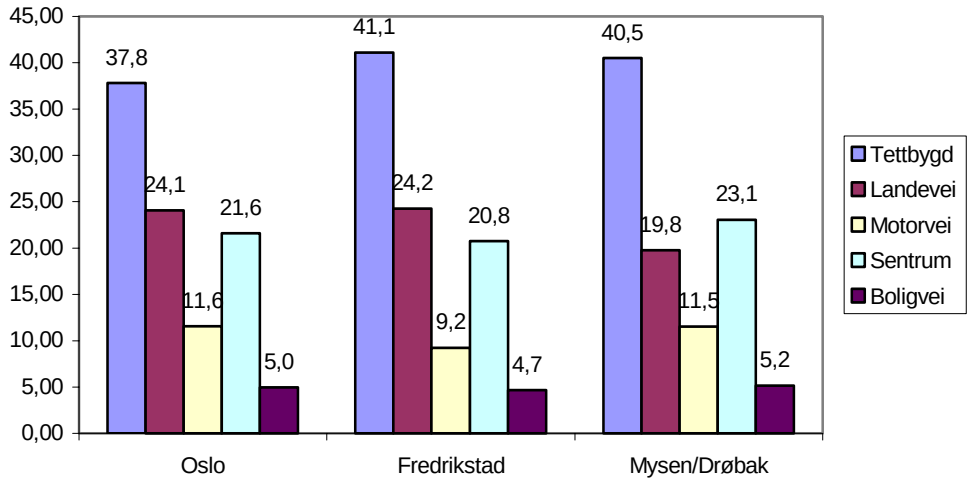
Blant de utenlandske kandidatene er det godt over 60 prosent som stryker. Blant de norske er det motsatt; over 60 prosent består prøven. Det viser seg at 7 av 8

med utenlandsk bakgrunn kommer fra Oslo, og dette er dermed også en grunn til at Oslo-kandidatene samlet gjør det svakere på prøven enn kandidatene fra de andre områdene.[2]

3.1.5 Fordeling av feil

Selv om kandidatene totalt ikke skiller seg så mye fra hverandre ut fra hva slags distrikt de kommer fra, kan det tenkes at fordelingen av feilene på trafikkmiljø er noe forskjellig. Det kan for eksempel tenkes at kandidatene fra Oslo er flinkere til å kjøre i by/tettbygd strøk enn kandidater fra mindre steder, og omvendt at de er relativt dårligere i landeveis- og motorveiskjøring.

Figur 3.8 viser fordelingen av vektet antall feil på trafikkmiljø for kandidater fra hhv. Oslo, Fredrikstad og Mysen/Drøbak.[3]



Kilde: TØI rapport 662/2003

Figur 3.8 Vektet antall feil til førerprøven fordelt etter trafikkmiljø og kandidatsted. Prosent.

Det er ikke store forskjeller i fordelingene av feilene ut fra hvor kandidatene kommer fra, men kandidatene fra Mysen/Drøbak har signifikant lavere andel feil på landevei enn kandidatene fra Oslo og Fredrikstad ($p = 0,01$). Kandidatene fra Oslo har også lavere andel feil i tettbygd strøk enn kandidatene fra Fredrikstad ($p = 0,078$).

Det er som sagt små forskjeller i fordelingen av feil, men i den grad det er forskjeller, går de i forventet retning; kandidatene fra Oslo har relativt flere feil på landevei og relativt færre feil i tettbebyggelse.

---

[2] I figur 3.7 er antall norske kandidater oppgitt til 164, og antall utenlandske kandidater oppgitt til 16. Disse antallene referer imidlertid til antall sensorbedømminger. I og med at hver kandidat er bedømt av to sensorer, blir antall sensorbedømminger dobbelt så stort som antall kandidater. Det er i alt 82 norske og 8 utenlandske kandidater i datasettet.

[3] Som nevnt i metodekapitlet er eventuelle pluss-merknader ikke med i vektet antall feil, mens dette er med i feilindeksen..

3.2 Sensorer

Hver sensor er både sensor 1 (hovedsensor, som avgjør prøven) og sensor 2 (ekstra sensor i baksetet). Tabell 3.2 viser gjennomsnittlig indeksskåre for hver sensor (som sensor 1 og som sensor 2), antall kandidater de har bedømt og antall som har stått til førerprøven.

Tabell 3.2 Sensorenes vurderinger som sensor 1 og sensor 2. Gjennomsnittlig indeksskåre, antall kandidater vurdert og antall bestått.

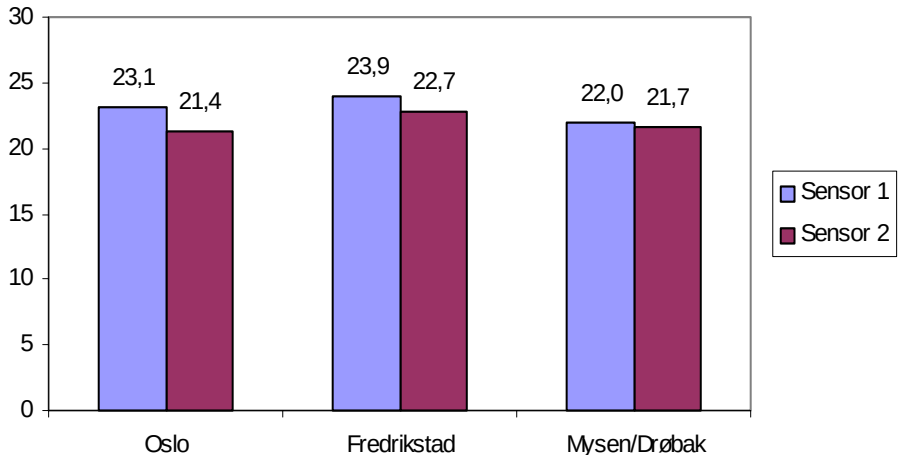
	Sensor 1			Sensor 2		
SENSOR NR	**Gj.snitt Indeks skåre**	**Antall kandidater**	**Antall bestått**	**Gj.snitt Indeks skåre**	**Antall kandidater**	**Antall bestått**
1	20,00	6	4	22,56	9	5
2	29,67	6	4	30,11	9	4
3	20,00	9	7	21,00	6	1
4	23,78	9	5	6,83	6	5
5	22,33	6	3	28,33	9	4
6	25,33	6	4	21,89	9	7
7	23,33	9	7	17,16	6	5
8	24,67	9	5	21,16	6	4
9	24,50	6	4	22,89	9	7
10	23,11	9	6	24,00	6	4
11	22,33	9	4	23,67	6	2
12	17,17	6	3	17,56	9	5
Total	22,99	90	56	21,92	90	53

Kilde: TØI rapport 662/2003

Stort sett er det små variasjoner i gjennomsnittlig indeksskåre mellom sensorene. Det største avviket finner vi for sensor 4 som har et gjennomsnitt på 6,8 som sensor 2, hvilket er svært mye bedre enn snittet for de andre sensorene, både som sensor 1 og som sensor 2. Vedkommende sensor har imidlertid også selv et helt annet gjennomsnitt som sensor 1, og det tyder på at han eller hun tilfeldigvis har hatt svært flinke kandidater som sensor 2. Det er meget store variasjoner mellom kandidatene, slik at det kan bli nokså store variasjoner mellom sensorenes gjennomsnittsskåre av rene tilfeldigheter.

