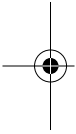


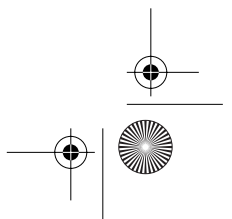
**Marit Kjærnsli, Svein Lie,  
Rolf Vegar Olsen og Astrid Roe**

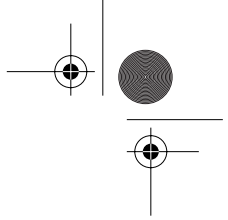
# **TID FOR TUNGE LØFT**

Norske elevers kompetanse i naturfag,  
lesing og matematikk i PISA 2006



Universitetsforlaget





© Universitetsforlaget 2007

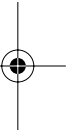
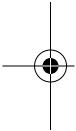
ISBN 978-82-15-01146-2

Materialet i denne publikasjonen er omfattet av åndsverklovens bestemmelser. Uten særskilt avtale med rettighetshaverne er enhver eksemplarfremstilling og tilgjengeliggjøring bare tillatt i den utstrekning det er hjemlet i lov eller tillatt gjennom avtale med Kopinor, interesseorgan for rettighetshavere til åndsverk. Utnyttelse i strid med lov eller avtale kan medføre erstatningsansvar og inndragning og kan straffes med bøter eller fengsel.

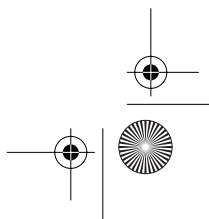
Henvendelser om denne utgivelsen kan rettes til:

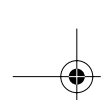
Universitetsforlaget AS  
Postboks 508 Sentrum  
0105 Oslo

[www.universitetsforlaget.no](http://www.universitetsforlaget.no)



Omslag: AIT Trykk Otta AS  
Sats: Laboremus Prepress AS  
Trykk og innbinding: AIT Trykk Otta AS  
Boken er satt med: Times 11/13 pkt.  
Papir: 90 g G-print





## Vedlegg 2

# Metodisk grunnlag

## Innledning

I stedet for å forklare underveis de ulike statistiske og psykometriske metodene og terminologien som er brukt, og hva de forskjellige resultatene betyr, har vi her samlet slik informasjon og noen steder i boka referert til dette under gjennomgangen av resultatene. I stor grad bygger dette på det som var kapittel 5 i rapporten fra PISA 2000 (Lie mfl. 2001). Alt etter nysgjerrighet og egne forutsetninger kan leseren selv vurdere behovet for eller ønsket om å forstå mer av hvordan resultatene er framkommet. Vedlegget starter med noen sentrale statistiske begreper som for mange er godt kjent. Videre kommer forsøk på å forklare på en enkel måte noen av de mer avanserte metodene som er brukt i PISA. Det dreier seg særlig om den såkalte Rasch-modellen, en metode for beregning av skårverdier, og om metoder for å beregne feilmarginer for de gjennomsnittsverdiene som regnes ut. Vedlegget utdyper også i noe detalj bakgrunnen for de prestasjonsnivåene som benyttes for å presentere resultater.

## Litt deskriptiv statistikk

### Kvartiler og prosentiler. Intervallvariabler

En viktig forutsetning for å kunne beregne et gjennomsnitt er at et slikt gjennomsnitt har mening. Dette krever at den aktuelle variabelen vi skal ta gjennomsnittet av, er en såkalt *intervallvariabel*. Dette betyr at hvis det er flere mulige verdier (flere punkter på skalaen) enn 2, så må det være like store avstander mellom punktene langs skalaen. For eksempel må det være like stor forskjell «i virkeligheten» mellom verdiene 0 og 1 som mellom 1 og 2. Det hender ofte at kravet til å være intervallvariabel bare nesten er oppfylt, og i slike tilfeller gjør vi da en tilnærming når vi går fram som om kravet er oppfylt. I denne rapporten er det flere slike eksempler på det vi kan kalle *kvasiintervallvariabler*. Typisk gjelder det for såkalt Likert-skala, som har svaralternativer av typen fra «svært uenig» via «uenig» og «enig» til «svært enig»; her kodes svarene fra 1 til 4. Dette har vi behandlet som intervallvariabler mange steder i rapporten uten videre kommentarer. Vi sammenlikner da grupper av elever ut fra hvilket gjennomsnitt de har



for denne variabelen. I andre mer problematiske tilfeller vil vi kommentere de tilnærminger vi gjør.

I noen tilfeller, for eksempel hvis kravet ovenfor ikke er oppfylt, brukes heller *medianen* som et mål. Dette er den verdien i et datamateriale som deler respondentene i to like store deler. I vårt tilfelle vil medianen være den måleverdien som er slik at halvparten av elevene har høyere og den andre halvparten lavere verdi enn nevnte måleverdi. På tilsvarende måte som for median kan man definere andre *prosentiler* ut fra den prosentandelen som har en lavere verdi enn den angitte. Den 5. prosentilen betyr altså at 5 % av personene har en lavere verdi enn den angitte. Hvis vi deler personene i fire like store deler, så vil de tre variabelverdiene som sørger for dette, være de såkalte *kvartiler*. Første og tredje kvartil er henholdsvis 25. og 75. prosentilen, mens den andre kvartilen er identisk med medianen, eller om vi vil, 50. prosentilen.

### Varians og standardavvik

De vanligste målene for spredningen i et datamateriale er standardavvik og varians. Med *varians* menes det gjennomsnittlige kvadratiske avviket fra gjennomsnittsverdien. For en målt variabel  $X$  kan vi finne variansen,  $s^2$ , ved hjelp av formelen:

