2022

# Prøver og internasjonale storskalaundersøkelser i Norge:

## HVA KJENNETEGNER DISSE MÅLINGENE OG HVORDAN HAR DE VIRKET I DET NASJONALE KVALITETSVURDERINGSSYSTEMET DE SISTE 20 ÅRENE?

JULIUS K. BJØRNSSON

# Forord

Denne rapporten er skrevet for kvalitetsutviklingsutvalget og er ment å gi en oversikt over nasjonale prøver og de internasjonale storskalaundersøkelsene som Norge deltar i. Det er en hel del forskning sitert og gjengitt her, men også en del meninger og synspunkter som kommer fra forfatterens lange erfaring med begge disse prøvesystemene. Dette må derfor leses med dette i mente og hvis dette oppleves som en nyttig sammenfatning, var det bryet verdt.

Julius K. Bjørnsson

# Innholdsfortegnelse

## Sammendrag

Denne rapporten er ment å gi en oversikt over både status og utvikling av de viktigste elementene i det nasjonale kvalitetssikringssystemet (NKVS) i Norge, de siste 20 årene. Hovedelementene i systemet er nasjonale prøver, kartleggingsprøver og internasjonale storskalaundersøkelser, i tillegg til elevundersøkelsen, som er bare så vidt nevnt her.

Rapporten introduserer noen nøkkelbegrep angående prøver og prøvekonstruksjon, som er nødvendige å ha med seg når de prøvene og undersøkelsene som systemet består av, er undersøkt. Nasjonale prøver og deres utvikling er beskrevet og noen hovedpunkter belyst. Internasjonale storskalaundersøkelser, deres metoder og bruken av disse undersøkelsene er beskrevet, med vekt på metoder og forskjeller fra mer konvensjonelle prøver som nasjonale prøver.

Til slutt blir det trukket noen konklusjoner fra denne gjennomgangen, med vekt på hva hvert enkelt prøvesystem kan og ikke kan brukes til.


## 1. Introduksjon

I mandatet til kvalitetsutviklingsutvalget står følgende:

> ***«Hvilke konsekvenser har innføringen av kvalitetsvurderingssystemet og de ulike elementene i systemet hatt på elevenes prestasjoner og lærernes pedagogiske praksis i skolen på 2000-tallet.»***

For å prøve å svare på dette, vil denne oppsummeringen av forskning og kunnskap på området behandle følgende emner:

1. Det nasjonale kvalitetsvurderingssystemet (NKVS) består av prøver og internasjonale storskalaundersøkelser. De viktigste elementene i systemet er:
   - Nasjonale prøver i lesing, regning og engelsk for 5., 8. og 9. trinn
   - Kartleggingsprøver i lesing og regning for 1. og 3. trinn
   - Internasjonale storskalaundersøkelser som PISA, TIMSS, TALIS, ICILS, ICCS, PIRLS
   - Elevundersøkelsen (i liten grad omtalt i denne rapporten)

   Første del beskriver noen generelle kjennetegn ved ulike prøver og internasjonale undersøkelser for å tydeliggjøre hvor ulike de ulike elementene i NKVS er. Når NKVS skal evalueres er det viktig å forstå hvilke formål og begrensninger ulike prøver har.

2. Nasjonale prøver og deres utvikling – hva er status for de nasjonale prøvene i dag. Denne delen beskriver bakgrunnen for dagens system med nasjonale prøver, i tillegg til en drøfting av prøvenes metodologiske egenskaper. Dessuten vil bruk av prøveresultatene bli nevnt, med referanser til noen relevante studier som eksemplifiserer hvordan prøveresultatene har blitt brukt eller brukes.

3. Internasjonale storskalaundersøkelser – hvilke undersøkelser deltar Norge i, hva er hensikten med undersøkelsene og hva kan de brukes (og ikke brukes) til. Hensikten er å gi en enkel oversikt over de ulike undersøkelsene (hva de måler, hvordan er operasjonalisert, hvordan de gjennomføres og organiseres).

4. Hvordan har de internasjonale storskalaundersøkelsene og de nasjonale prøvene påvirket læreres kompetanse og pedagogiske praksis? Denne delen handler om effekten av

målingene, hvordan lærere og andre har brukt (og ikke brukt) resultatene og hvilke effekter dette muligens har hatt på skolesystemet og læreres arbeid.

5. Sammenfatning og konklusjoner. Til slutt vil gjennomgangen av NKVS bli oppsummert og det vil bli gjort noen betraktninger om effekten av nasjonale prøver og ILSA-studier på skolen og hvilke konsekvenser NKVS har hatt for læreres praksis og elevenes skoleprestasjoner.

## 2. Noen innledende kommentarer

I mandatet til kvalitetsvurderingsutvalget, gjengitt ovenfor, ligger det en antakelse om at innføringen av nasjonale og internasjonale undersøkelser av elevprestasjoner har ført til en endring eller forbedring av prestasjonene, dvs. at det å ta en prøve eller å bli testet i visse emner vil føre til prestasjonsforbedringer. Det er lett å gjøre slik antakelser, kanskje spesielt fordi organisasjoner som OECD, på bakgrunn av evalueringer av utdanningssystemene i mange land, gir slike anbefalinger. På begynnelsen av 2000 tallet anbefalte OECD blant annet å innføre nasjonale prøver i Norge, med en antakelse om at dette på sikt ville føre til bedre resultater og/eller elevprestasjoner. Men, etter mitt syn, ble ikke selve mekanismen for at dette skulle realiseres nøyaktig beskrevet den gang. I etterkant av innføringen av det nasjonale kvalitetsvurderingssystemet, i 2011, konkluderte OECD blant annet slik:

> *"The national authorities should emphasise that the evaluation and assessment framework includes both formative and summative elements, and school internal as well as external components. For each of the key components of evaluation and assessment, the framework or strategic plan could provide links to the relevant reference standards and point to existing tools and professional learning opportunities. To make the system coherent, it is important that learning goals are placed at the centre of the framework and that all other elements align to work towards these goals.*
>
> *The successful implementation of an evaluation and assessment framework crucially depends on whether professionals in counties, municipalities and schools have the understanding and competencies to collect, analyse and interpret evaluative information with a view to improve practices. Embedding an evaluation culture in schools and municipalities across Norway is a large culture shift that requires further investment in professional learning opportunities, targeted to the needs of different stakeholder groups." (Nusche et al., 2011. S 10).*

OECD understreker med andre ord at prøver og evalueringer *i seg selv* ikke nødvendigvis fører til bedre elevprestasjoner eller endret undervisningspraksis, men utbyggingen av et sammenhengende kvalitetssystem og økt kompetanse hos de involverte på alle nivåer vil på sikt ville kunne føre til en bedre evalueringskultur som igjen potensielt ville kunne gi bedre resultater. Dette synet på kvalitetsvurdering er ganske konsekvent i alle rapporter som OECD senere har publisert. Mandatet til kvalitetsvurderingsutvalget handler derfor ikke om enkle sammenhenger eller årsakssammenhenger, men om kumulative effekter av evalueringskulturen og læringsvurdering i hele det nasjonale kvalitetsvurderingssystemet. Disse er vanskelige å belegge empirisk.

OECD-rapporten fra 2011 understreket også at det manglet en helhet og en sammenheng mellom de ulike elementene i kvalitetsvurderingssystemet i Norge. Senere har Blömeke og Olsen (2018) etterlyst tydeligere koblinger mellom de ulike delene i dagens kvalitetsvurderingssystem.