Det kan se ut til at det er en viss tendens til korrelasjon mellom sensorverdiene som sensor 1 og som sensor 2. Sensor nr. 2 ligger f. eks. på ca. 30 i begge tilfeller, og sensor nr 12 ligger på ca. 17 i begge tilfeller. Dette kan ha med ”personlig stil” å gjøre, og behøver ikke å bety at sjansen for å stå til prøven er forskjellig mellom disse to sensorene. Tabellen viser at begge sensorene har nøyaktig like stor andel som har bestått førerprøven.

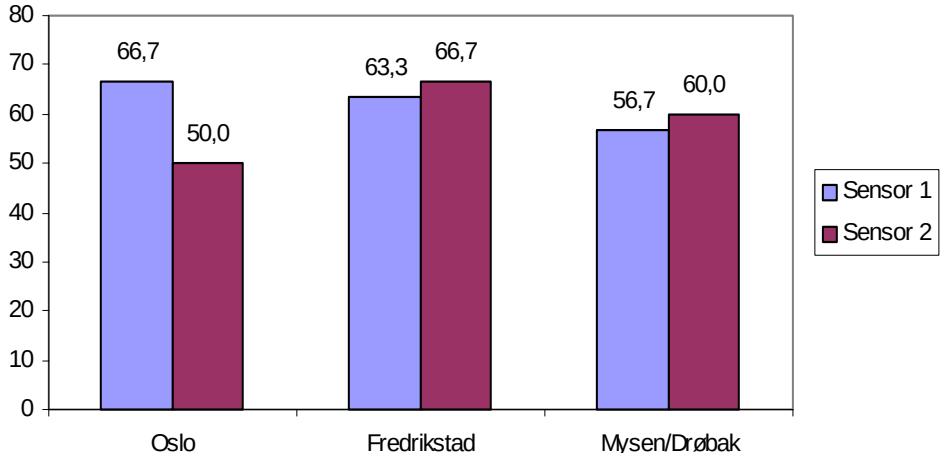
I figur 3.9 er gjennomsnittlig indeksskåre for sensorene fra de tre trafikkstasjonene vist både som sensor 1 og som sensor 2.



Kilde: TØI rapport 662/2003

Figur 3.9 Gjennomsnittlig indeksskåre for sensorer som sensor 1 og sensor 2 fordelt etter hvilken trafikkstasjon sensor kommer fra.

Det ser ikke ut til å være systematiske forskjeller i vurderingene mellom sensorene fra de ulike trafikkstasjonene. Gjennomsnittlig skåre på indeksen er svært lik. Samme tendens finner vi om vi studerer andelen stryk, jf. figur 3.10.



Kilde: TØI rapport 662/2003

Figur 3.10 Andel som har bestått førerprøven vurdert av sensor1 og sensor2 fordelt etter hvilken trafikkstasjon sensor kommer fra. Prosent.

Det er visse forskjeller i andelen som har bestått førerprøven etter hvilken trafikkstasjon sensor kommer fra, men figuren viser at det ikke er slik at den ene gruppen sensorer ligger systematisk høyere eller lavere både som sensor 1 og som sensor 2. For eksempel ser vi at sensorene fra Oslo har flest beståtte kandidater når de har vært sensor 1, men samtidig færrest når de har vært sensor 2.

Figurene 3.9 og 3.10 viser at det ikke er noen entydig tendens til at sensorene fra ett område er systematisk strengere eller mildere i sin vurdering enn sensorer fra andre områder. Sensorene fra Oslo har latt to av tre bestå prøven som sensor 1,

men bare en av to som sensor 2. At Oslo-sensorene har større forskjeller i vurderingene som sensor 1 og sensor 2 enn sensorene fra Fredrikstad og Mysen/Drøbak betyr antakelig bare at Oslo-sensorene har hatt noe svakere kandidater når de har vært sensor 2. Hadde det vært systematiske forskjeller i vurderingene mellom sensorene fra ulike områder, skulle vurderingene både som sensor 1 og som sensor 2 fra ett område ligge systematisk over eller under vurderingene til sensorene fra ett annet område. Vi ser at det er en viss tendens til at Fredrikstad-sensorene har større andeler bestått både som sensor 1 og som sensor 2 enn sensorene fra Mysen/Drøbak, men denne forskjellen er ikke statistisk signifikant.

3.2.1 Sensorenes vurdering av samme kandidat

I og med at hver kandidat vurderes av to sensorer, er det mulig å teste i hvilken grad sensorene vurderer en og samme oppkjøring likt. Dette er m.a.o. en test på hvor reliabel den praktiske førerprøven er.

Vi har sett at det i gjennomsnitt ikke er store forskjeller mellom sensorenes vurderinger når de er sensor 1 og sensor 2 og det er heller ikke store forskjeller mellom vurderingene ut fra hvor sensor kommer fra. Når vi undersøker to sensorers vurdering av en og samme kandidat, er det imidlertid klare forskjeller.

Tabell 3.3 Sensor1 og Sensor 2s vurdering av samme kandidater. Andel bestått. Prosent av total.