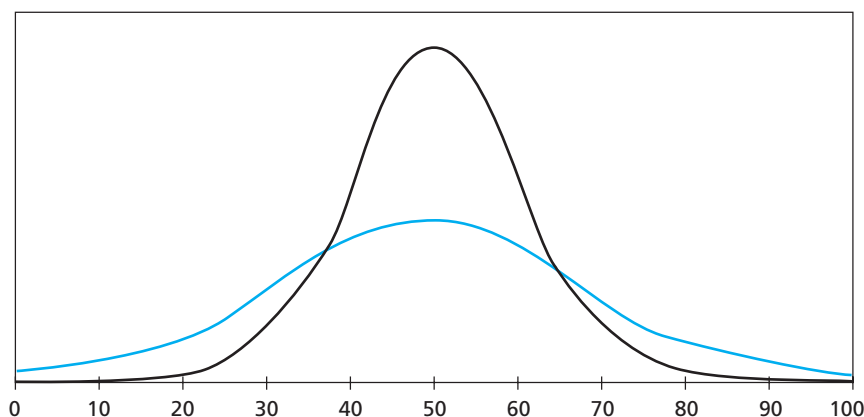
$$s^2 = \frac{(x_1 - \bar{X})^2 + (x_2 - \bar{X})^2 + \dots + (x_n - \bar{X})^2}{N}$$

Her står  $x_1, x_2, x_3$  osv. for de målte verdiene for person nummer 1, 2, 3 osv., mens  $\bar{X}$  står for gjennomsnittsverdien for alle personene. Antall personer er  $N$ . (At det i mange tilfeller er mer riktig å sette  $N - 1$  i nevneren i stedet for  $N$ , er i vår sammenheng helt uvesentlig.) Vi ser at jo større spredningen i dataene er, jo større vil avvikene fra gjennomsnittsverdien bli, og jo større blir variansen.

Kvadratrot av variansen er *standardavviket* ( $s$ ), og her ser vi grunnen til at vi bruker et kvadrattall som symbol for varians. Standardavviket er en størrelse som har samme dimensjon og måleenhet som  $X$  selv, og som derfor lett kan gis en grafisk fortolkning, se det følgende.

### Normalfordeling

Tilfeldighetenes lover bestemmer at når et stort utvalg av personer måles med en intervallvariabel med mange trinn, så vil fordelingen av verdier nærme seg en bestemt fordeling, den såkalte *normalfordelingen*. Fordelingen av antall personer langs skalaen følger av gjennomsnittet og standardavviket for variabelen. I stedet for å angi hele fordelingen er det derfor i mange



Figur 1: To normalfordelingskurver

sammenhenger tilstrekkelig å bare oppgi disse to verdiene. På figur 1 er det vist to eksempler på normalfordelingskurver. De to fordelingene har begge et gjennomsnitt på 50, men de har tydelig ulik spredning. For den svarte kurven er standardavviket 10, mens den blå kurven har standardavviket 20.

To viktige trekk ved normalfordelingene skal nevnes spesielt her. For det første ser vi at fordelingen er symmetrisk om gjennomsnittsverdien. Og videre er det slik at uavhengig av hvor stort standardavviket er, så vil det være en bestemt del av personene som faller innenfor et bestemt intervall målt med standardavviket som enhet. Dersom vi går ett standardavvik ut til begge sider fra gjennomsnittet (for den svarte fra 40 til 60, for den blå fra 30 til 70), så dekker vi omtrent  $2/3$  av personene. På tilsvarende måte faller omtrent 95 % av personene innenfor et intervall på to standardavvik i hver retning (for den svarte fra 30 til 70, for den blå fra 10 til 90). Dette siste er av spesiell interesse siden vi ofte nyttiggjør oss det når vi oppgir såkalte 95 % konfidensintervaller eller i forbindelse med signifikanstesting (mer om dette senere).

### Standardisering

Siden alle normalfordelinger blir identiske når vi måler med standardavviket som mål, er det svært vanlig å gjøre nettopp dette: Målte skårverdier ( $X$ ) regnes om slik at vi i stedet oppgir hvor mange standardavvik ( $s$ ) verdien ligger over eller under gjennomsnittet ( $\bar{X}$ ). Dette kan vi gjøre ved en enkel transformasjon:

$$z = \frac{X - \bar{X}}{s}$$



Vi sier at vi har *standardisert* variabelen  $X$ , og at vi i stedet bruker den standardiserte verdien  $z$ . Gjennomsnittet er da satt til 0 og standardavviket til 1. En slik standardisering er brukt på de aller fleste samlevarene i PISA. Dette innebærer at de angitte verdiene ikke forteller noe direkte om hvilke svar elevene har gitt, siden standardiserte verdier bare har mening ved å fortelle hvordan elever har svart *i forhold til alle andre elever*. En negativ verdi for en standardisert holdningsvariabel betyr altså ikke nødvendigvis en negativ holdning, men snarere en holdning som er mindre positiv enn det gjennomsnittet av elevene har.

Når det gjelder skår for prestasjoner på de faglige testene i PISA, er det standardisert på en litt annen måte. I stedet for å standardisere til gjennomsnitt 0 og standardavvik 1 brukes det for faglig skår en standardisering til gjennomsnitt 500 og standardavvik 100. En skårverdi på 550 betyr da en skår som ligger et halvt standardavvik over gjennomsnittet. For å kunne sammenlikne landene langs samme skala er standardiseringen foretatt internasjonalt. Verdiene for gjennomsnitt og standardavvik er beregnet ved å la alle OECD-landene telle like mye, uavhengig av utvalgenes og hele populasjonenes størrelse. Landene utenfor OECD er i denne sammenheng ikke regnet med.

Standardavviket internasjonalt er altså satt lik 100, men for et enkelt land vil det aktuelle standardavviket avhenge av hvor stor spredning det er blant elevene i forhold til i andre land. Og gjennomsnittet for de nasjonale standardavvikene blir gjerne noe lavere enn 100 fordi det som regel er noe mindre spredning i ett og samme land enn internasjonalt. Dette forhindrer ikke at standardavviket for noen enkeltland ligger høyere enn 100, og disse landene framstår da med en forholdsvis høy spredning.

### Vanlig (bivariat) korrelasjon

Hittil i dette kapitlet har det dreid seg om *univariat* statistikk, om begreper for en og en variabel om gangen. Når vi beskriver sammenhengen mellom to variabler, snakker vi om *bivariat* statistikk. Noen ganger framstilles slike sammenhenger ved hjelp av en *krystabell* som viser fordelingen når det gjelder kombinasjoner av svar på de to variablene. Men i andre sammenhenger, særlig hvis begge variablene er intervallvariabler, er det gunstigere å kunne angi et tall som et uttrykk for i hvor stor grad de to variablene varierer sammen. Vi snakker da om graden av samvariasjon eller korrelasjon. En *korrelasjonskoeffisient* er et mål på i hvor stor grad de to variablene varierer «i takt», altså i hvor stor grad den ene variabelen har en høy verdi samtidig med (for eksempel for samme elev) når den andre har det, og omvendt. Den vanligste korrelasjonskoeffisienten er den såkalte Pearsons korrelasjonskoeffisient (ofte symbolisert med  $r$ ), og den måler i hvor stor grad de målte dataene faller langs en rett linje når de avtegnes i et koordinatsystem.

