NORGES HANDELSHØYSKOLE

Bergen, Våren 2013

Motivasjon og insentiver

Betydningen av oppgavens karakter for effekten av insentiver

Mastergradsutredning innenfor

Økonomisk Styring

av

Ole Johan Langenes

Veileder: Alexander Wright Cappelen

NORGES HANDELSHØYSKOLE

Sammendrag

Denne utredningen tar for seg betydningen av en oppgaves karakter for effekten av insentiver. Formålet er å undersøke om motivasjon for en oppgave kan ha betydning for hvordan insentivering fungerer. I teoridelen presenteres den klassiske Prinsipal-Agent-modellen, som stammer fra standard økonomisk teori. I standardmodellen snakker man kun om ytre motivasjon. I påfølgende del blir derfor begrepene *indre motivasjon* og *crowding-out-effekten* presentert. Med det som grunnlag blir det utarbeidet en utvidet Prinsipal-Agent-modell som inkluderer indre motivasjon og som viser hvordan crowding-out-effekten kan virke inn ved insentivering.

Ulike oppgaver gir forskjellig grad av indre motivasjon. Med dette til grunn presenteres et eksperiment som er gjennomført med 282 respondenter online gjennom Amazon Mechanical Turk (MTurk). I eksperimentet blir respondentene randomisert i fire treatments. Treatment 1 og 2 inneholder en IQ-test, mens treatment 3 og 4 består av finn-tall-oppgaver. Betalingsstrukturen er eneste forskjell mellom treatmentene som har samme oppgavetype. Mens man i T1 og T3 har en fast betaling på 3 USD for å løse oppgavene, får man i T2 og T4 i tillegg en bonus på 2 cent per korrekt svar. Med alle spørsmål korrekt besvart, vil bonusen summere seg til 30 cent eller 10 % av fastlønnen. Det er således en lav insentivlønn som introduseres.

Hovedhypotesen i utredningen er at insentiver gir økt innsats ved finn-tall-oppgavene, mens innsatsen for IQ-testen forblir uendret. Fra hypotesetesting og individuelle regresjoner finner man isolert sett signifikante effekter som støtter dette, mens man fra en samlet regresjonsanalyse ikke får tilsvarende signifikante resultat. Hovedfunnet er derfor at en ikke kan påstå at det er signifikant forskjell i påvirkningen fra insentivene. Resultatene viser likevel at det finnes en tydelig effekt i retning av at insentiver fungerer dårligere når respondentene har høy indre motivasjon for oppgaven. Det vil derfor være hensiktsmessig å ta hensyn til oppgavetypen ved utarbeidelse av bonussystem.

Forord

Allerede tidlig på bachelorgraden min ved NHH fattet jeg interesse for sammenhengen mellom bedriftsøkonomiske og strategiske fag. Våren 2011 tok jeg kurset *Økonomi og psykologi*, holdt av Alexander W. Cappelen og Bertil Tungodden. Sammen gjorde de kurset til et av de mest engasjerende jeg har tatt ved NHH. Gjennom kurset fikk man blant annet lære om eksperimentell metode gjennom forelesninger, artikler og en gruppebasert kursoppgave. På mastergraden ved NHH har jeg siden hentet mye av inspirasjonen til temaet motivasjon og insentiver fra kurs som *Styring av større foretak*, *Personalpolitikk og insentiver* og *Økonomisk organisasjonsteori*.

Masterutredningen er til min viten den første ved NHH som baserer seg på eksperimentell metode gjennom bruk av både Amazon Mechanical Turk (MTurk) og Qualtrics. Dette gjør den på mange måter litt annerledes enn andre eksperimentelle utredninger. Deltakerne er ikke studenter på et spesifikt studiested eller ansatte i en spesifikk bedrift, men en sammensetning av personer i alle aldre primært fra USA og India. Med hele 264 godkjente besvarelser er også datamaterialet robust. Selv om det gjerne ikke kommer direkte frem av utredningen, ligger det mye forarbeid bak å sette sammen et slikt eksperiment og få gjennomføringen til å gå knirkefritt. Mye tid har gått med til å lese seg opp på MTurk gjennom ulike artikler og tidligere eksperimenter. Til gjengjeld har prosessen vært både spennende og interessant – det har vært gøy å forklare andre hva jeg har arbeidet med.

Jeg vil takke veilederen min, Alexander W. Cappelen, for en god tone og presise tilbakemeldinger. Etter hvert veiledningsmøte har jeg returnert med enda mer engasjement for utredningen. Jeg vil også takke for at han gjorde gjennomføringen mulig gjennom økonomisk støtte. Videre vil jeg takke PhD-student Bjørn-Atle Reme for et hyggelig møte med utveksling av idèer. Til slutt vil jeg takke Kaia og Kjersti for gjennomlesning av utredningen i en hektisk periode.

Bergen, 18. juni 2013

Ole Johan Langenes

Innholdsfortegnelse

SAMMENDRAG ... **3**
FORORD ... **5**
FIGURLISTE ... **8**
TABELLISTE ... **9**
1. INNLEDNING ... **10**
1.1 BAKGRUNN OG PROBLEMSTILLING ... 10
1.2 GJENNOMFØRING ... 11
1.3 HOVEDFUNN ... 11
2. TEORI ... **12**
2.1 AVGRENSNING ... 12
2.2 STANDARD ØKONOMISK TEORI ... 12
2.3 DEN KLASSISKE PRINSIPAL-AGENT-MODELLEN ... 12
2.4 INDRE MOTIVASJON ... 14
2.5 CROWDING-OUT-EFFEKTEN ... 17
2.5.1 Priseffekten og crowding-out-effekten ... 17
2.5.2 Økonomi og psykologi ... 18
2.5.3 Skjulte kostnader ... 19
2.6 UTVIDET PRINSIPAL-AGENT-MODELL ... 21
2.7 FELTEKSPERIMENTER ... 23
3. METODE ... **25**
3.1 EKSPERIMENTELL METODE ... 25
3.1.1 Amazon Mechanical Turk ... 27
3.1.2 Qualtrics ... 28
3.2 DESIGN ... 29
3.2.1 Oppgavetyper ... 30
3.2.2 Belønningsstruktur ... 31
3.2.3 Spørsmål til respondentene ... 32
3.2.4 Muligheter ... 33
3.3 STYRKER OG SVAKHETER VED MTURK ... 34
3.4 GJENNOMFØRING ... 38
3.5 HOVEDHYPOTESE ... 40

4. RESULTATER 42
4.1 DESKRIPTIV ANALYSE 42
4.1.1 Rekruttering, dropouts og utbetaling 42
4.1.2 Kontrollspørsmål 43
4.1.3 Randomisering 43
4.1.4 Score 44
4.1.5 Fordeling av tidsbruk 46
4.1.6 Normalitetstest 48
4.1.7 Analyse 48
4.2 REGRESJONSANALYSER 50
4.2.1 Regresjonsmodell for IQ-oppgaver 52
4.2.2 Regresjonsmodell for finn-tall-oppgaver 53
4.2.3 Regresjonsmodell for hele datasettet 54
4.2.4 Interaksjonseffekter 57
5. DISKUSJON 60
6. AVSLUTNING OG KONKLUSJON 63
APPENDIKS 65
1. Hvordan oppgavene ble oppfattet 65
2. Tidspunkt for dropouts 67
3. Kjønnsfordeling 68
4. Aldersfordeling 68
5. Etnisitet og bosted 69
6. Utdanning 70
7. Inntekt 71
8. Fordelingen av score 72
9. Residualplott og Cook-avstander 73
10. Regresjonsmodeller for interaksjonseffekter 77
11. Fullstendig oversikt over eksperimentets innhold 79
LITTERATURLISTE 91

Figurliste

Figur 1: Jobbkarakteristika-modellen 16
Figur 2: Design av eksperimentet..... 30
Figur 3: Påstander respondentene måtte ta stilling til 32
Figur 4: Beskrivelse av oppgaven som arbeiderne fikk presentert i MTurk..... 39
Figur 5: Gjennomsnittlig score i hver treatment.. 45
Figur 6: Gjennomsnittlig tidsbruk i sekund for hver treatment..... 46
Figur 7: Fordelingen av tidsbruk for hver treatment. 47
Figur 8: Fordelingen og normalitetsplott av tidsbruk..... 48
Figur 9: Fordeling av hvordan oppgavene ble oppfattet 66
Figur 10: Kjønnsfordeling totalt..... 68
Figur 11: Aldersfordeling - originale og reviderte grupper..... 69
Figur 12: Etnisitet..... 69
Figur 13: Bosted..... 70
Figur 14: Utdanningsgrad..... 70
Figur 15: Inntektsfordelingen i USA og India 71
Figur 16: Fordelingen av score for hver treatment..... 72
Figur 17: Residualplott, Cook-avstander og normalitetsplott for regresjon IQ-oppgaver 73
Figur 18: Residualplott, Cook-avstander og normalitetsplott regresjon finn-tall-oppgaver .. 74
Figur 19: Residualplott, Cook-avstander og normalitetsplott for regresjonsmodell 1..... 75
Figur 20: Residualplott, Cook-avstander og normalitetsplott for regresjonsmodell 2..... 76

Tabelliste

Tabell 1: Fordeling av fullførte besvarelser i datasettet basert på kontrollspørsmål..... 43
Tabell 2: Fordelingen i hver treatment for spesifikke egenskaper ved respondentene. 44
Tabell 3: Gjennomsnittstid, 95% konfidensintervall og standardavvik for treatments..... 47
Tabell 4: Regresjonsmodell for respondenter som løste IQ-oppgaver..... 52
Tabell 5: Regresjonsmodell for respondenter som løste finn-tall-oppgaver 53
Tabell 6: Regresjonsmodeller for alle respondenter..... 55
Tabell 7: Regresjonsmodell som viser interaksjonseffekten «KvinneInsentiv». 57
Tabell 8: Regresjonsmodell som viser interaksjonseffekten «LavinntektInsentiv»..... 58
Tabell 9: Besvarelsenes snittverdier fra påstander oppdelt for oppgavetype..... 60
Tabell 10: Oversikt over dropouts..... 67
Tabell 11: Kjønnsfordeling i hver treatment 68
Tabell 12: Inntektsfordeling..... 71
Tabell 13: Regresjonsmodell for interaksjon mellom kjønn og «IQInsentiv». 77
Tabell 14: Regresjonsmodell for interaksjon mellom inntektsnivå og «IQInsentiv». 78

1. Innledning

1.1 Bakgrunn og problemstilling

Økonomistyring er et bredt tema som skal si noe om «*hvordan verdiskapingen i en virksomhet kan måles, analyseres og styres på en best mulig måte*» (Norges Handelshøyskole, 2013). I den sammenheng er det er viktig å forstå hvem menneskene man arbeider med er, og hvordan adferden deres kan forklares eller forespeiles. Det reiser seg da noen interessante spørsmål: Virker det fornuftig at man i dagens arbeidsmarked bruker mer variabel lønn? Hva har oppgavens karakter å si? Hvordan kan den ansattes motivasjon ha noe å si? Hvordan bør man optimalt sett betale den ansatte?

Den tradisjonelle økonomiske antagelsen er at personer handler rasjonelt og profittmaksimerende. De siste årene har imidlertid adferdsøkonomi fått stort gjennomslag. Fagfeltet, som ser på forholdet mellom psykologi og økonomi, viser at mennesker ofte har andre og mer komplekse motiver bak handlingene sine (Cappelen og Tungodden, 2012). For eksempel betyr dette at insentiver, i form av en pengemessig bonus, kan ha motstridende effekter i forhold til standardteoriens prediksjon. Det har også blitt gjennomført en del forskning på temaet og det er i senere tid publisert flere vitenskapelige artikler. Et viktig utgangspunkt for arbeidet med denne utredningen var artikkelen til Gneezy og Rustichini (2000a), som undersøker forholdene mellom økonomi og psykologi nærmere. Fagfeltet er meget interessant, og i tiden fremover vil det trolig bli enda mer aktuelt for vårt kunnskapsbaserte samfunn.

Målet med et belønningssystem bør være at det skal komme både ansatte og bedriften til gode. Det er i begge parters interesse at belønningssystemet er hensiktsmessig utformet. Utredningen vil derfor undersøke hvordan den ansattes motivasjon for oppgaven kan ha betydning for om, og eventuelt hvordan, vedkommende bør belønnes. Dette gjøres gjennom å studere hvordan insentiver påvirker den ansatte avhengig av oppgavetypen. Utredningens problemstilling er derfor:

Hva har oppgavens karakter å si for effekten av prestasjonslønn?

Gjennom denne problemstillingen ønsker jeg å sette standard økonomisk teori, som en økonom typisk lærer, opp mot alternative antagelser. Målet er å finne resultater som kan

utdype eller utfordre standardteorien om belønning. Sagt på en annen måte utfordrer dermed utredningen tanken om at mennesker enkelt kan gis en «*gulrot*» for å øke innsatsen på jobben.

1.2 Gjennomføring

Utredningen baserer seg på et relativt stort, kvantitativt datamateriale som ble hentet inn over internett gjennom en kombinasjon av Amazon Mechanical Turk og programvaren Qualtrics. Denne internettbaserte metoden er relativt ny og gir et godt representativt utvalg av befolkningen. Samtidig innebærer et beskjedent lønnsnivå lave gjennomføringskostnader. Man har derfor mulighet til å inkludere et stort antall respondenter på et relativt lavt budsjett. For å kunne vise årsakssammenhenger, eller kausalitet, ble respondentene randomisert i fire grupper eller treatments. Randomiseringen sikrer at gruppene kan ses på som identiske der eventuelle forskjeller vil komme av den forskjellige behandlingen de får. For å belyse problemstillingen, skilte jeg mellom to oppgavetyper og to belønningssystemer slik at jeg fikk et 2x2-design. Ved å variere mellom fastlønn og prestasjonslønn kunne jeg dermed se på effektene dette ga innenfor hver oppgavetype. Avslutningsvis stilte jeg både personlige spørsmål og spørsmål knyttet mer direkte til eksperimentet. Dette ble gjort for å kontrollere randomiseringen og for å kunne finne eventuelle interessante sammenhenger til analysen.

1.3 Hovedfunn

Fra hypotesetestingen og de isolerte regresjonene finner jeg at insentiver ikke påvirker innsats når oppgavene preges av høy indre motivasjon, mens insentivene gjør at innsatsen øker signifikant når oppgavene preges av lav indre motivasjon. Dette resultatet er i tråd med den nyere teorien. Gjennom to regresjonsmodeller, som inkluderer hele datasettet, finner jeg at forskjellen mellom de to funnene fra hypotesetestingen ikke er signifikant. Dette ser jeg ved hjelp av et interaksjonsledd. Hovedfunnet er dermed at man ut fra eksperimentet ikke kan påstå at det er signifikant forskjell i påvirkningen fra insentivene. I begge modellene gir bonus økt innsats uavhengig av oppgavetype. Resultatene tilsier likevel at det finnes en tydelig effekt i retning av at insentiver fungerer dårligere når respondentene har høy indre motivasjon for oppgaven.

2. Teori

2.1 Avgrensning

Det finnes mye litteratur og forskning som beskriver belønning og motivasjon. Et tema som utmerker seg i nyere forskning, er skillet mellom ytre og indre motivasjon, samt relasjonen til oppgavetype og insentiver. Med insentiver menes her et belønningssystem som har til hensikt å få den ansatte til å yte en viss innsats på en eller flere oppgaver. Presentasjonen av teori vil ta utgangspunkt i standard økonomisk teori og den klassiske Prinsipal-Agent-modellen. Deretter vil begrepene *indre motivasjon* og *crowding-out-effekten* gjennomgås. Mot slutten av teorigjennomgangen vil det bli presentert en utvidet prinsipal-agent-modell som inkluderer indre motivasjon.

2.2 Standard økonomisk teori

Standard økonomisk teori antar at ansatte opptrer rasjonelt og kun er motivert av sin økonomiske egeninteresse. Ansatte må derfor motiveres gjennom penger eller goder for å yte innsats. Dette fenomenet kalles gjerne *ytre motivasjon* og er motivasjon som kommer av ytre faktorer som lønn, bonuser og frynsegoder. Også det å lokke med forfremmelser kan være en form for ytre motivasjon. Ryan og Deci (2000) beskriver det som å gjøre noe «*fordi det fører til et separerbart utfall*». I et slikt perspektiv er jobben i seg selv ikke en motiverende faktor – grunnen til at ansatte jobber er motivasjonen som kommer av belønningen. Uavhengig av hvilken type oppgave en ansatt står overfor, sier standard økonomisk teori at innsatsen er lav, dersom vedkommende ikke får ekstra belønning basert på høyere innsats. Med denne tankegangen til grunn har man utviklet mange belønningssystemer der fokuset er å få ansatte til å jobbe godt gjennom ytre motivasjon. Et utgangspunkt for belønningssystemene etter standard økonomisk teori finner man i den klassiske Prinsipal-Agent-modellen.

2.3 Den klassiske Prinsipal-Agent-modellen

Prinsipal-Agent-modellen, eller PA-modellen, er et kjent redskap innenfor økonomi for å forklare blant annet et arbeidsforhold. Prinsipalen vil i den sammenheng være arbeidsgiver, mens agenten er arbeidstaker. PA-modellen er slik at en eller flere agenter handler på

prinsipalens vegne (Ross, 1973), slik at det agentene gjør har mye å si for prinsipalen. Ofte snakker man om en *interessekonflikt* mellom prinsipalen og agenten. Arbeidsgiver vil typisk at arbeidstaker skal løse sine oppgaver best mulig for bedriften med høyest mulig innsats til lavest mulig pris. Den ansatte er «lat» og har en personlig kostnad eller et ubehag ved å yte innsats og vil derfor ønske høyest mulig lønn samtidig som man yter lavest mulig innsats. En slik tankegang blir representert ved at den ansatte maksimerer sin nyttefunksjon.

I sammenheng med PA-modellen snakker man ofte om to sentrale problemer. Disse kalles *ugunstig utvalg* (adverse selection) og *moralsk hasard* (moral hazard). Ugunstig utvalg-problemet går ut på at prinsipalen ikke kan være sikker på hvilken person arbeidssøkeren faktisk er. Vedkommende har skjult informasjon om sine karakteristika før ansettelsen. Dersom man ikke tar hensyn til denne problematikken, vil man kunne tiltrekke seg uønskede arbeidere. Moralsk hasard-problemet går ut på at det er vanskelig for prinsipalen å kunne si hvor høy innsats den ansatte faktisk har lagt ned i arbeidet. Dette fordi man kun observerer resultatene og ikke innsatsen direkte. Agenten har med andre ord privat informasjon om egne handlinger under ansettelsesforholdet.

På grunn av asymmetrisk informasjon og interessemotsetninger får man agentkostnader som består av overvåkningskostnader og kostnader grunnet for lav innsats fra ansatte. Direkte overvåkning av de ansatte vil være kostbart, i tillegg til at det i seg selv kan virke demotiverende. En fullstendig kontrakt mellom prinsipal og agent er dermed ikke et reelt alternativ. Løsningen PA-modellen tilbyr, er å utvikle et optimalt insentivsystem som gjør behovet for overvåkning mindre, samtidig som den ansatte yter ønsket innsatsnivå. Agentkostnadene blir på den måten redusert.

I den klassiske lineære PA-modellen, der man kun konsentrerer seg om en oppgave med innsats e, kan agentens nytte skrives som en funksjon av fastlønn, variabel lønn og kostnad ved å yte innsats. Fastlønnen kan betegnes α. Den variable inntekten, kalt insentiver, består gjerne av en pengemessig bonus β per enhet multiplisert med et prestasjonsmål z(e), mens agentens kostnad ved å yte innsats kan skrives som c(e). Nytten av innsats kan dermed skrives:

$$\mathbf{U(e) = \alpha + \beta * z(e) - c(e)}$$

Det første og siste leddet er alltid tilstede, mens den variable inntekten kun er tilstede ved insentivering. Agenten, eller den ansatte, vil maksimere sin forventede nytte av jobben man

står overfor – det er altså nyttefunksjonen som bestemmer innsatsen man vil yte. Prinsipalen vil at innsatsen skal være høy, men må samtidig foreta en avveining med tanke på lønnskostnadene forbundet med det. Agentens innsats vil øke med økt fastlønn og eventuelt økt insentivlønn som følge av høyere nytte. Jo høyere ubehag eller kostnad ved innsats agenten har, jo lavere nytte vil man få ved å ta jobben. Øker kostnadsleddet, vil derfor innsatsen isolert sett reduseres.

Fra modellen ser man at agentens innsats blir sett på som en funksjon av pengemessig inntekt og personlig ubehag i form av kostnad ved innsatsen man yter. Modellen ser på ansatte som individer der ytre motivasjon har alt å si for utfallet. Det er lett å forstå at en slik forenklet tenkemåte kan være problematisk i den virkelige verden.

2.4 Indre motivasjon

Indre motivasjon er motivasjon som kommer av indre faktorer som interesse for jobben, en følelse av å utrette noe og en tilfredshet i det å gjøre jobben. I nyere tid har betydningen av dette blitt stadig mer vektlagt. Ryan og Deci (2000) beskriver indre motivasjon som å gjøre noe fordi det er «*i seg selv interessant eller morsomt*». Folk flest vil ønske å løse oppgaver man synes er interessant. Sannsynligheten er derfor stor for at de fleste ansatte har en indre motivasjon i seg for jobben de har tatt. Indre motivasjon er personlig og varierer mellom ansatte. Personer har gjerne høy indre motivasjon for visse oppgaver, mens de har lav indre motivasjon for andre. Siden variasjonen kan være stor, er det ofte dumt å ta ansatte for gitt og gi alle samme belønningssystem. En alternativ løsning er å undersøke motivasjonen til den enkelte for oppgaven vedkommende står overfor, for deretter å tilpasse belønningsstrukturen.

Det finnes etter hvert mye litteratur som omtaler indre motivasjon i ulike former. En del av teoriene har sin rot i behovsteoriene. Disse sier at motivasjon kommer av at visse behov er dekket. For eksempel skriver Pink (2009) om i) autonomi, ii) mestring og iii) mening. Dersom disse tre faktorene er til stede, mener Pink at både prestasjoner og personlig tilfredsstillelse øker.

Autonomi går ut på at man får påvirke hvordan egne arbeidsoppgaver skal løses. Pink hevder at autonomi må være til stede dersom man vil ha engasjement for en jobb, og ikke bare en etterlevelse om å gjøre den. I en Youtube-video knyttet til boken (RSA Animate, 2010),

nevner Pink det australske programvareselskapet Atlassian som eksempel. Hvert kvartal gir selskapet de ansatte 24 timer til å gjøre hva de vil i arbeidstiden. Det eneste kravet som stilles er at man viser resultatet av arbeidet etterpå. Fremgangsmåten har, i følge Pink, resultert i både feilrettinger av eksisterende programvare og utarbeidelse av helt nye programmer. Standardteorien tilsier derimot at de ansatte skulle ytt minst mulig, for deretter å slappe av eller gjøre andre ting. Dette eksempelet viser at man ikke nødvendigvis trenger å gi provisjonslønn for innovasjon; autonomi alene kan gi gode resultater.

Mestring beskrives som «*trangen til å bli bedre i ting*». Tankegangen her er at det å bli bedre til å løse en oppgave, i seg selv er tilfredsstillende. Pink nevner Linux og Wikipedia, som begge er utviklet på fritiden av høyt utdannede mennesker. Både Linux og Wikipedia er gratis for brukerne. Det er dermed tydelig at mestring, og ikke penger, har vært motivet for utviklerne.

At arbeidet har *mening* er i følge Pink viktig for at det skal føles bedre for ansatte å komme på jobb. Samtidig sier han at en jobb med mening vil tiltrekke seg talent. Å gi arbeidet mening strider derfor ikke nødvendigvis med en profitabel drift; profittmaksimering og «*meningsmaksimering*» trenger ikke være ulike ting.

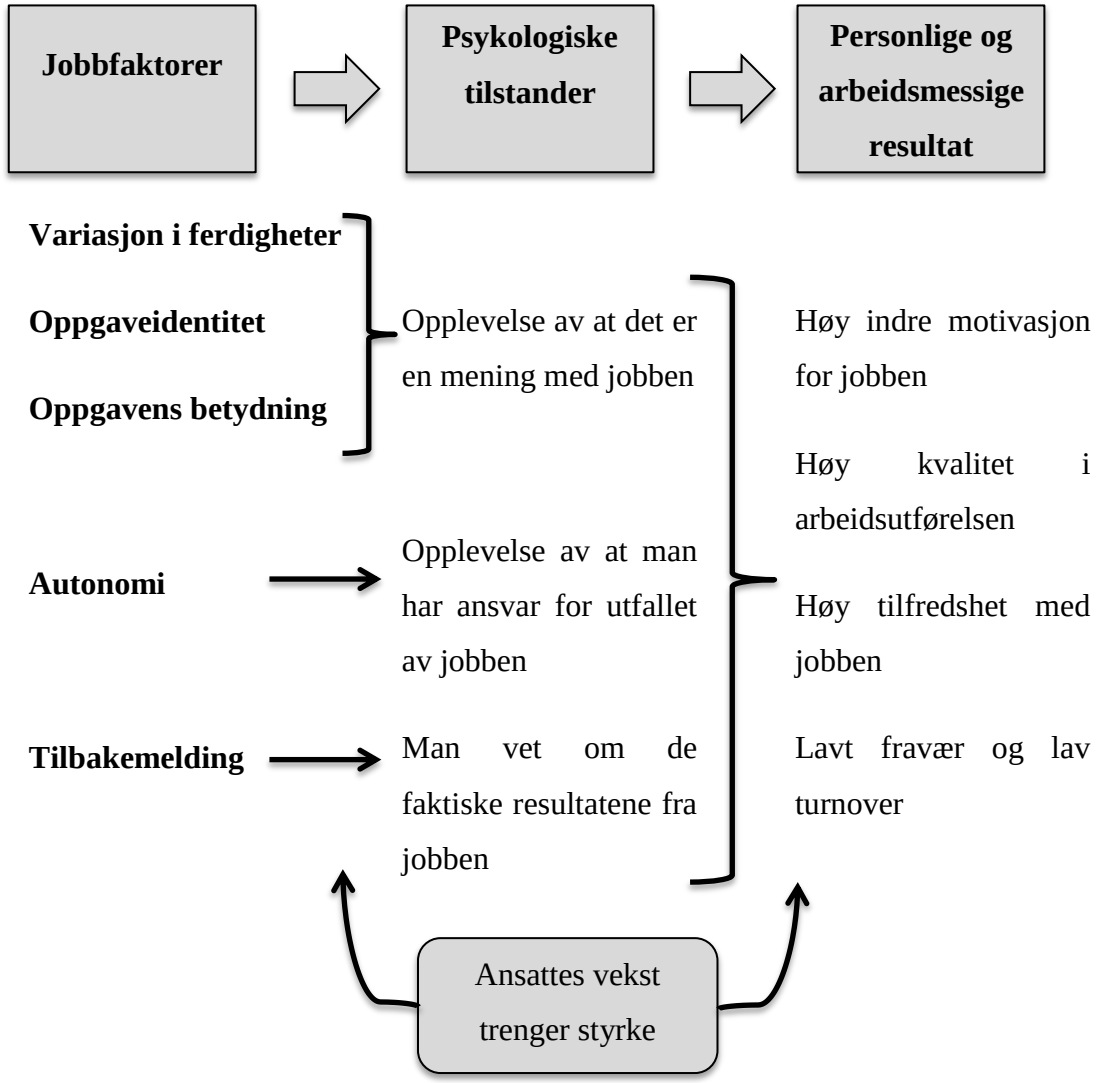
Ut fra dette ser man at innsatsen kan påvirkes uten å benytte den tradisjonelle tankegangen om å henge en «*gulrot*» foran nesen til den ansatte. Ved å heller ta hensyn til autonomi, mestring og mening i arbeidet, kan man enklere påvirke den indre motivasjonen, og dermed innsatsen, i positiv retning.