For å kunne svare på utvalgets mandat vil vi derfor ikke bare måtte se på ulike former for prøver og evalueringer, men også på hvordan alle nivåer i utdanningssystemet har utviklet sin

evalueringskompetanse de siste 20 årene: lærere, skoleledere, skoleeiere og de som styrer den nasjonale utdanningspolitikken, samt hos de fagmiljøene som utvikler prøver. Sist, men ikke minst, må vi også undersøke om, og i så fall hvordan, innføringen av et kvalitetsvurderingssystem har påvirket elevenes læring, trivsel og prestasjoner.

Det er derfor ingen enkel oppgave å besvare utvalgets mandat eller peke på årsakssammenhenger om hvorvidt nasjonale og internasjonale prøver og undersøkelser fører til bedre prestasjoner eller har endret praksisfeltet. Dette er en ganske mye større oppgave enn hva denne korte rapporten kan dekke, men som det vil framgå av de følgende avsnittene kan kanskje noe av kunnskapsgrunnlaget i denne rapporten legge et grunnlag for hva som skal til for å øke sammenhengen mellom de ulike delene dagens system. Videre vil jeg derfor legge vekt på å beskrive og evaluere de delene av NKVS som består av prøver og internasjonale, komparative undersøkelser.

## 2.1. Hva karakteriserer en prøve?

Før jeg beskriver de ulike elementene i det nasjonale kvalitetsvurderingssystemet, er det viktig å ha klart for seg noen grunnleggende forhold som gjelder alle prøver og målinger av elevprestasjoner og elevenes trivsel i tillegg til målinger av skolemiljø.

I enhver måling må man velge hva man ønsker å finne ut blant mange mulige kompetanser og tilstander i skolesystemet. En prøve måler altså alltid kun et utvalg av mange mulige kunnskaper, kompetanser, holdninger eller ferdigheter. Dette er en grunnleggende egenskap ved alle prøver og målinger, noe vi alltid må huske på når vi evaluerer kvaliteten på prøver og undersøkelser. Det er også viktig å huske at utdanningsmålinger kan karakteriseres på veldig mange måter: først og fremst ut ifra deres hensikt og formål, men også ut ifra hvordan de er ment å bli brukt og forstås. Forskjellene kan beskrives på ulike måter, noe som kompliserer bildet, men det er kanskje lettere å bruke følgende dimensjonstenkning der det bør legges vekt på at dimensjonene ikke er enkle kategorier, men en kontinuerlig skala med en blanding av ulike løsninger. Videre kan vi bruke denne dimensjons tenkingen for å bedre forstå prøvene som inngår i NKVS.

Inntil for omtrent et tiår siden var den samlede kompetansen på prøver og storskalaundersøkelser i Norge på et lavt nivå, og begreper som psykometri var nokså ukjent begrep i skole-Norge. I USA og England hadde prøver og målinger gjennomgått en vitenskapelig utvikling i over 100 år. I Norge har man kanskje ikke vært like opptatt av dette, med den konsekvens at blant annet prøveresultater har vært behandlet på en ganske enkel måte: for eksempel har man gitt poeng for riktige svar og kandidatens prestasjon i en prøve ble oppsummert i antall poeng som ofte ble konvertert til en karakterskala. Dette gjøres fortsatt i dag, for eksempel på eksamener i noen fag, men også naturligvis i et stort antall prøver som lærere utvikler og gjennomfører selv på den enkelte skole. Det er i seg selv ikke noe galt å gjøre dette, men en slik praksis tilfredsstiller ikke de psykometriske kravene som ligger til grunn for de nasjonale prøvene og andre tilsvarende målinger.

Selv om den klassiske psykometrien ble videreutviklet i løpet av det siste århundret, brukes den i dag som regel sammen med «Item Response Theory» (IRT) i de fleste prøver og utdanningsmålinger.

1. *Klassisk psykometri----------Moderne IRT og skalering*
   Den første dimensjonen man kan bruke for å klassifisere prøver, er i spennet fra klassisk psykometri til moderne IRT og skaleringsmetoder, som sammen danner grunnlaget for å gjøre presise målinger i moderne prøver og, ikke minst, for de metoder man i dag bruker for å måle utvikling eller endring over tid. I praksis er dette aldri et spørsmål om å bruke det ene eller det andre, men som regel en blanding av begge testteorietiske tilnærminger.

2. *Formative*----------*Summative*
   Prøver er som regel klassifisert som enten formative eller summative, men i noen tilfeller, som for eksempel med de nasjonale prøvene i Norge, har man brukt elementer av begge tilnærminger. Hensikten med de nasjonale prøvene er både måle aggregerte elevprestasjoner og gi lærere informasjon om sine elever, slik at de kan bruke resultatene i den videre opplæringen. De nasjonale prøvene har i dag med andre ord et dobbelt formål. Dette kan skape problemer, for eksempel ved at formålene vektes ulikt av ulike aktører i utdanningssektoren eller at resultatene brukes på lite valide måter. En prøve med ett formål vil sannsynligvis måle mer målrettet.

3. *Systeminformasjon*----------*Individuelle resultater*
   Mange av dagens prøver laget for å gi informasjon om skolesystemet, for eksempel PISA, TIMSS og andre storskalaundersøkelser. Prøver som leverer individuelle resultater er vanligvis mindre egnet til å produsere systeminformasjon.

4. *Alle elever tar prøven*----------*Et utvalg tar prøven*
   Storskalaundersøkelser som PISA og TIMSS gir bare informasjon om skolesystemet, og ingen individuelle resultater, fordi prøvene bare gjennomføres av et representativt utvalg elever. Nasjonale prøver gjennomføres av omtrent alle elever. Siden de internasjonale storskalaundersøkelsene gjennomføres med et rotasjonsdesign, måler de bredere enn undersøkelser der alle elever besvarer de samme oppgavene. Nasjonale prøver har oppgaver som det tar 90 minutter å besvare, mens for eksempel TIMSS inneholder samlet oppgaver det tar 7-8 timer å besvare. Dette har også noen andre konsekvenser som blir beskrevet senere.

5. *Gjennomført hvert år*----------*Gjennomført med noen års mellomrom*
   Nasjonale prøver gjennomføres hvert år. Noen storskalaundersøkelser gjennomføres hvert tredje år (PISA), hvert fjerde (TIMSS) eller hvert femte år (PIRLS, ICILS). Slik sykluser er på mange måter fornuftig, ettersom endringer i systemet, eller endringer i prestasjoner hos hele elevkull, ofte er små og skjer gradvis. Det er derfor ikke nødvendig å måle hvert eneste år for å kunne monitorere skolesystemet og eller om det generelle kompetansenivået endres.

6. *«Low-stakes»* ---------- *«High-stakes»*
   Denne beskrivelsen av en prøve er velkjent. I Norge er de fleste statlige prøver «low-stakes»: prestasjon på en prøve har få eller ingen konsekvenser for eleven. Eksamen er i de fleste tilfeller en «high-stakes prøve», ettersom resultatet får konsekvenser for elevens muligheter i videre utdanning for eksempel etter videregående skole. De øvrige prøvene som inngår i NKVS er «low-stakes prøver».

7. *Enhetlig faglig innhold*----------*Mange fag eller kompetanser samtidig*
   Av prøvene i NKVS har de nasjonale prøvene et nokså enhetlig innhold (regning, engelsk eller lesing), mens de internasjonale undersøkelsene PISA og TIMSS måler kompetanser i to eller flere fag i den samme prøven. Både TIMSS og PISA tester matematikk og naturfag, mens i tillegg måler PISA elevenes lesekompetanse. PISA tester derfor tre ulike kompetanser i hver gjennomføring.