Korrelasjonskoeffisienten kan ha verdier fra -1 (perfekt negativ korrelasjon) via 0 (ingen korrelasjon) til 1 (perfekt positiv korrelasjon).

Hvis vi kvadrerer korrelasjonskoeffisienten, får vi et tall som forteller oss hvor stor andel av variansen i den ene variabelen som er felles med den andre variabelen. Ofte sier vi at dette kvadratet forteller oss hvor stor del av variansen i den ene variabelen som kan *forklares* ved den andre variabelen. Et eksempel vil illustrere dette: Hvis en korrelasjonskoeffisient på 0,50 er regnet ut mellom skår på en matematikktest og en variabel som måler positiv holdning til matematikk, kan vi si at holdningsvariabelen kan «forklare» 25 prosent (siden  $0,50^2 = 0,25$ ) av variansen til matematikkskåren. Språkbruken må her ikke oppfattes bokstavelig, idet «forklare» ikke skal oppfattes som at det nødvendigvis er snakk om en kausal sammenheng; årsak og virkning. Og selv om det skulle være det, så kan vi ikke si hva som i tilfelle er årsak, og hva som er virkning. Vi må huske på at det eneste vi vet, er at de to variablene til en viss grad varierer sammen. Noen ganger er det en tredje variabel som påvirker begge to og gjør at disse to korrelerer høyt. La oss ta et eksempel. Når vi finner en høy korrelasjon mellom matematikkskår og antall bøker hjemme, er det lite trolig at noen skårer høyt *fordi* de har alle disse bøkene hjemme, eller at de har alle bøkene *fordi* de er gode i matematikk. Å kjøpe flere bøker til bokhylla vil neppe i seg selv være til særlig hjelp i matematikk! I slike tilfeller er det naturlig å oppfatte familiens «kulturelle kapital» som en viktig underliggende tredje variabel som påvirker begge, altså at både antall bøker i bokhylla og elevens skoleprestasjoner stiger med økende kulturell kapital.

### Multipel korrelasjon og regresjon

På tilsvarende måte som for korrelasjon mellom to variabler kan vi også snakke om multipel korrelasjon mellom en (avhengig) variabel og en gruppe av andre (uavhengige) variabler. En *multipl korrelasjonskoeffisient*,  $R^2$ , angir hvor stor del av variansen til den avhengige variabelen som kan «forklares» ut fra alle de uavhengige variablene til sammen. I denne rapporten er vi både interessert i hvor stor del av variansen til de faglige skårene som kan «forklares» ut fra de ulike holdnings- og bakgrunnsvariablene hver for seg (vanlig korrelasjonskoeffisient) og samlet (multipel korrelasjon). Symbolet  $R^2$  brukes for multipl korrelasjonskoeffisient for å minne om at den svarer til vanlig korrelasjonskoeffisient,  $r$ , opphøyd i annen potens, nemlig andel av variansen som er «forklart» (se ovenfor). Det framgår også av symbolet  $R^2$  at en multipl korrelasjonskoeffisient alltid er positiv. Det følger av definisjonen at den ligger mellom 0 (ingen varians «forklart») og 1 (all varians «forklart»).

En gruppe av uavhengige variabler kan altså brukes sammen for å forklare en avhengig variabel. Dette kan de gjøre fordi de alle sammen hver



for seg korrelerer med den avhengige variabelen. Imidlertid korrelerer de også seg imellom. Dette er grunnen til at  $R^2$  ikke er lik, men mindre enn summen av alle de enkelte bidragene ( $r^2$ ) fra hver uavhengig variabel. Hvis vi ønsker å se detaljert på hvor mye de uavhengige variablene *hver for seg* bidrar til å «forklare» den avhengige variabelen, snakker vi om at vi foretar en multippel regresjon. Det viser seg imidlertid at hvor mye hver variabel «bidrar» til dette, i aller høyeste grad er avhengig av hvilke variabler som allerede er tatt med. Grunnen til dette er den samme som nevnt ovenfor, nemlig at de uavhengige variablene også korrelerer seg imellom. Hvilket sett av variabler vi har med i vår *regresjonsmodell*, vil i stor grad bestemme resultatet, og tolkningen av de individuelle bidragene vil være vanskelig. Derfor vil vi i denne rapporten stort sett nøye oss med å gjengi  $R^2$  for hver regresjonsmodell uten å diskutere de enkelte bidragene. I så måte er multippel regresjon og multippel korrelasjon det samme.

## Å slutte fra utvalg til populasjon

### Populasjon og utvalg

Når vi gjør en undersøkelse slik som i PISA, lar vi et *utvalg* av norske elever delta i undersøkelsen, men de deltakende elevene er trukket ut fra en *populasjon*, i vårt tilfelle alle 15-årige skoleelever i Norge. Elevene i utvalget representerer hele populasjonen, og de er trukket ut på en slik måte at alle elevene i populasjonen har en viss kjent sannsynlighet for å bli med i utvalget, men ikke nødvendigvis den samme sannsynligheten for alle. Vi er i utgangspunktet ikke spesielt interessert i resultatene for elevene i utvalget, men derimot i hvilken grad de kan fortelle oss om situasjonen for hele populasjonen. Når vi for eksempel finner at guttene skårer lavere enn jentene på leseforståelse, så er det et viktig spørsmål om vi kan slutte fra resultatene for utvalget til at guttene i gjennomsnitt har dårligere leseforståelse enn jentene *også i hele populasjonen*, altså for landet som helhet. Signifikans og konfidensintervall er i denne sammenhengen to viktige begreper.