Jobbkarakteristika-modellen

Hackman og Oldham (1976) utviklet allerede på 70-tallet en modell for å vise hva som måtte til for å sikre størst mulig grad av indre motivasjon i en arbeidssituasjon. Denne blir gjerne referert til som *jobbkarakteristika-modellen*. Modellen skiller mellom fem faktorer ved en jobb. Disse fører via psykologiske tilstander til personlige og arbeidsmessige resultater.

For at ansatte skal oppleve en mening med jobben, mener Hackman og Oldham at tre jobbkarakteristika eller faktorer må være på plass. For det første må jobben være variert og kreve en rekke forskjellige ferdigheter og kunnskaper. På den måten får den ansatte brukt sitt talent. Det andre punktet går på at oppgavens identitet må være slik at den krever involvering i arbeidsprosessen fra begynnelse til slutt. Forfatterne viser også til at det bør være et synlig utfall, slik at en ansatt ser at resultatet kommer av innsats vedkommende har lagt ned. Den

tredje faktoren har med oppgavens betydning å gjøre, og i hvilken grad jobben har en innvirkning på andre, i eller utenfor bedriften. Dersom en ansatt ser at det han eller hun gjør betyr noe for andre, mener Hackman og Oldham at vedkommende lettere ser en mening med jobben.



Figur 1: Jobbkarakteristika-modellen til Hackman og Oldham (1976)

Den neste psykologiske tilstanden handler om at ansatte opplever et ansvar for utfallet av arbeidet. Her spiller autonomi en viktig rolle. Det går i følge forfatterne ut på i hvilken grad jobben gir «*frihet, uavhengighet og skjønn*» i planlegging av arbeidet og ved bestemmelsen av løsningsprosedyrer. Dette tilsvarer Pinks (2009) definisjon om at det er å påvirke hvordan egne arbeidsoppgaver skal løses. Dersom ansatte får en viss kontroll over egne arbeidsoppgaver, vil de kunne føle seg mer betydningsfulle. Dette kan føre til høyere indre motivasjon.

Det siste punktet går på i hvilken grad ansatte får en detaljert tilbakemelding om hvordan hans eller hennes utførelse av arbeidet er gjennomført. At man vet hvordan man ligger an, er viktig. Samtidig vil positive tilbakemeldinger ha en motiverende effekt.

Hackman og Oldham (1976) mener at dersom de fem jobbkarakteristikkene eller faktorene er oppfylt, vil man gjennom de psykologiske tilstandene kunne få høyere indre motivasjon, kvalitet og tilfredshet. Samtidig vil fraværet og turnover-raten kunne gå ned.

Med et kritisk blikk på jobbkarakteristika-modellen er det lett å se at individuelle forskjeller ikke er tatt hensyn til. Modellen antar at alle ansatte vil reagere likt og at alle vil ønske variasjon og autonomi. I virkeligheten er folk forskjellige. Mange vil sette pris på faktorene beskrevet i modellen, men det er også noen som ønsker faste rutiner uten for mye involvering. En arbeidsgiver må derfor være observant på at ansatte ikke nødvendigvis kan motiveres på samme måte og i tråd med modellen.

Det er likevel interessant at man allerede på 70-tallet hadde en modell som pekte spesielt på indre motivasjon. At en mening med jobben er viktig, har siden blitt bekreftet av for eksempel Ariely et al (2008) gjennom artikkelen «*Man's search for meaning: The case of Legos*». Gjennom to tilsvarende eksperimenter, der det ene handler om å bygge Lego, finner de at reservasjonslønnen er signifikant høyere i grupper som får en mindre meningsfull versjon av eksperimentene. Med andre ord krever deltakerne høyere lønn når oppgavene de gjør oppfattes som mindre meningsfylte. Dette tilsier at innsatsen vil være høyere for en meningsfylt oppgave enn en lite meningsfylt oppgave, gitt at lønnen er lik for de to.

2.5 Crowding-out-effekten

2.5.1 Priseffekten og crowding-out-effekten

Frey og Jegen (2001) diskuterer i sin artikkel to motstridende effekter: *priseffekten* og *crowding-out-effekten*. Priseffekten, eller *crowding-in-effekten* som den gjerne kalles, går ut på at prestasjonslønn forsterker motivasjonen man har for oppgaven, slik at resultatet forbedres. Dette er i utgangspunktet i tråd med standard økonomisk teori, men forfatterne skriver at den ansatte også må oppfatte belønningen som støttende for at standardmodellen skal stemme. En slik priseffekt er med andre ord den effekten insentiv har i følge den klassiske Prinsipal-Agent-modellen.

Crowding-out-effekten går derimot ut på at prestasjonslønn, som en ytre motivasjon, driver ut indre motivasjon for oppgaven. Sagt på en annen måte ødelegger den ytre motivasjonen for den indre motivasjonen. Insentivene gjør at ansatte ikke lenger føler at de yter innsats grunnet en indre motivasjon. I stedet får de en følelse av at det er den ytre motivasjonen, i form av bonusen, de jobber for. Dette har negativ påvirkning på ansatte med indre motivasjon for oppgaven.

I følge Frey og Jegen påvirker crowding-out-effekten to psykologiske prosesser negativt slik at man får: i) «*Svekket selvbestemmelsesrett*» og ii) «*Svekket selvtillit*». Førstnevnte går ut på at man føler at prestasjonslønnen er kontrollerende. Dermed mister man den indre motivasjonen, eller gløden, for å utføre arbeidet. Mottakeren av prestasjonslønnen føler at jobben ikke lenger blir gjort fordi man liker den, men fordi man får betalt for den. Svekket selvtillit går ut på at man føler at ens kompetanse ikke blir satt tilstrekkelig pris på. Frey og Jegen viser til at crowding-out-effekten er en av de viktigste anomaliene innenfor økonomi.

2.5.2 Økonomi og psykologi

Gneezy og Rustichini (2000a) skriver i sin artikkel, *«Pay Enough - Or Don't Pay At All»*, om motsetningene til de tradisjonelle antagelsene mellom økonomi og psykologi. Denne artikkelen er utgangspunktet for arbeidet med utredningen. Forfatterne viser først til standard økonomisk teori, som sier at pengemessige belønninger vil øke resultatene på en hvilken som helst oppgave. Videre viser de til at resultater er positivt korrelert med innsats og at «*innsats er ubehagelig og penger er bra»*. Ut fra denne tankegangen vil resultatene kun bli bedre ved å innføre økonomiske insentiver. Dette er i tråd med det man typisk har lært innenfor økonomi på høyskoler og universiteter. I psykologien mener man derimot at det motsatte er tilfellet, nemlig at økonomiske insentiver kan ødelegge for resultatene. Det er dette forholdet Gneezy og Rustichini undersøker nærmere gjennom artikkelen.

Artikkelen beskriver to eksperimenter: en IQ-test og et eksperiment med bøssebærere. I begge eksperimentene antar man at oppgavene er av en karakter som tilsier at deltakerne har en indre motivasjon om å gjøre det bra. Resultatene er meget interessante og til dels i strid med standard økonomisk teori.

Det første eksperimentet beskrevet i artikkelen ble gjort med 160 studenter fra universitetet i Haifa. Hver student ble bedt om å ta en IQ-test med 50 reelle oppgaver. Alle fikk beskjed om at de fikk 60 NIS (israelske shekel) for å delta. Respondentene ble tilfeldig delt opp i fire

grupper på 40 deltakere, der hver gruppe fikk ulik kompensasjon. Dermed hadde man et design med fire treatments: Alle grupper fremstod som like med unntak av hvordan man ble kompensert for besvarelsene. Den første gruppen fikk ingen betaling for besvarelsene, mens gruppe to fikk 0,1 NIS per rett svar. Gruppe tre og fire fikk henholdsvis 1 NIS og 3 NIS per korrekt svar. På tiden eksperimentet ble gjennomført, tilsvarte 3,5 NIS rundt 1 USD (amerikanske dollar). Betalingen på cirka 3 cent i gruppe to fremstod dermed som lav.

Resultatet var et snitt på 28,40 rette svar for gruppen uten betaling, mens gruppen med lav insentivbetaling på 0,1 NIS per rett svar hadde et signifikant fall ned til 23,08 rette i snitt. Med andre ord førte lav insentivbetaling til et dårligere resultat enn tilfellet uten betaling. For gruppen med middels og høy insentivlønn fikk man signifikant bedre resultat enn uten betaling – derav tittelen på artikkelen: Betal nok eller ikke betal i det hele tatt.

I eksperimentet med bøssebærerne brukte man 180 elever fra videregående skole og fant på tilsvarende måte at innsamlet beløp signifikant gikk ned når deltakeren selv fikk en liten andel (1 %) av innsamlet beløp. Her fant man ingen forbedring når bøssebæreren fikk en stor andel av innsamlet beløp (10 %).

Resultatene til Gneezy og Rustichini (2000a) er dermed i strid med standard økonomisk teori og viser at svake økonomiske insentiver kan virke motsatt av det som opprinnelig var meningen. Dermed viser forfatterne at crowding-out-effekten eksisterer. Etterhvert har crowding-out blitt et mye brukt uttrykk innenfor adferdsøkonomi.

2.5.3 Skjulte kostnader

Weibel et al (2009) gjennomførte en metastudie som inkluderte 46 studier og viser at: i) Motivasjonen er nøkkelen til effekten av prestasjonslønn og ii) Prestasjonslønn er generelt dyrere enn det ser ut til på grunn av «*skjulte kostnader*». Sistnevnte skyldes i stor grad crowding-out-effekten, og forfatterne viser til at indre motivasjon i stor grad kan bli skadet av prestasjonslønn. De påstår at oppgaver som typisk blir skadelidende er oppgaver som er utfordrende, morsomme og meningsfylte. I artikkelen settes det opp i alt fire hypoteser, hvorav en forkastes og tre støttes gjennom funnene fra studiene.

Hypotesen som tilsvarer standard økonomisk teori om at «*prestasjonslønn øker prestasjonene uavhengig av oppgavetype*», forkastes. De finner derimot støtte for en hypotese som sier at «*prestasjonslønn øker prestasjonene for mindre interessante oppgaver*

og reduserer prestasjonen i interessante oppgaver». Dette er et interessant funn som tilsier at prestasjonslønn kun bør brukes for mindre interessante oppgaver.

Videre finner Weibel et al støtte for at prestasjonslønn påvirker sammenhengen mellom indre motivasjon og prestasjon negativt. Det betyr at indre motivasjon reduseres når lønnen i stor grad avhenger av prestasjon. Dette er crowding-out-effekten. Tilsvarende finner man støtte for at sammenhengen mellom prestasjonslønn og ytre motivasjon påvirkes positivt gjennom prestasjonslønn. Det betyr at lønn som i stor grad avhenger av prestasjon øker den ytre motivasjonen av innsats, og er det som kalles priseffekten eller crowding-in-effekten som Frey og Jegen (2001) skriver om.

I boken «*Not just for the Money*» skriver også Frey (1997) om skjulte kostnader eller «*hidden cost of reward»*. Han beskriver disse som indirekte negative konsekvenser som følge av crowding-out-effekten. Effekten vil i følge ham være sterkere dersom man i utgangspunktet har personlige relasjoner og interessante oppgaver. For å unngå en for sterk crowding-out-effekt må man ta hensyn til anerkjennelse, samt klare å differensiere mellom personer med ulik motivasjon. Dersom priseffekten overvinner crowding-out-effekten vil insentivene fungere som standardteorien tilsier. Frey viser derfor til at det er viktig å kjenne personens motivasjon for oppgaven.

Frey skriver også om «*spill-over*»-effekter som går på at endringer i indre motivasjon også påvirker andre områder enn de som er direkte påvirket av insentivene. Dermed kan prestasjonslønn drive ut mer motivasjon enn man skulle tro. Dette gjør at man bør være ekstra på vakt.

I tillegg til svekket selvbestemmelsesrett og svekket selvtillit, nevner Frey også den svekkede muligheten til å uttrykke seg, som en viktig psykologisk prosess. Dette handler om at personen er fratatt muligheten til å vise sin indre motivasjon overfor andre personer. Resultatet blir at man gir avkall på den indre motivasjonen og oppfører seg kun i henhold til den ytre motivasjonen i form av prestasjonslønn.

Frey viser til at jo sterkere en ekstern påvirkning anerkjenner personens indre motivasjon, jo mer fostrer den indre motivasjon. Ut fra dette bør man ved oppgaver som i preges av høy indre motivasjon bygge på det i stedet for å lansere ekstern påvirkning som prestasjonslønn eller andre insentiver.

2.6 Utvidet Prinsipal-Agent-modell

Den klassiske PA-modellen tar ikke hensyn til en eventuell indre motivasjon og at ytre motivasjon kan ødelegge for denne gjennom crowding-out-effekten. Sett i sammenheng med den nyere litteraturen på området, er dette en klar svakhet ved standardmodellen. En løsning er å utvide den klassiske PA-modellen for å vise effekten av indre motivasjon. Dette kan enkelt gjøres ved å legge til et ledd kalt *m* på slutten av den eksisterende modellen. Dette leddet er en funksjon av bonusen β og oppgavetypen *T*, og representerer den indre motivasjonen agenten har for å yte innsats *e* på oppgaven. I modellen er det for enkelhets skyld antatt at motivasjonsleddet *m* multipliseres med innsatsnivå *e*. Nytten er dermed lineær i innsats *e*.

$$\mathbf{U(e) = \alpha + \beta * z(e) - c(e) + m(\beta,T) * e}$$

Agentens nytte av innsats avhenger som tidligere av både fastlønn, eventuell insentivlønn og personlig ubehag i form av en innsatskostnad. Forskjellen er at det i den utvidede modellen også tas hensyn til indre motivasjon. I modellen vil motivasjonsleddet *m* avhenge av både eventuelle insentiver og oppgavetypen i seg selv. Ved å se bort fra insentiver vil en oppgave *T* preget av høy indre motivasjon isolert sett øke *m*, mens en oppgave preget av lav indre motivasjon vil redusere leddet.

Ved å innføre insentiver representert med $\beta * z(e)$ vil motivasjonsleddet *m* også påvirkes gjennom β. Dersom man kun fokuserer på insentivene i seg selv, kan det fra modellen virke som at også insentiver vil øke motivasjonsleddet *m* i modellen. Det er her interaksjonen mellom β og *T* i *m* spiller en viktig rolle. Ved lav indre motivasjon for oppgaven vil interaksjonen mellom β og *T* i *m* være som forventet. Insentiv β kan øke motivasjonsleddet slik at innsatsen øker sammenlignet med tilfellet uten insentiver.

Dersom den indre motivasjonen for oppgaven derimot er høy, vil insentivering i form av β kunne ødelegge for effekten en slik oppgavetype i utgangspunktet har. Resultatet er da at insentiveringen har drevet ut den indre motivasjonen slik at innsatsen går ned eller forblir uendret sammenlignet med tilfellet uten insentiver. Man får med andre ord den forklarte crowding-out-effekten. Insentivene har dermed økt lønnen som en ytre motivasjon (priseffekt), men redusert den indre motivasjonen (crowding-out-effekt). Resultatet avhenger av hvilken effekt som er størst, men den utvidede PA-modellen tilsier at innsatsen agenten

yter forblir uendret eller går ned. Dette bryter med standardmodellen, som tilsier at insentivene gir økt innsats grunnet $\beta*z(e)$ alene.

I eksperimentet som ble gjennomført i sammenheng med utredningen, var den ene oppgavetypen en IQ-test. Når man blir presentert for en IQ-test, vil det være naturlig å ha et ønske om å gjøre det bra. Dette ønsket kan gå både på å gjøre det bra for seg selv, men også på å fremstå som en person med høy IQ. Den indre motivasjonen for å løse oppgaven er derfor høy. I PA-modellen vises dette ved at *T* er positiv. Den andre oppgavetypen var derimot meget repeterende, kjedelig og lite utfordrende. Det taler for at den indre motivasjonen, representert ved *T*, er lav. Dersom oppgaven ikke bærer preg av noe indre motivasjon, vil *T* være lik 0. Den utvidede PA-modellen vil da tilsvare standardmodellen fra tidligere. Insentivene vil derfor, som i standardmodellen, øke innsatsen grunnet $\beta*z(e)$.

PA-modellen med indre motivasjon gir dermed samme resultat som standardmodellen når det gjelder oppgaver som ikke har indre motivasjon – insentiver vil gi høyere innsats. Dette blir testet gjennom å gi respondentene en «*finn-tall-oppgave*». For oppgaver som preges av høy indre motivasjon, predikerer modellen at insentiver gir uendret eller lavere innsats enn i tilfellet med kun fastlønn. Dette er i strid med standardmodellen og blir testet ved å gi respondentene en IQ-test.

I en undersøkelse med 1600 ansatte, gjort av arbeidsgiverorganisasjonen Virke høsten 2012, svarte 74 prosent at lønn er viktig (Normann, 2012). Et annet og kanskje mer interessant resultat fra den samme undersøkelsen, var at 93 prosent mente arbeidsoppgavene i seg selv var viktig når de skulle velge jobb. Med andre ord er lønn, som en ytre motivasjonsfaktor, viktig, men det er sjelden at lønn alene er motivasjonen for å søke på en jobb. Dette viser at indre motivasjon absolutt er til stede i arbeidsmarkedet. Å ha et ledd som viser påvirkningen av indre motivasjon i PA-modellen er derfor hensiktsmessig. Prinsipalens profitt avhenger av verdiskapingen til den ansatte fratrukket lønnen man betaler. Ut fra den utvidede PA-modellen ser man dermed at det er i bedriftens egeninteresse å undersøke hvordan ansattes motivasjon er og hvordan den påvirkes av insentivsystemer.

2.7 Felteksperimenter

Det har blitt gjennomført flere interessante felteksperimenter som viser crowding-out-problematikken i praksis. Titmuss (1970) gjennomførte et eksperiment med innføring av betaling til blodgivere. Eksperimentene viste at man ved betaling fikk inn mindre og dårligere blod. Med andre ord fikk man en crowding-out-effekt der folk ikke lenger følte at blodgivning var en forpliktelse man hadde overfor samfunnet, men i stedet noe man gjorde for betaling. Blodgivning kunne derfor bli sett på som å selge kroppen sin snarere enn å gjøre noe bra for samfunnet, og dette kun på grunn av at en innførte betaling.

Et annet eksempel er Gneezy og Rustichini (2000b) som undersøkte effekten av å innføre en straff, i form av en avgift, for foreldre som hentet barna sine for sent i barnehagen. Eksperimentet ble gjennomført i ti barnehager i Haifa i Israel. Med en straff på 3 USD for foreldre som kom mer enn 10 minutter for sent, skulle man tro at problemet med sene foreldre skulle bedre seg. Etter kun to uker viste derimot resultatene noe helt annet: Mer enn dobbelt så mange som tidligere hentet barna for sent. Nok en gang kan man vise til crowding-out-effekten – man rettferdiggjorde det å komme for sent ved å betale for seg. Den indre motivasjonen om å skulle hente barna til rett tid ble drevet ut av avgiften. At boten var relativt lav, førte for øvrig til at problemet vedvarte også etter avgiften ble tatt vekk igjen. En lav bot signaliserte nemlig at det var et lite problem, selv om problemet egentlig var stort for de ansatte som måtte bli værende lenger på jobb.

Frey og Oberholzer-Gee (1997) gjennomførte en interessant studie innenfor det som gjerne kalles «*Not In My Backyard*», eller NIMBY-problemet. De tok for seg plasseringen av et anlegg for atomavfall i Sveits. I den sammenheng gjennomførte de en spørreundersøkelse blant innbyggere som kunne bli berørt. I første omgang ga man en forklaring på at respondentens bosted var aktuell lokalisering for et atomavfallsanlegg, og spurte om vedkommende ville støttet en utbygging. Hele 50,8 prosent var med dette utgangspunktet støttende til en eventuell utbygging, mens 44,9 prosent var negative. De resterende 4,3 prosentene stilte seg nøytrale.

I følge standard økonomisk teori burde en økonomisk kompensasjon øke støtten til en eventuell utbygging. Forfatterne utvidet derfor undersøkelsen og skrev at det Sveitsiske parlamentet ville gi en kompensasjon på 5 000 franc per familiemedlem per år. Andelen som nå støttet utbyggingen falt fra over 50 prosent til 24,6 prosent. Dette er enda et eksempel på

crowding-out-effekten, der betaling gjør at støtten går ned fordi man ikke lenger «*tar en for laget*», men får en følelse av at man gjør det kun for pengene. I eksperimentet ble effekten vedvarende, med kun en respondent som endret holdning, selv om man økte kompensasjonen betraktelig. Forfatterne fant tilsvarende effekter da de gjentok eksperimentet på et annet sted i Sveits.

3. Metode

3.1 Eksperimentell metode

Eksperimenter er i dag en mye brukt metode innenfor adferdsøkonomi. Gjennom sin forskning og undervisning på 50-tallet blir ofte Vernon L. Smith sett på som en viktig grunn til dette (Cappelen og Tungodden, 2012). Eksperimentelle data er hentet inn under kontrollerte forhold for vitenskapelige formål. En stor fordel er dermed at eksperimentator har god kontroll over hva man henter inn av informasjon og hvordan man gjør det.

Man snakker ofte om to ulike typer eksperimenter i forskningen: *laboratorieeksperimenter* og *felteksperimenter*. Laboratorieeksperimenter kjennetegnes ved at de er fullt kontrollerbare langs de fleste variabler, slik at de lett kan etterprøves. For eksempel foregår slike eksperimenter ofte i et lukket rom på en høyskole eller universitet, slik at respondentene er fysisk isolert. Det er dermed svært liten sannsynlighet for at ytre påvirkninger vil ha innvirkning på besvarelsene. Situasjonen vil typisk være veldig uvant for respondenten. Et potensielt problem med eksperiment i en slik setting er derfor at respondentene finner det *for* kontrollerende og unaturlig. Videre krever det tilgang til både lokale og medhjelpere. I tillegg må respondentene rekrutteres til å fysisk møte opp på et gitt tidspunkt. Timingen av eksperimentet må dermed være god for at det skal passe inn i tidsskjemaet til flest mulig respondenter. Typisk gjennomføres eksperimenter med studenter. Fordelen er at studenter er lettere å rekruttere enn mange andre grupper, men generaliseringsmulighetene er diskuterbare. Det stilles derfor ofte spørsmål ved den eksterne validiteten til slike eksperimenter. Spørsmålet blir da om resultatene gjenspeiler virkeligheten korrekt eller ikke.

Felteksperimenter foregår, som navnet tilsier, i *felten*. Dette gir en mer naturlig situasjon der respondentene er i sine ordinære omgivelser. Respondentene er dermed i sitt naturlige miljø, og sjansen er derfor større for at resultatene gjenspeiler virkeligheten. Det største problemet for felteksperimenter er at man ikke har full kontroll over påvirkningene respondentene utsettes for. En respondent kan raskt bli forstyrret av noe i omgivelsene slik at besvarelsen påvirkes. Graden av kontroll er derfor noe lavere ved slike eksperimenter.

Innenfor eksperimentell metode finner man videre et viktig skille mellom *korrelasjon* og *kausalitet*. Korrelasjon betyr ganske enkelt at man ser at verdiene på to variabler samvarierer. Det betyr for eksempel at når den ene variabelen reduseres, så øker den andre.

Da har variablene noe til felles, men man kan ikke påstå at verdien på den ene er årsak til den andre variabelens verdi. Kausalitet betyr derimot at man har en påvist årsakssammenheng. Sagt på en annen måte betyr kausalitet at man kan påstå at en påvirkning av en variabel er skyld i at en annen variabel endres. Ofte er det nettopp sistnevnte effekt man ønsker å finne gjennom et eksperiment.

For å kunne påstå et årsak-virknings-forhold er det grunnleggende at gruppene man sammenligner er tilfeldige og uten noen form for mønster. Dette får man til gjennom randomisering. Ved å randomisere respondentene og dele dem tilfeldig inn i grupper vet man at det ikke er noen grunnleggende forskjell mellom gruppene, heller ikke når det gjelder ikke-observerbare egenskaper (Cappelen og Tungodden, 2012). Det betyr at dersom gruppene står overfor et helt likt eksperiment, vil man forvente at svarene er like. Dersom en av gruppene derimot får en ulik situasjon, kalt en «*treatment*», vil eventuell forskjell i adferd skyldes nettopp denne skapte forskjellen. I et randomisert eksperiment er altså gitt treatment, eller behandling, den avgjørende faktoren. Dette er en skapt ulikhet og vil være den eneste forskjellen mellom gruppene i analysen. På den måten kan man påvise et årsak-virkningsforhold eller kausalitet der den kontrollerte faktoren er den som forårsaker ulik adferd.