8. *Basert på en læreplan*----------*Ikke basert på læreplan*

Nasjonale prøver i Norge er basert på kompetansemål fra året før prøven blir gjennomført. Prøvene på 5. trinn måler et utvalg kompetansemål og beskrivelser av de grunnleggende ferdighetene fra 4. trinn, mens prøvene på 8. trinn er basert tilsvarende beskrivelser fra læreplanen for 7. trinn. De ulike ILSA-studiene har ulike tilnærminger. I grove trekk kan vi si at TIMSS i stor grad bygger på innholdet i den norske lærerplanen og prøvenes innhold er definert etter en felles, internasjonal gjennomgang av hvert deltakerlands lærerplaner for de aktuelle årskullene. PISA baserer seg derimot på ekspertvurderinger av hvilke kompetanser elevene trenger å ha for å kunne fungere bra som voksne.

9. *Papirprøver----------Elektronisk gjennomføring*
   De norske nasjonale prøvene har fra 2007 hatt sin nåværende form (først på papir),men siden 2014 har prøvene i engelsk og regning vært gjennomført elektronisk og fra og med 2016 også leseprøvene. ILSA-studiene var opprinnelig på papir, og fram til 2019 ble TIMSS ble gjennomført delvis på papir, delvis elektronisk. I 2015 ble PISA for første gang gjennomført i en digital prøveplattform. Slike endringer kan ha konsekvenser: måler vi den samme kompetansene på papir som i digitale gjennomføringssystemer? Slike endringer er det blitt skrevet en hel del om (Csapó et al., 2012). De digitale kartleggingsprøvene i lesing og regning ble introdusert i 2021, men de fantes i papirversjoner mange år før det. Obligatoriske kartlegginger (screening) var tidligere på 1. og 3. trinn, men er nå bare obligatoriske på 3. trinn.

10. *Numeriske resultater/karakterer----------Deskriptive resultater*
    Alle prøveprestasjoner må omdannes til et resultat, enten et talluttrykk som oppsummerer prestasjonen eller et deskriptivt resultat som beskriver elevens kompetanse. Det er viktig å forstå hvordan dette er gjort, og hvilke former for tall og skalaer som blir brukt. Dette finnes mye forskning på, for eksempel viser Tan & Michael (Tan & Michel, 2011) hvor viktig det er at skalerte skårer er beskrevet utførlig.

11. *Statens ansvar----------Individuelt ansvar*
    Alle prøvene i NKVS er et statlig ansvar, men det finnes også en god del andre prøver som skolene bruker. Disse er ofte konstruert og gjennomført av private aktører.

12. *Obligatorisk for alle elever----------Frivillig deltakelse*
    Mange av prøvene i NKVS er obligatoriske for elevene og/eller skolene, blant annet de nasjonale prøvene og ILSA-studiene (PISA; TIMSS og ICILS). Kartleggingsprøvene i lesing og regning er derimot frivillige på 1. trinn, men obligatoriske på 3. trinn.

13. *Gjennomført på et årstrinn----------Gjennomført på mange årstrinn*
    De fleste prøver er gjennomført på et klassetrinn. Blant annet gjennomføres nasjonale prøver på 5. og 8. trinn. Det utvikles nye, og unike, prøver hvert år for disse to trinnene, mens prøven for 8. trinn også tas av elever på 9. trinn i regning og lesing. PISA-undersøkelsen gjennomføres av 15 år gamle elever. Som regel går disse i 10. trinn, men ikke nødvendigvis. Dermed er PISA-undersøkelsen en av få undersøkelser som ikke trekker ut elever basert på årstrinn (som TIMSS og ICILS), men inkluderer alle 15- åringer uansett hvilket klassetrinn de går på. TIMSS-, ICILS- og PIRLS-undersøkelsene inkluderer alle elever på bestemte klassetrinn, og utelater elever fra samme alderskohort som går på andre klassetrinn. Denne dimensjonen varierer ganske mye i de ulike prøvene, noe som også har en effekt på hvordan resultatene kan tolkes og forstås.

14. *«Screening» ---------- Hele kompetansen*
    «Screeningprøver» eller kartleggingsprøver som de har uheldigvis blitt kalt i Norge, er prøver som ikke måler hele ferdigheten (som navnet kanskje antyder), men er prøver som konsentrerer seg om å identifisere elever på de laveste ferdighetsområdene av en skala. Dette er elever som kanskje sliter eller trenger ekstra opplæring, hjelp eller støtte. I kartleggingsprøver vil vi dermed ikke få noe som helst informasjon om velfungerende elever, og det er heller ikke meningen å måle disse elevenes ferdigheter. De nasjonale prøvene måler derimot måler langs hele ferdighetsskalaen, men med en lavere presisjon i målingen øverst og nederst på skalaen.  Mer om dette senere.

15. *Prøven gir rettigheter----------Ingen rettigheter eller konsekvenser*
    Prøvene i NKVSgir vanligvis ingen rettigheter, og de har som regel ingen konsekvenser for elevene. Kontrasten til dette ville være eksamen som gir rettigheter, eller andre prøver som fører til sertifisering eller andre kvalifikasjoner, som for eksempel svenneprøven på yrkesfaglig utdanning i videregående opplæring.

## 2.2. Det nasjonale kvalitetsvurderingssystem (NKVS)

NKVS er blitt beskrevet mange ganger og av ulike forfattere, men en av de nyere beskrivelsene og diskusjonene om systemet finnes i Blömeke & Olsen (Blömeke & Olsen, 2018), som også retter et kritisk blikk mot innholdet i NKVS og foreslår noen grunnleggende endringer. De understreker at systemet kjennetegnes av å ikke være bygget opp på en systematisk måte, slik at det kan dannes et helhetlig bilde av utviklingen over tid og, ikke minst, hvordan de forskjellige elementene i systemet er relatert til hverandre. NKVS ble også evaluert i 2009 av Allerup et. al. (Allerup et al., 2009), og de nasjonale prøvene i 2013, (Seeland et al., 2013) og i noen mindre rapporter gjennom årene. For eksempel har Høst skrevet om kvalitet i fag- og yrkesopplæringen (Høst, 2015), mens flere artikler og bøker i det såkalte PraDa-prosjektet undersøker hvordan skoler og lærere bruker resultater fra prøver og undersøkelser (Mausethagen et al., 2018).

# 3. Nasjonale prøver

Nasjonale prøver ble innført i 2004 og ble da gjennomført med klassiske testteoretiske metoder, der poengsummer ble rapportert og konvertert over til mestringsnivåer som så ble rapportert til utdanningsmyndighetene. De første prøvene var ikke vellykket og møtte mye motstand (Lie et al., 2005). Etter dette ble det besluttet å ta en pause med prøvene, og prøvene startet opp på nytt i 2007 etter en grundig revisjon. Noe av dette arbeidet er beskrevet i Jensen et al., (2020). Siden da har prøvene stort sett vært uendret, med unntak av at hele grunnmetodologien ble endret i 2014. Prøvene mellom 2007 og 2013 var nye hvert år, og i denne perioden var det ikke mulig å følge med utviklingen over tid. Resultatene fra prøvene var også nokså upresise, og rapporteringen fra dem var først og fremst av mestringsnivåer som ble bestemt ut ifra en prosentvis fordeling av elever, foretatt på nytt hvert år, for hele kohorten.