### Signifikante forskjeller mellom gjennomsnittsverdier

Når vi skal sammenlikne grupper av elever for å se hvem som har høyest gjennomsnittsverdi for en eller annen variabel, spør vi ofte om forskjellen vi finner, er statistisk *signifikant* eller ikke. Vi kan være interessert i om norske elever skårer signifikant lavere i lesing enn svenske, eller om gutter har signifikant mer positiv holdning til matematikk enn jenter. Nøkkelen til å forstå dette begrepet ligger i å innse at enhver uttrekking av utvalget innebærer et element av tilfeldighet. Hadde uttrekkingen av utvalget i PISA blitt foretatt på nytt, ville andre elever blitt med, og resultatene ville vært annerledes enn

de ble. Men statistikkens lover sørger for at det er en grense for hvor forskjellig de enkelte utvalgene kunne ha blitt, og denne grensen er lavere jo flere som er med i utvalget. Vi knytter gjerne en bestemt sannsynlighet til overveielene nevnt over. Ofte bruker vi 5 % sannsynlighet som et kriterium for vårt *signifikansnivå*, og det har vi brukt i denne boka også. For å vende tilbake til PISA og kjønnsforskjeller i leseforståelse: Når jentene skårer signifikant høyere enn guttene, mener vi med det at det er liten sannsynlighet for at denne forskjellen bare skyldes tilfeldigheter ved de uttrukne elevene.

### Feilmargin, konfidensintervall og standardfeil

Noen ganger ønsker vi å sammenlikne flere elevgrupper på en gang, og vi vil gjerne angi den *feilmarginen* vi har når det gjelder å anslå gjennomsnittsverdier for hele populasjonen ut fra gjennomsnittet for utvalget. Et 95 % *konfidensintervall for gjennomsnitt* angir det intervallet som gjennomsnittet for hele populasjonen med 95 % sannsynlighet ligger innenfor. Andre prosent forekommer, for eksempel 99 %, men i denne boka bruker vi konsekvent 95 %.

Ved hjelp av diagrammer med konfidensintervaller kan vi visuelt få et raskt inntrykk av hvilke forskjeller som tydelig er signifikante, og hvilke som åpenbart ikke er det. Hvis konfidensintervallene ikke dekker hverandre, er det et tegn på at forskjellene er signifikante. Tilsvarende, hvis de i stor grad overlapper hverandre, er forskjellene ikke signifikante. (Det kan være svak overlapping og likevel signifikant forskjell.)

Når vi skal bestemme hvor store feilmarginer vi må operere med (eller størrelsen på konfidensintervallene), brukes ofte begrepet *standardfeil*. Vi tenker oss at vi mange ganger trekker et tilfeldig utvalg fra en og samme populasjon. Gjennomsnittsverdien vil ikke bli den samme hver gang, men den vil variere litt rundt gjennomsnittet for hele populasjonen. Jo større utvalgene er, jo mindre spredning vil det bli for disse gjennomsnittene. Disse gjennomsnittene danner selv en fordeling, og standardavviket til denne fordelingen kalles *standardfeilen* (egentlig standardfeilen til gjennomsnittet) og forkortes ofte *SE*, på engelsk Standard Error (of the mean). Standardfeilen (til gjennomsnittet) kan uttrykkes ved hjelp av standardavviket  $s$  (til fordelingen) og antall personer i utvalget ( $N$ ):

$$SE = \frac{s}{\sqrt{N}}$$

Vi ser altså at jo større utvalget er, jo mindre blir standardfeilen og dermed konfidensintervallet. Ved å ha et stort utvalg kan vi oppnå stor presisjon ved bestemmelse av populasjonsgjennomsnittet. Derimot inngår ikke



populasjonens størrelse i dette uttrykket. Dette er grunnen til at for å oppnå samme grad av presisjon i PISA er det ikke nødvendig å ha et større utvalg elever i store land enn i små.

Siden to standardavvik i hver retning dekker 95 % av tilfellene, vil konfidensintervallet (95 %) for gjennomsnittet bestå av det intervallet vi får når vi går to standardfeil i hver retning ut fra gjennomsnittsverdien. Når standardfeilene til gjennomsnittsverdier er gitt, kan vi altså lett selv regne ut feilmarginene: Feilmarginene utgjør alltid to standardfeil.

I resultatdelen av denne rapporten har vi stort sett gitt feilmarginer for de oppgitte gjennomsnittsverdiene. Men for å unngå en altfor omstendelig presentasjon når det gjelder dette, har vi ofte nøyd oss med litt summariske kommentarer om hvor store feilmarginer vi må regne med for en gruppe av variabler og elever sett under ett.

### Standardfeil og konfidensintervall for prosenttall

Når vi skal slutte fra utvalget av elever til hele populasjonen, vil det, som beskrevet ovenfor, bli en usikkerhet som skyldes tilfeldigheten ved trekking av skoler og elever. Gjennomsnittsverdiene av de ulike skårverdiene blir derfor angitt med standardfeil (eller feilmarginer). På samme måte forholder det seg med prosenttall. I slike sammenhenger er det en enkel sammenheng mellom standardfeilen til prosenttallet og antall elever:

$$SE = \sqrt{\frac{p(100-p)}{N}}$$

Der  $p$  er prosentandelen og  $N$  er antall elever.

Et eksempel vil belyse dette. Hvis vi trekker 400 elever tilfeldig fra en populasjon, og 70 % av dem svarer riktig på en oppgave, vil standardfeilen i dette tilfellet bli:

$$SE = \sqrt{\frac{70 \cdot 30}{400}} = 2,3$$

Feilmarginen blir dobbelt så stor, eller 4,6 prosentpoeng (her brukes ikke prosent, for det kan misforstås), og det betyr at konfidensintervallet går omtrent fra 65 % til 75 %. Med 95 % sannsynlighet ligger prosenttallet som kan svare riktig i hele populasjonen, mellom disse grensene. Slike feilmarginer for prosenttall har vi flere steder oppgitt, ofte nokså summarisk, for ikke å bli fristet til å legge vekt på forskjeller i prosentandeler som ikke er signifikante.



### Signifikante vs. store forskjeller. Effektstørrelse

Som nevnt må målte forskjeller være signifikante for å være interessante. Men at målte forskjeller mellom elevgrupper er signifikante, er i seg selv ikke noen garanti for at de er store, interessante eller viktige. Jo flere som deltar i en undersøkelse, jo mindre blir feilmarginene, og jo lettere er det at målte forskjeller er signifikante. Men de blir ikke større av den grunn. Målte forskjeller kan være signifikante fordi de er store, og/eller fordi mange er med i undersøkelsen. I en storskala undersøkelse som PISA, med flere tusen elever i utvalget, blir selv små og ubetydelige forskjeller signifikante i statistisk forstand. Det ligger derfor en viss fare forbundet med å rapportere funn som i og for seg er signifikante, men som fra et pedagogisk eller skolepolitisk synspunkt er totalt uvesentlige.