		Sensor 2	
Sensor 1:	**Bestått**	**Ikke bestått**	**I alt**
Bestått	46,7 (42)	15,6 (14)	62,2 (56)
Ikke bestått	12,2 (11)	25,6 (23)	37,8 (34)
I alt	58,9 (53)	41,1 (37)	100,0 (90)

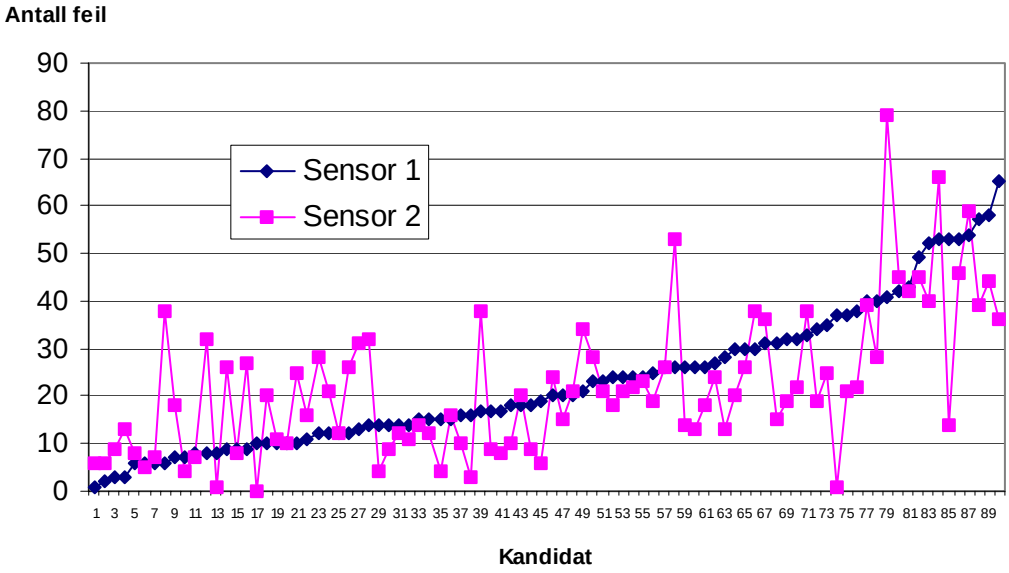
Kilde: TØI rapport 662/2003

Tabell 3.3 viser at 42 av kandidatene har bestått prøven i henhold til begge sensorenes vurderinger. Dette utgjør i alt 46,7 prosent av alle kandidatene. 23 kandidater eller 25,6 prosent er vurdert til stryk av begge sensorer, mens det er i alt 25 kandidater eller 28 prosent der sensorene er uenige om kandidaten skal stå eller stryke. Dette er en forholdsvis høy andel. Det betyr at mer enn hver fjerde førerprøve avgjøres av kjennetegn ved sensor og ikke av kandidatens prestasjoner.

Et mye benyttet reliabilitetsmål er koeffisienten ”Kappa”. Kalkulerer vi den for disse resultatene får vi verdien 0,419.[4] Dette indikerer en klar samvariasjon, men det er likevel relativt svak reliabilitet i henhold til vanlige kriterier (Murphy og Davidshofer 1998, Raaheim 2000).

[4] Kappa = (total overensstemmelse – tilfeldig overensstemmelse)/(1-tilfeldig overensstemmelse). Her er total overensstemmelse = 0,467+0,256 = 0,723. Tilfeldig overensstemmelse = [(0,622 x 0,589)+(0,378 x 0,411)] = 0,522.

At det er nokså store forskjeller i hvordan sensorene vurderer en og samme kandidat, avspeiles også av avstanden i indeksverdiene til sensor 1 og sensor 2, jfr. figur 3.11.



Kilde: TØI rapport 662/2003

Figur 3.11 Indeksskåre i følge vurderingene til sensor 1 og sensor 2 på samme kandidat.

Figur 3.11 viser indeksskåre for hver kandidat i henhold til vurderingene til sensor 1 og sensor 2. Kandidatene er sortert i stigende rekkefølge ut fra vurderingene til sensor 1. Indeksverdiene til sensor 1 varierer fra 1 til 65. Indeksverdiene til sensor 2 varierer fra 0 til 79. I de tilfellene der begge sensorene har vurdert en og samme kandidat likt, ligger punktene oppå hverandre. Det gjelder for eksempel kandidat nr. 20, 25 og 57. Ideelt sett skulle indeksverdiene til sensor 1 og 2 være identiske for alle kandidatene, og kurvene ville da vært sammenfallende. Vi ser at det er langt fra tilfellet, og avstandene er til dels store. For eksempel har kandidat nr. 74 fått verdi 38 av sensor 1, mens sensor 2 ikke har gitt denne kandidaten verdi 0. Og den kandidaten som har størst verdi, 79 på indeksskåren til sensor 2, har ikke mer enn verdi 41 på indeksskåren i henhold til sensor 1 sin bedømming.

Avstanden i indeksverdiene til sensor1 og sensor 2 varierer mellom 0 og 39. I gjennomsnitt er avstanden på 9,7. Det betyr at i gjennomsnitt har den ene sensoren registrert 10 små feil mer enn den andre, eller 5 større feil. Selv om dette virker mye, er indeksverdiene til sensor 1 og sensor 2 klart positivt korrelerte (Pearsons r = 0,62, p = 0,001). Det er med andre ord en forholdsvis klar tendens til at de kandidatene som har fått en høy verdi i henhold til den ene sensorens vurderinger også har fått en høy verdi i henhold til vurderingene fra den andre sensoren.

Raaheim (2000) gjengir flere undersøkelser av såkalt inter-reliabilitet ved eksamenbesvarelser på universitetet, dvs. hvor samstemte ulike sensorer som retter de samme oppgavene er. Resultatene varierer i stor grad, men gjennomgående ligger korrelasjonene noe over det nivået vi har funnet her.

For å vurdere om sensorene fra samme distrikt vurderer kandidatene mer likt enn sensorer som kommer fra forskjellige distrikter, har vi undersøkt hvor stor gjennomsnittlig avstand det er mellom indeksverdien til sensor 1 og sensor 2, når de er fra henholdsvis samme og forskjellige trafikkstasjoner. Det viser seg at avstanden i indeksverdiene er noe større når sensorene kommer fra forskjellige trafikkstasjoner enn de er når sensorene kommer fra samme stasjon, hhv. 7,8 og 10,1. M.a.o. dersom både sensor 1 og sensor 2 er fra Oslo, er det en tendens til at de vurderer en kandidat mer likt enn om sensor 1 er fra Oslo og sensor 2 f. eks. er fra Fredrikstad. Denne forskjellen er imidlertid ikke statistisk signifikant.

Tabell 3.4 viser hvor samstemte sensorene er m.h.t. om kandidaten skal bestå prøven eller ikke, fordelt etter om sensorene kommer fra samme trafikkstasjon og ut fra om sensorene er av samme kjønn.