I 2014 ble hele analysegrunnlaget for prøvene endret og IRT ble introdusert som grunnmetode for å bestemme oppgavers vanskegrad og diskriminering (hvor godt oppgavene skiller mellom svake og flinke elever) og hvor mye informasjon om elevene på ferdighetsskalaen en prøve gir. I tillegg ble det introdusert et ankerdesign, der ankeroppgaver gjorde det mulig å måle utvikling over tid. Hele det metodologiske grunnlaget er beskrevet i et dokument publisert av utdanningsdirektoratet i 2015

(Bjørnsson, 2015). Dette designet har vært brukt helt fra 2014 til 2021, men fra og med høsten 2022 ble et nytt design tatt i bruk, med et nytt ankerdesign som blir beskrevet senere.

De viktigste egenskapene ved metodene fra 2014, er de følgende:

- Alle oppgaver er pilotert 2-3 ganger før gjennomføring. Tillater kun å inkludere oppgaver som fungerer godt.
- Eliminerer at elever som besvarer like mange oppgaver, men med ulik vanskegrad, får samme resultat.
- Analysen av oppgavene og elevenes besvarelser benytter en 2-parameter IRT modell.

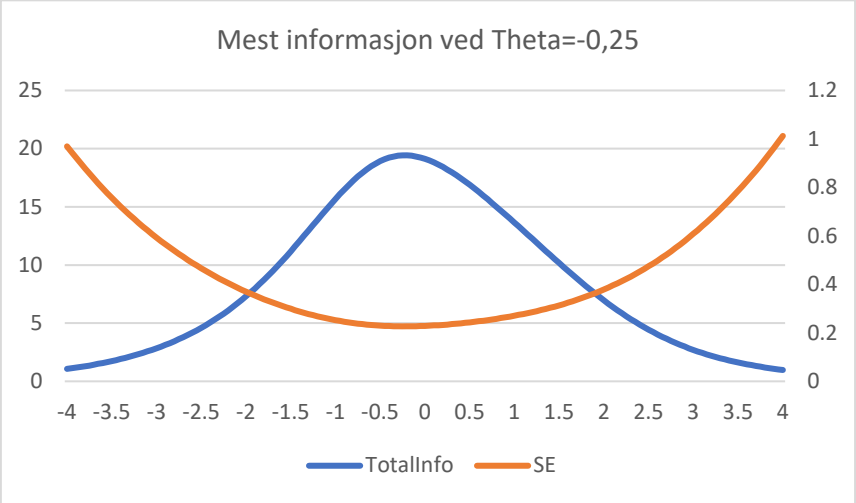
$$\boldsymbol{P(\theta) = \frac{e^{(\theta - b)}}{1 + e^{Da(\theta - b)}}}$$

- Ankeroppgaver er inkludert i hver gjennomføring av en prøve. Ankeroppgaver besvares av omtrent 6 prosent av alderskohorten, valgt på en tilfeldig måte med et såkalt NEAT-design («Non-Equivalent Anchor Test»). I tillegg til er ankerprøven konstruert med en såkalt «miditest» som betyr at den har en spredning i vanskegrad som går fra -1 til 1 i Theta-verdier. Dette er en vel ansett og robust metode for konstruksjonen av en ankerprøve, ifølge Singaray & Holland (Sinharay & Holland, 2006).
- Alle oppgaver i både engelsk og regning er delt i fire testsett, hvorav tre av settene inneholder alle oppgaver, men i forskjellig rekkefølge. Dette motvirker sekvenseffekter. Det fjerde testsettet inneholder ankeroppgavene. I de nasjonale prøvene i regning vil dtte testsettet inneholde 50 oppgaver, der 30 er kohortoppgaver og 20 er ankeroppgaver.
- Lesing har et annet design enn engelsk og regning. Bakgrunnen for dette er at leseprøvene inneholder tekster, og hver tekst følges av 5-7 oppgaver. Det betyr at hvis en ankertekst blir byttet ut, må alle oppgaver tilhørende teksten også byttes ut. Dette gjør det nødvendig med fire ulike ankersett i leseprøven.
- Alle ankeroppgaver fornyes, slik at ankeret er helt nytt på fem år, men dette blir gjort delvis gjennom perioden slik at trenden holder seg. I leseprøvene på 8./9. trinn er ankerprøven fornyet i løpet av syv år.
- Alle elevbesvarelser er skåret ut ifra hver enkelt elevs svarmønster med en såkalt EAP-prosedyre (der EAP betyr «expected a posteriori»). For nærmere forklaringer, se for eksempel Embretson & Reise (Embretson & Reise, 2013).
- Alle resultater er standardisert og omregnet til såkalte skalapoeng, som er en klassisk t-skåre med gjennomsnitt 50 og standardavvik 10.
- Prøvene har høy reliabilitet, ofte over 0,8, noe som antyder at IRT-metodens antakelse om en endimensjonal ferdighet holder mål. Dette er en særdeles viktig egenskap, ettersom høy reliabilitet er en forutsetning for at prøven skal kunne være valid.
- Et rammeverk for nasjonale prøver er utviklet og vedlikeholdt av Utdanningsdirektoratet (Utdanningsdirektoratet, 2022). Det holdes oppdatert slik at innholdet i prøvene og prøvenes egenskaper er i tråd med, og operasjonalisert etter, gjeldende rammeverk og konstruktbeskrivelser.

Alt dette er beskrevet i (Bjørnsson, 2015).

Til og med 2021 varte de nasjonale prøvene i 90 minutter, med unntak av prøven i engelsk som varer 60 minutter. Fram til 2021 fikk flesteparten av elevene (94%) de samme oppgavene. Dette er endret i 2022: Alle elever besvarer nå 1 ankeroppgave hver. Men alle prøver av denne typen hvor alle elever svarer på de samme oppgavene er en måling som er begrenset øverst og nederst i ferdigheten. Dette

innebærer også at presisjonen av målingen er ganske liten øverst og nederst på skalaen. Figur 1 viser måleegenskapene til en nasjonal prøve, hvor mye informasjon den leverer på forskjellige steder på ferdighetsskalaen og kondisjonal målefeil over hele skalaen.




Figur 1. Fisher-informasjon, summert for en hel prøve og

avhengig standard målefeil for hele skalaen (Bjørnsson, 2015)

## 3.1. Endring over tid.

Fra 2014 (for engelsk og regning, og fra 2016 for lesing) og fram til i dag har det ikke vært signifikante endringer i gjennomsnittlig prestasjon på nasjonale prøver. Den manglende endringen kan muligens forklares med at det enten ikke har skjedd noen endringer eller at prøvene ikke fanger opp de endringer som faktisk har skjedd på nasjonalt nivå. I mindre enheter, som for eksempel for skoler eller kommuner, har det vært endringer fra ett år til det neste, men dette er å forvente i mindre grupper uten at det nødvendigvis bør tolkes som en trend. Det at en prestasjon ikke endrer seg over tid er mulig, men kanskje ikke så veldig sannsynlig. Det er derfor kanskje mer sannsynlig at det er egenskaper ved selve målingen som gjør at endringer ikke fanges opp.