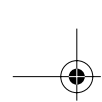
Det er derfor ofte naturlig å også kunne angi et mål for hvor «store» forskjeller mellom elevgrupper er. Hva som er «stor» forskjell for en bestemt variabel, avhenger av hvor stor spredning det er i materialet som helhet. En vanlig måte er å angi forskjellene som *standardiserte forskjeller*, også kalt *effektstørrelse*: hvor stor forskjellen er i forhold til ett standardavvik.

Fordelen ved å måle for eksempel kjønnsforskjeller i effektstørrelse er at et slikt mål kan brukes til å sammenlikne resultater for helt ulike variabler på en meningsfull måte. Hvis jentene skårer 10 poeng høyere enn guttene på en lesetest, sier det oss ingen ting om hvor «stor» denne forskjellen egentlig er. For å kunne vurdere det måtte vi vite hvordan fordelingen av poengsummer er. Hvis standardavviket for denne lesetesten er 20 poeng, sier vi at effektstørrelsen er på  $10/20 = 0,5$ . Tilsvarende, hvis guttene skårer 10 poeng høyere på en matematikktest der standardavviket er på 50 poeng, blir effektstørrelsen i guttenes favør 0,2, altså mye lavere. Kjønnsforskjellen, og dermed den pedagogiske betydningen av resultatet, er mye større i det første tilfellet.

### Signifikante vs. store korrelasjoner

Også for korrelasjoner mellom to variabler kan vi snakke om signifikans. En positiv korrelasjon mellom to variabler innebærer at det er en tendens til at høy verdi for den ene variabelen og høy verdi for den andre variabelen opptrer for de samme elevene. Men dette gjelder for elevene som var med i utvalget. Hvis denne korrelasjonskoeffisienten er *signifikant positiv*, så betyr dette at den med stor (minst 95 %) sannsynlighet også er positiv for hele populasjonen. Den positive korrelasjonen kan da ikke med rimelig sannsynlighet tilskrives bare en tilfeldighet ved utvalget som ble trukket ut.

På samme måte som når det gjelder forskjeller mellom gjennomsnittsverdier, kan korrelasjonene godt bli signifikant positive selv om de er små ( $< 0,10$ ), bare vi har mange elever med i undersøkelsen. Hva som regnes som store og små korrelasjoner, må vurderes i hvert tilfelle. Den pedago-



giske eller politiske betydningen av en korrelasjon avhenger både av dens størrelse og av hva det saklig sett dreier seg om.

### Gruppeutvelging og designeffekt

De vurderingene vi hittil har presentert når det gjelder standardfeil og dermed konfidensintervaller for gjennomsnitt og prosenttall, forutsetter egentlig at alle elevene er trukket ut enkeltvis. Slik er det imidlertid ikke i PISA, og slik er det nesten aldri i andre undersøkelser i skolen heller. Uttrekkingen i PISA har foregått ved en såkalt gruppeutvelging, ved at det først er trukket ut skoler, og deretter er 30 elever trukket ut på hver deltaker-skole. Dersom det er færre enn 30 aktuelle ved en skole, er alle med. Elever som går på samme skole, har imidlertid – i større grad enn om de hadde vært trukket enkeltvis – noen egenskaper til felles ved at de blant annet bor på samme sted, kommer fra liknende hjemmebakgrunn og har opplevd samme type undervisning. Spredningen av verdier for de enkelte variablene blir derfor kunstig lav, noe som gjør at vanlige utregninger fører til for lave standardfeil og feilmarginer. Uten at vi vil gjøre noe stort nummer av dette i denne rapporten, har vi derfor måttet korrigere for denne såkalte *designeffekten* ved å øke feilmarginene noe.

Prosedyrene for å gjøre dette på en rigorøs måte er svært kompliserte og omstendelige. I PISA er det brukt en metode som kalles «Balanced Repeated Replication Method». Uten å gå i for mange detaljer kan vi beskrive denne metoden slik: På kunstig måte lages det mange alternative utvalg fra det originale utvalget av skoler. Dette gjøres ved å gi hver skole ulik vekt hver gang. I alt er det et sett med 80 forskjellige vekter for hver skole. Korrekte beregninger av standardfeil tar altså utgangspunkt i 80 beregninger av gjennomsnittsverdier og hvordan disse varierer. For en detaljert beskrivelse av dette, se OECD (2005a, kap. 3).

Hvor store disse designeffektene er for vårt utvalg av elever, avhenger av hvilke variabler det dreier seg om. De gjør seg aller sterkest gjeldende for spørsmål som har med forholdene på skolen som helhet å gjøre, siden elever på samme skole her i stor grad har samme erfaring. Designeffekten kan i slike tilfeller komme helt opp i en faktor 4, og det betyr at mens det virkelige antallet elever som svarte på et spørsmål, er ca. 4000, så svarer det til at vi har et *effektivt* utvalg på bare 1000. Feilmarginene for gjennomsnittsverdier og prosenttall for slike variabler må da justeres opp med en faktor  $4^{1/2} = 2$ . Også for skårverdier for kunnskap har vi måttet justere standardfeilene, men utslagene er ikke like store for slike variabler. I vårt land dreier det seg typisk om å justere opp standardfeilene med en faktor 1,5. Men i land der forskjellene mellom skoler er større, særlig der det er «streaming» mellom skoleslag på det aktuelle alderstrinnet, kan designeffekten bli vesentlig større.



## Utvalgssannsynligheter og vekting

I alle resonnementene ovenfor er det antatt at elevene ble trukket ut med like stor sannsynlighet, men så enkelt er det ikke. De elevene som deltok i undersøkelsen, skulle representere alle norske elever i hele populasjonen. Men for at de skulle kunne gjøre det på beste måte, måtte vi tillegge hver elev *en vekt som forteller hvor mange elever i hele populasjonen eleven representerer*. Den vekten en elev tillegges, avhenger av hvilken sannsynlighet eleven ble trukket ut, og av hvor mange andre uttrukne elever på den samme skolen og i samme stratum som faktisk deltok. For de fleste elevene er vektene nokså like. De er gjennomsnittlig omkring 12, siden det er omtrent 12 ganger så mange elever i hele populasjonen som i vårt utvalg. For mer informasjon om vekting av elever i PISA, se OECD (2005a, kap. 2).