Tabell 3.4 Sensorenes vurdering av om kandidaten skal bestå eller stryke. Prosent

Sensor 1 og sensor 2	Lik vurdering	Ulik vurdering	N
Fra samme trafikkstasjon	77,8	22,2	18
Fra ulike trafikkstasjoner	70,8	29,2	72
Med samme kjønn	73,9	26,1	46
Med forskjellig kjønn	70,5	29,5	44

Kilde: TØI rapport 662/2003

Tabell 3.4 viser at dersom begge sensorer kommer fra samme trafikkstasjon, dvs. at begge er fra Oslo, eller fra Fredrikstad eller fra Mysen/Drøbak, har de vært enige om at kandidaten skulle stå eller stryke i 77,8 prosent av tilfellene, og uenige i 22,2 prosent. Når sensorene har kommet fra forskjellige trafikkstasjoner har de vært litt mer uenige om kandidaten skulle stå eller stryke. Det er altså igjen en tendens til at sensorer fra samme trafikkstasjon vurderer kandidatene mer likt enn sensorer fra forskjellige stasjoner, men denne forskjellen er ikke statistisk signifikant.

Det er en ørliten tendens til at sensorene er mer samstemte i sine vurderinger når begge er av samme kjønn enn når de har forskjellig kjønn. Denne tendensen er imidlertid ikke signifikant. Det ser imidlertid ut til å være en samspillseffekt mellom kjønn på sensor og kjønn på kandidat for vurderingene om kandidaten skal bestå eller ikke. Tabell 3.5 viser at andelen som består er høyere for gutter enn for jenter blant mannlige sensorer, og klart lavere blant kvinnelige sensorer.

Tabell 3.5 Andel som har bestått og strøket fordelt etter kjønn på kandidat og kjønn på sensor. Prosent.

Sensor	**Mann**			**Kvinne**		
Kandidat:	Gutt	Jente	I alt	Gutt	Jente	I alt
Bestått	67,7	56,4	62,5	51,5	63,0	56,7
Ikke bestått	32,3	43,6	37,5	48,5	37,0	43,3
N = 100 %	65	55	120	33	27	60

Kilde: TØI rapport 662/2003

Forskjellen i andelen gutter som har bestått mellom mannlige og kvinnelige sensorer er på om lag 16 prosentpoeng, og den er nesten signifikant (p=0,059 ensidig test). Signifikanstesten tester bare om den høyere andelen bestått blant gutter med mannlig sensorer er statistisk pålitelig. Vi ser imidlertid at *samtidig* er andelen jenter som har bestått høyere for kvinnelige sensorer. Denne forskjellen (mellom andelen jenter som har bestått m/mannlig vs. kvinnelig sensor) er heller ikke signifikant, men signifikanstesten fanger ikke opp samspillseffekten, dvs. at jentekandidater i større grad består med kvinnelig sensor og guttekandidater i større grad med mannlig sensor.[5]

3.2.2 Analyse av forskjellene i sensorvurderingene

Vi har sett foran at den praktiske førerprøven til en kandidat vurderes til dels svært forskjellig av de to sensorene som er satt til å bedømme prøven. For å undersøke om det er bestemte faktorer som forklarer hvorfor vurderingene til tider er så forskjellige, har vi gjennomført en lineær regresjonsanalyse der indeksavstanden mellom vurderingen til sensor 1 og sensor 2 er brukt som avhengig variabel. Som tidligere nevnt varierer denne indeksavstanden mellom 0 og 39.

Vi har benyttet kjennetegn ved kandidat og sensorkombinasjoner som uavhengige variabler. Vi finner ikke signifikante utslag av noen av de uavhengige variablene (SENSKOMB d.v.s. om sensorene er fra samme sted eller ikke, SEXKOMB, d.v.s om sensorene har samme kjønn, ALDER, d.v.s. om kandidaten er over eller under 19 år, LANDBK d.v.s. om kandidaten er norsk eller utenlandsk, eller kandidatsted d.v.s. om kandidaten er fra Oslo, Fredrikstad eller Mysen/Drøbak[6]). Den variabelen som er nærmest til å være signifikant (p = 0,127) er LANDBK. B-verdien for denne variabelen er på 5,4 d.v.s. at i dette datamaterialet er indeksavstanden vel fem enheter større når kandidatene er utenlandske enn når de er norske. Det betyr at i gjennomsnitt er forskjellene i sensorenes vurderinger 5 poeng større når kandidaten er utenlandsk enn når han eller hun er norsk. At en såpass stor forskjell ikke er signifikant skyldes at det bare er 8 kandidater som er av utenlandsk opprinnelse i dette materialet.

Koeffisienten for landbakgrunn blir imidlertid redusert fra 5 til 2 når vi inkluderer kjøretimer og startpunkt for privat øvelseskjøring i analysen, og ingen av de uavhengige variablene er signifikante. Det viser seg at utenlandske kandidater i gjennomsnitt har hatt 32 kjøretimer utenom den obligatoriske undervisningen, mens de norske kandidatene har hatt 14. De utenlandske kandidatene har også startet å øvelseskjøre mye senere enn de norske. Selv om kjøretimer og øvelseskjøring dermed ”forklarer” mye av effekten av landbakgrunn, kan resultatene bety at sensorene faktisk vurderer utenlandske kandidaters prestasjoner mer forskjellig enn norske kandidaters prestasjoner, men at datamaterialet her er for lite til at disse forskjellene blir signifikante.

---

[5] Signifikanstesten tester bare om 67,7 % er signifikant høyere enn 51,5 % og om 63,0 % er høyere enn 56,4 %. Den tar ikke hensyn til at det er mannlig sensor/mannlig kandidat og kvinnelig sensor/ kvinnelige kandidat som utgjør de høyeste andelene bestått.

[6] Kandidatsted er kodet om til tre dummy-variabler.

3.3 Multivariate analyser av resultater på praktisk prøve

Vi har sett at en rekke kjennetegn ved kandidatene som påvirker både antall feil og om man har bestått førerprøven eller ikke, er ulikt fordelt mellom kandidater fra de ulike stedene. Vi har for eksempel sett at det er flere eldre kandidater fra Oslo enn fra Fredrikstad og Drøbak/Mysen, samtidig som de eldre kandidatene gjennomgående gjør det dårligere på førerprøven.

For å isolere effektene av de enkelte variablene har vi derfor gjennomført to sett med multivariate analyser. For det første har vi gjennomført et sett med logistiske regresjonsanalyser med bestått/ikke bestått til førerprøven som avhengig variabel og ulike kjennetegn ved kandidater og sensorer som uavhengige variabler. For det andre har vi også gjennomført tre sett med lineære regresjonsanalyser, med vektet antall feil totalt, vektet antall feil på landevei/motorvei og vektet antall feil i tettbebyggelse/by som avhengige variabler.