De internasjonale studiene har vist noen endringer i perioden fra 2014 til 2021, men disse endringene har også vært relativt små. De internasjonale undersøkelsene er mange ganger bredere enn nasjonale prøver. Og ettersom ILSA-studiene har vist små endringer og nasjonale prøver få eller ikke noen endringer over tid, er det ganske sannsynlig at nasjonale prøver ikke har målt akkurat det samme innholdet som gjenspeiler disse endringene. Men dette mangler vi rett og slett forskning på. Det har også vært hevdet at en mulig forklaring kunne være at skolenes undervisning ikke er den eneste effekten her, det vil si at det som skjer i klasserommet har en relativt liten effekt eller en effekt som muligvis er sterkt påvirket forhold utenfor klasserommet. Det er også blitt nevnt som en mulig forklaring at endring over tid ville vise seg bedre hvis man målte de faktiske skolefagene, men ikke generelle kompetanser som nasjonale prøver kanskje gjør. Foreløpig har vi dessverre ikke god forskning på dette, og det trengs i årene framover, både med bedre lenking mellom nasjonale prøver og læreplanen og med bedre kartlegging av hvordan ILSA studiene måler det som læreplanene faktisk definerer som det elevene skal mestre.

## 3.2 Lenkefeil

I alle målinger som lenkes eksisterer det en målefeil som rett og slett skyldes lenkingen. Denne målefeilen må signifikans-testes når vi undersøker om det er endringer fra en prøve til en annen eller fra et år til et annet. Det finnes metoder for å kvantifisere denne lenkefeilen. Dette ble testet i nasjonale prøver i perioden fra 2014 til 2021, og det viste seg at lenkefeilen ikke var ubetydelig (Björnsson, 2018), men den varierte avhengig av hvilke metoder som ble benyttet. Björnsson (2018) foreslo å inkludere lenkefeilen i evalueringene av endring over tid, noe litteraturen anbefaler, men det er hittil ikke blitt gjort av Utdanningsdirektoratet. I det nye ankerdesignet, som ble tatt i bruk for første gang fra høsten 2022, kan vi regne med at lenkefeilen blir noe mindre, ettersom både lenkingen blir sterkere fordi ankeret er større og ankeroppgavene besvares av vesentlig flere elever og fordi kalibreringen av oppgavene bruker flere tidsserier med resultater, dvs. samkalibrering av minst to år hver gang. Dette vil vi forhåpentligvis se tydeligere i 2023 og i årene som kommer.

## 3.3. Bredden i målingen.

Som tidligere nevnt, er bredden i målingen på nasjonale prøver er ganske liten sammenliknet med ILSA-studiene. Dette er nødvendigvis en konsekvens av å teste alle elever med nøyaktig de samme oppgavene, og en generell egenskap ved prøver og prøvesystemer av denne typen. Lineære prøver har ofte færre oppgaver, og dermed også høyere målefeil i øvre og nedre del av ferdighetsskalaen. En konsekvens av dette kan være at prøvene ikke klarer å fange opp endringer blant de lavpresterende eller de høytpresterende elevene. Etter mitt syn er det nettopp resultatene for disse elevgruppene vi kan forvente er mest følsomme for endringer i opplæringssituasjonen (som vi opplevde i mars 2020) og/eller forstyrrelser utenfra (som for eksempel lærerstreik). Slike faktorer kan være med på å forklare hvorfor det har vært små endringer i gjennomsnittet over tid på nasjonale prøver, men vi mangler empirisk forskning, og dette er kun spekulasjoner og kvalifiserte gjetning fra min side.

## 3.4. Ankringsmetoden.

Utdanningsdirektoratet har testet ut, i samarbeid med ekstern kvalitetssikrer for de nasjonale prøvene, forskjellige kalibreringsmetoder og forskjellige ankerdesign for prøvene. Det har vist seg (upublisert) at metoden som ble brukt i perioden 2014-2021 er robust og fullt sammenliknbar med mer andre metoder, som for eksempel samkalibrering. Resultatene er i tråd med den internasjonale forskningen (Kim, 2006). Den såkalte «fixed-parameter»-metoden benyttet i trendanalysene av nasjonale prøver i perioden fra 2014 til 2021 gir nesten helt like resultater som en samkalibrering av to eller flere år, selv om samkalibreringen muligens er en mer robust metode fordi datagrunnlaget blir dobbelt så stort. Dette er den metoden som de flere av internasjonale undersøkelsene bruker (blant annet PISA, TIMSS).

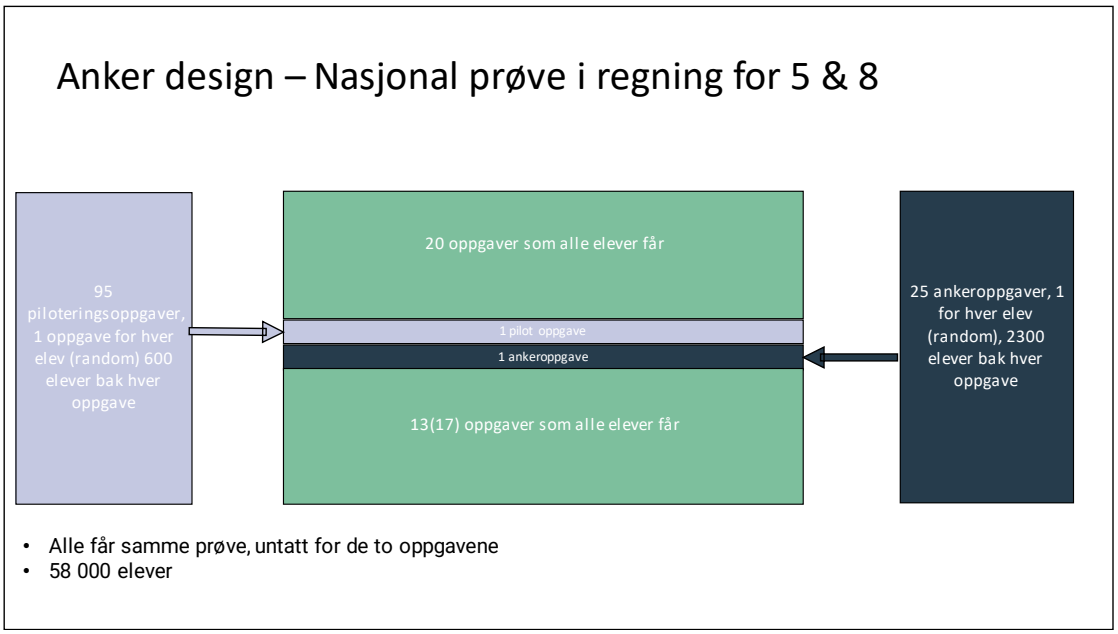
## 3.5. Endringer i 2022

Fra og med høsten 2022 er det hovedsakelig to store endringer i ankerdesignet for nasjonale prøver. For det første blir resultatene fra prøvene samkalibrert med resultatene fra flere tidligere gjennomføringer («Concurrent calibration»). I tillegg blir ankerdesignet endret slik at hver elev får én ankeroppgave og én piloteringsoppgave under hovedgjennomføringen (i regning og engelsk. I leseprøvene får alle elever én tekst med et cluster ankeroppgaver). Dette gjør at tallgrunnlaget for ankeroppgavene blir vesentlig større enn i forrige design. I tillegg er det en klar fordel at det piloteres nye oppgaver under reelle forhold (og i en reell prøvesituasjon) i engelsk og regning. Det bør også nevnes at nasjonale prøver (og eksamen) rulles ut i nytt prøvegjennomføringssystem fra og med 2022.