Som beskrevet i vedlegg 1 ble populasjonen først delt inn i fire strata, og deretter ble elever trukket fra hvert stratum for seg. For å effektivisere utvalget (for ikke å få for mange skoler med få elever) ble stratum 3 (små skoler) og 4 (videregående skoler) *undersamlet*. De uttrukne elevene i disse strataene hadde derfor små sjanser til å bli trukket ut. Under utregningene må de til gjengjeld tillegges høyere vekt fordi de representerer så mange andre elever. Selv innen samme stratum vil elevene på ulike skoler ikke ha samme sannsynlighet for å bli trukket ut. Trekkingen innen hvert stratum foregikk på den måten at store skoler hadde større sannsynlighet for å bli trukket ut enn små. Og videre: Hvis en uttrukket elev var fraværende under testen, måtte de andre deltakende elevene på skolen tillegges en litt høyere vekt siden de hver for seg skal representere litt flere.

Ved beregning av gjennomsnittsverdier og prosenttall i denne rapporten er det hele tida brukt vektete verdier for elevdata. For Norge blir resultatene om vi bruker vektete eller uvektete data (alle elevene teller like mye) nokså like, og unntatt i noen få tilfeller innebærer det ingen vesentlig forskjell når det gjelder kvalitativ tolkning av resultatene. I andre land som har trukket elever annerledes, kan imidlertid vekting av dataene være helt avgjørende for å få meningsfulle landsgjennomsnitt.

## Konstrukter som samlevariabler. Reliabilitet

Både PISA-testen og spørreskjemaene inneholder en rekke enkeltvariabler. Noen av disse er ment å fungere separat (som for eksempel kjønn), men de fleste av dem inngår i en *samlevariabel* (eller *indeks*), og vi sier at denne samlevariabelen representerer skårverdier for et *konstrukt*. Vi beregner skårverdier for prestasjoner ved hjelp av en rekke oppgaver for derved å oppnå en tilstrekkelig høy *reliabilitet*. Hadde vi bare brukt noen få oppgaver, ville det vært altfor store tilfeldigheter med hensyn til hvor godt opp-



gavene passet den enkelte elev eller elevgruppe. Slik er det også når det gjelder konstrukter av typen hjemmets sosioøkonomiske status eller elevens selvoppfatning. Alle slike konstrukter er målt ved hjelp av et sett av variabler, og det er et ufravikelig krav at disse variablene må støtte opp om hverandre, at de viser en rimelig høy indre konsistens. Jo lavere konsistens (eller om vi vil, jo mer forskjellig enkeltspørsmålene er), jo flere spørsmål må vi ta med for å få tilstrekkelig høy reliabilitet. Dette er grunnen til at vi i en faglig test trenger mange oppgaver, mens det holder med noen få (typisk 3–5) spørsmål for et holdningskonstrukt. Denne boka inneholder mange eksempler på konstrukter. I hvert tilfelle vil vi angi hvilke konkrete spørsmål til elevene som inngår i konstruktet. For alle konstruktene som er brukt i denne rapporten, er reliabiliteten beregnet til å være god nok (alfa rundt eller over 0,70, se nedenfor).

Men hva er «god nok» reliabilitet? For å svare på det vil vi først gi en kvalitativ beskrivelse av en reliabilitetskoeffisient i form av det som kalles *Cronbachs alfa*. Vi deler først testen (eller spørsmålene) i to deler og lager en samlevariabel for hver del. Så beregner vi korrelasjonskoeffisienten mellom de to delene. Denne inndelingen i to deler kan vi gjøre på mange måter, og vi får derfor mange korrelasjonskoeffisienter. Gjennomsnittet av alle disse (korrigert for at halvdelene er kortere enn hele testen) gir oss alfa. Vi kan også si at alfa forteller oss hvor stor del av variansen til en variabel som virkelig representerer konstruktet, og hvor mye som simpelthen er tilfeldigheter i valg av oppgaver/enkeltspørsmål. En høy alfa betyr at resultatet for enkeltelever i liten grad bestemmes av nøyaktig hvilke oppgaver/spørsmål som er med i samlevariabelen, så da ville resultatene blitt omtrent de samme om vi byttet ut en oppgave med en annen. En verdi på 0,70 for alfa regnes i mange sammenhenger som en nedre grense for et konstrukt som skal brukes til å sammenlikne grupper av elever. En slik verdi forteller oss at 70 % av variansen (som representerer den informasjonen samlevariabelen gir oss) er «sann varians», mens resten (30 %) er «feilvarians». Begrepet «feilvarians» indikerer ikke at noe er gjort feil, men at det representerer noe annet enn det som er felles for variablene som inngår. Populært sagt: Vi har 70 % sann varians og 30 % «bingo». (Når vi summerer opp slike konstrukter på gruppenivå, vil imidlertid dette være en tilfeldig feilkilde. Gjennomsnitt for en gruppe vil derfor være mer stabilt enn det «bingo»-metaforen indikerer.)

I tilfeller der man tilstreber å sammenlikne enkeltpersoner og grupper av personer med høy presisjon, ligger alfa vanligvis mye høyere enn 0,70. For de ulike skalaene i PISA-testen ligger de over 0,80. Også i tester eller eksamener som får store konsekvenser for enkeltpersoner, bør det være et viktig mål at alfa er høy.