Valget av uavhengige variabler er dels gjort ut fra hva slags data vi har tilgang til og dels ut fra hva vi har funnet av sammenhenger i tabellanalysene presentert foran.

3.3.1 Andel bestått førerprøven som avhengig variabel

Tabell 3.6 Logistisk regresjon, bestått til førerprøven er avhengig variabel. Trinnvis prosedyre, oddsrater.

	Modell						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Kandidatsted [1]							
Oslo	0,57	0,58	0,58	0,95	1,15	0,90	0,97
Fredrikstad	0,80	0,84	0,83	0,95	0,92	0,72	0,754
Sensorsted [1]							
Oslo		1,04	1,09	0,96	0,98	1,01	0,99
Fredrikstad		1,35	1,34	1,34	1,36	1,34	1,42
Sensor (mann= 1)			1,21	1,30	1,26	1,23	1,09
Alder (18=1, 19+=0)				3,30***	3,34***	2,63**	2,36**
Landbakgrunn (norsk = 1)					2,60	1,31	1,53
Antall kjøretimer						0,96**	0,97
Øvelseskjøring							0,85
Konstant	2,00***	1,77*	1,53	0,55	0,22*	1,07	1,22
-2 log likelihood	239,17	238,46	238,32	227,32	224,78	202,5	195,4

[1] Drøbak/Mysen er referansekategori
*** signifikant på 1% nivå, ** signifikant på 5 % nivå, * signifikant på 10% nivå.

Kilde: TØI rapport 662/2003

Tabell 3.6 viser at hvor kandidaten kommer fra og hvor sensor kommer fra ikke har signifikant betydning for sannsynligheten for å stå eller stryke til den praktiske førerprøven. Modell 1 viser at kandidater fra Oslo bare har 0,57 ganger så høy sjanse (odds) for å bestå prøven som en kandidat fra Drøbak/Mysen, noe

som samsvarer med fordelingen som ble vist i figur 3.1, men forskjellen er ikke signifikant. Etter hvert som vi kontrollerer for andre variabler i de påfølgende modellene, blir forskjellen redusert. Dette skyldes at det er flere eldre og flere utenlandske kandidater fra Oslo.

I modell 6 gir alder og antall kjøretimer signifikante utslag. Sjansen for å bestå prøven er 2,6 ganger høyere for en 18-åring enn for en som er 19 år eller eldre. Denne forskjellen er betydelig og forklaringen er antakelig at det er de som har lettest for å lære å kjøre som i størst grad tar førerkortet når de er 18 år.

Landbakgrunn gir ikke signifikante effekter i disse modellene, men vi ser at oddsraten i modell 5 er på 2,6. Det innebærer at sjansen for å bestå er 2,6 ganger høyere for en norsk kandidat enn for en utenlandsk kandidat, etter kontroll for blant annet alder. Grunnen til at en såpass sterk effekt ikke er signifikant (p=0,12) er at det som nevnt bare er 8 kandidater med utenlandsk bakgrunn i utvalget.

Vi ser at effekten av både alder og landbakgrunn blir kraftig redusert når vi tar antall kjøretimer inn i modellen. Det betyr at de eldste kandidatene og de utenlandske kandidatene har hatt flest kjøretimer.[7]

Resultatene viser også at jo flere kjøretimer man har, desto større er sjansen for å stryke til førerprøven. Dette kan virke paradoksalt, men forklaringen er antakelig også her at dette fanger opp de som har vanskelig for å lære å kjøre bil. De trenger mange timer, men de kjører opp før de er utlært. Både aldersvariabelen og kjøretimer-variabelen fanger dermed opp noe av den samme egenskapen; de som har vanskeligheter med å lære å kjøre bil har mange kjøretimer og kjører ofte opp når de er 19 år eller mer. De stryker også signifikant oftere enn de med færre timer og som kjører opp når de er 18 år.

Det viser seg at når vi legger inn omfanget av privat øvelseskjøring (målt i antall måneder man har kjørt før førerprøven) i modell 7 sammen med antall kjøretimer, forsvinner langt på vei effekten av kjøretimer.[8] Det betyr at vår tolkning over er riktig; de med mange timer på skole har hatt lite privat øvelseskjøring og er derfor ikke så flinke når de kjører opp.

Det viser seg at de som startet å øvelseskjøre mindre enn 3 måneder før de kjørte opp i gjennomsnitt hadde 29 timer ved kjøreskole, men kun 25 % besto den praktiske prøven. Til sammenligning hadde de som startet å øvelseskjøre mer enn 12 måneder før oppkjøring i gjennomsnitt 14 timer ved kjøreskole, og hele 70 % besto den praktiske prøven.

I tillegg til variablene som er testet i modellene i tabell 3.6 har vi også foretatt analyser der flere variabler knyttet til kandidat og sensor er tatt med. Om kandidaten er mann eller kvinne bidrar ikke signifikant i noen modeller, og det er heller ingen bivariat sammenheng mellom kjønn og andelen som har bestått, slik at modeller med denne variabelen er ikke presentert i tabell 3.6. Kjennetegn ved

---

[7] De som er 18 år når de kjører opp har i gj.snitt 13,8 timer på skole; de som er 19 har 20,3 timer i tillegg til den obligatoriske undervisningen.

[8] Variabelen har fem verdier; 1 = Mer enn 18 mnd siden startet å øvelseskjøre, 2 = mellom 12 og 18 mnd, 3 = mellom 6 og 12 mnd, 4 = mellom 3 og 6 mnd og 5 = mindre enn tre mnd siden startet å øvelseskjøre. Høy verdi på variabelen innebærer m.a.o. lite privat øvelseskjøring, og følgelig innebærer oddsrater < 1 at lite øvelseskjøring reduserer sjansen for å bestå førerprøven.

sensor som sensors alder, erfaring eller om han/hun har sensorkurs har heller ikke effekter i disse modellene. Mer detaljerte analyser viser imidlertid en viss tendens til at i de tilfellene der kandidater stryker med en kvinnelig sensor har sensor kortere erfaring enn i de tilfellene der kandidater står med kvinnelig sensor (T-verdi = 1,58, p=0,12 tosidig test). Det ser m.a.o. ut til at det er en viss tendens til at sensors erfaring påvirker resultatene med kvinnelige sensorer, men ikke med mannlige.

3.3.2 Vektet antall feil som avhengig variabel

I tabell 3.7 - 3.9 er resultatene av ulike lineære regresjonsmodeller vist, med hhv. vektet antall feil totalt, vektet antall feil i på landevei/motorvei, og vektet antall feil i tettsted/by som avhengige variabler.