Forarbeidet til et eksperiment er viktig. Når man på forhånd har tenkt nøye gjennom hva som skal undersøkes, vil analysen av datamaterialet bli lettere. Man kan da raskere finne de sammenhenger man ønsker å studere nærmere. Det kan også være hensiktsmessig å prøve ut designet på en mindre gruppe før man gjennomfører selve eksperimentet. Videre er det viktig å kontrollere at randomiseringen fungerer.

Et viktig spørsmål er hvilke personer en ønsker at respondentene skal være. Ansatte i en bedrift vil ofte være vanskelige å rekruttere, samtidig som kostnadene kan bli store. Studenter vil derimot være lette å rekruttere til en lavere kostnad, men generaliserbarheten kan da, som nevnt over, være diskuterbar. Et annet spørsmål går på hvor mange respondenter man ønsker å ha med. For få respondenter vil typisk føre til støy i besvarelsene. Dersom man har et stort antall, kan dette føre til unødvendig store gjennomføringskostnader, selv om betalingen per respondent er lav. Det gjelder derfor å finne en balansegang.

Hawthorne-effekten (Cappelen og Tungodden, 2012) eller *eksperimentereffekten* er også et sentralt problem i eksperimenter. Effekten går ut på at respondenten endrer adferden på en

måte som gjør at man opptrer slik en tror lederen av eksperimentet ønsker, nettopp fordi de blir studert. En slik adferdsendring kommer av at man vil tilfredsstille eksperimentator. For å unngå uønskede effekter kan det derfor være lurt å ikke røpe formålet direkte for respondentene. Det kan også være hensiktsmessig å spørre respondentene på slutten av eksperimentet hva de tror meningen med eksperimentet var. Dersom de forstår hensikten, kan dette påvirke besvarelsen de gir.

3.1.1 Amazon Mechanical Turk

Amazon Mechanical Turk, eller MTurk, er en relativt ny metode for å gjennomføre eksperimenter. Eksperimentatoren oppretter en profil som arbeidsgiver eller *requester* og logger seg inn via internettsiden www.mturk.com. Her får man tilgang til en stor gruppe potensielle respondenter, og kostnadene er relativt lave. Metoden er dermed et felteksperiment som er billig, og der man unngår noen av de typiske ulempene knyttet til ordinære eksperimenter i lab. Nettsiden viser tydelig hvilke muligheter som åpner seg gjennom å bruke internett. Konseptet er interessant og egner seg godt til eksperimenter, selv om dette ikke var planen for utviklerne. Man får effektivt testet ut effekter som ofte vil gjøre seg gjeldende i det generelle arbeidsmarkedet. I tillegg er det spådd en kraftig økning i online-jobbing i årene som kommer (Kaspersen, 2013), slik at forskning på plattformen er interessant selv om det ikke nøyaktig gjenspeiler en ordinær arbeidssituasjon i dag. Kaspersen skriver at norske bedrifter allerede i dag har tatt i bruk arbeidskraft fra India, USA og Pakistan over internett. Man får blant annet fordeler som rask oppgaveløsning, god pris på arbeidskraft og lettere opp- eller nedskalering av bedriften.

En enkel forklaring for hva MTurk er, blir gitt av Mason og Suri (2011): «*Et online arbeidsmarked der en arbeidsgiver legger ut jobber og arbeidere velger hvilke jobber de vil gjøre for en spesifikk betaling*». MTurk er med andre ord en online jobbportal der en arbeidsgiver eller anmoder legger ut en jobb han eller hun ønsker å få gjort av en uidentifisert gruppe personer. Howe (2006) kaller dette «*crowdsourcing*». Anmoder kan være en bedrift, enkeltperson eller eksperimentator. Arbeiderne får så opp en lang liste med tilgjengelige jobber hvor både betaling, en kort beskrivelse av jobben og omtrentlig tidsforbruk er presentert. Den store forskjellen fra å annonsere en jobb for eksempel gjennom finn.no, er at jobben på MTurk som regel kan gjøres umiddelbart fra ens egen datamaskin og ofte er lite tidkrevende. Det eneste som kreves er tilgang til internett. Når man er ferdig med

jobben, står man dessuten fritt til å velge en annen arbeidsgiver til neste jobb. Betalingen er i form av amerikanske dollar, indiske rupi eller gavekort på Amazon.com (mturk.com, 2013).

Utgangspunktet da Amazon laget Mechanical Turk, var at det skulle formidle oppgaver for Amazon selv. Det var snakk om oppgaver som krevde menneskelige ressurser, men samtidig var relativt enkle (Mason og Suri, 2011). Eksempel på dette var å hente ut data fra bilder, transkribere lydfiler og filtrere voksent materiale. Amazon fant ut at dette var en god måte å la andre enn dem selv få løst sine oppgaver mot betaling på, og samtidig tjene penger selv. De lanserte derfor nettsiden www.mturk.com. Amazon tar selv et honorar på minimum 10 % av alle utbetalinger fra arbeidsgiver i avgift for å levere tjenestene. Siden lanseringen av nettsiden i 2005 har antallet oppgavetyper økt og nettsiden brukes i dag også til å gjennomføre ulike former for eksperimenter (Huang et al, 2010).

Et besøk på www.mturk.com 30. januar 2013 viser over 413 000 såkalte HITs eller «*Human Intelligence Tasks*». Dette er oppgavene som er tilgjengelig for å løses, og det er dermed tydelig at nettsiden blir aktivt brukt av bedrifter eller andre arbeidsgivere. Det er vanskelig å finne et eksakt tall for antall brukere eller arbeidere på MTurk per dags dato, men Amazon skrev selv i 2011 at siden da hadde flere enn 500 000 brukere fordelt på over 190 land (Amazon Web Services, 2011). Det er grunn til å tro at tallet er enda høyere i dag.

3.1.2 Qualtrics

Qualtrics er et verktøy tilgjengelig online fra www.qualtrics.com for å utvikle spørreundersøkelser eller surveys. Fra oppstarten i 2002 og frem til i dag har Qualtrics blitt verdens ledende leverandør av datainnsamlings- og analyseprogramvare (Tanner, 2013). På hjemmesiden deres står det videre at antallet organisasjoner som brukte Qualtrics i 2012 økte til over 5 000, inkludert 1 300 universiteter og college. Antallet surveyer som ble gjennomført i 2012 passerte en milliard. Det er med andre ord liten tvil om at verktøyet er mye brukt både av bedrifter og utdanningsinstitusjoner.

En viktig grunn for å bruke Qualtrics er at det oppleves enkelt fordi det har et oversiktlig brukergrensesnitt. Likevel har man utallige muligheter både ved utforming og analyse av spørreskjema. Når man logger inn på nettsiden deres, får man mulighet til å utvikle, redigere og distribuere surveyer. I utviklingsfasen kan man velge å sette opp spørsmålene manuelt i hvilken form man måtte ønske, eller man kan benytte seg av flere maler for hvordan typiske spørsmål og svaralternativer kan se ut. Oppsettet kan være alt fra å krysse av for ett eller

flere riktige alternativ, besvare spørsmål gjennom tekst, eller dra en markør for å besvare ved hjelp av gradering.

Ved distribusjonen velger man om surveyen skal være åpen for alle som klikker på den spesifikke linken som lages, eller om kun inviterte personer skal få lov å svare. Man får deretter full oversikt over hvor mange respondenter man har til enhver tid, og kan enkelt bestemme hvor lenge surveyen skal pågå. Når datainnhentingen er ferdig, kan man lage rapporter av resultatene på nettsidene eller enkelt eksportere resultatene til Excel eller statistikkprogrammet SPSS for videre analyse. Ved integrasjon med andre ofte brukte programpakker er dermed Qualtrics en smart og trygg måte å samle inn data på.

3.2 Design

Etter å ha studert mulighetene nærmere, kom jeg frem til at en kombinasjon av MTurk og Qualtrics ville være det beste alternativet for arbeidet med utredningen. Rekrutteringen og betalingene skjedde dermed gjennom MTurk, mens selve eksperimentet ble gjort i Qualtrics. På den måten ble det beste fra to verdener satt sammen. For å få en slik kombinasjon til å fungere, var det en del utfordringer. Heldigvis fantes det personer som hadde brukt kombinasjonen i andre sammenhenger tidligere, og som gjerne delte sine erfaringer på ulike forum på internett. Løsningen ble å legge ut «jobben» på MTurk med en link til Qualtrics.

Eksperimentet bestod av fire hoveddeler. Første del var en kort innledning, mens andre del var selve oppgaveløsningen. I del tre fikk respondentene personlige spørsmål. Siste del representerte avslutningen der respondentene måtte svare på spørsmål rundt det å delta i eksperimentet. For å få en full oversikt over oppgavene og spørsmålene som ble gitt, henvises det til punkt 11 i appendiks. Der er eksperimentet gjengitt i sin helhet.

Respondentene ble, ved hjelp av en randomiseringsfunksjon, tilfeldig fordelt på en av fire treatments i Qualtrics. To av treatmentene var en IQ-test der eneste forskjell mellom de to var betalingsstrukturen. De to andre treatmentene inneholdt oppgaver hvor man skulle finne antall 5-tall i en matrise. Denne oppgavetypen er omtalt som finn-tall-oppgaver. Også her var eneste forskjell betalingsstrukturen. Dermed fikk eksperimentet et 2x2-design med variasjon mellom oppgavetype og belønningsstruktur som vist i figur 2. Et slikt design er til min kunnskap ikke gjennomført tidligere i ett og samme eksperiment, og det var derfor interessant hvorvidt man kunne påvise eller avvise noen effekter gjennom det.

		Oppgavetype	
		IQ-test	**Finn-tall-oppgaver**
Belønningsstruktur	**Fast betaling**	*Treatment 1*	*Treatment 3*
	Insentivbetaling	*Treatment 2*	*Treatment 4*

Figur 2: Design av eksperimentet – variasjon mellom oppgavetype og belønningsstruktur gir fire treatmentgrupper.

3.2.1 Oppgavetyper

Eksperimentet ble satt opp med to vidt forskjellige oppgavetyper, en IQ-test og en repeterende finn-tall-oppgave. Respondentene visste på forhånd ingenting direkte om hvilke oppgavetyper de ville møte, og de fikk heller aldri vite at det var ulike oppgavetyper i samme eksperiment. På den måten unngikk man effekter som kunne komme av eventuelle urettferdighetsoppfatninger.

På samme måte som Gneezy og Rustichini (2000a) brukte oppgaver som tilsvarte en IQ-test, brukte jeg en IQ-test laget for egen underholdning (Gale og Skitt, 1994). IQ-testen bestod av 15 varierte oppgaver der man måtte bruke hodet og være kreativ. Med variasjon skulle dermed motivasjonen for oppgavene øke i tråd med jobbkarakteristika-modellen til Hackman og Oldham (1976).

Noen av oppgavene var enkle, mens andre var vanskeligere. På den måten var det lagt opp til at respondentene skulle få en mestringsfølelse underveis, slik Pink (2009) mener er viktig. Respondentene fikk beskjed om at de ut fra svarene ville få oppgitt en estimert IQ, i tillegg til fullstendig oversikt over svarene i slutten av eksperimentet. I tråd med jobbkarakteristika-modellen fikk man dermed både et synlig utfall og en detaljert tilbakemelding. Ved å ha en slik struktur var meningen at den indre motivasjonen skulle være høy for oppgaven. IQ-testen skulle derfor tilsvare en jobb med varierte arbeidsoppgaver preget av indre motivasjon.

Den andre oppgavetypen var av en mye mer repeterende karakter. Her fikk respondentene 15 nesten identiske oppgaver der man skulle telle antall 5-tall i en matrise på 8 kolonner og 16 rekker. Denne oppgavetypen ble laget for å være repeterende og lite engasjerende for respondentene. Målet var at oppgaven skulle reflektere en jobb etter samlebåndsmetoden, der hver ansatt har en klart definert arbeidsoppgave som er enkel og gjentagende. Dette skulle tilsi at den indre motivasjonen for oppgaven var lav for de fleste. Under denne oppgavetypen fikk respondentene en tilbakemelding på oppnådd score.

På forhånd ble det beregnet at de to oppgavetypene totalt sett burde ta like lang tid å løse. Begge oppgavesettene bestod av 15 oppgaver, men med mer varierende oppgaver i IQ-testen. Oppgaver som tok lang tid å løse i IQ-testen, ble oppveid av oppgaver som skulle ta kort tid å løse. Hver finn-tall-oppgave var nærmest identisk og skulle gjennomgående ta like lang tid å løse. På den måten burde sluttresultatet bli det samme når det gjaldt tidsbruk for de to oppgavetypene.

Problemet med uoppmerksomhet, som Berinsky et al (2012) nevner, ble testet gjennom et spørsmål i hver treatment der svaret ble oppgitt i oppgaveteksten. Dersom respondenten leste spørsmålene underveis hadde man derfor ingen grunn til å svare feil. Et feil svar kunne derfor tyde på uoppmerksomhet. Gjennom dette kunne jeg også avsløre eventuelt misbruk av systemet fra respondenter eller bots. Respondenter som ikke svarte korrekt på kontrollspørsmålene ble tatt ut av datasettet ved analysen.

3.2.2 Belønningsstruktur

Belønningsstrukturen varierte mellom fastlønn og insentivlønn. Variasjonen ble lagt inn i starten av hver treatment slik at ingen av respondentene var klar over at det eksisterte forskjeller. På den måten unngikk man effekter som kom av respondenter som synes det var urettferdig at andre fikk en annen kompensasjon enn dem selv for samme oppgaver. Fastlønnen bestod av en sum på 3 USD for å fullføre eksperimentet. Insentivlønnen bestod av oppmøtekompensasjon på 3 USD pluss 0,02 USD per korrekt svar. Maksimal bonus var dermed på 30 cent eller 10 % av fastlønnen. Dette tilsvarer nivået Gneezy og Rustichini (2000a) brukte i sin studie (5/60 = 8,33 %).

Med en beregnet tidsbruk på 20-30 minutter ville lønnsnivået være på mellom 6 og 9 USD i timelønn uten bonus. Det kan anses som et nokså vanlig nivå i MTurk for tilsvarende oppgaver. På den måten var det meningen at jobben skulle ses på som ordinær, uten å skille

seg spesielt ut verken positivt eller negativt fra andre på nettstedet. Det var også fullt mulig å gjennomføre surveyen raskere ved å legge mindre tid i tenkning, og heller velge et alternativ basert på gjetning. På den måten ville den effektive timelønnen kunne økes betraktelig. Det ble presisert at svarene ikke ville påvirke betalingen. Motivasjonen og ønsket om å gjøre det bra ville derfor være avgjørende for innsatsen til respondenten.

3.2.3 Spørsmål til respondentene

Etter selve oppgaveløsningen fikk alle respondentene noen personlige spørsmål om kjønn, etnisitet, alder, bosted, utdanning, inntekt, interesse for matematikk og hvor ofte de tok IQ-tester. Slike spørsmål kunne for det første gi interessant input til analysen hvor man kunne undersøke om variablene hadde betydning for innsatsen eller resultatet. For det andre kunne man ved hjelp av svarene kontrollere at randomiseringen fungerte som ønsket.

I siste del fikk respondentene først spørsmål om hvor gøy de synes oppgavene var. På dette spørsmålet skulle man besvare ved å dra en markør mellom fem ulike nivå. Deretter fikk respondentene en rekke påstander de måtte ta stilling til. Her besvarte man gjennom en likert-skala på 7 punkter fra sterkt uenig til sterkt enig (se figur 3). Påstandene gikk på hvordan man oppfattet oppgavene man hadde løst i eksperimentet, og ble lagt inn for å kontrollere om oppgavene ble oppfattet som forventet.

Please consider each statement below.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
The tasks were difficult	○	○	○	○	○	○	○
I have a feeling of mastery after solving the tasks	○	○	○	○	○	○	○
The tasks were meaningful	○	○	○	○	○	○	○
I had to think hard to solve the tasks	○	○	○	○	○	○	○
The tasks felt repeating	○	○	○	○	○	○	○
The tasks were interesting	○	○	○	○	○	○	○

Figur 3: Påstander respondentene måtte ta stilling til mot slutten av eksperimentet

Den første påstanden omhandlet i hvilken grad man oppfattet oppgavene som vanskelige. I følge nyere teori vil en passelig vanskelighetsgrad være motiverende, mens for enkle oppgaver kan virke demotiverende. For å undersøke dette nærmere, ble det også spurt om mestringsfølelsen man satt igjen med etter å ha løst oppgavene. En høy følelse av mestring kan ha ført til høyere motivasjon, slik Pink (2009) hevder. Pink mener også at følelsen av at oppgaven har en mening, gir høyere motivasjon. Det ble derfor gitt en påstand rundt dette,

samt hvor hardt man måtte tenke for å løse oppgavene. Respondentene skulle også ta stilling til i hvilken grad oppgavene ble oppfattet som repeterende. Repeterende oppgaver kunne redusere motivasjon ut fra jobbkarakteristika-modellen. Den siste påstanden omhandlet i hvilken grad respondenten følte at oppgavene var interessante.

Risikoaversjon, innsats og rettferdighet ble undersøkt gjennom tre separate spørsmål. For å vurdere risikoaversjon, samt tro på egne ferdigheter, ble respondentene bedt om å ta stilling til et belønningssystem for en hypotetisk jobb. Valget stod mellom en fastlønn på 1 000 USD eller en variabel lønn mellom 500 USD og 1 500 USD basert på prestasjoner. Svaret på dette spørsmålet ville også si noe om hvilke preferanser vedkommende hadde til belønningssystem.

Den enkeltes selvoppfattede innsats ble spurt om gjennom å bruke en skala fra 0 til 10, der 10 var høyeste mulige innsats. Dette spørsmålet ble inkludert fordi det kunne være interessant å ikke bare sammenligne prestasjonene, men også innsatsen man legger ned ved ulike belønningssystem. Et annet interessant spørsmål gikk på hvordan vedkommende trodde man hadde gjort det i forhold til gjennomsnittet når det gjaldt å løse oppgavene.

Det nest siste spørsmålet gikk ut på at respondenten skulle si sin mening om hvor rettferdig betalingen for deltakelsen i eksperimentet føltes. På dette spørsmålet kunne man besvare alt fra en til fem stjerner, hvor en stjerne representerte «veldig urettferdig» og fem stjerner var «veldig rettferdig». Helt til slutt fikk respondenten mulighet til å skrive inn hva han eller hun trodde formålet med oppgavene og spørsmålene var. Dersom en respondent mot formodning skulle forstå formålet, kunne det påvirke besvarelsen uheldig. Ingen av respondentene klarte i dette tilfellet å avsløre det fulle formålet med eksperimentet.

3.2.4 Muligheter

Gjennom å variere både belønningsstrukturen og oppgavetypene, samt stille relevante spørsmål knyttet til respondenten, åpnet det seg flere muligheter for analysen. For det første gikk det an å studere eventuelle forskjeller i innsats (tidsbruk) og prestasjoner (score) innenfor en oppgavetype når belønningsstrukturen ble endret. Dette ble gjort gjennom ordinære t-tester. Videre kunne jeg studere eventuelle forskjeller mellom oppgavetypene når det gjaldt hvordan belønningsstrukturen påvirket innsatsen eller prestasjonene. Dette ble gjort gjennom regresjonsmodeller. Med personlige spørsmål knyttet til respondenten kunne

jeg så sammenligne hvordan personens karakteristika eventuelt påvirket utfallet eller ga opphav til andre interessante sammenhenger.

3.3 Styrker og svakheter ved MTurk

Med et stort antall brukere som er innstilt på å jobbe via internett, er det lett å få nok respondenter til å delta i «jobber» man legger ut. Dette har gjort at flere forskere også har fått øynene opp for MTurk. I den sammenheng har det de siste årene blitt gjennomført flere studier som evaluerer hvordan dette online arbeidsmarkedet fungerer sammenlignet med et vanlig arbeidsmarked.

Ekstern validitet

Et viktig poeng ved eksperimenter er om utvalget man har med er representativt for resten av befolkningen, slik at resultatene kan generaliseres til å gjelde også utenfor studien. Dette kalles ekstern validitet. Berinsky et al (2012) tar opp problemstillingen i sin artikkel der de vurderer om MTurk er et godt verktøy for å gjennomføre eksperimenter. De tester den eksterne validiteten på tre måter ved hjelp av et utvalg av respondenter bosatt i USA. For det første sammenligner de karakteristikkene til utvalg fra MTurk med utvalg brukt i tidligere eksperimenter. Videre repliserer de noen utvalgte eksperimenter i MTurk for å se om treatmenteffektene er de samme som utenfor MTurk. Til slutt vurderer de hvorvidt det at respondentene i MTurk er vant til å svare på undersøkelser, gjør dem mindre representative for andre populasjoner.

På første punkt konkluderer Berinsky et al at det finnes noen signifikante forskjeller mellom den amerikanske populasjonen og utvalget fra MTurk, men at disse er små og av liten betydning. Utvalget er hverken perfekt representativt eller langt fra å være det. Det er svært vanlig å bruke studenter i eksperimenter, og et viktig poeng fra forfatterne er at man ved å benytte MTurk alltid vil få et større mangfold i utvalget. Dette er et meget positivt aspekt ved MTurk.

Forfatterne repliserte videre tre eksperimenter: i) En undersøkelse om velferdsordninger, ii) «*The Asian disease problem*» og iii) En undersøkelse om risiko i forhold til framing. I alle tre fikk man resultater i MTurk som var svært like de tidligere publiserte resultatene.

For å teste representativiteten la Berinsky et al ut syv eksperimenter mellom januar og april 2010 for å se hvor mange av respondentene som gikk igjen. Hver arbeider på MTurk har et individuelt ID-nummer slik at dette lett lar seg undersøke. Forfatterne finner at hele 70 % kun deltok i ett av eksperimentene, mens kun 2 % deltok på fem eller flere eksperimenter. Dermed finner de at det er lite som tilsier at respondentene tar med seg stimuli på tvers av eksperimenter. De fant heller ingen forskjell i besvarelsene når de tok hensyn til de få respondentene som gjennomførte flere av undersøkelsene. Disse var fremdeles påvirket av treatmentene, på samme måte som de andre respondentene.

Mason og Suri (2011) viser til at arbeiderne i MTurk har en mangfoldig bakgrunn i tillegg til at man har et stort spekter i alder, etnisitet, økonomisk status, språk og hjemland. I artikkelen sin viser de også til at det er mange studier som viser samsvar i adferd mellom arbeidere på MTurk og arbeidere i andre offline eller online eksperimenter. Dette er momenter som taler for at representativiteten på MTurk i forhold til den generelle befolkning er bedre enn for eksempel ved bruk av studenter på et bestemt lærested.

Intern validitet

Intern validitet handler om hvorvidt man har god kontroll på eksperimentet, designet og dataanalysen, slik at vi kan si at det eksisterer et årsak-virkningsforhold eller kausalitetsforhold. Spørsmålet er altså om funnene kommer av eget forskningsdesign eller om det kan være andre årsaker. Berinsky et al kommer i den sammenheng inn på to viktige tema. Det første går på om det er mulig for respondentene å delta mer enn en gang i eksperimentet, og dermed ødelegge for resultatene. I MTurk er det lagt inn en sperre slik at brukere som allerede har gjort en HIT, ikke kan ta denne om igjen. Berinsky et al finner at kun 2,4 % av besvarelsene kommer fra samme IP-adresse i deres eksperimenter. Det viser at gjentagelse av samme eksperiment ikke representerer et stort problem. Forfatterne viser også til at en årsak til gjentagende IP-adresse kan være at flere ulike personer tar undersøkelsen fra for eksempel en internettkafe eller bedrift.

For det andre er det spørsmål om uoppmerksomhet spiller en stor rolle i MTurk. Dersom respondentene er uoppmerksomme, vil man ha problemer med å gjennomføre stimuli gjennom ulike treatments. Berinsky et al viser til at brukere av MTurk faktisk er ekstra oppmerksomme fordi en arbeidsgiver kan avstå fra å betale lønn dersom vedkommende ikke er fornøyd med arbeidet som er utført. Arbeidsgiver kan også blokkere ansatte som ikke er ønsket til fremtidige jobber. På bakgrunn av disse forholdene vil arbeiderne lett søke etter

arbeidsgivers hensikt og svare slik man tror arbeidsgiver ønsker. Forfatterne argumenterer derfor for at man ikke bør avsløre formålet med eksperimentet for tidlig.

Tid og kostnad

To viktige fordeler med MTurk er tid og kostnader. For å sette opp et ordinært eksperiment kreves både nøye planlegging og tilgang til lokale, hjelpere og respondenter. Gjennom å bruke MTurk gjør man rekrutteringsprosessen raskere og ikke minst billigere. Man får enkelt tilgang til en stor, stabil og mangfoldig gruppe. Berinsky et al (2000) viser til at de rekrutterte over 300 deltakere på en dag ved å betale mellom 50 og 75 cent for en oppgave som tok to til fire minutter. Å rekruttere studenter til et eksperiment i lab vil være langt mer krevende. Ved å sette opp et eksperiment med NHH-studenter er det i tillegg vanlig praksis å kompensere studentene med 100 kroner i oppmøtekompenasjon. Med 280 studenter ville dette tilsvare 28 000 kroner før insentiveringen under selve eksperimentet startet. Til sammenligning ligger betalingene typisk mellom to og tre amerikanske dollar (USD) per respondent for et slikt tidsbegrenset eksperiment i MTurk. Der er et slikt lønnsnivå vanlig praksis, og virker derfor ikke demotiverende for arbeiderne slik man gjerne skulle tro. Som Mason og Suri (2011) skriver, er kostnaden ved å delta også betydelig mindre i MTurk siden man unngår reisekostnader og man kan delta når det passer best for en selv. Ved å bruke MTurk kan man gjennom å betale mindre, velge å ha betydelig flere respondenter for samme budsjett, eller spare inn betydelige summer ved å ha like mange respondenter som ved et ordinært eksperiment.