I siste revisjon av rammeverket for nasjonale prøver står det følgende om måling av utvikling over tid og ankerprøver:

> *To eller flere prøver som er konstruert på helt samme måte, vil aldri ha nøyaktig samme vanskegrad. Derfor er det nødvendig å foreta en lenking av prøvene mellom år. Det gjøres ved å bruke ekvivaleringsmetoder som sikrer at samme tall alltid beskriver samme ferdighetsnivå. I nasjonale prøver blir dette gjort ved å bruke IRT-metodologi til å kalibrere hver oppgave i prøvene, og sette dem sammen til en prøve som beskriver ferdigheten til hver elev med en skalert skåre. Samme tall betyr samme ferdighetsnivå hver gang en ny prøve blir gjennomført. Dette er mulig å gjøre med et såkalt ankerdesign der et antall oppgaver blir gjentatt hvert år ved at hver elev får en til to ankeroppgaver som en del av prøven. Ved å bruke de samme ankeroppgavene hvert år, kan vi derfor lenke sammen prøver fra ett år til det neste.*
>
> *Denne ankermetoden, der resultatene hvert år blir satt på samme skala, gjør det mulig for skoler og skoleeiere å vurdere utvikling i fordeling på mestringsnivåer og endring i gjennomsnitt over tid. Måling av utvikling over tid startet i 2014 for prøvene i regning og engelsk, og i 2016 for prøven i lesing. Fra 2022 starter målingen av utvikling over tid på nytt.*
>
> *Prøveutviklerne utvikler med tanke på ankerdesignet spesielle ankeroppgaver som representerer ferdigheten så langt det er mulig. Ankeroppgavenes vanskegrad skal minimum ligge på +/- 1 standardavvik fra gjennomsnittet i ferdighet. Omtrent 20 prosent av ankeroppgavene skal byttes ut hvert år, slik at hele ankeret blir fornyet hvert femte år. Ankerdesignet til prøvene i regning og engelsk er likt. Ankeroppgavene tilsvarer omtrent 40 prosent av den totale prøvelengden for prøvene i engelsk og regning, men kun en liten del av prøven for den enkelte elev (1-2 oppgaver).*
>
> *Ankerdesignet til prøvene i lesing er annerledes. I lesing bygger alle oppgavene i prøven på 5–7 forskjellige tekster. Det betyr at hvis vi bytter ut en tekst, så bytter vi ut mange oppgaver samtidig. Konsekvensen er at ankerdesignet må organiseres annerledes enn i engelsk og regning. I leseprøvene er det et integrert anker med et enkelt blokkdesign, der alle elever vil få en prøve som består av én ankertekst med tilhørende oppgaver i tillegg til et visst antall kohorttekster. Dette gir i tillegg mulighet for testing av en ny ankertekst/blokk hvert år. Dette gjør det mulig å ha samme variasjon og bredde i ankertekstene som i kohorttekstene, og å bytte ut én til to ankertekster hvert år uten å miste lenken mellom år. (Utdanningsdirektoratet, 2022).*

Dette betyr som allerede nevnt, at datagrunnlaget for kalibreringen av ankeroppgavene blir dobbelt så stort som før, ettersom to år blir kalibrert sammen og alle elever får en ankeroppgave. Nå brukes ikke lenger et NEAT («Non-Equivalent Anchor Test»)-design for ankeroppgavene, men alle elever vil få minimum én ankeroppgave. Designet inkluderer også mulighet for utprøving (pilotering) av nye oppgaver i hele populasjonen/alderskohorten. I praksis vil dette fungere slik at alle elever får to oppgaver, som ikke blir rapportert. Designet vil, slik eksemplet fra prøven i regning viser, ser slik ut:




Figur 2. Ankerdesign fra 2022

Designet i engelsk blir tilsvarende, mens det i lesing blir noe mer komplisert fordi leseprøven inneholder oppgave-cluster som er tilknyttet tekstene i prøvene. Leseprøvene vil bruke 5 roterte prøvesett på 5. trinn, 7 på 8. trinn, ettersom ankerprøven skal være like lang som kohortprøven. Det vil med andre ord distribueres én ankertekst med ankeroppgaver i hvert prøvesett, hver og femte elev i 5. trinn, hver syvende elev i 8. trinn, får de samme oppgavene i prøven. Alle elevene får på denne måten én tekst med et cluster med ankeroppgaver, i tillegg til fire eller seks tekster med de øvrige kohortoppgavene på tvers av de ulike prøvesettene på henholdsvis 5. trinn og 8. trinn.

## 3.6. Bruk av resultatene – læreres og skolers erfaringer-konsekvenser

Som tidligere antydet finnes det ikke mye forskning på nasjonale prøvers metodologiske egenskaper utover det som allerede er nevnt her. Men det foreligger noe forskning på bruk av prøveresultatene og av deres effekt på lærere og skoleledere.

I «Temanummer om prøver i skolen: Nasjonale prøver og eksamener i norsk og svensk grunnopplæring» av Acta Didactica Norden i 2018, (Olsen et al., 2018) ble det beskrevet en del forhold angående prøver i Norge og Sverige, men ikke så mye om nasjonale prøver utover det som allerede nevnt her. Redaktørene av temanummeret la vekt på ulike perspektiver på prøver, blant annet:

- De overordnete rammene for prøver eller eksamen i Norge og Sverige, hvor det ble poengtert at de fleste prøvesystemer må forstås i en historisk, utdanningspolitisk og testteoretisk ramme, noe som alt påvirker elevenes prestasjoner og skolenes kvalitet.
- Prøver må være likeverdige og rettferdige, som går på etisk bruk av prøver og rettferdighet mot elevene som gjennomfører dem.
- Prøver må være pålitelige (reliable), noe som er en grunnleggende egenskap for alle prøver som skal bli valide og måle relevante ferdigheter hos elevene.
- Standardsetting og lenking er belyst som viktige aspekter av prøveutvikling, noe som muliggjør sammenlikninger over tid og trendmålinger og i tilfellet standardsetting dreier dette seg om koblinger mellom tallmessige resultater og faktuelle beskrivelser og definisjoner av de kompetansene som prøvene måler.

I temanummeret ble det også beskrevet hvordan prøvene har innvirket på skolenes arbeid. I en studie av Ole Petter Vestheim (Vestheim, 2018), ble det undersøkt hva skoler som hadde oppnådd

gode resultater på nasjonale prøver gjorde for å oppnå disse resultatene. Studien konkluderer med at skolene ikke ser ut til å legge stor vekt på eksterne vurderinger eller praktisere «teach to the test», men at de i større grad la vekt på en bred læringsorientering både i lærerkollgiet og i arbeidet med elevene, og det var disse praksisene som førte til gode resultater på nasjonale prøver. Vestheims funn er helt i tråd med OECDs anbefalinger og intensjonene med NKVS omtalt tidligere.

Arntzen (Arntzen et al., 2019) beskriver også en situasjon hvor elevene opplever nasjonale prøver som motiverende og interessante i seg selv, selv om de ikke får noen uttelling for prestasjonen sin. Arntzen peker på at det mangler forskning og forståelse av hvordan elevene opplever leseprøvene, som var emnet i denne studien. Grandemo (Grandemo, 2017) beskriver en liknende situasjon hvor skoleledere i høyt presterende skoler ble spurt om sitt arbeid med prøvene, og hvor de attribuerer de gode resultatene på nasjonale prøver til skolenes avslappede forhold til prøvene, og systematisk arbeid over tid ved å forbedre skolens arbeidsbetingelser, kompetanse og tillit mellom alle involverte, lærere, ledere og elever. Igjen er dette i tråd med det som OECD fremhever bør være effekten av økt evaluerings- og vurderingskompetanse i skolesektoren. Det er også blitt skrevet noen masteroppgaver om dette emnet. En av dem (Landmark, 2018) konkluderer med at effekten av nasjonale prøver stort sett er positiv for lærere og elever, og han/hun finner at forekomsten av øving og overdrevet fritak fra dem for å heve gjennomsnittsresultatet forkommer ganske sjeldent.