Tabell 3.7 Multivariat regresjon. Avhengig variabel = vektet antall feil totalt. Trinnvis prosedyre. B-koeffisienter.

	Modeller							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Kandidatsted [1]								
Oslo	3,65	3,67	3,48	0,57	-0,68	2,67	1,87	1,98
Fredrikstad	-3,62	-3,46	-3,39	-4,18	-3,95	-1,18	-1,15	-0,46
Sensorsted [1]								
Oslo		0,67	-0,15	0,57	0,44	-0,04	-0,61	-0,57
Fredrikstad		1,72	1,74	1,92	1,79	1,56	1,76	5,59*
Sensor (mann= 1)			-3,32	-3,54	-3,38	-3,92*	-3,26	-2,84
Alder (18=1, 19+=0)				-7,06***	-6,98***	-4,79*	-3,60	-4,12
Landbakgrunn (norsk = 1)					-6,49*	-0,09	-2,84	-3,80
Antall kjøretimer						0,35***	0,28**	0,33**
Øvelseskjøring							1,17	0,25
Sensorerfaring								-0,38**
Konstant	23,28***	22,43***	24,95***	30,96***	37,15***	23,66***	23,27***	18,19**
R^2	0,04	0,04	0,06	0,10	0,11	0,17	0,15	0,19

[1] Drøbak/Mysen er referansekategori
*** signifikant på 1% nivå, ** signifikant på 5 % nivå, * signifikant på 10% nivå.

Kilde: TØI rapport 662/2003

Modell 1 i Tabell 3.7 viser at kandidater fra Fredrikstad har færre feil enn kandidater fra Mysen/Drøbak, mens kandidatene fra Oslo har flere feil. Dette så vi også i figur 3.2. Forskjellene er imidlertid ikke statistisk signifikante (p=0,166 (Oslo) og p=0,17 (Fredrikstad)).

Hvilken trafikkstasjon sensor kommer fra bidrar ikke signifikant i modellene 2-7, men i modell 8 synes det plutselig å være en klar tendens til at sensorer fra Fredrikstad registrerer flere feil enn sensorer fra andre trafikkstasjoner. I modell 8 er antall år som sensor lagt inn i tillegg til de andre uavhengige variablene. Det fører til at koeffisienten for sensorsted Fredrikstad blir sterkt positiv (5,59) og signifikant på 10 prosents nivå. Forklaringen er at Fredrikstad-sensorene

gjennomgående har langt større erfaring enn sensorene fra de andre to trafikk-stasjonene. Vi må være varsomme med å tolke denne effekten substansielt i og med at det er såpass få sensorer fra hver trafikkstasjon.

Effekten av sensors kjønn er mer konsistent, og denne variabelen er signifikant (p=0,099) i modell 6. Tolkningen er at mannlige sensorer registrerer noe færre feil enn kvinnelige sensorer. I gjennomsnitt har kvinnelige sensorer registrert 25,5 feil (vektet), mens mannlige sensorer har registrert 22 feil (vektet). Forskjellen er faktisk større når vi kontrollerer for ulike egenskaper ved kandidat.

I modell 5 er det en signifikant effekt av landbakgrunn, men den forsvinner når vi kontrollerer for antall kjøretimer i modell 6. Det skyldes som tidligere nevnt at utenlandske kandidater nesten har dobbelt så mange kjøretimer som norske kandidater. Tabell 3.7 viser også at alder og særlig antall kjøretimer gir signifikante effekter. Vi ser at effekten av alder ikke svekkes nevneverdig når landbakgrunn tas inn i modell 5, men blir klart svekket når kjøretimer tas inn i modell 6. Det betyr at de eldste kandidatene har tilnærmet samme fordeling på landbakgrunn som de yngre, men at de gjennomgående har flere kjøretimer på kjøreskole.

Hvor lenge man har drevet med privat øvelseskjøring bidrar ikke signifikant i de to modellene der denne variabelen er med. Det skyldes i stor grad at variabelen er lagt inn *etter* at antall kjøretimer er med i modellen. Om vi i stedet benytter en modell der antall kjøretimer ikke er med, får vi klare og signifikante effekter av privat øvelseskjøring. Kjøretime-variabelen kamuflerer m.a.o. mye av effekten av privat øvelseskjøring. De som har hatt mye privat øvelseskjøring har om lag halvparten så mange kjøretimer på skole og halvparten så mange feil ved oppkjøring som de med minst privat øvelseskjøring.

Sensors erfaring målt i antall år som sensor viser seg å ha signifikant negativ effekt i modell 8. Det betyr at sensorer med lang erfaring krysser av for færre feil på skjemaet enn sensorer med kortere erfaring. I og med at vi ikke fant noen signifikant betydning av sensors erfaring på sjansen for å bestå prøven, er det nærliggende å tolke resultatet fra tabell 3.7 som at erfarne sensorer i mindre grad bruker skjemaet i sine vurderinger enn mer uerfarne sensorer.

R^2 er på 0,19 i den beste av modellene i tabell 3.7. Det innebærer at de uavhengige variablene forklarer om lag 19 % av variasjonen i den avhengige variabelen, noe som er relativt beskjedent. En viktig grunn til det er at indeksskårene til sensor 1 og sensor 2 for en og samme kandidat varierer forholdsvis mye noe som tyder på at mye av variasjonen i antall feil er knyttet til variasjoner mellom sensorene som vi ikke har variabler som fanger opp.

Tabell 3.8 viser tilsvarende regresjonsmodeller som tabell 3.7, men nå med vektet antall feil på landevei/motorvei som avhengig variabel. Det er i stor grad de samme variablene som er utslagsgivende for antall feil ved kjøring på landevei/motorvei som for antall feil totalt. Det er også i stor grad de samme endringene som skjer i b-koeffisientene etter hvert som vi kontrollerer for flere uavhengige variabler.

Det er imidlertid to interessante forskjeller i forhold til resultatene fra tabell 3.7. For det første ser vi her at Oslokandidatene har signifikant flere feil enn kandidater fra Mysen/Drøbak (og også enn kandidater fra Fredrikstad, men det fremkommer ikke av tabellen) i de fleste modellene. Det er i tråd med hva man kunne forvente; kandidater fra Oslo har antakelig mindre trening i landeveiskjøring enn kandidater fra mindre steder.

Tabell 3.8 Multivariat regresjon. Avhengig variabel = vektet antall feil på landevei/motorvei. Trinnvis prosedyre. B-koeffisienter.