Mason og Suri (2011) understreker også fordelen ved at MTurk har betalingsmekanisme innebygget. De skriver at man gjennom å slippe eksterne betalingsmekanismer som PayPal får en bedre responsrate i eksperimentene. Respondentene er også sikret å få betaling for en god jobb gjennom at arbeidsgiver må ha overført lønnen til Amazon før arbeidet starter. En grunnregel i MTurk er at jo bedre man betaler, jo raskere får man respondenter.

Datakunnskaper

MTurk kan kreve en del datakunnskaper. Dette oppleves kanskje som en hindring for mange forskere. Riktignok er det laget mange maler for hvordan man setter opp en undersøkelse eller oppgave i MTurk, men de er begrenset til å gjelde helt enkle ting som å skrive inn en nettadresse eller tagge et bilde. Også korte surveyer er dekket, men med en gang mer avanserte undersøkelser skal lages, krever dette god datakunnskap gjennom å bruke kommandolinjer og programmering. Et godt alternativ er da muligheten til å kunne legge inn

en link til en ekstern side hvor selve eksperimentet gjennomføres. På den måten kombinerer man det beste fra to verdener gjennom at MTurk brukes til rekruttering og betaling av arbeidere, mens en ekstern side med bedre funksjonalitet brukes til å lage gode undersøkelser og eksperimenter.

Misbruk

Siden MTurk er internettbasert, har man hele tiden en fare for misbruk. Såkalte *bots*, programmer som automatisk besvarer undersøkelser, kan true validiteten i svarene man får. Det kan også være folk som forsøker å ta flest mulig HITs for å tjene penger og derfor klikker seg vilt gjennom undersøkelser uten å faktisk ta stilling til spørsmålene de blir stilt. Som regel vil dette være relativt lett å avsløre når man ser igjennom besvarelsene. For eksempel vil et kontrollspørsmål, der svaret er oppgitt i oppgaveteksten, enkelt avsløre om besvarelsene er genuine og respondenten har fulgt med.

Kontroll

I avsnittet om eksperimentell metode ble det fremhevet fordelen for lederen av eksperimentet av å ha full kontroll. Når man bruker nettbaserte løsninger mister man noe av oversikten man ville hatt i lab. Selvfølgelig vil kontrollen over spørsmål og treatments fremdeles være gjeldende, men mer direkte overvåkning over respondenten forsvinner. Dette betyr at man ikke har innsikt i om respondenten gjør noe annet samtidig som eksperimentet, eller om vedkommende har forstyrrende omgivelser. Siden eksperimentet blir gjennomført over internett, har man heller ikke garanti for at respondentens internettforbindelse holder eller at datamaskinen ikke bryter sammen. Dersom respondenten synes eksperimentet blir for kjedelig eller tar for lang tid, kan han eller hun enkelt droppe ut ved å krysse ut nettleservinduet. Dette vil normalt sett være lettere enn å avbryte et eksperiment i lab. Dermed er det en reell risiko å få mange halvferdige responser over internett.

Reliabilitet

Reliabilitet går ut på hvorvidt resultatene hadde blitt de samme om man gjentok eksperimentet. Det er med andre ord snakk om påliteligheten til eksperimentet og i hvilken grad man kan ha eventuelle målefeil. Tidspunktet der «jobben» blir lagt ut vil ha betydning for dette (se neste punkt).

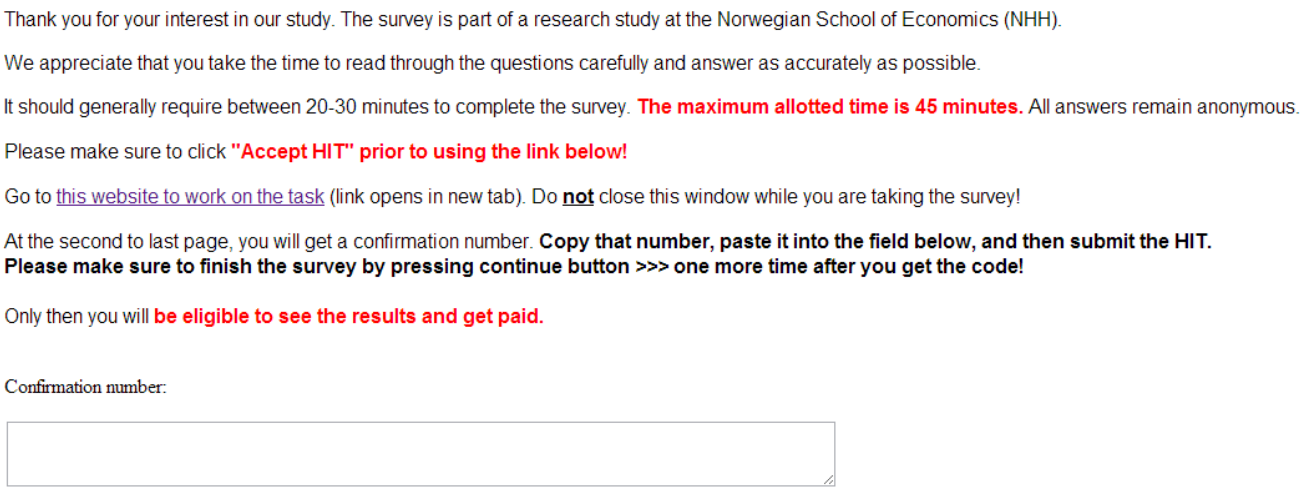
3.4 Gjennomføring

Eksperimentet ble først nøye satt sammen i Qualtrics. For deretter å teste ut det tekniske brukte jeg MTurk Requester Sandbox (requestersandbox.mturk.com), som er identisk med MTurk bortsett fra at ingen betalinger finner sted. Her publiserte jeg HITen for så å teste den ut selv gjennom MTurk Worker Sandbox (workersandbox.mturk.com). Det ble deretter gjort en testkjøring med betalinger i MTurk for å kontrollere at designet og betalingsløsningene fungerte som ønsket. Oppgaven ble publisert på samme måte som eksperimentet senere, men med plass til kun fire respondenter. I hovedeksperimentet ble det kontrollert at ingen av de tidligere testdeltakerne hadde forsøkt å delta.

Selve HITen ble publisert mandag 11. mars 2013 klokken 18.00 norsk tid. Tidspunktet var nøye gjennomtenkt på forhånd. USA og India er de to nasjonene som har flest arbeidere på MTurk. Grunnet sommertid i California var klokken da 10.00 der, mens klokken i India var 22.30. Dette skulle dermed tilsi at jeg fikk flest amerikanere, som sannsynligvis ligner oss nordmenn mer enn innbyggerne i India. Samtidig ønsket jeg ikke å utelukke respondenter fra India helt, da det kunne være interessant å studere eventuelle forskjeller nærmere. HITen ble begrenset til 280 deltakere, som skulle gi rundt 70 deltakere på hver av fire treatments. Sammenlignet med Gneezy og Rustichini (2000a) er dette hele 120 flere respondenter.

I MTurk la jeg ut jobben med tittel «*Take part in a research study – between 20-30 minutes*» og tilhørende beskrivelse «*You are invited to participate in a research study. You will be asked to solve tasks and answer some questions. It will require about 20-30 minutes to complete*». Gjennom dette håpte jeg å vekke interesse for jobben uten å avsløre formålet. Oppsettet var dermed i tråd med anbefalingen fra Berinsky et al (2012).

For å sikre kvaliteten, valgte jeg å la oppgaven kun være tilgjengelig for arbeidere som hadde løst minimum 100 HITs tidligere, og som hadde fått godtatt minst 90 % av arbeidet sitt. Når potensielle arbeidstakere klikket seg inn på jobben, fikk de en nærmere beskrivelse som vist i figur 4.


Thank you for your interest in our study. The survey is part of a research study at the Norwegian School of Economics (NHH).

We appreciate that you take the time to read through the questions carefully and answer as accurately as possible.

It should generally require between 20-30 minutes to complete the survey. **The maximum allotted time is 45 minutes.** All answers remain anonymous.

Please make sure to click **"Accept HIT" prior to using the link below!**

Go to this website to work on the task (link opens in new tab). Do **not** close this window while you are taking the survey!

At the second to last page, you will get a confirmation number. **Copy that number, paste it into the field below, and then submit the HIT. Please make sure to finish the survey by pressing continue button >>> one more time after you get the code!**

Only then you will **be eligible to see the results and get paid.**

Confirmation number:

Figur 4: Beskrivelse av oppgaven som arbeiderne fikk presentert i MTurk

Ved å akseptere HITen, eller jobben, fikk man så 45 minutter tilgjengelig for å fullføre eksperimentet. Dermed kunne respondenten ta seg forholdsvis god tid om ønskelig. Det ble antydet at besvarelsen ville ta mellom 20 og 30 minutter. Fordelingen var da tenkt som rundt 20 minutter på selve oppgavene og 10 minutter på besvarelsen av de påfølgende spørsmålene. Det ble også kunngjort for respondentene at besvarelsene deres ble holdt anonyme. I beskrivelsen av oppgaven ble det oppfordret til at man aksepterte HITen før man klikket på linken til surveyen. Når arbeiderne gjorde dette, ble det mulig for meg knytte brukernavnet og besvarelsen sammen i Qualtrics. I tillegg fikk alle generert en unik autorisasjonskode på slutten av eksperimentet, som de skulle kopiere inn i MTurk. Sammen fungerte da både brukernavnet og oppgitt kode som bevis for at eksperimentet var gjennomført. Å bruke begge deler ga også en større sikkerhet når man skulle knytte sammen besvarelsene med senere utbetalinger.

I MTurk må man ha betalt inn tilstrekkelig med midler for å kunne dekke alle respondentene før publisering er mulig. I tillegg ble det lagt inn en beskjed om at respondentene automatisk ville få betalt etter en frist på 7 dager dersom jeg ikke skulle rekke å gå gjennom besvarelsene før det. På den måten var respondentene sikret penger med mindre jeg aktivt avslo en for dårlig jobb innen fristen.

Dersom det skulle oppstå problemer eller være spørsmål, ble det også opprettet en egen e-postadresse, «*mturk.nhh@hotmail.com*», som respondentene kunne kontakte meg på. Etter at alle besvarelsene var fullført, ble dataene organisert og analysert i statistikkprogrammet SPSS.

3.5 Hovedhypotese

Hovedformålet med utredningen er å si noe om betydningen av oppgavens karakter for effekten av insentiver. For å svare på problemstillingen «*Hva har oppgavens karakter å si for effekten av prestasjonslønn?*» er det fornuftig å sette opp en hypotese, og teste denne. På grunnlag av presentert teori, tidligere eksperimenter og annen forskning har jeg satt opp følgende hovedhypotese for effekten av insentiver:

> **Hovedhypotese:** ***Insentiver gir økt innsats ved finn-tall-oppgaven, mens innsatsen for IQ-testen forblir uendret.***

Hypotesen baserer seg på at insentiver fungerer bra ved repeterende oppgaver med lav indre motivasjon, mens man får en crowding-out-effekt der insentivene ikke fungerer som ønsket når oppgavene kjennetegnes av høy indre motivasjon. Dersom IQ-testen gir respondentene høy indre motivasjon, tilsier de nye teoriene at insentiver *ikke* medfører at respondentene øker innsatsen. Med andre ord sier påstanden at standardteorien om belønning tar feil når indre motivasjon er høy. Siste del av hypotesen er basert på Gneezy og Rustichini (2000a) sin artikkel, men jeg legger til grunn at priseffekten og crowding-out-effekten oppveier hverandre i stedet for at sistnevnte effekt dominerer. Gjennom MTurk blir mitt eksperiment gjennomført med flere respondenter og en mer representativ gruppe for befolkningen (se Berinsky et al, 2012) enn Gneezy og Rustichini, der respondentene er studenter. Det er derfor interessant å se om man får tilsvarende crowding-out-effekter eller ikke.

Dersom den indre motivasjonen er lav, er hypotesen i tråd med standardteorien som tilsier at insentivbetaling gir høyere innsats. Finn-tall-oppgavene i eksperimentet er av repeterende karakter og vil derfor trolig bli oppfattet som lite motiverende. Hypotesen sier at insentiver gir en priseffekt som overstiger crowding-out-effekten slik at innsatsen øker ved insentivering. En slik tankegang stemmer også sett i lys av nyere forskning og teorier.

For å teste hypotesen må man vurdere hva som er et hensiktsmessig mål på innsats. Både score og tidsbruk fremstår som aktuelle kandidater. Å basere seg på score ville bety at man vurderte resultatet respondentene oppnådde som det beste målet på innsats. I hver treatment var det kun 15 oppgaver, slik at variasjonene neppe ville bli særlig store. Samtidig er nivået generelt høyt i MTurk. Arbeidsmassen består av respondenter som har gjort lignende oppgaver tidligere. Et godt resultat kunne også skyldes flaks, dersom respondenten kun gjettet seg til svarene gjennom tilfeldig klikking. I et slikt tilfelle ville innsatsen bli tolket

som høy selv om den i virkeligheten var noe helt annet. Motsatt kunne også forekommet – noen som la ned mye innsats i å løse HITen kunne få dårlig score rett og slett fordi de ikke var flinke nok.

Et mer hensiktsmessig mål for innsats vil derfor være tidsbruk. Tiden en respondent bruker på å løse oppgavene vil i stor grad vise hvor mye innsats vedkommende legger ned i arbeidet. Det ble på forhånd satt en maksimumstid på 45 minutter, men respondentene stod fritt til å bruke så liten tid de ville. Variasjonen i tidsbruk kunne således bli stor. Beregnet tidsbruk med normalt tempo var mellom 20 og 30 minutter. Ved å bruke tid som mål på innsats kunne jeg også avsløre de som gjennom flaks hadde scoret bra. Samtidig fikk jeg sett respondentene som satte seg nøye inn i oppgavene, men som likevel ikke scoret veldig høyt.

4. Resultater

4.1 Deskriptiv analyse

4.1.1 Rekruttering, dropouts og utbetaling

Eksperimentet ble, som tidligere beskrevet, lagt ut mandag 11. mars 2013 kl. 18.00 norsk tid. Rekrutteringen gikk meget bra – etter kun 20 minutter hadde 280 respondenter tatt på seg oppgaven med å delta i eksperimentet. Totalt fikk jeg 312 besvarelser i Qualtrics, der 282 var fullstendig utfylte. To personer hadde besvart alt, men glemte å klikke seg videre til siste side slik at det ble registrert som fullført. Som beskrevet har man en risiko for å få mange dropouts ved internettbaserte eksperiment. Differansen mellom 312 og 282 betyr at kun 30 respondenter (9,6 %) droppet ut og returnerte HITen. Dette bør betegnes som et godt resultat. Flest dropouts finner man i treatment 1 og treatment 2, som begge bestod av IQ-testen. Dette er interessant, men vil ikke bli analysert videre. Fordelingen av de som ikke fullførte, vises i en egen tabell under punkt 2 i appendiks.

Når en respondent droppet ut og returnerte HITen ble den igjen tilgjengelig for en ny person. På den måten var jeg hele tiden sikret å få 280 respondenter gjennom MTurk. I løpet av 1 time og 22 minutter hadde jeg fått alle de 282 fullførte responsene i Qualtrics, inkludert de to som glemte å klikke seg videre.

Etter å ha fått inn alle besvarelsene ble de nøye gjennomgått for å kontrollere MTurk-IDer og autorisasjonskoder. Av alle besvarelsene var det kun tre tilfeller som ikke stemte overens og som derfor var besvarelser som ikke var mulig å spore opp for betaling. Det var også en respondent som ikke hadde klart å skrive inn koden der han skulle, og dermed ikke kunne bevise å ha gjort eksperimentet. Jeg valgte likevel å betale vedkommende. Ingen av respondentene hadde tidligere deltatt i testkjøringen.

Fastlønnen på 3 USD ble betalt ut til 280 respondenter. Dette utgjorde totalt 840 USD. Deretter beregnet og betalte jeg bonuser på 2 cent per korrekt svar i tråd med den oppnådde scoren for treatment 2 og 4. For IQ-testen med insentivbetaling utgjorde dette 13,84 USD, mens det for finn-tall-oppgaven summerte seg opp til 18,00 USD. Med 10 % i kommisjon til Amazon, ble utgiftene totalt 972 USD eller 5 675 NOK, inkludert testkjøring.

4.1.2 Kontrollspørsmål

Av de 282 fullførte responsene svarte 264 (93,6 %) korrekt på kontrollspørsmålet som var lagt inn i hver treatment. Dette er et godt resultat og tyder på at respondentene leste spørsmålene som ble gitt nøye. Fordelingen av respondentene på hver treatment er vist i tabell 1.

Treatment	*Bestått besvarelser*		*Besvarelser med stryk*	
	Antall	*Prosent*	*Antall*	*Prosent*
T1	59	22,3 %	6	33,3 %
T2	68	25,8 %	3	16,7 %
T3	68	25,8 %	6	33,3 %
T4	69	26,1 %	3	16,7 %
Sum	**264**	**100 %**	**18**	**100 %**

Tabell 1: Fordeling av fullførte besvarelser i datasettet basert på om kontrollspørsmålet er bestått

Randomiseringsfunksjonen fungerte bra i forhold til antall respondenter, men den tok ikke hensyn til dropouts. Dermed fikk man en noe lavere andel respondenter i treatment 1 blant de bestått besvarelsene, da 8 personer droppet ut der. I tillegg strøk 6 respondenter på kontrollspørsmålet i T1. Det var på forhånd tatt høyde for at slike ting kunne skje, og med så mange respondenter hadde det ingen betydning for den videre analysen. Alle påfølgende analyser inkluderer kun de 264 besvarelsene som bestod kontrollspørsmålet.

4.1.3 Randomisering

Tabell 2 viser fordelingen av respondentene på de ulike treatmentene. Med fire treatmenter betyr det at en randomisering vil gi rundt 25 % av hver egenskap på T1 til T4. Fra tabellen ser man at randomiseringen i eksperimentet har fungert utmerket og fordelingene ligger rundt 25 % i hver treatment. Laveste verdi er 17,2 % i treatment 1 for alder under 25 år, mens høyeste verdi er 31,7 % i treatment 2 for respondenter av asiatisk opprinnelse.

	Treatment	*T1*	*T2*	*T3*	*T4*	*Sum (100%)*
Kjønn	*Menn*	30 (21,0 %)	35 (24,5 %)	37 (25,9 %)	41 (28,6 %)	143
	Kvinner	29 (24,0 %)	33 (27,3 %)	31 (25,6 %)	28 (23,1 %)	121
Etnisitet	*Hvite*	31 (19,7 %)	37 (23,6 %)	45 (28,7 %)	44 (28,0 %)	157
	Asiatiske	20 (25,3 %)	**25 (31,7 %)**	17 (21,5 %)	17 (21,5 %)	79
	Andre	8 (28,6 %)	6 (21,4 %)	6 (21,4 %)	8 (28,6 %)	28
Alder	*Under 25 år*	**10 (17,2 %)**	16 (27,6 %)	15 (25,9 %)	17 (29,3 %)	58
	25 til 34 år	32 (23,9 %)	34 (25,4 %)	38 (28,3 %)	30 (22,4 %)	134
	Over 35 år	17 (23,6 %)	18 (25,0 %)	15 (20,8 %)	22 (30,6 %)	72
Bosted	*USA*	39 (20,2 %)	48 (24,9 %)	52 (26,9 %)	54 (28,0 %)	193
	India	16 (25,0 %)	19 (29,7 %)	16 (25,0 %)	13 (20,3 %)	64
Utdanning	*Lav*	24 (21,4 %)	34 (30,4 %)	28 (25,0 %)	26 (23,2 %)	112
	Høy	35 (23,0 %)	34 (22,4 %)	40 (26,3 %)	43 (28,3 %)	152
Inntekt	*Lav*	25 (24,0 %)	24 (23,1 %)	25 (24,0 %)	30 (28,9 %)	104
	Middels	20 (25,0 %)	20 (25,0 %)	23 (28,8 %)	17 (21,2 %)	80
	Høy	14 (17,5 %)	24 (30,0 %)	20 (25,0 %)	22 (27,5 %)	80

Tabell 2: Fordelingen i hver treatment for spesifikke egenskaper ved respondentene. Perfekt randomisering tilsier 25 % av hver egenskap i hver treatment. Laveste verdi på 17,2 % og høyeste verdi på 31,7 % er uthevet.

Nærmere informasjon om kjønnsfordeling, aldersfordeling, etnisitet, bosted, utdanning og inntekt er gitt i punkt 3 til 7 i appendiks.

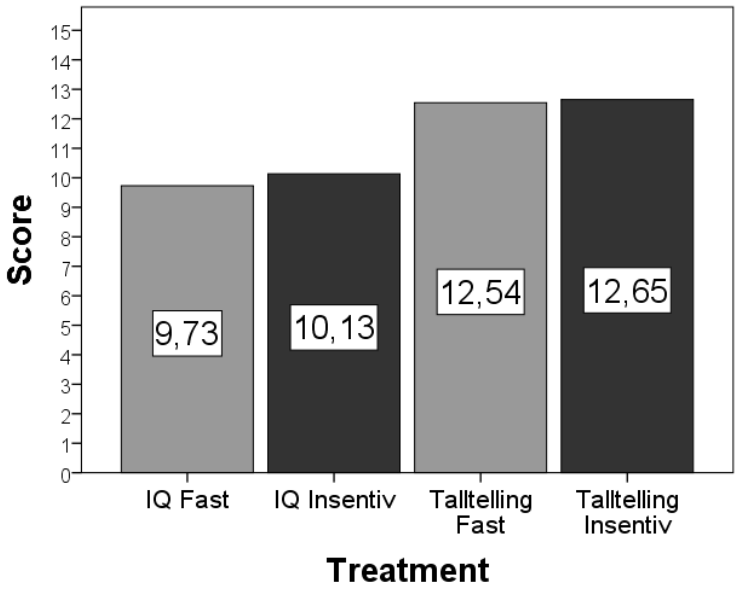
4.1.4 Score

Selv om det er tidsbruken som er det beste målet for innsats, er det interessant å studere resultatet i form av score. Scoren ble målt på en skala fra 0 til 15 poeng. For hvert riktige svar i oppgavedelen fikk man ett poeng. For de 264 respondentene var snittet 11,32 poeng,

med 4 poeng som laveste score og 15 poeng som det beste resultatet. Kun 27,7 % scoret 9 poeng eller lavere totalt sett. Dette viser at respondentene var godt kvalifisert for å løse oppgavene med få feil.

Ved å skille på oppgavetype, finner man at snittet for de 127 respondentene som gjorde IQ-testen er en score på 9,94 poeng. Her scoret 46,5 % av respondentene 9 poeng eller lavere. Tilsvarende tall for de 137 respondentene som gjorde finn-tall-oppgavene var 12,60 poeng og 10,2 % med 9 poeng eller lavere.

Man kan videre dele dataene opp for hver treatment. Ved å gjøre det ser man at scoren er nokså lik innenfor hver oppgavetype uavhengig av belønningsstruktur, mens man har en differanse på rundt 2,5 poeng mellom oppgavetypene. Det viser at IQ-testens oppgaver sannsynligvis var vanskeligere enn finn-tall-oppgavene, samtidig som insentivene ikke betyr noe for resultatet. Sistnevnte blir bekreftet med en t-test som gir p-verdi på 0,381 for IQ-oppgavene og p-verdi på 0,806 for finn-tall-oppgavene. Figur 5 viser gjennomsnittlig score for hver treatment. Fordelingen av score kan studeres i punkt 8 i appendiks.

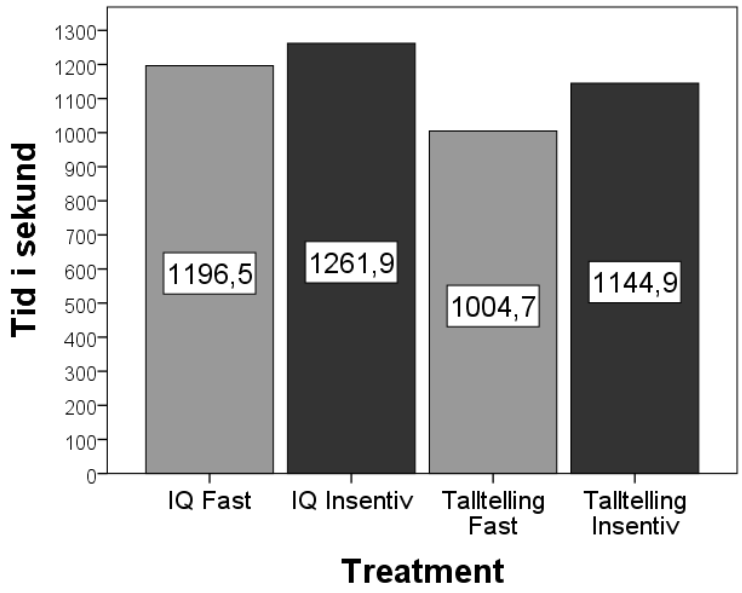


Figur 5: Gjennomsnittlig score i hver treatment. Maksimum score er 15 poeng.

4.1.5 Fordeling av tidsbruk

Tiden respondentene bruker er, som tidligere argumentert, et godt mål på innsatsen hver enkelt legger ned i arbeidet. Det er en del variasjoner i datasettet når det kommer til tidsbruk. Den raskeste respondenten brukte kun 6 minutter og 9 sekunder, mens den lengste tiden var på 44 minutter og 53 sekunder. Gjennomsnittlig tidsbruk var for alle respondentene 19 minutter og 10 sekunder, noe som stemmer bra med det jeg på forhånd estimerte. Gjennomsnittlig tidsbruk isolert sett for respondentene med IQ-test er 20 minutter og 31 sekunder, mens finn-tall-oppgavene gir en gjennomsnittlig tid på 17 minutter og 55 sekunder.