Nytten av nasjonale prøver har også blitt undersøkt av Hovdehaugen et al. i noen små og store kommuner i Norge i et forskningsprosjekt (Hovdhaugen et al., 2017). Denne studien finner at nytten for små kommuner og små skoler er vesentlig mindre enn for de store. Dette følger naturligvis av det forhold at statistikken, gjennomsnitt i fleste tilfeller, blir mer og mer usikker når gruppene de er utregnet fra er små. Derfor er utvikling over tid en meget usikker målestokk for små kommuner og små skoler, og de må bruke andre målestokker for å følge med på sin egen utvikling, selv om selve resultatene fra nasjonale prøver sikkert er nyttig for individuelt formativt arbeid og for videre opplæring for enkeltelever også på små skoler. I tillegg så er det velkjent at naturlige svingninger forekommer i små grupper. Alle som har jobbet på skolen kjenner til at ulike årskull har ulike forutsetninger. Men det foreligger også forskning som understreker at det å øke skolens kvalitet gjennom bruk av nasjonale prøver, har fortsatt en vei å gå (Gunnulfsen, 2017).

Tidligere ble det foretatt en evaluering av prøvene fram til 2013 (Seland et al., 2013), og der ble det konkludert at de ble opplevd som nyttigere på store skoler, at skolene oppfattet dem stort sett som nyttige, at skoleeiere og skoleledere opplevde dem som nyttige, men at men at det opplevdes som problematisk at resultatene ble rapportert på aggregert nivå og var ganske usikre. Konklusjonen til denne omfattende studien var:

> *«Vår evaluering tyder på godt utbytte fra nasjonale prøver for skoleeiere, skoleledere, lærere, elever, foresatte og de nasjonale utdanningsmyndighetene fremmes av at kommunen har forholdsvis mange skoler/elever og et aktivt forhold til rollen som skoleeier, der lærerne på den enkelte skolen opplever å være del av et kollektiv som samarbeider med ledelsen om den pedagogiske utviklingen av skolen. Her kan prøvene ha stor betydning både for arbeidshverdagen og for hvordan skolen arbeider langsiktig med å forbedre elevenes læringsresultater. Siden skole-Norge viser stor variasjon i disse faktorene, vil likevel prøvenes bruk og betydning være forskjellig på tvers av skoler og kommuner.» (s. 10).*

Til slutt i denne oppsummeringen er det verdt å nevne en studie som ser nærmere på hvordan noen lærere bruker data fra nasjonale prøver (Werler & Færevaag, 2017). Denne kvalitative undersøkelsen konkluderer med at lærerne føler seg usikre og opplever seg for lite kvalifisert til å bruke resultatene fordi de ikke behersker de grunnleggende metodene, for eksempel IRT.

Kvalitative undersøkelser, basert på få respondenter, er vanskelige å generalisere ut ifra. Det er derfor helt klart at vi trenger flere kvantitative undersøkelser, blant representative utvalg med lærere, skoleledere og skoleeiere, som evaluerer effekten og nytten av nasjonale prøver. I tillegg er det, etter mitt syn, stor mangel på forskning om de metodologiske aspektene ved de nasjonale prøvene. Det er kun gjennom systematisk forskning at vi finner svar på om de fungerer etter intensjonen. For eksempel burde det gjennomføres standardsetting-øvelser for å bedre kunne begrunne innholdet i prøvene. Rammeverkene for de nasjonale prøvene burde også revideres, spesielt for å styrke lenken mellom de nye læreplanene (LK20) og prøvenes konstrukter og innhold. Sist, men ikke minst, må vi gi lærere og skoleledere mer støtte i å tolke og bruke resultatene til det beste for elevene, spesielt når vi vet at det lille som er blitt gjort er blitt godt mottatt. Se for eksempel Astrid Roe et. al.s bok om hvordan nasjonale prøver i lesing kan brukes i den videre opplæringen (Roe et al., 2018).

# 4. Kartleggingsprøver

Kartleggingsprøver i lesing og regning gjennomføres i første og tredje klasse i grunnskolen. Hensikten med disse prøvene er å finne de elevene som trenger mer hjelp og støtte, og prøvene gir derfor lite informasjon om elevene over den såkalte bekymringsgrensen. Prøvene ble heldigitale fra og med 2022. Prøvene måler kun rundt den såkalte «bekymringsgrensen» som er normert til et kuttskår på 20 prosent i den nedre del av ferdighetsskalaen for å kunne finne de svakeste elevene i en alderskohort.  Prøvene er frivillige for 1. trinn og obligatoriske for 3. trinn. Selv om prøvene er frivillige, viser gjennomføringstall at omtrent tre fjerdedeler av skolene valgte å gjennomføre prøvene for 1. trinn. (muntlig kommunikasjon fra prøveutviklerne).

De nye kartleggingsprøvene er såkalt blokk-adaptive, det vil si at vanskelighetsgraden på oppgavene tilpasser seg elevens nivå. I lesing får elevene denne tilpasningen etter at de har besvart en blokk med felles oppgaver innledningsvis. I regning får elevene en blokk med ekstra oppgaver etter en felles innledende del. Selv om prøvene omtales som adaptive så er de det kun på en begrenset måte, ettersom de anvender poengsumme for å bestemme om elever er under eller over en bekymringsgrense.  En reell adaptiv prøve ville bruke Theta estimering, dvs en ferdighetsestimering basert på IRT analyse av alle oppgavene og en «maximum likelihood» eller liknende estimering av elevferdigheter

Gjennomføring av kartleggingsprøver tar omtrent 40 minutter, men prøvene har ingen øvre tidsbegrensing. Når skolen og lærerne får resultatene, kan de identifisere de elevene som faller inn under bekymringsgrensen eller det såkalte «oppfølgningsområdet». Dette er elever som læreren må være oppmerksomme på, og bestemme om de skal ha ekstra hjelp og støtte i lese- eller regneopplæringen, eller om de skal henvises videre til for eksempel PPT for videre utredninger.

UDIR har, i samarbeid med prøvekonstruktørene av kartleggingsprøvene, laget støtteressurser med opplysninger til lærere og foreldre, skoleledere og skoleeier med anvisninger om hvordan rektor bør følge opp resultatene for sin skole. Støtteressursene gir også informasjon om hva som må skje før, under og etter gjennomføringen. Skolen eier sine egne resultater, og det er ingen rapportering fra kartleggingsprøvene til UDIR eller andre sentrale utdanningsmyndigheter.

Lese- og regneprøven utvikles av to ulike fagmiljøer: kartleggingsprøvene i lesing utvikles av Lesesentret, ved Universitetet i Stavanger, mens kartleggingsprøven i regning utvikles av Matematikksenteret i Trondheim, ved Institutt for lærerutdanning og skoleforskning, NTNU.

Foreløpig foreligger det lite forskning på kartleggingsprøvene. Det viktigste er om bekymringsgrensen fungerer etter hensikten, og hvorvidt resultatene fra kartleggingsprøvene fanger opp de elevene som

faktisk sliter. Prøvene bør heller ikke over- eller underidentifiserer elevener i ferdighetsområdet rundt bekymringsgrensen. Kartleggingsprøven i lesing har blitt planlagt og utviklet over flere år, og det finnes en artikkel i *Acta Didactica Norden* som beskriver grunnideene bak dagens prøver og hvordan forfatterne tenker at den vil virke (Walgermo et al., 2018). Forberedelsene til innføringen av kartleggingsprøven i regning er i noen grad dokumentert i (Forsbakk & Nortvedt, 2021). I årene framover blir det viktig å dokumentere dette arbeidet empirisk, og vi bør kunne forvente evalueringer av hvordan de nye adaptive tilnærmingene fungerer.