	Modeller							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Kandidatsted [1]								
Oslo	2,32**	2,29**	2,18**	1,24	0,59	2,33**	2,39**	2,57**
Fredrikstad	-0,70	-0,65	-0,61	-0,87	-0,75	0,55	0,61	0,83
Sensorsted [1]								
Oslo		0,82	0,32	0,55	0,49	-0,06	-0,49	-0,12
Fredrikstad		1,14	1,15	1,21	1,14	1,03	1,12	2,32**
Sensor (mann= 1)			-2,00**	-2,07**	-1,99**	-2,16***	-1,87**	-1,6*
Alder (18=1, 19+=0)				-2,29**	-2,25**	-1,37	-1,14	-1,3
Landbakgrunn (norsk = 1)					-3,38**	-0,82	-0,27	-1,14
Antall kjøretimer						0,15***	0,13***	0,11**
Øvelseskjøring							0,46	0,19
Sensorerfaring								-0,13*
Konstant	7,28***	6,62***	8,15***	10,09***	13,31***	7,28***	5,65**	7,63**
R^2	0,06	0,06	0,09	0,12	0,15	0,25	0,23	0,25

[1] Drøbak/Mysen er referansekategori
*** signifikant på 1% nivå, ** signifikant på 5 % nivå, * signifikant på 10% nivå.

Kilde: TØI rapport 662/2003

For det andre ser vi at kjønn på sensor har signifikant betydning i alle modellene. Mannlige sensorer registrerer signifikant færre feil enn kvinnelige sensorer ved landeveiskjøring. Noe av kjønnseffekten henger sammen med at mannlige sensorer gjennomgående har mer erfaring enn kvinnelige sensorer (hhv.10 og 5 års erfaring); effekten av sensors kjønn svekkes når sensorerfaring legges inn i modellen. Vi ser igjen at dette fører til at det blir en signifikant positiv effekt av sensorsted Fredrikstad. Vi må ta de samme forbeholdene her som foran når det gjelder disse effektene.

Tabell 3.9 viser tilsvarende modeller som tabell 3.7 og 3.8, men med vektet antall feil i tettbebyggelse/by som avhengig variabel. Resultatene viser i stor grad det samme mønsteret som tabell 3.7 og 3.8. Til forskjell fra tabell 3.8 med feil på landevei/motorvei som avhengig variabel, får vi ingen signifikante effekter av kandidatsted. Det betyr at Oslo-kandidatene, som hadde signifikant flere feil enn de andre under landeveiskjøring, er relativt sett flinkere under bykjøring. Vi ser også at kandidatene fra Fredrikstad har færre feil enn kandidatene fra Mysen/Drøbak. Disse resultatene er ikke overraskende i og med at kandidatene fra Fredrikstad (og Oslo) antakelig har mer trening i bytrafikk enn kandidatene fra Mysen/Drøbak.

Forskjellen mellom kandidatstedene er imidlertid ikke lenger signifikant når vi kontrollerer for antall kjøretimer. Igjen bidrar antall kjøretimer signifikant, og effekten er i litt sterkere enn når feil på landevei/motorvei ble benyttet som avhengig variabel.

Tabell 3.9 Multivariat regresjon. Avhengig variabel = vektet antall feil i tettbebyggelse/by. Trinnvis prosedyre. B-koeffisienter.

	Modeller							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Kandidatsted [1]								
Oslo	1,20	1,21	1,11	-0,68	-1,22	0,31	-0,50	-0.52
Fredrikstad	-2,63	-2,59	-2,55	-3,04*	-2,94*	-1,66	-2,05	-1,23
Sensorsted [1]								
Oslo		0,15	-0,26	0,19	0,13	-0,30	-0,46	-0,39
Fredrikstad		0,46	0,47	0,59	0,53	0,24	0,35	2,86
Sensor (mann= 1)			-1,64	-1,78	-1,71	-2,05	-1,77	-1,59
Alder (18=1, 19+=0)				-4,36***	-4,33***	-3,20*	-2,40	-2,77
Landbakgrunn (norsk = 1)					-2,78	0,26	-2,67	-2,86
Antall kjøretimer						0,19**	0,14	0,19*
Øvelseskjøring							0,64	0,04
Sensorerfaring								-0,25**
Konstant	14,80***	14,58***	15,82***	19,54***	22,19***	15,24***	17,22***	18,52***
R^2	0,03	0,03	0,04	0,08	0,08	0,11	0,11	0,15

[1] Drøbak/Mysen er referansekategori
*** signifikant på 1% nivå, ** signifikant på 5 % nivå, * signifikant på 10% nivå.

Kilde: TØI rapport 662/2003

Når effekten av kandidatsted reduseres når vi kontrollerer for antall kjøretimer, innebærer det sannsynligvis at kandidatene fra Fredrikstad har hatt bedre forutsetninger og/eller ferdigheter i utgangspunktet, før de startet med kjøring på kjøreskole. Vi ser at variabelen sensorerfaring igjen bidrar signifikant (p=0,044) og bidrar til å styrke effekten av sensorsted Fredrikstad slik som i tabell 3.7 og 3.8.

3.4 Diskusjon

Vi finner systematiske forskjeller mellom kandidatene, og til en viss grad støtte for at kandidatenes ferdigheter varierer med hvor de kommer fra. Vi finner imidlertid ikke at kandidatene fra Oslo er flinkere enn de andre, snarere tvert om. I dette materialet har kandidatene fra Mysen/Drøbak og Fredrikstad signifikant færre feil ved landeveis- og motorveiskjøring enn Oslo-kandidatene. Kandidatene fra Fredrikstad er flinkere enn kandidatene fra Mysen/Drøbak ved bykjøring.

De viktigste årsakene til at Oslo-kandidatene har dårligere resultater en de andre er det er flere eldre kandidater fra Oslo, og flere kandidater med innvandrerbakgrunn. Begge deler øker antall feil, og dermed også sjansen for å stryke. At landbakgrunn har såpass klar betydning er interessant. Andelen ungdom med utenlandsk bakgrunn har økt betydelig de senere år, og dette kan være en av grunnene til at også strykprosenten har økt de senere år.

Den eneste variabelen som slår signifikant ut i alle analysene er antall kjøretimer på kjøreskole. Det ser ut til at jo flere kjøretimer kandidaten har, jo flere feil har han og jo høyere strykprosent. Grunnen til at antall kjøretimer så entydig er assosiert med flere feil og høyere strykprosent er at de med mange timer gjennomgående har dårligere ferdigheter (mindre trening) og at de kjører opp tidligere enn de strengt tatt burde. Den enkelte elev kan selv bestemme når han/hun vil gå opp til prøven, og en del fristes antakelig til å gå opp før de har de nødvendige ferdighetene. Sagberg (2002) finner at de med flest kjøretimer har høyest risiko for uhell etter avlagt førerprøve, noe som peker i samme retning.