Tidsbruken isolert sett for hver treatment kommer frem av figur 6 og tabell 3. Gjennomsnittstid for treatment 1 er 1196,5 sekunder eller 19 minutter og 57 sekunder. Ved innføring av insentiv under IQ-testen (treatment 2) øker gjennomsnittstiden til 21 minutter og 2 sekunder (1261,9 sekunder). Dette er en økning på 1 minutt og 5 sekunder. I finn-tall-oppgavene er gjennomsnittlig tidsbruk med fastlønn 16 minutter og 45 sekunder (1004,7 sekunder) i treatment 3, mens den ved innføring av insentiver i treatment 4 øker til 19 minutter og 5 sekunder. Økningen er da på hele 2 minutter og 20 sekunder. Ut fra dette kan det derfor virke som om innsatsen i mye større grad blir påvirket av insentiver ved finn-tall-oppgavene enn i IQ-testen. At 95 % konfidensintervallene overlapper mellom T3 og T4, taler imot en slik konklusjon.

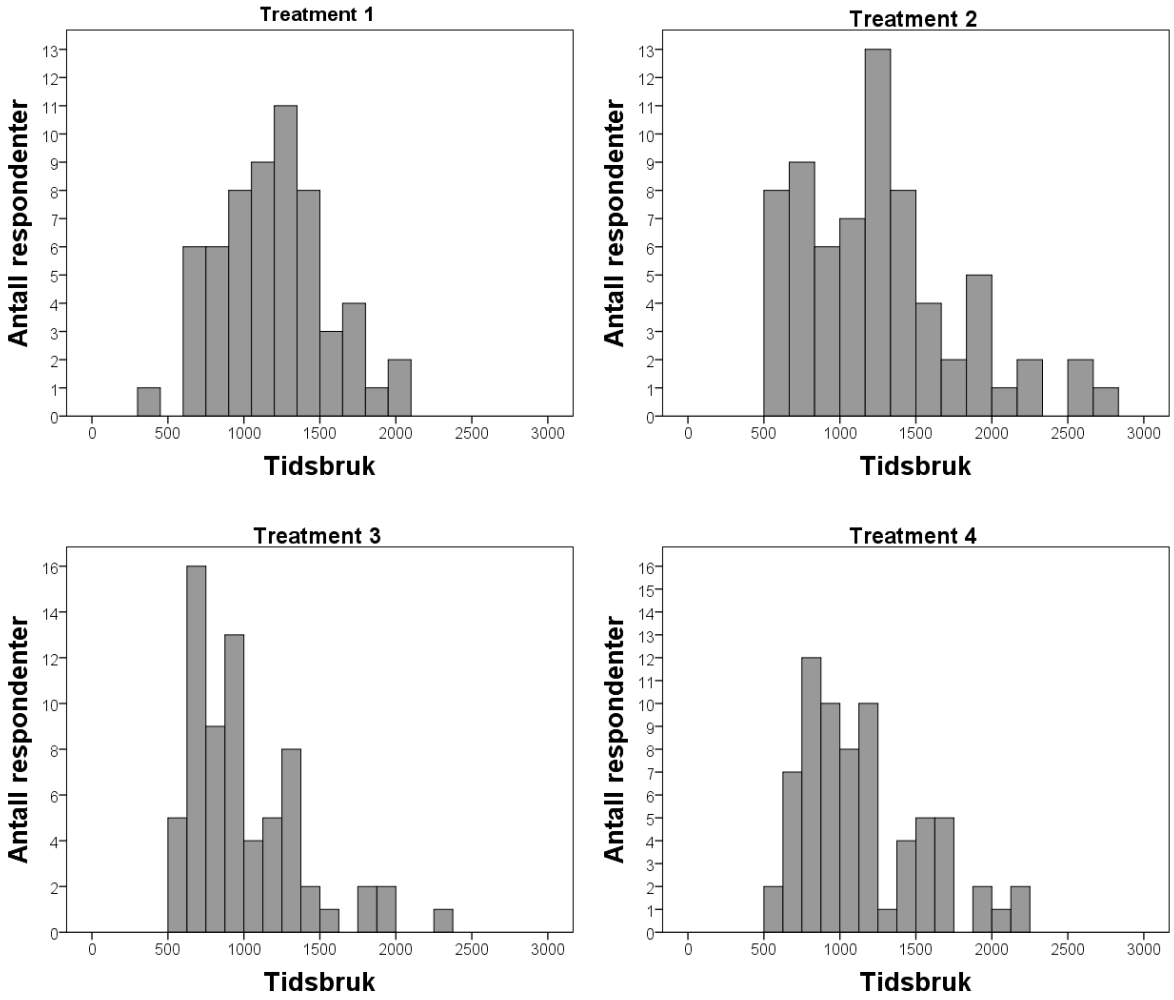


Figur 6: Gjennomsnittlig tidsbruk i sekund for hver treatment.

Treatment	*Gjennomsnittstid*	*95 % konfidensintervall*	*Standardavvik*
T1	1196,5	(1103,0:1290,0)	358,7
T2	1261,9	(1137,3:1386,5)	514,8
T3	1004,7	(915,9:1093,5)	366,9
T4	1144,9	(1047,0:1242,8)	407,6

Tabell 3: Gjennomsnittstid, 95% konfidensintervall og standardavvik for hver treatment. Alle tall er oppgitt i sekund.

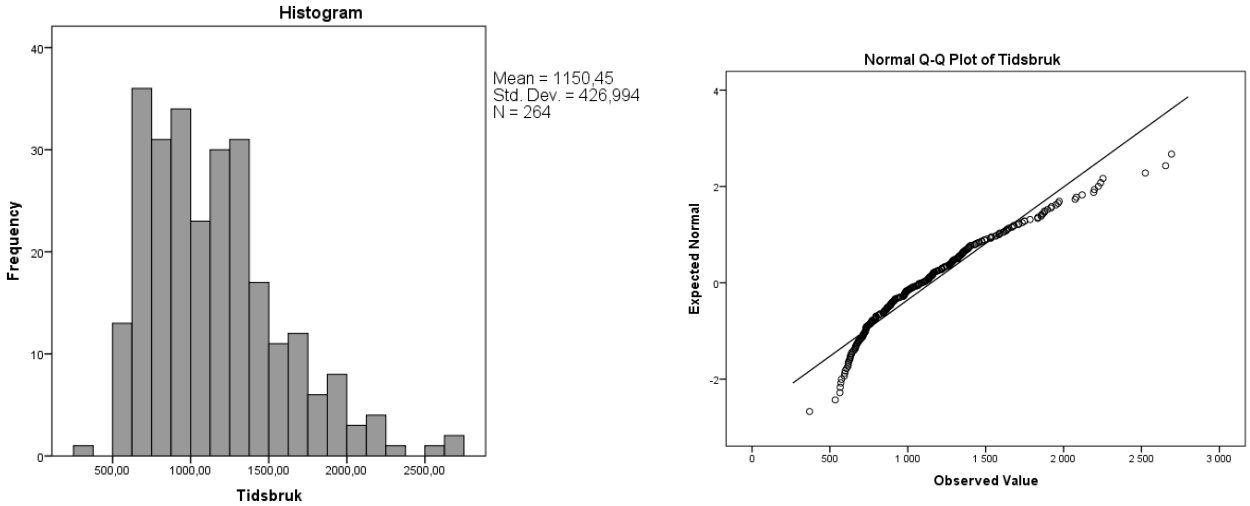
Fordelingen av tidsbruk i hver treatment er vist i figur 7. Fra figuren ser man at en større del av respondentene bruker mindre tid i treatment 3 og 4 sammenlignet med treatment 1 og 2. Det kan også virke som om fordelingen flyttes mot høyre (økt tid) i treatment 4 sammenlignet med treatment 3. Tilsvarende skift er vanskelig å påstå mellom treatment 1 og 2 i figuren.



Figur 7: Grafisk fremstilling av fordelingen av tidsbruk (x-aksen) og antall respondenter (y-aksen) for hver treatment.

4.1.6 Normalitetstest

For å bestemme hvilke tester man kan bruke i analysen av datasettet, er det nyttig å kjenne til om dataene er normalfordelte. Nullhypotesen om at dataene er normalfordelte vil forkastes dersom p-verdien ved en Shapiro-Wilk-test er under 0,05. Normalitetstest av variabelen tidsbruk, som er målet på innsats, gir en p-verdi på < 0,001. Shapiro-Wilk-testen er derfor klar på at dataene ikke er normalfordelte. I utgangspunktet tilsier det at en Mann-Whitney U-test bør brukes. Frekvensen av tidsbruk (i sekund) er vist i figur 8. Dersom dataene var normalfordelte, ville histogrammet hatt en klokkeform med toppunkt i midten. Plottet til høyre ville da også vært bedre tilpasset til linjen.



Figur 8: Fordelingen og normalitetsplott av tidsbruk. Ved normalfordeling skulle histogrammet til venstre hatt en klokkeform, mens punktene i normalitetsplottet til høyre skulle fulgt linjen.

Selv om normalitetstesten tilsier bruk av Mann-Whitney U-test, vil likevel en tradisjonell t-test bli benyttet i analysene. I motsetning til førstnevnte som sammenligner medianverdiene, sammenligner t-testen gjennomsnittene. Det er mer hensiktsmessig for arbeidet med dette datamaterialet. Som en kontroll er alle testene også gjennomført med Mann-Whitney U-test. Konklusjonene er de samme uavhengig av hvilken metode som blir benyttet.

4.1.7 Analyse

For å teste hypotesen om at insentiver gir uendret innsats (tidsbruk) i IQ-testen, mens det øker innsatsen ved finn-tall-oppgavene, brukte jeg en tradisjonell t-test. Testen

sammenligner gjennomsnittet i to uavhengige utvalg og har som nullhypotese at utvalgene har samme snitt. En p-verdi over 0,05 betyr derfor at man kan beholde nullhypotesen og konkludere med at man ikke kan påstå at gruppene er signifikant forskjellige. Dersom testen gir en p-verdi under 0,05 vil derimot nullhypotesen forkastes og man kan påstå at det er signifikant forskjellig snitt i de to utvalgene. I denne testen betyr det at p-verdien må være under 0,05 for at insentivene, som er eneste forskjellen mellom T1/T3 og T2/T4, skal ha hatt innvirkning på innsatsen.

Ved å dele opp hovedhypotesen får man to hypoteser som kan testes hver for seg:

> *Hypotese A:* Insentiv gir uendret tidsbruk i arbeidet med IQ-testen – altså uendret tidsbruk mellom T1 og T2.
>
> *Hypotese B*: Insentiv gir økt tidsbruk i arbeidet med finn-tall-oppgaven – altså økt tidsbruk fra T3 til T4.

Ved å teste hypotese A får man en p-verdi på 0,403 (p-verdi 0,839 med Mann-Whitney U-test). Det betyr at man beholder hypotesen om at det ikke er forskjell i tidsbruk mellom T1 og T2. Man kan ut fra dette ikke påstå at insentivene påvirker innsatsen signifikant når respondenten løser en IQ-test. Ved å foreta en tilsvarende test for hypotese B får man en p-verdi på 0,036 (p-verdi 0,023 med Mann-Whitney U-test). En slik p-verdi betyr dermed at tidsbruken er signifikant forskjellig mellom T3 og T4. Det finnes derfor også støtte for hypotese B, og man kan påstå at insentiv gir signifikant økt innsats ved finn-tall-oppgaver. Gitt at IQ-testen oppfattes som en oppgave med høy indre motivasjon, og at finn-tall-oppgavene oppfattes som det motsatte, finner jeg dermed støtte til den nyere teorien om indre motivasjon. Ut fra de isolerte t-testene kan jeg påstå at insentivene ikke påvirker innsats når oppgaven preges av høy indre motivasjon, mens insentivering fungerer og gir høyere innsats når respondenten har lav indre motivasjon for oppgaven.

Dersom man deler opp hovedhypotesen uten å ta hensyn til interaksjonseffekter, finner man altså støtte for den. En grundigere analyse med interaksjonseffekter er likevel ønskelig, og dette blir gjort gjennom påfølgende regresjonsmodeller.

4.2 Regresjonsanalyser

En regresjonsanalyse gir et bilde av samvariasjonen mellom en avhengig variabel og uavhengige variabler. Hensikten til en regresjonsmodell er å tilby en best mulig tilpasset linje som kan forklare dataene man har. Man kan se på en regresjonsanalyse som en mer avansert analyse enn en t-test. Det er derfor hensiktsmessig å foreta ytterligere analyser på datasettet ved hjelp av regresjonsmodeller. For å forklare innsatsen, i form av tidsbruk, lar man innsats være den avhengige variabelen i en regresjonsmodell. Tidsbruken kan da uttrykkes ved hjelp av et sett uavhengige variabler eller forklaringsvariabler. Typisk uttrykkes en lineær regresjon med formelen:

$$\mathbf{y = \alpha + \beta x + \varepsilon}$$

I formelen er y den avhengige variabelen, som i dette tilfellet er tidsbruk. Det er ikke et krav at den avhengige variabelen må være normalfordelt, noe tidsbruk heller ikke er i eksperimentet. α er et konstantledd, mens β kalles for regresjonskoeffisienten og betegner hvor mye tidsbruken øker når variabelen x stiger med en enhet. Sagt på en annen måte viser β stigningen på regresjonslinjen. Med $\beta > 0$ har man en positiv sammenheng mellom x og y, mens $\beta < 0$ representerer en negativ sammenheng. Tilfeldig støy blir betegnet av ε i modellen, og skal ha samme spredning, samtidig som den må være tilnærmet normalfordelt.

Hvert ledd i en multippel regresjonsmodell blir testet med utgangspunkt i en nullhypotese som sier at $\beta = 0$. Dersom p-verdien til en regresjonskoeffisient er under 0,05 forkaster man nullhypotesen og påstår at den uavhengige variabelen har signifikant påvirkning på den avhengige variabelen.

Når man lager regresjonsmodeller er det vanlig å kontrollere at modellen passer til observasjonene. For det første bør man kontrollere at modellen passer med hensyn til den avhengige variabelen. Dette gjøres gjennom å plotte de standardiserte residualene mot regresjonslinjen. Dersom residualene ligger fint fordelt langs x-aksen, vil man godta modellen. Det er også vanlig å se om noen av residualene har standardiserte verdier som ligger utenfor +/-3.

Den andre kontrollen går ut på om modellen passer til forklaringsvariablene. I den sammenheng bruker man Cook-avstander til å vurdere om datapunkter som ligger langt utenfor de andre kan ha for stor innvirkning. Dersom det er tilfellet, vil det kunne gjøre

estimatene ustabile. En vanlig grenseverdi for å vurdere Cook-avstander er 1,0. Med andre ord vil verdier under 1,0 være tilstrekkelig til å si at modellen i den sammenhengen passer bra.

For å utføre en regresjonsanalyse på datasettet ble flere av svarene omkodet til 0 eller 1 variabler, såkalte dummy-variabler. En rekke forklaringsvariabler ble kodet ut fra de personlige spørsmålene i eksperimentet. Etter flere analyser ble en del av dem tatt bort som følge av at man ikke kunne påvise en signifikant sammenheng mellom dem og den avhengige variabelen tidsbruk. Dette gjaldt i utgangspunktet forklaringsvariablene kjønn, utdanning, inntekt, matematikkinteresse og hvor ofte man tar IQ-tester. De gjenværende forklaringsvariablene er «*Insentiv*», «*IQ_test*», «*Asiatisk*», «*Alder_25til34år*» og «*Alder_over35år*» . Siden kjønn anses som interessant å ha med i modellene, har dette blitt tatt med som variabelen «*Kvinne*» på tross av manglende signifikans.

4.2.1 Regresjonsmodell for IQ-oppgaver

Ved å kun betrakte respondentene som fikk IQ-testen, får man følgende regresjonslinje:

Forklaringsvariabel	
Insentiv	57,6 (70,46)
Kvinne	102,8 (72,9)
Asiatisk	461,9*** (73,8)
Alder_25til34år	71,6 (93,0)
Alder_over35år	73,4 (106,3)
Konstant	929,4*** (95,1)
Antall observasjoner	127
R^2	0,257

Standardfeil i parantes. (*: p-verdi < 0,1, **: p-verdi < 0,05, ***: p-verdi < 0,01)

Tabell 4: Regresjonsmodell for respondenter som løste IQ-oppgaver. Alle tall er oppgitt i sekund og avrundet til en desimal.

Modellen har en forklaringsgrad R^2 på 0,257 og konstantleddet gir et utgangspunkt på hele 15 minutter og 29 sekunder (929 sekunder). Insentiver øker innsatsen med 58 sekunder, men dette er ikke signifikant. Resultatet tilsvarer dermed resultatet fra test av hypotese A. Heller ikke variablene kvinne eller alder, som begge øker tidsbruken, har signifikant påvirkning i modellen. Etnisitet er derimot sterkt signifikant med en økning i tid på hele 7 minutter og 42 sekunder (462 sekunder) for respondenter av asiatisk opprinnelse. Cook-avstander og residualplott er vist i punkt 9 i appendiks.

4.2.2 Regresjonsmodell for finn-tall-oppgaver

Ved å kun betrakte respondentene som fikk finn-tall-oppgavene, får man følgende regresjonslinje:

Forklaringsvariabel	
Insentiv	137,1**
	(61,3)
Kvinne	-39,5
	(63,6)
Asiatisk	359,5***
	(71,4)
Alder_25til34år	176,3**
	(76,6)
Alder_over35år	247,9**
	(89,6)
Konstant	779,6***
	(76,1)
Antall observasjoner	137
R^2	0,216

Standardfeil i parantes. (*: p-verdi < 0,1, **: p-verdi < 0,05, ***: p-verdi < 0,01)

Tabell 5: Regresjonsmodell for respondenter som løste finn-tall-oppgaver. Alle tall er oppgitt i sekund og avrundet til en desimal.

Modellen har en forklaringsgrad på R^2 på 0,216, noe som er litt lavere enn modellen for IQ-oppgavene. Konstantleddet gir her et utgangspunkt på kun 13 minutter (780 sekunder). Innføring av insentiv gir en signifikant (p-verdi = 0,027) økt tidsbruk på 2 minutter og 16 sekunder. At insentiv påvirker tiden signifikant for finn-tall-oppgavene, stemmer overens med resultatene fra test av hypotese B. Dersom respondenten er kvinne, reduserer dette tidsbruken noe ut fra modellen. Med en p-verdi på 0,536 er leddet ikke signifikant i denne modellen heller. Som i modellen for IQ-testen er også etnisitet sterkt signifikant i modellen for finn-tall-oppgavene. Her øker tidsbruken med 5 minutter og 59 sekunder dersom respondenten er asiatisk. Økningen er noe lavere enn for IQ-modellen. Alderen har

signifikant betydning for tidsbruken – jo eldre respondenten er, jo lenger tid bruker vedkommende (p-verdi = 0,023 for Alder_25til34år og p-verdi = 0,006 for Alder_over35år). Cook-avstander og residualplott er vist i punkt 9 i appendiks.

4.2.3 Regresjonsmodell for hele datasettet

For å gjennomføre regresjoner på hele datasettet ble det i tillegg til de nevnte forklaringsvariablene laget et interaksjonsledd, mellom «*Insentiv*» og «*IQ_test*». Dette blir kalt «*IQInsentiv*» og består av forklaringsvariablene Insentiv og IQ_test multiplisert med hverandre. Interaksjonsleddet sier noe om at effekten av «*Insentiv*» på tidsbruk gjerne endres avhengig av variabelen «*IQ_test*». Et signifikant interaksjonsledd IQInsentiv i en regresjonsmodell for tidsbruk betyr at insentivene (bonusen) signifikant påvirker respondentene ulikt avhengig av oppgavetype (IQ-test eller ikke). Dersom interaksjonsleddet derimot ikke er signifikant, kan jeg ikke påstå at forskjellen på en signifikant og en ikke-signifikant effekt (som funnet i t-testen og egne regresjonsmodeller tidligere) er signifikant. Sagt på en enklere måte betyr det at man ikke kan påstå at forskjellen som blir funnet gjennom testing av hypotese A og B er signifikant.

Regresjonslinjene er vist i tabell 6. Den første regresjonsmodellen tar for seg sammenhengen mellom oppgavetype, bonus og interaksjonsleddet mellom disse. Regresjonsmodell 2 er mer komplett og inkluderer flere forklaringsvariabler.

Modell 1 har en forklaringsgrad R^2 på 0,050, noe som er lavt. Modellen tas likevel med for å være en ren og enkel modell som kun viser virkningen av insentiver og oppgavetype. «*Insentiv*» har her en p-verdi på 0,051 slik at den akkurat ikke er signifikant på et 5%-signifikansnivå.

Fra modellen ser man at bonus har mindre effekt på tidsbruken under IQ-oppgavene (fra T1 til T2) enn under finn-tall-oppgavene (fra T3 til T4). Denne forskjellen får man på grunn av interaksjonsleddet, som sier at tiden reduseres med 75 sekunder når man gir bonus til en IQ-oppgave sammenlignet med finn-tall-oppgave. Siden interaksjonsleddet har en p-verdi på 0,470 er det ikke signifikant. Fra regresjonsmodellen kan man derfor ikke påstå at bonusen virker signifikant forskjellig på IQ-testen og finn-tall-oppgavene. Basert på det klare resultatet i t-testen tidligere, vil jeg likevel presisere at effekten går i retningen som predikert av hovedhypotesen. Modellen passer bra til de observerte data. Se appendiks punkt 9 for plott av standardiserte residualer og Cook-avstander.

Forklaringsvariabel	**Modell 1**	**Modell 2**
Insentiv	140,2*	142,7**
	(71,5)	(64,3)
IQ_test	191,8**	145,4**
	(74,5)	(65,9)
IQInsentiv	- 74,8	-79,0
	(103,2)	(92,7)
Kvinne		24,4
		(48,0)
Alder_25til34år		134,0**
		(59,5)
Alder_over35år		171,3**
		(68,8)
Asiatisk		- 406,3***
		(51,1)
Konstant	1004,7***	779,3***
	(50,7)	(66,1)
Antall observasjoner	264	264
R^2	0,050	0,252

Standardfeil i parantes. (: p-verdi < 0,1, **: p-verdi < 0,05, ***: p-verdi < 0,01)*

Tabell 6: Regresjonsmodeller for alle respondenter som bestod kontrollspørsmålene uavhengig av oppgavetype. Modell 1 er rendyrket og inneholder kun insentivering, oppgavetype og interaksjonsleddet mellom disse. Modell 2 er mer komplett og inkluderer i tillegg flere andre forklaringsvariabler. Alle tall er oppgitt i sekund og avrundet til en desimal.

Modell 2 har en forklaringsgrad R^2 på 0,252. At modellen forklarer rundt 25 % av innsatsen i form av tidsbruk er bra, men betyr at det er flere utenforstående faktorer som har betydning. Bortsett fra «*IQInsentiv*» og «*Kvinne*» er alle forklaringsvariablene statistisk signifikante med en p-verdi på under 0,05. Etnisiteten, representert ved forklaringsvariabelen «*Asiatisk*», er sterkt signifikant med en p-verdi < 0,001.

Fra regresjonen kan man lese flere interessante resultater. Konstantleddet sier at utgangspunktet er en tidsbruk på 779 sekunder eller 12 minutter og 59 sekunder. Dersom respondenten så får insentiver (T2 eller T4) øker tidsbruken signifikant med 2 minutter og 23 sekunder (143 sekunder) (p-verdi = 0,027).

I tilfellet der oppgavetypen er en IQ-test, viser forklaringsvariabelen «*IQ_test*» at tidsbruken signifikant øker med 2 minutter og 25 sekunder (145 sekunder) (p-verdi = 0,031). Det betyr at innsatsen øker når oppgaven preges av indre motivasjon.

De resterende forklaringsvariablene i regresjonsmodellen har jeg fått frem gjennom de personlige spørsmålene som ble stilt. Dersom respondenten er kvinne, sier modellen at tidsbruken øker med 24 sekunder. Dette er dog ikke signifikant (p-verdi = 0,611), og kjønn vil neppe ha særlig betydning for utfallet.

Etnisiteten viser seg derimot å ha stor betydning for tidsbruken. Dersom respondenten er asiatisk kan man legge til hele 6 minutter og 46 sekunder (406 sekunder). Resultatet er sterkt signifikant med en p-verdi < 0,001. Dette er interessant og viser at asiatisk opprinnelse medfører økt innsats i form av tidsbruk. En mulig delforklaring på den store økningen er at lavere engelskkunnskaper medfører at man bruker lengre tid på å sette seg inn i spørsmålene.

Respondentens alder har også betydning for tidsbruken. Dersom en respondent er 24 år eller yngre, påvirkes ikke tiden i regresjonslinjen av det. Med en alder på 25 til 34 år finner jeg derimot en signifikant økning (p-verdi = 0,025) på 2 minutter og 14 sekunder (134 sekunder). Er respondenten over 35 år, er økningen på hele 171 sekunder eller 2 minutter og 51 sekunder (p-verdi = 0,013). Det viser seg altså at innsatsen i form av tidsbruk øker jo eldre respondenten er.

Fra interaksjonsleddet «*IQInsentiv*» kan man lese at insentivering gir 79 sekunder (1 minutt og 19 sekunder) lavere tidsbruk ved IQ-test sammenlignet med om oppgaven er av typen finn tall. Sagt på en annen måte virker insentiveringen bedre når oppgaven er av repeterende karakter. Som i modell 1 kan jeg heller ikke nå påstå at interaksjonsleddet er signifikant (p-verdi = 0,394). Modellen gir derfor ikke grunnlag for å signifikant påstå at bonus virker forskjellig på de to oppgavetypene. Likevel ser man at interaksjonsleddet har negativt fortegn slik at retningen på effekten er i samsvar med resultatet fra hypotesetestingen og de isolerte regresjonene vist tidligere. Også modell 2 passer bra til de observerte data. Se appendiks punkt 9 for plott av standardiserte residualer og Cook-avstander.

4.2.4 Interaksjonseffekter

Det er også undersøkt om det finnes andre interaksjonseffekter som kan være signifikante i datasettet. For hver egenskap ble det laget et interaksjonsledd med insentiv for å undersøke om insentivene påvirket respondentene ulikt avhengig av egenskapen. Regresjonsmodeller som inneholdt «*Insentiv*», egenskapen og et interaksjonsledd mellom insentiv og egenskapen ble så satt opp. Interaksjonsleddene «*KvinneInsentiv*» (p-verdi = 0,081) og «*LavinntektInsentiv*» (p-verdi = 0,036) ble funnet til å være signifikante på henholdsvis 10 % og 5 % signifikansnivå. Regresjonsmodellene er gitt i det følgende.