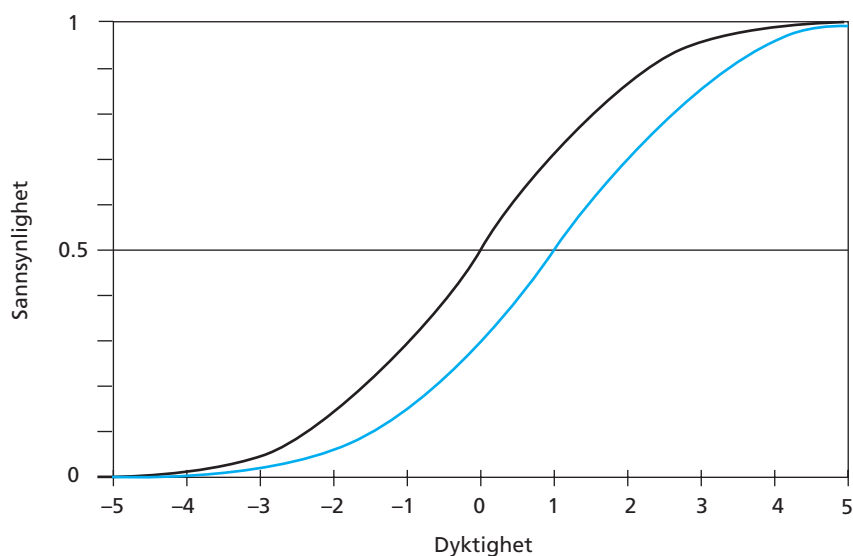
## Rasch-modell

### Internasjonale rapporteringsskalaer for prestasjoner

Som beskrevet i de foregående kapitlene er det beregnet kognitive skårverdier for elevene innenfor seks områder, en for hver av lesing, matematikk og naturfag, og i tillegg en for hver av de tre kompetansene i naturfag. I PISA-prosjektet ble alle skårverdiene beregnet ved hjelp av en såkalt Rasch-modell (oppkalt etter dansken Georg Rasch som utviklet modellen) ut fra antall oppnådde poeng innenfor hvert fagområde. Siden oppgaveheftene er svært forskjellige, går det ikke an å bruke antall riktige svar (råskår) til å sammenlikne elevene direkte. Den klassiske måten å sammenlikne på ville innebære å lage en standardisert skår ved å standardisere råskårene for hvert hefte for seg.

I denne delen vil noen sentrale aspekter ved Rasch-modellen bli forklart, spesielt hvordan den er brukt i PISA-sammenheng. For lesere som vil sette seg mer inn i temaet, finnes det mye spesiellitteratur. I manualen for dataanalyse (OECD 2005a) gir kapittel 4 en god innføring.

I Rasch-modellen bestemmer man hver oppgaves vanskegrad  $v$  og hver elevs dyktighet  $d$  ved hjelp av et sett av likninger som knytter disse to settene av variabler sammen. Modellen består av at det antas at det er en enkel sammenheng mellom en elevs dyktighet innen et fagområde og sannsynligheten for at en bestemt oppgave skal løses riktig. En oppgaves vanskegrad  $v$  er definert som den dyktigheten  $d$  en elev må ha for å ha 50 %



Figur 2: Karakteristiske kurver for to oppgaver

sannsynlighet for å få riktig svar. Fordelingen av sannsynligheter for alle andre verdier for  $d$  er antatt å følge kurver som vist på figur 2, såkalte *karakteristiske kurver* for oppgaver. Sannsynligheten for riktig svar ligger mellom 0 og 1 og øker med økende dyktighet.

Den matematiske formelen for en karakteristisk kurve er:

$$P = \frac{e^{(d-v)}}{1 + e^{(d-v)}}$$

I matematikken kalles en slik sammenheng en *logistisk* funksjon. Som det framgår av formelen, er det altså differansen mellom  $d$  og  $v$  som bestemmer hvor stor sannsynligheten er for riktig svar. For en svært flink elev er  $d$  høy, og eksponenten får en stor positiv verdi, slik at sannsynligheten for at eleven svarer riktig,  $P$ , blir nesten 1. For en svært svak elev får eksponenten en stor negativ verdi, og  $P$  blir nær 0. For en elev med dyktighet lik vanskegraden blir eksponenten lik 0, og sannsynligheten for riktig svar 0,5 eller 50 %. I eksemplet ovenfor (figur 2) kan vi se at en elev med dyktighet lik 0 har sannsynlighet 0,5 for å svare riktig på den «svarte» oppgaven, som dermed har en vanskegrad på 0. Den samme personen har sannsynlighet omtrent 0,25 for å svare riktig på den «blå» oppgaven. En person med dyktighet lik 1 vil ha sannsynlighet omtrent 0,75 for å skåre riktig på den «svarte» oppgaven, mens sannsynligheten for riktig svar på den «blå» oppgaven er 0,5. I henhold til dette sier vi at den «svarte» oppgaven har vanskegrad lik 0 og den «blå» vanskegrad lik 1.

En oppgaves vanskegrad er altså den dyktigheten en person må ha for å ha en sannsynlighet på 0,5 for å svare riktig. På tilsvarende måte er dyktigheten til en person den vanskegraden som en oppgave må ha for at personen skal ha 0,5 sannsynlighet for riktig svar. De to størrelsene  $d$  og  $v$  inngår altså i prinsippet på en symmetrisk måte, og de to sammenhengene gir mening til og definerer begge størrelsene samtidig.

De absolutte verdiene langs  $x$ -aksen kan velges tilfeldig, da det bare er de innbyrdes plasseringene som er vesentlige. Forskjellen i vanskegrad mellom de vanskeligste og de letteste oppgavene blir i praksis omtrent fem–seks enheter, mens nullpunktet kan velges fritt. Omtrent det samme blir forskjellen i dyktighet mellom de beste og de svakeste elevene.

I praksis bestemmer man seg for å standardisere og transformere vanskegrader og dyktigheter slik at personene får en bestemt gjennomsnittsverdi og et bestemt standardavvik. I PISA har man valgt å bruke henholdsvis 500 og 100 når alle OECD-lands data behandles sammen. Dette betyr for eksempel at en kan forvente at omtrent 2/3 av elevene internasjonalt har skårverdi mellom 400 og 600 poeng (mindre enn ett standardavvik fra

gjennomsnittet). Tilsvarende vil omtrent 95 % av elevene ha skår mellom 300 og 700 poeng.

Ved hjelp av slike skårverdier er gjennomsnittsverdier sammenliknet mellom land og mellom grupper av elever innad i vårt eget land. Det er verdt å merke seg at alle poengsummene er relative og bare har mening i forhold til den ovenfor nevnte standardiseringen av gjennomsnitt og spredning. En gjennomsnittsverdi for et land på 450 poeng betyr at elevene i landet i gjennomsnitt er beregnet til å ligge et halvt standardavvik lavere enn det internasjonale gjennomsnittet. Det er også verdt å merke seg at en skår på 0 ikke angir at ingen oppgaver er riktig, men derimot en skår som ligger fem standardavvik under gjennomsnittet. En så lav verdi vil ingen elev få, selv om alle oppgavene er galt besvart! Vi kan altså si at tallet 0 i denne sammenheng ikke har noen viktig betydning, og vi har derfor valgt å unngå å framstille skårverdier grafisk med søyler som begynner på 0.