Vi har funnet nokså store forskjeller i hvordan to sensorer vurderer en og samme kandidat. Sensorene er uenige i om kandidaten skal bestå eller stryke i 28 prosent av tilfellene, og det er også relativt store avstander mellom indeksverdiene til to sensorer for samme kandidat.

Analysene viser at det er et visst mønster som avtegnes når det gjelder å forklare disse forskjellene. For det første er det nær signifikante tendenser til samspill mellom kjønn på sensor og kjønn på kandidat. Mannlige kandidater består førerprøven i større grad med mannlig sensor enn med kvinnelig sensor, og den motsatte tendensen finner vi for kvinnelige kandidater. Generelt finner vi også at mannlige sensorer ikke registrerer like mange feil under prøven som det kvinnelige sensorer gjør. Noe av denne forskjellen kommer av erfaring; de mest erfarne sensorene registrerer færre feil. Samtidig har mannlige sensorer i gjennomsnitt dobbelt så lang erfaring som kvinnelige sensorer i denne undersøkelsen. Vi finner imidlertid ikke at disse tendensene til å registrere færre feil blant mannlige sensorer og erfarne sensorer fører til at de lettere lar kandidatene bestå prøven. Det er riktignok en svak tendens i den retning, men den er ikke signifikant. Det kan bety at mannlige sensorer sensorer/erfarne sensorer ikke noterer så mye underveis og bruker mer skjønn i vurderingen av kandidatene.

Vi har imidlertid funnet en viss tendens til at erfaring samvarierer med resultatene på førerprøven blant kvinnelige sensorer. Der kandidater har strøket med en kvinnelig sensor har sensor gjennomgående hatt mindre erfaring enn i de tilfellene der kandidater har bestått med en kvinnelig sensor. Og denne forskjellen i erfaring som vi bare finner blant kvinnelige sensorer er nær signifikant.

Gjennomgående er som nevnt de kvinnelige sensorene mindre erfarne enn de mannlige sensorene i denne undersøkelsen. Vi finner også som nevnt en nokså systematisk tendens til at kvinnelige kandidater i større grad står med en kvinnelig sensor, og mannlige kandidater i større grad står med en mannlig sensor. Vi vet fra tidligere undersøkelser (Sagberg 2002) at mannlige kandidater som regel har mer kjøreerfaring før de kjører opp enn det kvinnelige kandidater har. Det kan følgelig tenkes at det er en forskjell mellom sensorene knyttet både til kjønn og til erfaring når det gjelder hva slags kjøreatferd de aksepterer/premierer. Tendensene vi finner kan tyde på at erfarne/mannlige sensorer aksepterer/premierer en

kjørestil som er mer i overensstemmelse med ”normal” kjøreatferd i trafikken og som de mannlige kandidatene tjener på, mens uerfarne/kvinnelige sensorer i større grad premierer kjøreatferd som i større grad er i overensstemmelse med regelverket og som de kvinnelige kandidatene tjener på.

Det er viktig å understreke at våre funn bare vise tendenser som kan samsvare med en slik tolkning. Om det faktisk er slik kan vi ikke gi et godt svar på med de data vi har tilgjengelig her, så det må eventuelt nye undersøkelser besvare.

4 Konklusjon

Resultatene viser store forskjeller i kandidatenes resultater til førerprøven. Det er en viss tendens til at kandidater fra Oslo har relativt flere feil ved landeveiskjøring og at kandidater fra Mysen/Drøbak har flere feil ved bykjøring. Kandidatenes alder, landbakgrunn og antall kjøretimer på trafikkskole er i tillegg variabler som har stor betydning både for antall feil og for sannsynligheten for å stå eller stryke.

Vi finner temmelig store forskjeller i sensorenes vurderinger av en og samme kandidat. I mer enn hver fjerde førerprøve er de to sensorene uenige i om kandidaten skal stå eller stryke, og i gjennomsnitt registrerer den ene sensoren 10 mindre feil mer enn den andre. Med såpass store forskjeller i sensorenes vurderinger, er det rimelig å konkludere med at førerprøven *ikke* er tilstrekkelig reliabel slik den gjennomføres i dag.

Forskjellene i vurderinger synes imidlertid ikke å være knyttet til om sensor kommer fra en trafikkstasjon i storby eller på landet. I den grad det er mønster i disse forskjellene finner vi at både kjønn og erfaring har betydning. Det er en temmelig klar tendens til at mannlige sensorer og sensorer med lang erfaring registrerer færre feil enn kvinnelige sensorer og sensorer med kortere erfaring. Det ser imidlertid ikke ut til at dette fører til at mannlige/erfarne sensorer lar flere bestå prøven.

Vi har imidlertid funnet en påfallende tendens til at mannlige kandidater lettere består med mannlig sensor, og kvinnelige kandidater lettere består med kvinnelig sensor. Hva grunnen til dette er vet vi ikke. En mulig forklaring er at de kvinnelige sensorene gjennomgående er mindre erfarne, og at de premierer en kjøreatferd som er mer korrekt i henhold til regelverket enn det mannlige/erfarne sensorer gjør.

Vi har også funnet tendenser til at sensorene vurderer utenlandske kandidater mer forskjellig enn norske kandidater. Vi vet ikke hva forklaringen på dette er, men det kan tenkes at det for norske sensorer er vanskeligere å tolke utenlandske kandidaters atferd, men at dette varierer mellom sensorene.

Analysene peker i retning av flere mulige forklaringsfaktorer bak de relativt store variasjonene i sensorenes vurderinger. Dessverre har ikke dette datamaterialet vært tilstrekkelig til å konkludere om de nevnte hypotesene stemmer, men det vil forhåpentligvis kunne bli gjort i senere undersøkelser.

5 Referanser

Murphy K. R. og Davidshofer C. O. (1998): *Psychological Testing. Principles and Applications.* USA, New Jersey, Prentice Hall, New Jersey, USA.

Raaheim A. (2000): En studie av inter-bedømmer reliabilitet ved eksamen på psykologi grunnfag. *Tidsskrift for Norsk Psykologiforening* 37, 203-213.

Rismark M., Norberg P. F., Stenøien J. M. og Sitter S. (2003): *Kulturkollisjoner bak rattet*. Trondheim, VOX.

Sagberg F. (2002): *Mengdetrening, kjøreerfaring og ulykkesrisiko*. Oslo, Transportøkonomisk institutt, TØI rapport 566/2002.