Kjønn og insentiver

Forklaringsvariabel	
Insentiv	26,5 (70,8)
Kvinne	- 43,3 (75,1)
KvinneInsentiv	183,2* (104,5)
Konstant	1114,3*** (51,6)
Antall observasjoner	264
R^2	0,031

Standardfeil i parantes. (*: p-verdi < 0,1, **: p-verdi < 0,05, ***: p-verdi < 0,01)

Tabell 7: Regresjonsmodell som viser at interaksjonseffekten «KvinneInsentiv» er signifikant på et 10 %-nivå. Positivt fortegn betyr at kvinner i dette tilfellet øker tidsbruken mer enn menn når de får insentiv. Alle tall er oppgitt i sekund og avrundet til en desimal.

Fra interaksjonsleddet «*KvinneInsentiv*» i regresjonsmodellen ser man at respondenter som er kvinne og får insentiv øker innsatsen med 3 minutter og 3 sekunder. Med andre ord betyr dette at kvinnene som deltar i eksperimentet blir påvirket av insentiver i større grad enn de mannlige respondentene. Flere studier (se for eksempel Eckel og Grossman (2008)) har konkludert med at kvinner er mer risikoaverse enn menn. En risikoavers person vil foretrekke fastlønn fremfor insentiver. Basert på det skulle man kanskje tro at kvinner

reagerte negativt på insentivene i eksperimentet. At regresjonsmodellen viser at kvinner signifikant (10%-nivå) øker innsatsen i form av tidsbruk når man får insentiver er derfor interessant. Hverken «*Insentiv*» eller «*Kvinne*» er signifikant i denne modellen.

En videre analyse ble gjennomført for å undersøke om kjønn hadde en effekt på interaksjonsleddet «*IQInsentiv*» fra hovedmodellene. Det ble derfor laget et nytt interaksjonsledd kalt «*Kvinne_IQInsentiv*». Dersom dette leddet var signifikant, ville det bety at kvinner reagerte annerledes enn menn på insentivene når de skulle løse IQ-testen. En ny regresjonsmodell, som vist i punkt 10 i appendiks, ble satt opp for å undersøke dette. Interaksjonsleddet «*Kvinne_IQInsentiv*» er ikke signifikant med en p-verdi på 0,260. Det kan derfor ikke påstås at kjønn har en effekt på interaksjonsleddet «*IQInsentiv*». Kjønn har derfor ikke signifikant effekt for hvordan respondentene reagerer på insentiver når de løser IQ-oppgavene.

Inntektsnivå og insentiver

Forklaringsvariabel	
Insentiv	196,5**
	(65,9)
Lavinntekt	258,1***
	(75,6)
LavinntektInsentiv	- 221,9**
	(105,0)
Konstant	992,2***
	(47,5)
Antall observasjoner	264
R^2	0,059

Standardfeil i parantes. (*: p-verdi < 0,1, **: p-verdi < 0,05, ***: p-verdi < 0,01)

Tabell 8: Regresjonsmodell som viser at interaksjonseffekten «LavinntektInsentiv» er signifikant på et 5 %-nivå. Modellen predikerer at respondenter med lav inntekt reduserer sin tidsbruk når de får insentiver. Alle tall er oppgitt i sekund og avrundet til en desimal.

Både «*Insentiv*» (p-verdi = 0,003) og «*Lavinntekt*» (p-verdi = 0,001) er signifikant i regresjonsmodellen for inntektsnivå og insentiver. Insentiver øker tidsbruken med 3 minutter og 17 sekunder (197 sekunder), mens lav inntekt øker tidsbruken med 4 minutter og 18

sekunder (258 sekunder). Dette er som forventet. Interaksjonsleddet «*LavinntektInsentiv*» er også signifikant med en p-verdi på 0,036. Leddet sier at respondenter som har lav inntekt og får insentiv reduserer innsatsen med 3 minutter og 42 sekunder (222 sekunder). Med andre ord reduserer insentivene den store økningen i innsatsen som følge av lav inntekt. Reduksjonen er stor nok til at respondenter med lav inntekt yter en lavere innsats med insentiver (992 + 258 = 1250 sekunder) enn uten (992 + 258 + 196 – 222 = 1224 sekunder). Det er derfor en tydelig effekt der insentivene ødelegger for innsatsen uavhengig av oppgavetype for respondentene med lav inntekt. Man skulle tro at insentivene, i form av pengemessig bonus, var viktig for de med lav inntekt. Derfor er resultatet gjerne motsatt av hva man skulle forvente, noe som gjør det til et interessant funn.

Igjen ble det gjort en ytterligere analyse for å undersøke om inntektsnivået kunne påvirke interaksjonsleddet «*IQInsentiv*» fra hovedmodellene. I punkt 10 i appendiks er det presentert en modell som inkluderer interaksjonsleddet «*Lavinntekt_IQInsentiv*». Med en p-verdi på 0,716 er leddet ikke signifikant, og man kan ikke påstå at inntektsnivået har effekt for hvordan respondentene reagerer på insentiver når man løser IQ-oppgavene.

5. Diskusjon

Utgangspunktet for eksperimentet var PA-modellen og PA-modellen utvidet med indre motivasjon. Fokuset har vært crowding-out-effekten og om den virker inn slik nyere forskning og teori hevder. Målet var å skape et design der to ulike oppgavetyper skulle representere to motsetninger når det gjaldt indre motivasjon, for deretter å undersøke om bonus ville gi en crowding-out-effekt.

For å kontrollere at oppgavene ble oppfattet slik de var ment, testet jeg gjennomsnittsoppfatningen ved hjelp av en t-test med to uavhengige utvalg. Svarene for hvor gøy oppgavene ble oppfattet gikk på en skala fra 1 til 5, mens de andre spørsmålene hadde en skala fra 1 til 7 der 1 var sterkt uenig og 7 var sterkt enig. Det vil si at jo høyere snitt et spørsmål har i svaret, jo mer enig var respondentene. I tabell 9 er snittene til hvert spørsmål representert. Fordelingen for hver påstand er vist i punkt 1 i appendiks.

Spørsmål/påstand	*Snitt IQ*	*Snitt finn-tall*
Hvor gøy hadde du det ved løsningen av oppgavene?	3,91	3,60
Oppgavene var vanskelige	5,87	3,16
Jeg har en mestringsfølelse etter å ha løst oppgavene	5,09	4,89
Oppgavene var meningsfulle	5,70	4,34
Jeg måtte tenke hardt for å løse oppgavene	6,31	3,50
Oppgavene var interessante	6,39	4,85
Oppgavene var repeterende	3,87	6,73

Tabell 9: Besvarelsenes snittverdier fra påstander oppdelt for oppgavetype. Jo høyere verdi, jo mer enig er respondentene i påstanden. Skala fra 1 til 5 for «gøy» og skala fra 1 til 7 for alle andre påstander.

Resultatene viser for det første at IQ-testen oppfattes som gøyere (p-verdi 0,032) enn finn-tall-oppgavene. Videre er det sterkt signifikante resultat på at IQ-testen oppfattes vanskeligere, mer meningsfull og mer interessant (p-verdi < 0,001 for alle). I tillegg rapporterer respondentene som tok IQ-testen at de måtte tenke hardere sammenlignet med de

som fikk finn-tall-oppgavene (p-verdi < 0,001). Jeg finner også støtte for at finn-tall-oppgaven oppfattes som mer repeterende enn IQ-testen (p-verdi < 0,001). Alt dette er som forventet og ønsket på forhånd. Det ble også forespeilet at IQ-testen skulle gi økt mestringsfølelse, men t-testen gir ikke indikasjon på at dette er tilfellet (p-verdi 0,336). Man kan derfor ikke påstå at det er statistisk signifikant forskjell, selv om snittet fremstår som noe høyere for IQ-testen.

Finn-tall-oppgavene bestod av sterkt standardiserte og repeterende oppgaver som skulle oppleves som svært lite motiverende i seg selv. Resultatene viser at nettopp dette var tilfellet. På spørsmålet rundt hva respondentene trodde meningen med eksperimentet var, svarte en respondent for eksempel «*Too see how boring you could make it?*», mens en annen svarte «*To see how long it took before people got bored and just started guessing*». At oppgavene var repeterende ble bekreftet av en tredje respondent som svarte «*I'm honestly not sure, but I will notice every "5" I see for the rest of the day!*».

IQ-testen hadde varierte oppgaver og det var meningen at den skulle være «*i seg selv interessant eller morsom*», slik Ryan og Deci (2000) beskriver indre motivasjon. En indikasjon på at jeg har fått dette til, er følgende kommentar fra en deltaker: «*This one was a useful and interesting survey. Made me to think after a long time*».

For å få til det tydelige skillet mellom oppgavetypene, ble jobbkarakteristika-modellen til Hackman og Oldham (1976) fulgt under utviklingen av IQ-testen. For det første var oppgavene som ble gitt i stor grad varierte. Videre fremhevet jeg betydningen av at man løste dem – oppgavene var en del av et «*reasearch project*». En slik mening med arbeidet ble også gitt på grunnlag av Pink (2009) sine punkter.

Autonomien, som både jobbkarakteristika-modellen og Pink fremhever, ble ivaretatt i den grad at respondentene selv var ansvarlig for utfallet i form av score og tidsbruk. Respondentene stod fritt til å bruke hvilken som helst løsningsmetode for å komme frem til riktige svar, men hadde maksimum 45 minutter tilgjengelig. Det var ikke lagt opp til noe større påvirkningsmulighet for løsning av oppgavene, men respondentene fikk, som vist over, uttrykke hva de trodde eksperimentet handlet om. Siden IQ-oppgavene var mer varierte, kan dette ha ført til at autonomien føltes mer fremtredende. Det var ellers ikke lagt opp til mye autonomi, grunnet tekniske begrensninger.

Det siste relevante punktet i jobbkarakteristika-modellen var å gi tilbakemelding på arbeidet. For å sikre høyest mulig motivasjon, fikk respondentene en reell estimert IQ basert på besvarelsene de hadde gjort. I finn-tall-oppgavene fikk man ganske enkelt kun oppgitt en score. Dette har sannsynligvis vært med på å øke den indre motivasjonen for IQ-testen sammenlignet med finn-tall-oppgavene, slik resultatene viser.

Crowding-out-effekten, som er hovedtemaet til Gneezy og Rustichini (2000a), spiller inn i delen med IQ-oppgavene under hypotesetestingen og de separate regresjonene. Her oppveier priseffekten og crowding-out-effekten hverandre, slik at det ikke er forskjell i innsats når jeg gir respondentene insentiver. Jeg finner derimot ikke at crowding-out-effekten dominerer priseffekten, slik Gneezy og Rustichini finner i sin studie.

Når det gjelder de skjulte kostnadene med insentiver, blant annet omtalt av Weibel et al (2009), finner jeg fra hypotesetestingen og de separate regresjonsmodellene støtte for at insentiver ikke gir økt innsats for IQ-oppgavene. Med andre ord er insentiveringen her bortkastede penger som kun gir økte kostnader, uten noen form for økt innsats. Jeg finner derimot, ut fra disse metodene, støtte for uttalelsene til Frey (1997) om at man i situasjoner med indre motivasjon heller bør bygge videre på dette enn å tilby prestasjonslønn. Dette skyldes crowding-out-effekten. Når det gjelder regresjonene for hele datasettet, er crowding-out-effekten ikke like klar. Selv om interaksjonsleddet ikke er signifikant, er det likevel tydelig at effektene går i den samme retningen som under de isolerte t-testene og regresjonene for hver oppgavetype. Det er derfor tydelig at man bør ta hensyn til crowding-out-effekten. Med et stort antall respondenter og god kontroll på gjennomføringen, er dette er robust resultat.

Resultatene stemmer også bra med prediksjonene til PA-modellen som inkluderer indre motivasjon. Den utvidede modellen predikerer at innsatsnivået i treatment 2 forblir uendret eller går ned sammenlignet med treatment 1. I standardmodellen tror man derimot at innsatsnivået øker som følge av insentivene som blir gitt. Fra hypotesetestingen og de isolerte regresjonene viser resultatene at innsatsen forblir uendret mellom treatment 1 og 2. Dette stemmer derfor godt med den utvidede PA-modellen. Når oppgavetypen har lav indre motivasjon, gjør dette at innsatsnivået øker i tråd med standardmodellens prediksjon. Dette gjelder når man beveger seg fra treatment 3 til treatment 4, hvor resultatene stemmer bra med det standardmodellen tilsier. Som Bragelien (2005) skriver, har bonusen her ført til at de ansatte jobber «hardere» i form av høyere innsats.

6. Avslutning og konklusjon

Totalt sett ser man fra resultatene at jeg tydelig har klart å skille de to oppgavetypene når det gjelder motivasjon. I stor grad kan dette tilskrives jobbkarakteristika-modellen og forarbeidet med utformingen av oppgavene rundt denne. Gjennomføringen av eksperimentet i MTurk fungerte meget godt, og respondentene ble jevnt fordelt på de fire treatmentene gjennom randomisering. Målet med utredningen var å besvare problemstillingen som ble presentert innledningsvis:

Hva har oppgavens karakter å si for effekten av prestasjonslønn?

Jeg stilte også innledningsvis flere interessante spørsmål relatert til arbeidsmarkedet. For det første stilte jeg spørsmål om det virker fornuftig at man i dagens arbeidsmarked bruker mer variabel lønn. Jeg spurte deretter hva oppgavens karakter har å si og hvordan motivasjonen spiller inn. Ut fra resultatene ser man at insentiver fungerer godt når oppgavene er standardiserte og av repeterende karakter. Hypotesetestingen og de isolerte regresjonene er derimot klare på at insentiver ikke fungerer når oppgavene preges av høy indre motivasjon. Ut fra regresjonsanalysene for hele datasettet kan man på tross av dette ikke konkludere klart med at effekten av insentiver er signifikant forskjellig for de to oppgavetypene. Likevel viser resultatene at *retningen* på effekten av insentiver stemmer overens med den nyere teorien som er presentert. Dersom oppgaven er repeterende og preges av lav motivasjon, virker det derfor fornuftig å bruke variabel lønn. Hvis oppgaven er interessant og gir den ansatte en høy grad av indre motivasjon, kan det være mer problematisk med variabel lønn. Oppgavens karakter og motivasjonen spiller dermed inn. Det er dermed ikke fornuftig å ukritisk gi variabel lønn på varierte arbeidsoppgaver som preges av indre motivasjon. I dagens arbeidsmarked er mange oppgaver gjerne av sistnevnte karakter. En eventuell insentivering bør derfor være nøye gjennomtenkt på forhånd.

Med dette som grunnlag er problemstillingen besvart – *effekten av prestasjonslønn er avhengig av om oppgaven preges av høy eller lav indre motivasjon. Dersom oppgaven er av en slik karakter at den ansatte har lav indre motivasjon for å løse den, vil insentiver kunne hjelpe på innsatsen. Om oppgaven derimot er av en slik karakter at den ansatte har høy indre motivasjon for å løse den, vil effekten av insentiver være mer uklar. I sistnevnte tilfelle risikerer man derfor å gi bedriften en unødvendig ekstrakostnad.*

Det siste spørsmålet jeg stilte handlet om hvordan man optimalt bør betale den ansatte. Sett i sammenheng med resultatene finnes det ikke et fasitsvar på dette. Bragelien (2005) skriver om 10 bonustabber i sin artikkel, men ingen av dem går direkte på oppgavetype. En viktig bonustabbe vil etter min mening være det å ikke ta hensyn til oppgavetypen den ansatte står overfor. Min anbefaling er derfor å undersøke hvordan oppgavens karakter er, og hvordan den ansattes motivasjon for den er. Ved å gjøre det, er man et stort skritt nærmere svaret for hvordan vedkommende optimalt sett bør kompenseres.

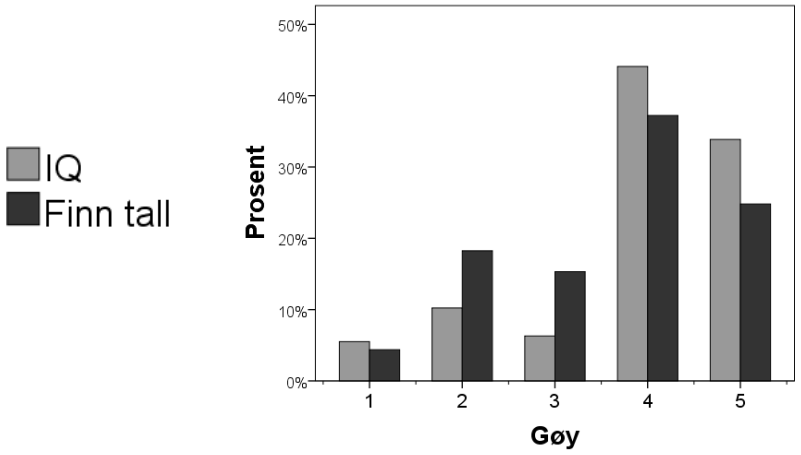
Med et begrenset antall oppgaver var det lite hensiktsmessig å studere forskjeller i score eller prestasjon. For fremtidig forskning kunne det derfor være interessant å lage et tilsvarende eksperiment der man tok med et større antall oppgaver. Med enda større variasjon i vanskelighetsgrad og en større avsatt tidsbruk, vil det kunne være rom for interessante resultater også når det gjelder prestasjoner i de ulike treatmentene.

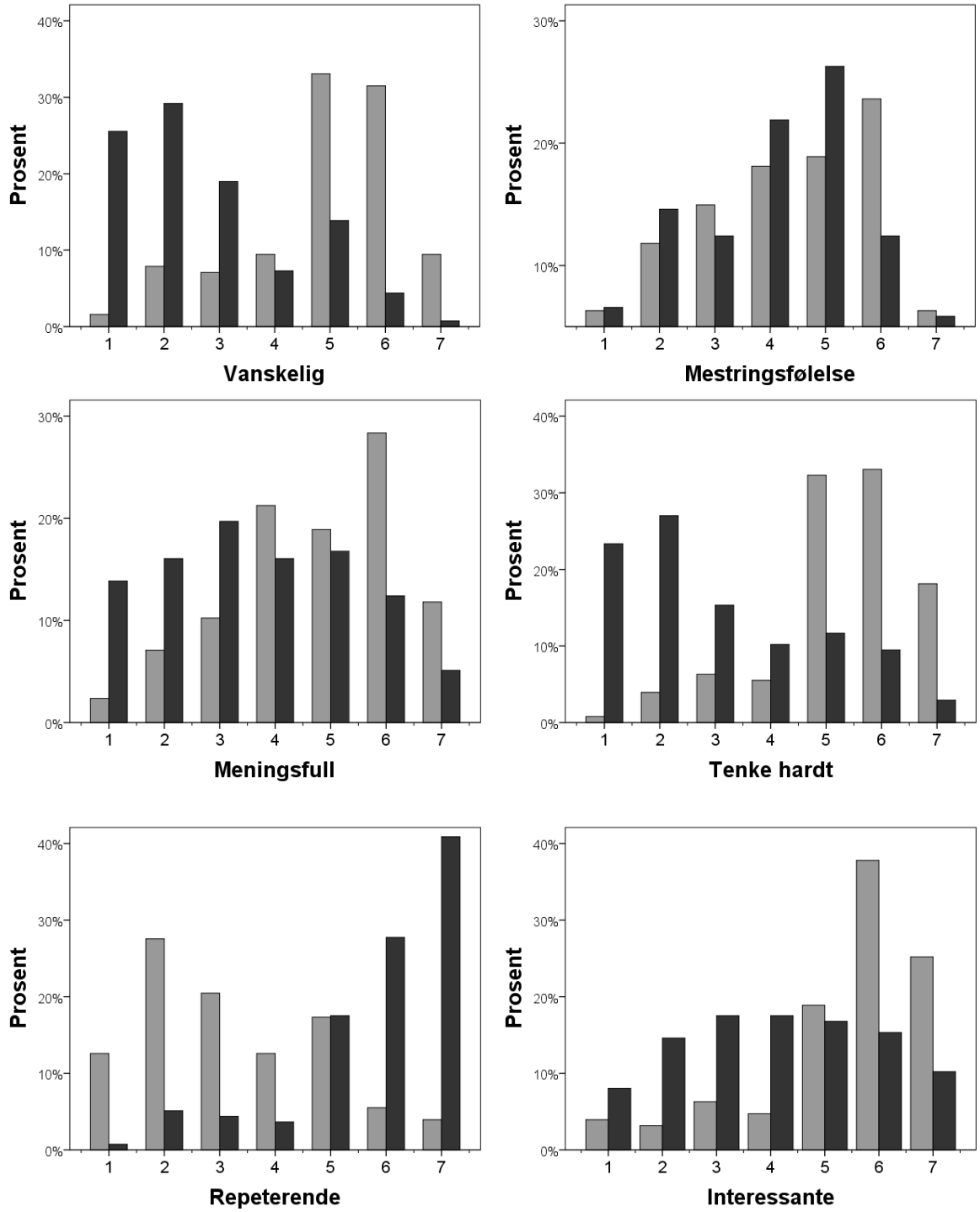
Appendiks

Utredningens appendiks er omfattende og delt inn i 11 deler. I utredningens tekst er det referert til hvilket punkt i appendiks som hører til. Punkt 1 til 8 viser ulike resultater fra datasettet. Punkt 9 viser residualplott og cook-avstander til de fire hovedmodellene i regresjonsanalysene. I punkt 10 finner man to regresjonsmodeller som undersøker om henholdsvis kjønn eller inntektsnivå påvirker interaksjonsleddet IQInsentiv gjennom å lage et ytterligere interaksjonsledd. Siste punkt i appendiks, punkt 11, gjengir eksperimentet som ble produsert og publisert i sin helhet. Ut fra dette ser man dermed hvordan eksperimentet fremstod for hver av enkelt treatment.

1. Hvordan oppgavene ble oppfattet

I histogrammene er andelen av besvarelsene fra respondentene i IQ-testen og i finn-tall-oppgavene fremstilt i søyler. Jo høyere tall på x-aksen, jo mer enig er respondentene i påstanden. For eksempel betyr 7 at man er veldig enig i at oppgavene var repeterende på spørsmålet om det, mens 1 betyr at man er helt uenig i påstanden.





Figur 9: Fordeling av hvordan oppgavene ble oppfattet, oppdelt for IQ-test og finn-tall-oppgaver.

2. Tidspunkt for dropouts

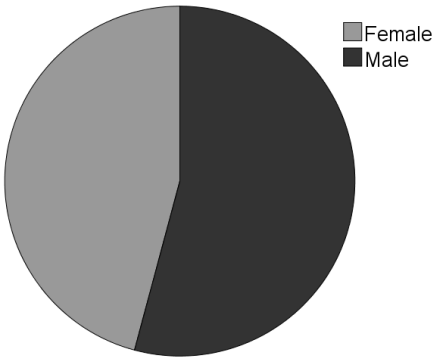
I tabell 10 er det vist tidspunkt for de 30 respondentene som droppet ut etter å ha sagt seg villig til å delta på eksperimentet. Flest respondenter droppet ut under arbeid med treatment 1 og treatment 2. Fem respondenter krysset ut eksperimentet umiddelbart slik at de ikke fikk sett innledningen. Totalt sett anses dropoutraten som lav med kun 9,6 %. Dette er et godt resultat for en online undersøkelse.

Tidspunkt for dropout	*Antall*	*Prosent*
Under arbeid med T1	8	26,7 %
Under arbeid med T2	6	20,0 %
Under arbeid med T3	3	10,0 %
Under arbeid med T4	4	13,3 %
I innledning	4	13,3 %
Før innledning ble presentert	5	16,7 %
Sum	**30**	**100 %**

Tabell 10: Oversikt over når respondenter droppet ut av eksperimentet

3. Kjønnsfordeling

Kjønnsfordelingen i datasettet er meget jevn. Av de 264 godkjente respondentene er 143 personer (54,2 %) menn, mens 121 (45,8 %) er kvinner. Figur 10 viser den totale kjønnsfordelingen i et paidiagram, mens tabell 11 viser kjønnsfordelingen i hver treatment.



Figur 10: Kjønnsfordeling totalt

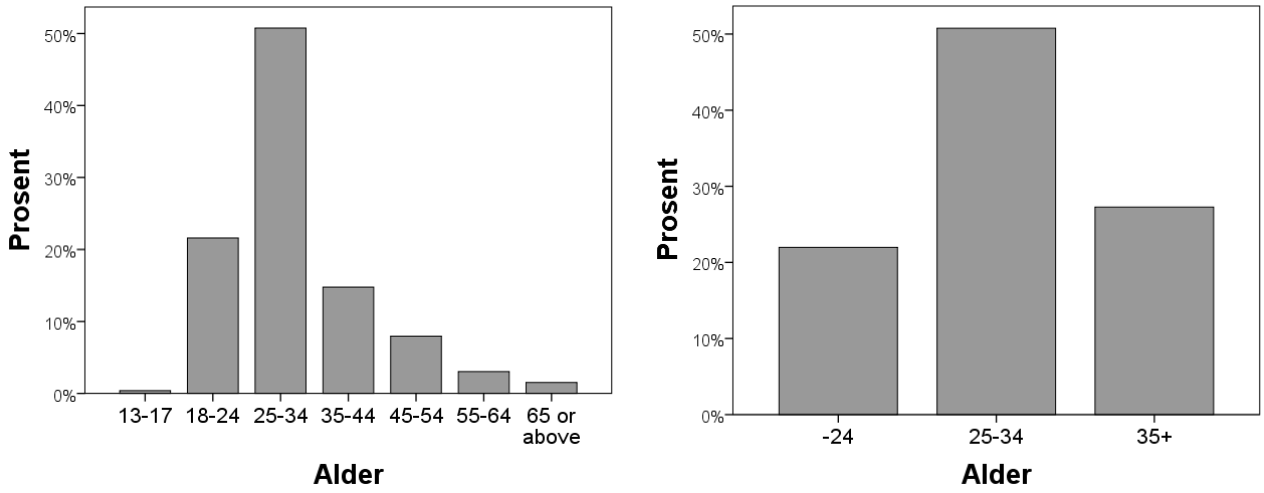
Treatment	*Menn*		*Kvinner*	
	Antall	*Prosent*	*Antall*	*Prosent*
T1	30	50,8 %	29	49,2 %
T2	35	51,5 %	33	48,5 %
T3	37	54,4 %	31	45,6 %
T4	41	59,4 %	28	40,6 %

Tabell 11: Kjønnsfordeling i hver treatment

I tabellen ser man fordelingen av menn og kvinner i de fire treatmentene. Også her er fordelingen nokså jevn i hver treatment. Dette understøtter at randomiseringen har fungert godt. Fordelingen samsvarer også med for eksempel Kaufmann et al (2011), som fant at 55 % av respondentene deres på MTurk var menn.