### Rasch-skala for holdninger og bakgrunnsdata

Ved måling av holdninger i PISA er det gjennomgående brukt en såkalt Likert-skala fra 1 (svært uenig) via 2 (uenig) og 3 (enig) til 4 (svært enig). Når det gjelder bakgrunnsdata, er det også brukt andre graderte skalaer. Men felles for mange av disse variablene er at de inngår i samlevariabler eller konstrukt (se ovenfor). For hvert konstrukt beregnes elevenes «skår» ut fra svarene på de variablene som inngår. Også for slike konstrukt er utregningene av skårverdier gjort ved hjelp av en tillempet versjon av Rasch-modellen. Dette har blant annet den fordel at det ellers ville vært vanskelig å håndtere elever som ikke har svart på alle spørsmålene som inngår i konstruktet. Å ta gjennomsnittet av de besvarte spørsmålene ville ikke bli helt riktig siden det ikke er like lett å være enig i hvert av spørsmålene. Med Rasch-modell får elevene en skår ut fra de spørsmålene de har svart på, uten at dette blir «skjevt». Alle disse skårverdiene er standardisert internasjonalt til gjennomsnitt 0 og standardavvik 1. En ulempe er da at skårverdiene bare har mening i forhold til det internasjonale gjennomsnittet og ikke direkte i forhold til ordlyden i spørsmålene og svaralternativene. For bedre å illustrere hvordan elevene faktisk svarte, har vi for de fleste konstruktene gitt informasjon om hvordan elevene svarte på hvert spørsmål for seg.

### Nivåer for dyktighet

Ved å bruke Rasch-modellen til beregning av elevenes dyktigheter og oppgavens vanskeligheter er alle oppgaver og elever plassert langs samme skala, som vi her kaller dyktighets- eller prestasjonsskala («proficiency scale»). Ved hjelp av dette kan man inndele skalaen etter *nivåer* (egentlig intervaller) av dyktighet og bruke kjennetegn for oppgavene til å beskrive

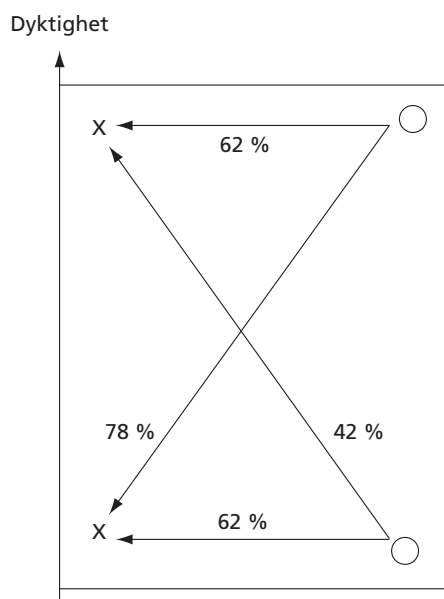


hva typiske elever på hvert nivå faktisk kan. Det er flere aktuelle framgangsmåter for å få til dette; et viktig skille går etter om nivåene fastsettes først og kompetansebeskrivelsene utledes fra oppgavene etterpå, eller om kompetansebeskrivelsene i hovedsak er gitt, og at skillet mellom nivåene blir satt til å passe med dette. I PISA brukes den første metoden, og den skal kort bli forklart her.

Grensene mellom nivåene er basert på to viktige premisser. For det første må det spesifiseres et fast punkt. Det har vært en omfattende diskusjon om hva som burde settes som en slags nedre grense (for det laveste nivået, nivå 1). Verdien 334,5 er bestemt ut fra en detaljert diskusjon av kompetansene som kreves for å løse de aller letteste oppgavene. Elever som ligger under dette, vil for hver eneste oppgave ha lav sannsynlighet for å løse den riktig. (At slike elever likevel svarer riktig på noen veldig få oppgaver, er en annen sak.) Disse elevene, som utgjør omtrent 5 % i OECD, sies å ligge under nivå 1. Deretter er skalaen (i praksis så langt opp som de vanskeligste oppgavene tilsier) delt inn i like store intervaller, og antall slike intervaller bestemmer hvor brede de er. I naturfag er det seks nivåer (nivå 1–6), og for hvert av disse er det laget en tekst som beskriver hva som kjennetegner elevenes dyktighet eller kompetanse. Disse beskrivelsene er gjengitt i delkapittel 2.9.

Prinsippet for å plassere elever i nivåer er slik: Hver elev plasseres i det høyeste nivået der han eller hun etter sin totalskår forventes å svare riktig på over halvparten av oppgavene. I en tenkt prøve bestående av en tilfeldig fordeling av oppgaver med vanskegrad innenfor et bestemt nivå, vil elever med lavest dyktighet innenfor dette nivået forventes å ha akkurat 50 % riktige svar. Elever helt på toppen av dette nivået forventes å få til en høyere prosentdel av oppgavene.

En konsekvens av denne måten å gjøre det på er vist i figur 3 som illustrerer situasjonen innenfor ett nivå. To oppgaver er vist med kryss og to elever med små sirkler. En elev nær toppen av nivået har 62 % sannsynlighet for å klare den vanskeligste oppgaven og 78 % sannsynlighet for den letteste oppgaven som hører til nivået. Tilsvarende tall for en elev nær bunnen av dette nivået er henholdsvis 42 % og 62 %. Resonnementet ovenfor forutsetter at oppgavene ikke lenger er plassert der de etter Rasch-modellen egentlig hører hjemme, etter 50 % sannsynlighet for riktig svar (se tidligere). Alle oppgavene er derimot flyttet nedover til det stedet der det er 62 % sannsynlighet for riktig svar. Denne endringen er en konsekvens av at alle elevene på nivået skal ha minst 50 % sannsynlighet for å klare oppgaver midt i nivået.



Figur 3: Skjematisk framstilling av sannsynligheter for riktig svar for elever (sirkler) på oppgaver (kryss) innen et nivå