4. Aldersfordeling

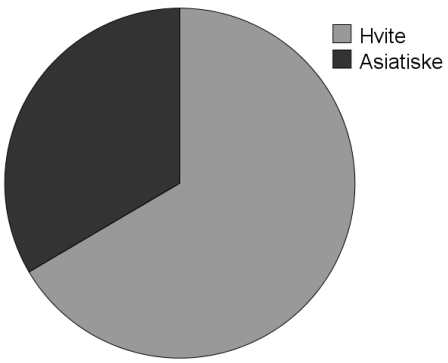
Aldersfordelingen i datasettet er interessant. Det ble her lagt inn syv kategorier fra 13 år til over 65 år. Som vist i figur 11 var over halvparten (50,8 %) av respondentene i aldersgruppen 25-34 år. Et slikt resultat samsvarer med det andre har funnet (se for eksempel Mason og Suri, 2011 eller Kaufmann et al, 2011). Det er videre verdt å merke seg at alle aldersgruppene er representert i datasettet. I analysen ble respondentene delt opp i tre ulike aldersgrupper: opp til 24 år, mellom 25 og 34 år og over 35år.



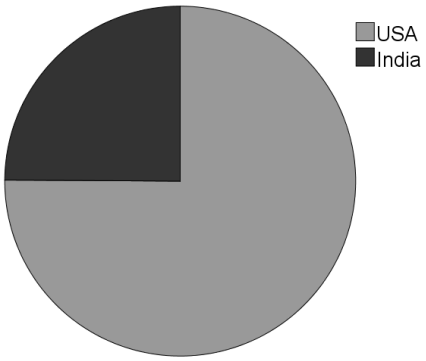
Figur 11: Aldersfordeling - originale og reviderte grupper

5. Etnisitet og bosted

Respondentene fikk også spørsmål om etnisitet og bosted. I datasettet er det to grupper som skiller seg ut når det gjelder etnisitet: 157 personer (59,5 %) er hvite/kaukasiske og 79 personer (29,9 %) er av asiatisk opprinnelse. De resterende 28 respondentene (10,6 %) fordeler seg på fem andre alternativer. Ved å ta ut respondentene med annen etnisitet ble det laget en variabel som viste om respondentene var hvite eller asiatiske. Fordelingen er da 66,5 % hvite og 33, 5 % asiatiske som vist i figur 12.



Figur 12: Etnisitet

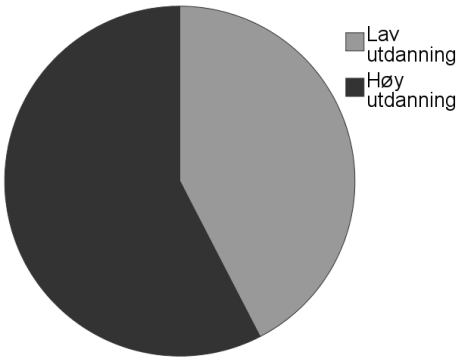


Figur 13: Bosted

Når det gjelder bosted finner jeg som forventet at de to landende som er høyest representert er USA med 193 respondenter (73,1 %) og India med 64 respondenter (24,2 %). En respondent (0,4 %) var fra Canada, mens seks respondenter (2,3 %) var fra andre land. Igjen er det hensiktsmessig med en variabel som viser fordelingen mellom de største gruppene som vist i figur 11. USA har da 75,1 % av respondentene, mens 24,9 % bor i India. Dette er vist i figur 13. En viktig grunn til at fordelingen viser et stort flertall av amerikanske respondenter, er sannsynligvis tidspunktet HITen ble publisert. Dette var en bevisst handling, som diskutert under *gjennomføring* i punkt 3.4.

6. Utdanning

På spørsmålet om utdanning var det hele åtte alternativ. På den måten var sjansen stor for at respondenten fant et mest mulig passende svar. Resultatene viser at alle alternativene ble brukt. For analyseformål har svarene blitt kategorisert i lav og høy utdanning. Med lav utdanning menes her to års college eller lavere, mens høy utdanning kategoriseres som bachelorgrad eller høyere. En slik kategorisering gjør at 112 av respondentene (42,4 %) har lav utdanning, mens 152 personer (57,6 %) har høy utdanning (se figur 14). Dette stemmer godt overens med Kaufmann et al (2011), som fant at 64,3 % av respondentene hadde bachelor eller høyere.



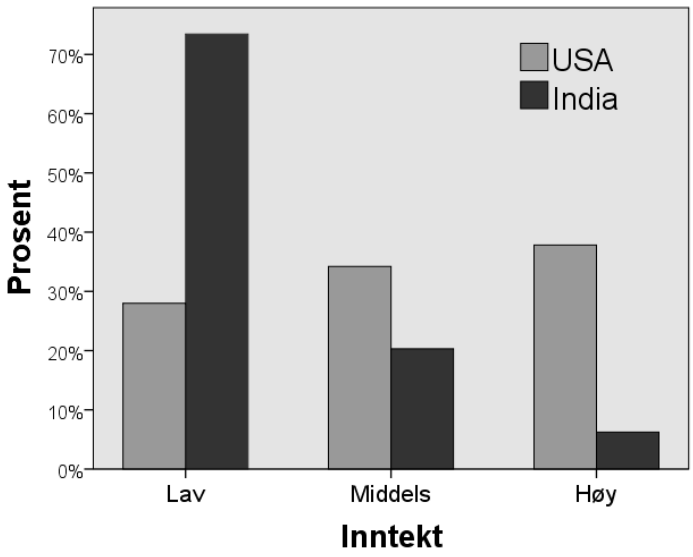
Figur 14: Utdanningsgrad

7. Inntekt

Inntekten har blitt delt i tre grupper: *Lav*, *Middels* og *Høy*. Lav inntekt definerer jeg som under 14 999 USD eller cirka 86 000 kroner i året. Middels inntekt er definert som 15 000 til 39 999 USD eller cirka 229 000 kroner, mens høy inntekt er alt over dette. Gjennomsnittlig inntekt per amerikaner var på rundt 28 000 USD i 2012 (United States Census Bureau, 2013), og ligger midt i gruppen for middels inntekt. Respondentene fordeler seg totalt sett med 104 personer (39,4 %) i gruppen med lav inntekt og 80 personer (30,3 %) i hver av gruppene for middels og høy inntekt. Dette er en relativt jevn fordeling, men med noen flere i kategorien lav inntekt. Ved å ta hensyn til bosted finner jeg at hele 73,4 % av de med bosted i India havner i gruppen for lav inntekt, mens det tilsvarende tallet for USA-boende er 28,0 % (se figur 15). Det er altså betydelige forskjeller mellom respondentenes inntekt i landene. En detaljert presentasjon er gitt i tabell 12.

Inntekt	*Totalt*		*USA*		*India*	
	Antall	*Prosent*	*Antall*	*Prosent*	*Antall*	*Prosent*
Lav	104	39,4 %	54	28,0 %	47	73,4 %
Middels	80	30,3 %	66	34,2 %	13	20,3 %
Høy	80	30,3 %	73	37,8 %	4	6,3 %
Sum	**264**	**100 %**	**193**	**100 %**	**64**	**100 %**

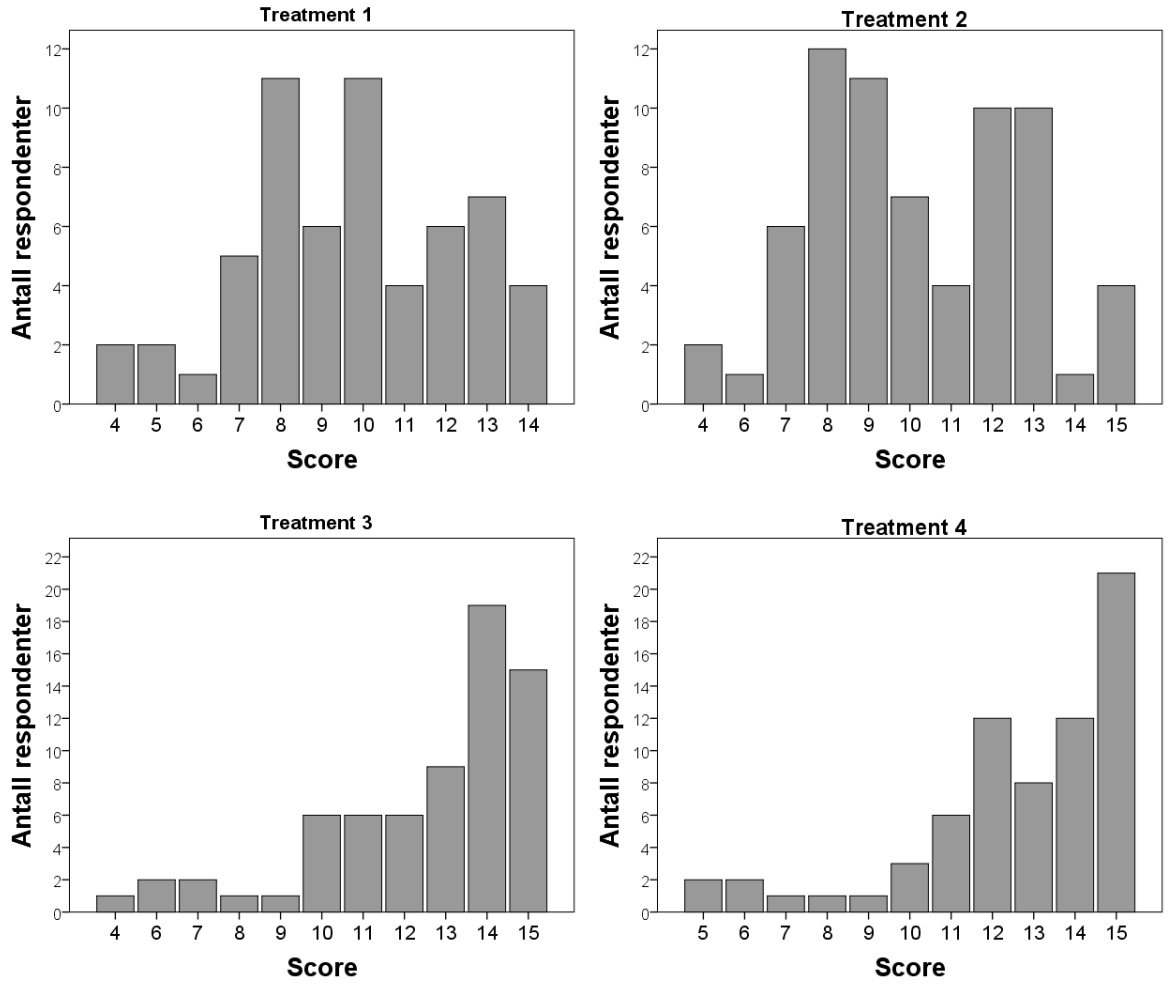
Tabell 12: Inntektsfordeling



Figur 15: Inntektsfordelingen i USA og India

8. Fordelingen av score

Fordelingen av score for hver treatment er vist i figur 16. Fra figuren ser man at scoren er jevnest fordelt i treatment 1 og 2, som består av IQ-testen. I treatment 1 finner man flest respondenter med score 8 og 10 poeng (11 respondenter), mens toppunktet i treatment 2 finnes ved score på 8 poeng (12 respondenter). For treatment 3 og 4, som begge har finn-tall-oppgave, ser man at scoren ligger høyere på skalaen. I treatment 3 er toppunktet ved en score på 14 poeng (19 respondenter), mens man i treatment 4 har et klart toppunkt på 15 poeng (21 respondenter).



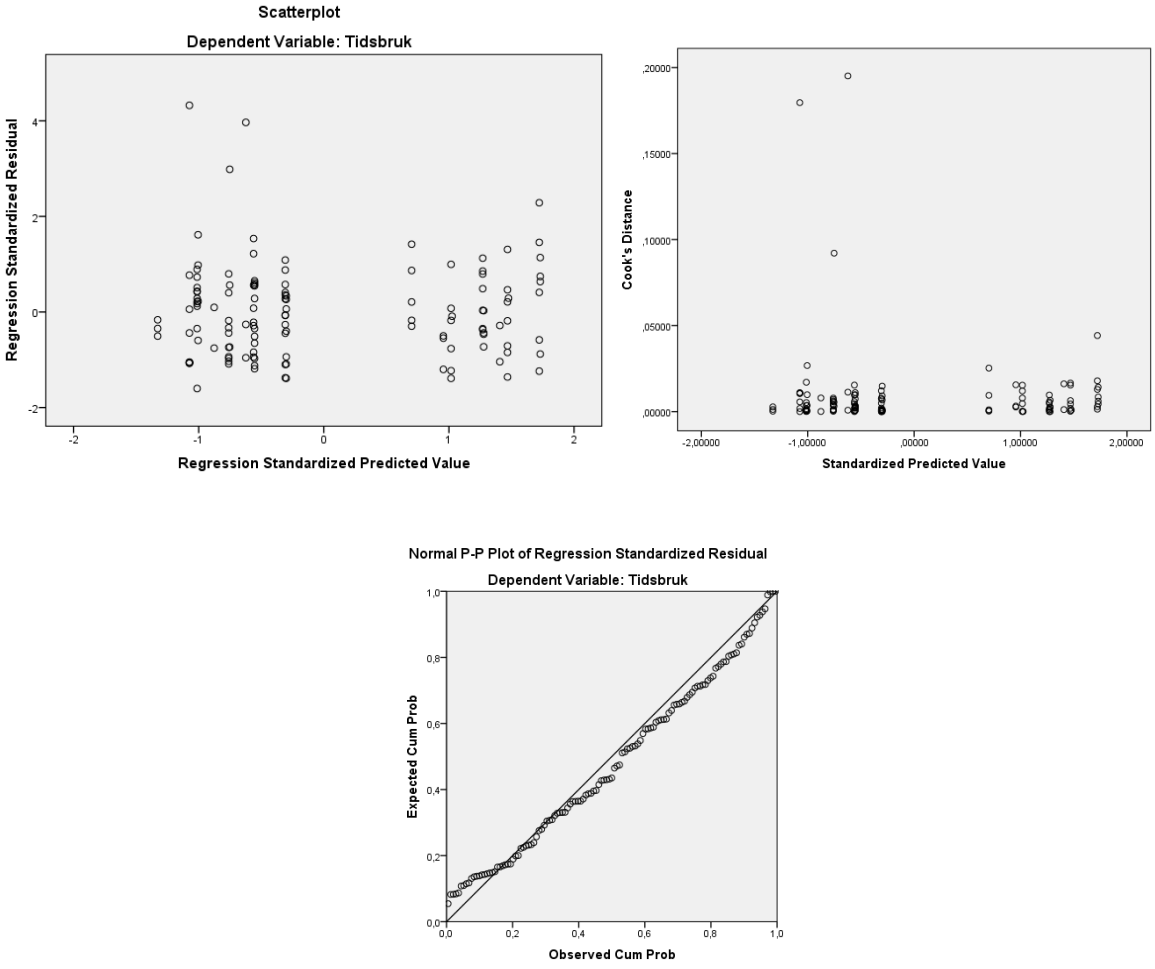
Figur 16: Fordelingen av score for hver treatment

9. Residualplott og Cook-avstander

I det følgende er det gjengitt residualplott, Cook-avstander og normalitetsplott for de fire hovedmodellene i utredningen.

Regresjonsmodell for IQ-oppgaver

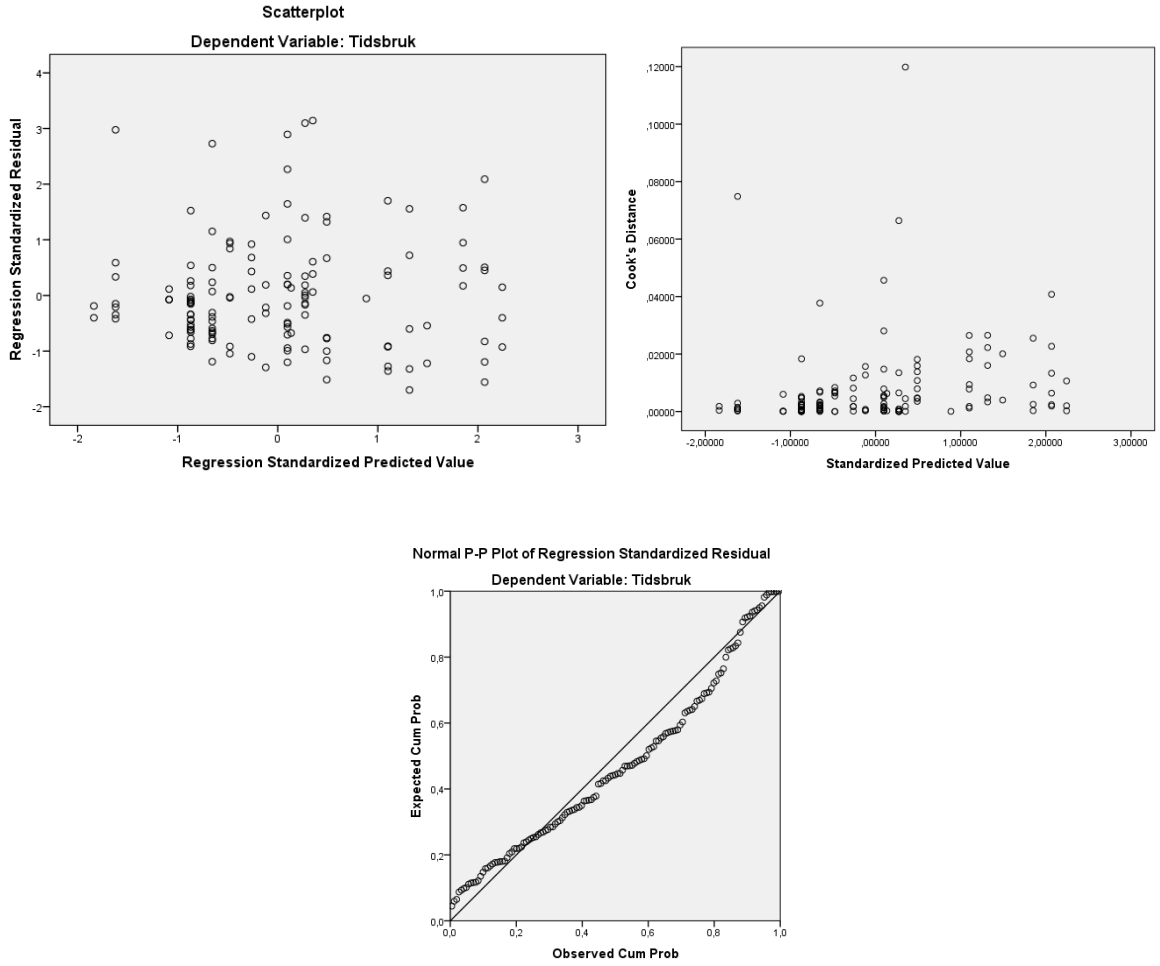
Figuren 17 viser ingen spesielle mønstre i residualplott, men man har to outliers over 3. Cook-avstandene viser to tilfeller med verdier over 1,000. Modellen godtas på tross av dette. Fordelingen følger linjen i normalitetsplottet nokså bra, og det er derfor naturlig å godta en antagelse om normalfordeling.



Figur 17: Residualplott, Cook-avstander og normalitetsplott for regresjonsmodell for IQ-oppgaver

Regresjonsmodell for finn-tall-oppgaver

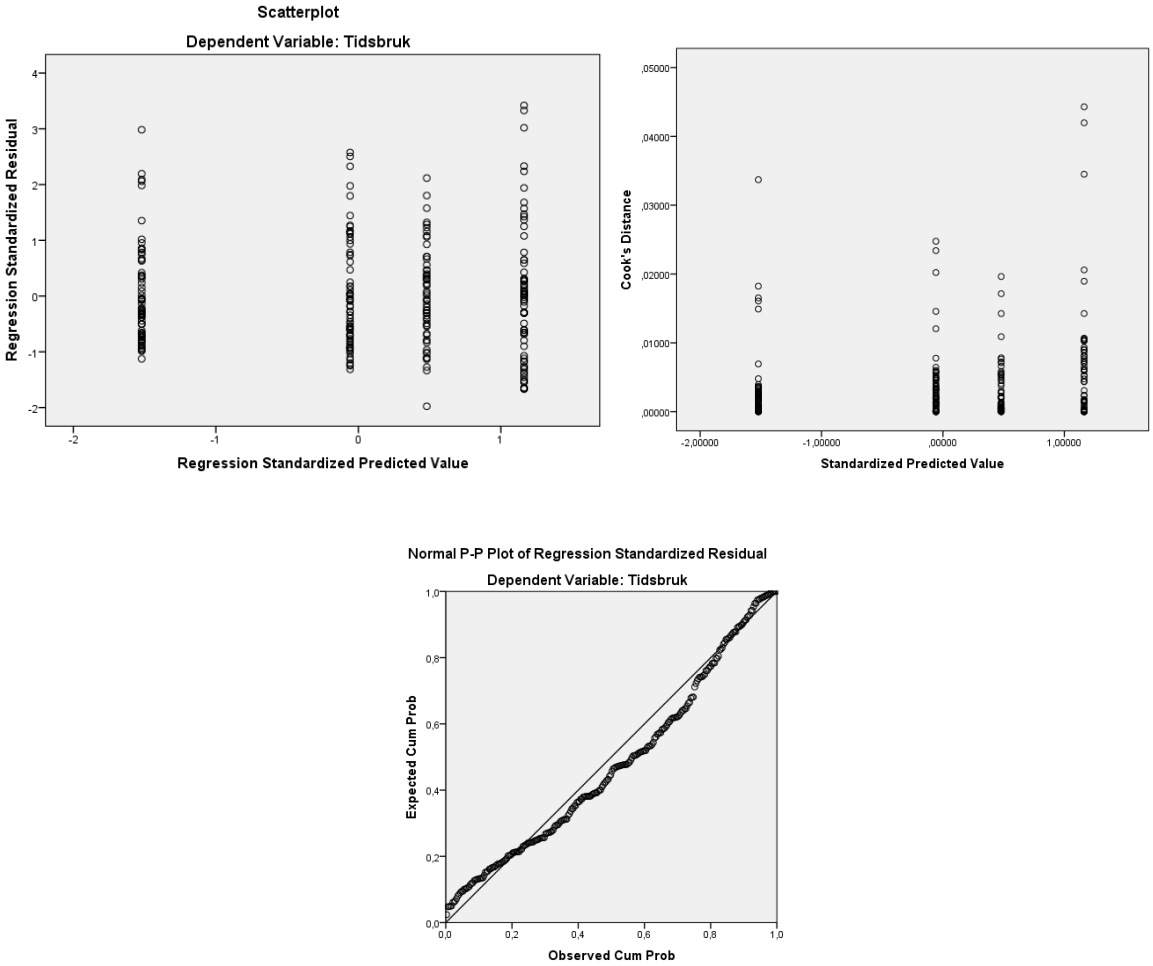
Figuren 18 avslører ingen spesielle mønstre i residualplott, men man har to verdier som ligger i overkant av 3. Cook-avstandene viser ett tilfelle med verdi over 1,000. Modellen godtas på tross av dette, og man antar normalfordeling da fordelingen følger linjen nokså bra i normalitetsplottet.



Figur 18: Residualplott, Cook-avstander og normalitetsplott for regresjonsmodell for finn-tall-oppgaver

Regresjonsmodell 1

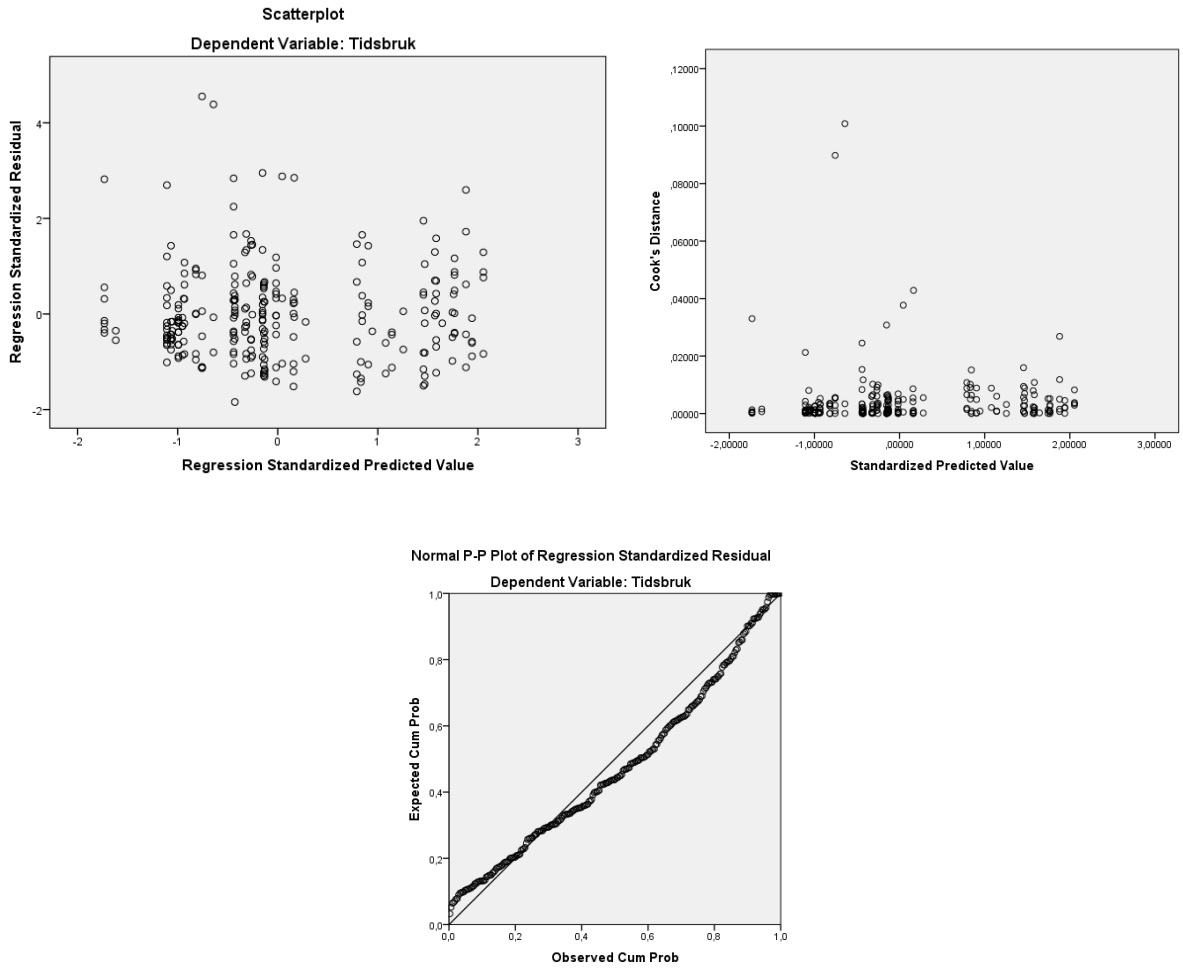
Begge plottene har fire søyler som representerer de fire kategoriene (T1-T4) som er med i regresjonsmodellen. Residualplottet viser ingen spesielle mønstre, men har tre outliers som er over 3. Plottet over Cook-avstandene har ingen verdier over 1,0 og vil dermed godtas. Totalt sett vil jeg derfor påstå at regresjonsmodellen passer bra til de observerte data. En kontroll for normalfordeling av residualer er vist under. Fordelingen følger linjen nokså bra, og det er derfor naturlig å godta en antagelse om normalfordeling.



Figur 19: Residualplott, Cook-avstander og normalitetsplott for regresjonsmodell 1.

Regresjonsmodell 2

Også i modell 2 er det tre outliers med verdi over 3 i residualene. Det er dog ingen spesielle mønstre, og det virker derfor rimelig å anta at residualene har samme varians. Når det gjelder Cook-avstandene, er alle bortsett fra en godt under grenseverdien på 1,0. Figur 20 viser at fordelingen av residualene virker å være nokså normalfordelt. På grunnlag av kontrollen kan man påstå at modellen passer rimelig bra for de observerte data.



Figur 20: Residualplott, Cook-avstander og normalitetsplott for regresjonsmodell 2.

10. Regresjonsmodeller for interaksjonseffekter

I det følgende er det presentert to regresjonsmodeller som vedlegg til punkt 4.2.4 i hoveddelen. Regresjonsmodellene er en videre analyse gjort for å undersøke om kjønn eller inntektsnivå hadde en effekt på interaksjonsleddet «*IQInsentiv*» fra hovedmodellene. I modellene er det derfor laget et interaksjonsledd med henholdsvis «*Kvinne*» og «*Lavinntekt*» og «IQInsentiv».

Kjønn og insentiver

Interaksjonsleddet «*Kvinne_IQInsentiv*» er ikke signifikant med en p-verdi på 0,260. Det kan derfor ikke påstås at kjønn har en effekt på interaksjonsleddet «*IQInsentiv*».

Forklaringsvariabel	
Insentiv	140,7** (71,6)
IQ_test	191,5** (74,5)
Kvinne	8,6 (60,2)
IQInsentiv	- 139,8 (117,4)
Kvinne_IQInsentiv	133,3 (118,1)
Konstant	1000,8*** (57,7)
Antall observasjoner	264
R^2	0,058

Standardfeil i parantes. (*: p-verdi < 0,1, **: p-verdi < 0,05, ***: p-verdi < 0,01)

Tabell 13: Regresjonsmodell for interaksjon mellom kjønn og «IQInsentiv».

Inntektsnivå og insentiver

Interaksjonsleddet «*Lavinntekt_IQInsentiv*» er ikke signifikant med en p-verdi på 0,716. Det kan derfor ikke påstås at inntektsnivå har en effekt på interaksjonsleddet «*IQInsentiv*».

Forklaringsvariabel	
Insentiv	129,9* (70,9)
IQ_test	183,1** (73,7)
Lavinntekt	154,4** (60,3)
IQInsentiv	- 37,9 (112,6)
Lavinntekt_IQInsentiv	- 44,2 (121,1)
Konstant	947,9*** (54,9)
Antall observasjoner	264
R^2	0,078

Standardfeil i parantes. (*: p-verdi < 0,1, **: p-verdi < 0,05, ***: p-verdi < 0,01)

Tabell 14: Regresjonsmodell for interaksjon mellom inntektsnivå og «IQInsentiv».

11. Fullstendig oversikt over eksperimentets innhold

NHH

English

Innledning

Thank you for participating in this study.

The survey is part of a research study at the Norwegian School of Economics (NHH). Please read through the questions carefully and answer them as accurately as possible.
We advice you to keep track of time - you have a maximum of 45 minutes to complete the HIT.

All responses remain anonymous.

Treatment 1: IQ-test med fast betaling

You will now be answering an IQ-test (Intelligence Quotient), which will require about 20 minutes to complete. The test consists of 15 questions. Your answers will not affect the payment or approval. At the end you will get an estimated IQ based on your answers.

Good luck!

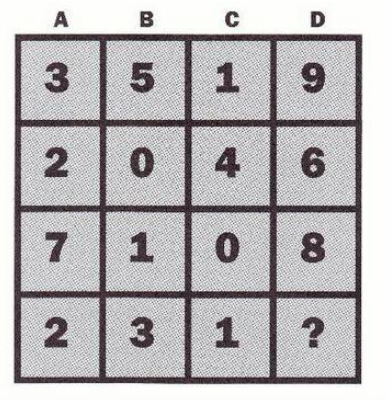
Treatment 2: IQ-test med bonusbetaling

You will now be answering an IQ-test (Intelligence Quotient), which will require about 20 minutes to complete. The test consists of 15 questions. At the end you will get an estimated IQ based on your answers.

You will be rewarded with an additional 0.02 USD for each correct answer. This will be given as a bonus payment in mTurk.

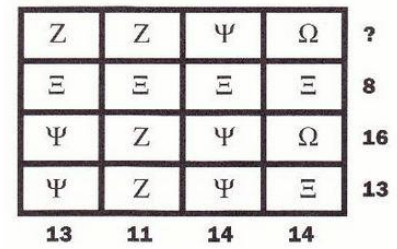
Good luck!

1. Which number should replace the question mark?



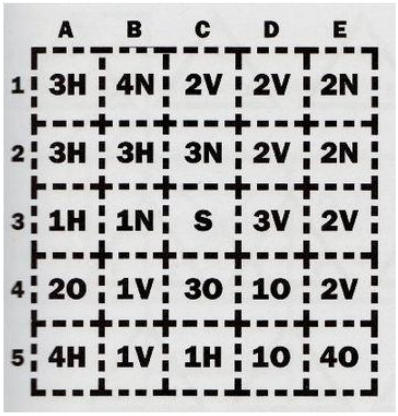
- 8
- 2
- 9
- 1
- 4
- 6

2. Each symbol represents a number. Find the number that should replace the question mark.



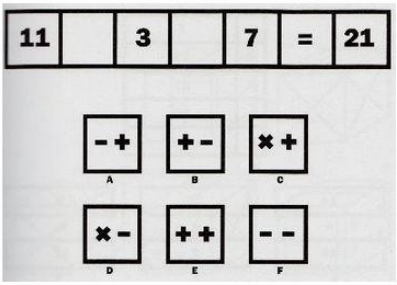
- 6
- 12
- 7
- 15
- 20
- 9

3. Here is an unusual lock. To open it, each button is pressed once in order. The last button is labeled with S. Your task is to find the first button. To check that you are reading the instructions, you are now told that the answer to this task is option 2C. Please check the correct alternative and continue to the next task.



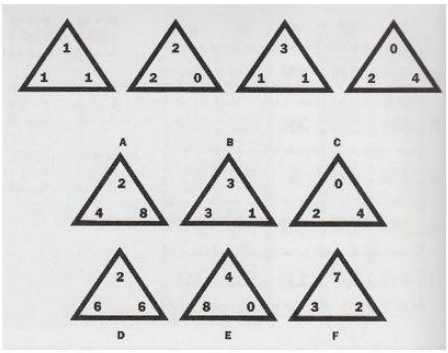
- 5D
- 3C
- 1A
- 4E
- 1B
- 2C

4. Which alternative should be inserted in the blank boxes?



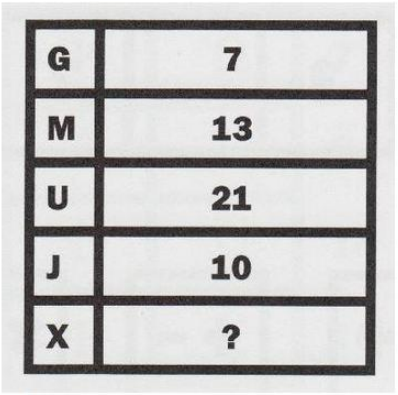
- A
- B
- C
- D
- E
- F

5. Which triangle continues the sequence?



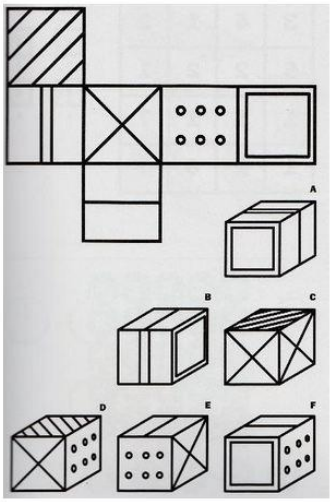
- A
- B
- C
- D
- E
- F

6. Find the relationship between the letters and numbers. Which number should replace the question mark?



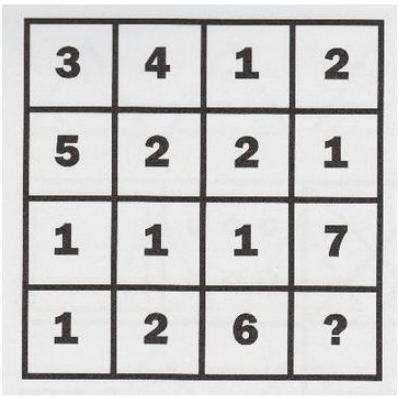
- 14
- 24
- 9
- 26
- 2
- 11

7. Which of the finished boxes can **not** be created from this figure?



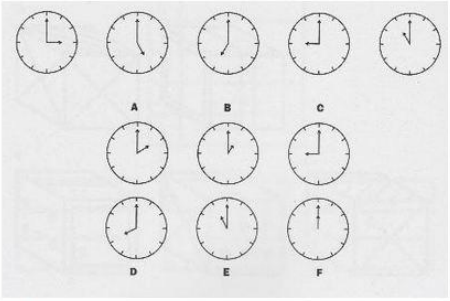
- A
- B
- C
- D
- E
- F

8. Which number should replace the question mark?



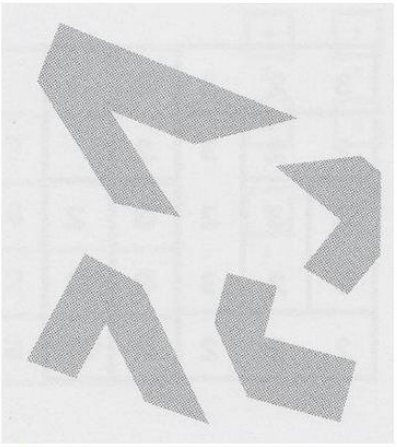
- 3
- 5
- 1
- 6
- 2
- 4

9. Which clock continues the sequence?



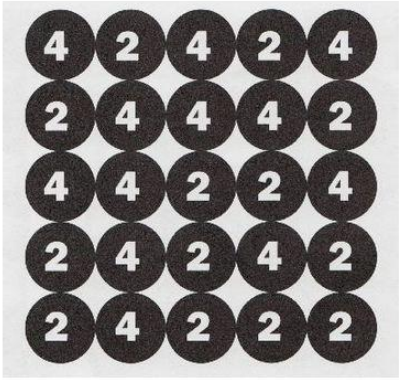
- A
- B
- C
- D
- E
- F

10. Which number can be found by assembling the shapes?



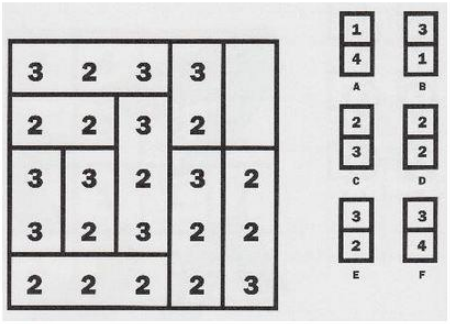
- 2
- 5
- 7
- 6
- 4
- 9

11. Start in the lower left corner and make your way through nine numbers to get to the upper right corner. You can only move vertically and horizontally. Find the route that provides the highest overall sum. What is the highest sum?



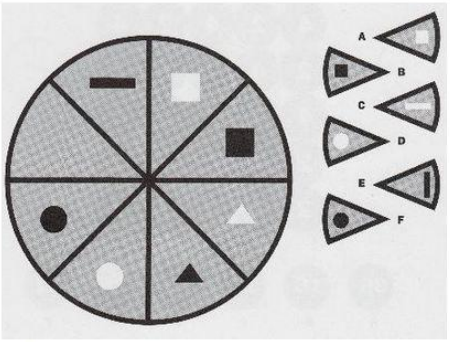
- 36
- 34
- 18
- 45
- 29
- 32

12. This square is made up of a logical pattern. Which piece completes the square?



- A
- B
- C
- D
- E
- F

13. Which piece of pie is missing?



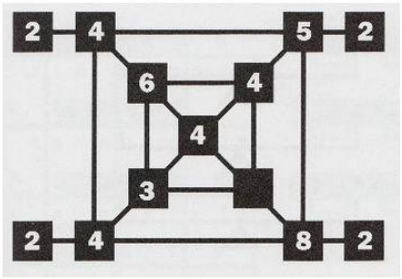
- A
- B
- C
- D
- E
- F

14. All rows, columns, and diagonals should have a sum of 20. Which number should replace the question mark?

5	2		2	5
1		?		1
5	8	4		3
	2	2	2	8
3	2	2	10	3

- 4
- 1
- 3
- 6
- 5
- 2

15. Start in one of the corners and follow the lines. Collect four additional numbers, and add together the five numbers. One of the alternatives underneath should fill the empty box. If you choose the right option, one of the routes that include the empty box gives the sum of 28. Which number should be written in the blank box?



- 4
- 3
- 6
- 1
- 5
- 8

You have now completed the IQ-test. Your score is: ${gr://SC1/Score}

By comparing the score with the table below you can find your estimated IQ. After finishing the survey you will be able to have a look at each answer in detail.

Score	Estimated IQ
15	130
14	125
13	122
12	117
11	115
10	112
9	108
8	105
7	100
6	95
5	90

Treatment 3: Finn tall med fast betaling

In the following, your task is to count the appearance of a specific number in a box. You are then asked to select the correct answer. Your answers will not affect the payment or approval. The results will be given at the end.

Good luck!

Treatment 4: Finn tall med bonusbetaling

In the following, your task is to count the appearance of a specific number in a box. You are then asked to select the correct answer. The results will be given at the end.

You will be rewarded with an additional 0.02 USD for each correct answer. This will be given as a bonus payment in mTurk.

Good luck!

1. How many "5" (five digits) are there in the box underneath?

8	7	7	6	3	2	6	7
6	3	4	9	5	8	7	8
2	6	3	1	6	5	3	8
3	5	4	9	4	4	9	2
9	6	5	6	7	8	7	4
6	9	4	9	9	3	6	3
6	7	8	3	6	1	3	8
6	9	2	8	9	8	5	4
9	2	8	1	3	3	1	3
1	1	8	4	4	6	2	6
2	9	9	1	6	5	1	9
3	7	8	5	1	9	3	1
1	3	1	4	6	1	4	4
8	1	2	8	6	8	9	7
9	3	6	5	8	2	5	8
1	2	2	9	7	3	3	4

- 5
- 7
- 9
- 11
- 13
- 15

6. How many "5" (five digits) are there in the box underneath? (The correct answer is 9)

8	3	6	6	6	5	8	4
2	3	2	4	3	9	2	5
1	7	9	2	1	4	3	6
1	2	1	9	4	4	2	2
4	3	1	6	1	1	7	9
8	7	5	2	3	9	4	6
2	3	3	8	7	8	7	2
7	5	2	7	9	9	7	6
4	3	7	2	4	2	3	8
9	5	7	4	7	6	7	6
7	9	2	6	2	3	6	5
4	2	2	8	6	9	8	4
4	1	7	7	6	9	9	6
2	7	2	6	9	2	2	8
9	4	5	7	5	4	2	3
2	9	5	8	9	9	3	7

- 5
- 7
- 9
- 11
- 13
- 15

Resterende spørsmål i denne del tilsvarer spørsmål 1 (med andre tall).

Your score on this task is ${gr://SC1/Score} out of 15.

After finishing the survey you will be able to study each answer in detail.

Please continue the survey.

Personlige spørsmål

You will now be asked a few personal questions. Please answer them truthfully. The answers will **not** be linked to your profile and will only be used in a pooled analysis with all respondents.

What is your gender?
- Female
- Male

What is your ethnicity?
- White/Caucasian
- African American
- Hispanic
- Asian
- Native American
- Pacific Islander
- Other

How old are you?
- Below 13
- 13-17
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 or above

Where do you live?
- USA
- India
- Great Britain
- Canada
- Other

What is your highest level of completed education?
- Less than High School
- High School / GED
- Some College
- 2-year College Degree
- 4-year College Degree
- Bachelors Degree
- Masters Degree
- Doctoral Degree
- Professional Degree (JD, MD)

What is your annual income range?
- Below $999
- $1,000 - $4,999
- $5,000 - $9,999
- $10,000 - $14,999
- $15,000 - $19,999
- $20,000 - $29,999
- $30,000 - $39,999
- $40,000 - $49,999
- $50,000 - $59,999
- $60,000 - $69,999
- $70,000 - $79,999
- More than $80,000

How interested are you in Math?

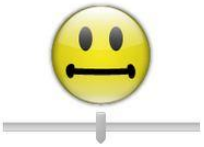
	Not interested at all	Very little interested	Neutral	Somewhat interested	Very interested
My interest in Math	○	○	○	○	○

How often do you take IQ-tests?

	Never	About one time a year	2-4 times a year	5-8 times a year	9-11 times a year	Every month	A few times a month	Every week	More than once a week
I normally take IQ-tests	○	○	○	○	○	○	○	○	○

Avslutning

You are almost done. The remaining questions focus on your perception of the survey. Please answer them honestly. There are no correct answers in this part.



How much fun did you have solving the tasks in this survey? Please use the sliding scale above to answer.

Please consider each statement below.

	Strongly disagree	Disagree	Somewhat disagree	Neutral	Somewhat agree	Agree	Strongly agree
The tasks were difficult	○	○	○	○	○	○	○
I have a feeling of mastery after solving the tasks	○	○	○	○	○	○	○
The tasks were meaningful	○	○	○	○	○	○	○
I had to think hard to solve the tasks	○	○	○	○	○	○	○
The tasks felt repeating	○	○	○	○	○	○	○
The tasks were interesting	○	○	○	○	○	○	○

Which of the reward systems below would you choose for a hypothetical job?

- ○ Fixed income of 1 000 USD regardless of performance
- ○ Income between 500 USD and 1 500 USD depending on performance

How much effort did you have to put in solving the tasks in this questionnaire? Please use the sliding scale below to answer.
(0 = very low effort, 10 = very high effort). Be honest!

Do you think you did better, equal or worse than the average person in solving the tasks in the first part of this survey?

- ○ Better
- ○ About the Same
- ○ Worse

How fair do you consider the pay for this task? (1 star = very unfair, 5 stars = very fair)

Fairness of pay ★★★☆☆ **3**

What do you think the purpose of this survey was?

Thank you for participating in this study.

Your validation code for mTurk is *${e://Field/mTurkCode}*

Please write down this code, as well as copying it into the confirmation code box in Amazon Mechanical Turk. It is your proof of completion.

Please press continue button >>> one more time to finish and be eligible for payment!
You will then also be able to see your results.

If you have any questions, please contact us on: *mturk.nhh@hotmail.com*

Litteraturliste

Amazon Web Services, 26. januar 2011. *MTurk CENSUS: About how many workers were on Mechanical Turk in 2010?* Tilgjengelig fra: <https://forums.aws.amazon.com/thread.jspa?threadID=58891> [Nedlastet 30. januar 2013]

Ariely, Dan, Emir Kamenica, Drazen Prelec. 2008. *Man's search for meaning: The case of Legos.* Journal of Economic Behaviour & Organization, 67 (2008): side 671-677.

Berinsky, Adam J., Gergory A. Huber og Gabriel S. Lenz. 2012. *Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk*. Oxford University Press: Political Analysis, 2. mars 2012, side: 351-368.

Bragelien, Iver. 2005. *10 bonustabber – Hvordan lære av teori og praksis?* Praktisk økonomi & finans, nr. 2, 2005. Side: 25-35.

Cappelen, Alexander Wright og Bertil Tungodden. 2012. *Adferdsøkonomi og økonomiske eksperimenter*. Magma, nr. 5, 2012: side 26-30.

Eckel, Catherine C. og Philip J. Grossman. 2008. *Men, women and risk aversion: experimental evidence.* Handbook of Experimental Economics Results, Volume 1, 2008. Side: 1061-1073.

Frey, Bruno S. 1997. *Not just for the Money – An Economic Theory of Personal Motivation*. Cheltenham: Edward Elgar Publishing Limited.

Frey, Bruno S. og Reto Jegen. 2001. *Motivation Crowding Theory*. Journal of Economic Surveys, Vol. 15, No. 5: side 589-611.

Frey, Bruno S. og Felix Oberholzer-Gee. 1997. *The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out*. The American Economic Review, vol. 87, no. 4, september 1997, side: 746-755.

Gale, Harold og Carolyn Skitt. 1994. *Mensa Boost Your IQ.* Carlton Books Limited, 1994.

Gneezy, Uri og Aldo Rustichini. 2000a. *Pay Enough Or Don't Pay At All*. The Quarterly Journal of Economics, august 2000: side 791-810

Gneezy, Uri og Aldo Rustichini. 2000b. *A Fine is a Price*. Journal of Legal Studies, vol. XXIX, januar 2000: side 1-18.

Hackman, Richard J. og Greg R. Oldham. 1976. *Motivation through the Design of Work: Test of a Theory*. Academic Press, Organizational Behavior and Human Performance 16, 1976, side: 250-279.

Howe, Jeff. 2006. *The Rise of Crowdsourcing*. Wired Magazine, Issue 14.06, Juni 2006: side 1-5

Huang, Eric, Haoqi Zhang, Dacid C. Parkes, Krzysztof Z. Gajos og Yiling Chen. 2010. *Toward Automatic Task Design: A Progress Report*. ACM.

Kaspersen, Line. 2013. *Boom i online-jobbing – og dette er bare starten*. DN.no, 11. februar 2013. Tilgjengelig fra: <http://www.dn.no/karriere/article2560563.ece> [Nedlastet 13. februar 2013]

Kaufmann, Nicolas, Thimo Schulze og Daniel Veit. 2011. *More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk*. Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, Michigan August 4.-7. 2011.

Mason, Winter og Siddharth Suri. 2011. *Conducting behavioral research on Amazon's Mechanical Turk*. Psychonomic Society, Inc. 2011.

Mturk.com. 2013. *Worker Web Site FAQs*. Amazon Mechanical Turk. Tilgjengelig fra: <https://www.mturk.com/mturk/help?helpPage=worker> [Nedlastet 31. januar 2013]

Norges Handelshøyskole, NHH. 2013. *Økonomisk styring/Business Analysis and Performance Management (BUS)*. Tilgjengelig fra: <http://www.nhh.no/Default.aspx?ID=3089> [Nedlastet 20. januar 2013]

Normann, Terje. 2012. *Oppgaver og kolleger viktigere enn lønn*. e24 Jobb, 06. november 2012. Tilgjengelig fra: < http://e24.no/jobb/undersoekelse-oppgaver-og-kolleger-viktigere-enn-loenn/20296061> [Nedlastet 11. februar 2013]

Pink, Daniel H. 2009. *Drive – The surprising truth about what motivates us*. Penguin Group, New York.

Ross, Stephen A. 1973. *The Economic Theory of Agency: The Prinsipal's Problem*. The American Economic Association, vol. 63, nr. 2, mai 1973: side 134-139.

RSA Animate. 2010. *Drive: The surprising truth about what motivates us*. YouTube, 1. april 2010. Tilgjengelig fra <http://www.youtube.com/watch?v=u6XAPnuFjJc> [Nedlastet 8. februar 2013]

Ryan, Richard M. og Edward L. Deci. 2000. *Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions*. Contemporary Educational Psychology 25, 2000, side: 54-67.

Tanner, Liz. 2013. *Qualtrics Crushes 2012 with Record Client Growth and more than a billion surveys served*. Qualtrics, 31. januar 2013. Tilgjengelig fra: <http://www.qualtrics.com/qualtrics-crushes-2012> [Nedlastet 4. februar 2013]

Titmuss, Richard M. 1970. *The gift relationship: from human blood to social policy*. Allen & Unwin, London.

United States Census Bureau. 2013. *State @ County Quick Facts*, 14. mars 2013. Tilgjengelig fra: <http://quickfacts.census.gov/qfd/states/00000.html> [Nedlastet 11. april 2013]

Weibel, Antoinette, Katja Rost og Margit Osterloh. 2009. *Pay for Performance in the Public Sector – Benefits and (Hidden) Costs*. Oxford University Press, 27. mai, 2009: side 387-412