# 5. Internasjonale storskalaundersøkelser (ILSA)

De store internasjonale utdanningsundersøkelsene som Norge deltar i, eller har deltatt i, er en sentral og meget viktig del av kvalitetsvurderingssystemet. Helt generelt, kan vi si at disse undersøkelsene er basert på best mulig teknologi og psykometri, i tillegg til å støtte seg til verdens fremste fageksperter i utarbeidelsen av innholdet i målingene.

Før vi ser på de ulike delene som til sammen utgjør en internasjonal storskalaundersøkelse, er det viktig å peke på at det finnes et utall artikler og bøker om emnet fra de siste 20 årene. Nylig ble det publisert en ny oversikt i «Handbook of Comparative Large-Scale Studies in Education» (Nilsen et al., 2022). Denne er redigert av blant annet en norsk forsker, Trude Nilsen ved ILS, UIO, sammen med den tyske forskeren Agnes Stancel-Piatak og professor Jan-Eric Gustafsson ved Gøteborgs universitet. Flere norske forskere bidrar med kapitler i boken, og håndboken er kanskje en av de mest omfattende antologiene om internasjonale storskalaundersøkelser i nyere tid. Et av kapitlene i boken slår fast at det er økt kompetanse og kunnskap om storskalaundersøkelser i Norge, og at flere norske forskere er blant de fremste på dette området i verden. Den er fritt tilgjengelig for alle som er interessert (Open source). Håndboken har, ifølge innledningen, følgende intensjoner og ambisjoner:

> *"A large body of knowledge and competence has accumulated since the first international large-scale assessment (ILSA) was implemented in the 1960s. Moreover, ILSA inspired numerous valuable debates about education, policy, assessment, and measurement over this period. Since the first ILSA, the number of publications using ILSA data for research has increased near exponentially. Given the important role of the ILSAs for education, policy, practice, and research, there is a need to synthesize all this knowledge. This handbook synthesizes the knowledge that has emerged from the ILSAs, the debates on ILSAs, theories underlying the ILSAs, historical and political perspectives, the methodology pertaining to ILSAs, and the findings from the studies using ILSA data. " (s.14).*

Alle som vil vite mer om ILSA-studiene, er herved henvist til denne håndboken.

De neste avsnittene vil beskrive de generelle metodene de fleste av disse studiene anvender, ettersom de har en felles grunnleggende metodikk. Dette er et av det fremste kjennetegnet på disse undersøkelsene. Det er selvsagt variasjoner og forskjeller, men disse ligger først og fremst i valg målgruppe, i undersøkelsenes innhold og tilnærming til dette innholdet, men selve målemetodene, analysekravene og analysemetodene er i grunnen nokså like.

## 5.1. Metoder og generell organisering av undersøkelsene

De fleste av ILSA-studiene som måler skoleprestasjoner er opprinnelig basert på NAEP-prøvene («National Assessment of Educational Progress») i USA som ble igangsatt 1960-tallet. Da var internasjonale, komparative undersøkelser vanskelige og den første studien av denne typen som Norge deltok var TIMSS 1995, selv om Norge hadde deltatt i noen studier tidligere, blant annet "FIMS-First mathematical and science study", "SIMS-second mathematical and science study" og deretter "TIMSS, third mathematical and science study". Den sistnevnte endret senere navn til

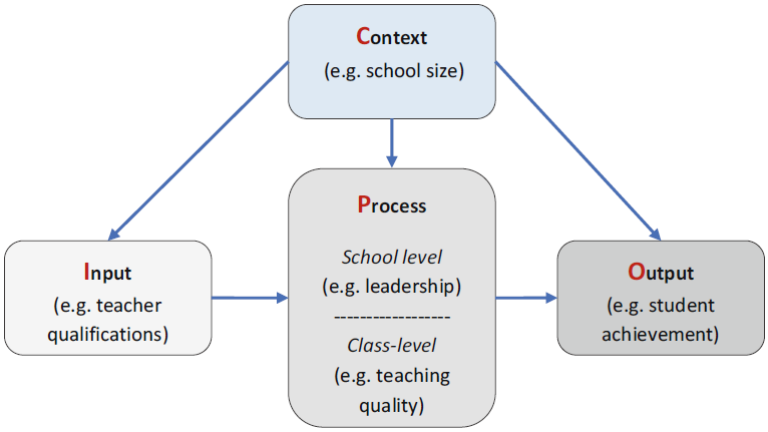
«Trends in science and mathematics study», slik vi kjenner den i dag. Norge deltok også i noen andre studier, bl.a. studier i lesing/literacy, men jeg velger å ikke omtale disse her. De som er interessert i historikken, anbefales Part III, kapittel 8, i den tidligere nevnte håndboken (Nilsen et al., 2022).

Alle ILSA-studiene bruker mange eksperter fra forskjellige fagområder. Et typisk trekk ved organiseringen, er at en ekspertgruppe vanligvis har ansvar for det faglige innholdet (som for eksempel lesing-matematikk-naturfag-IKT-kompetanse), mens en annen ekspertgruppe er ansvarlig for målemetodene og de psykometriske analysene. Alle studiene har også utvalgseksperter som sikrer at utvalgene blir representative for hvert land som deltar, og sammenliknbare på tvers av land etter gjennomføring. Det sistnevnte er ingen liten jobb.

I tillegg har hver studie en overordnet prosjektleder som leder arbeidet ved det som ofte kalles et internasjonalt studiesenter, hvor det vanligvis er mange mennesker som arbeider med, og koordinerer, aktivitetene i hvert enkelt deltakerland. I PISA er det i tillegg et styre i regi av OECD, det såkalte PISA Governing Board, som godkjenner alle aktivitet og innhold i studien. Utdanningsdirektoratet deltar i dette arbeidet på vegne av Norge. Hos IEA («International Association for the Evaluation of Educational Achievement») som gjennomfører TIMSS, PIRLS, ICCS, ICILS og flere mindre studier, har man også en såkalt «General Assembly» som tar policy-beslutninger, men som ikke er involvert i selve gjennomføringen av studiene. Dette skiller IEA fra OECD , selv begge organer utformer generell policy på utdanningsfeltet.

## 5.2. Konseptuelle modeller og forskjellige nivåer i skolen

Mange av studiene bygger på noen slags «Context-Input_Process Output»-modell. Den mest siterte er beskrevet av Scheerens (1990).




Figur 3. CIPO modellen, (Scheerens, 1990)

En enkel modell av denne typen kan hjelpe oss med å definere hva studiene skal måle, og hvordan, og modellen viser tydelig at resultatene fra undersøkelsene er avhengige av ulike nivåer. Dette må det tas hensyn til når resultatene skal analyseres og brukes for å konkludere om sammenhenger mellom alle de variablene som samtidig virker på elevenes prestasjoner. Figuren fra Scheerens viser tre av disse nivåene.

Generelt er altså studiene basert på å undersøke tilstander og prestasjoner på ulike nivåer i skolesystemet: dette dreier seg for eksempel om elevenes kompetanse og prestasjoner i lesing, naturfag, matematikk, IKT-ferdigheter, samt et spørreskjema til elevene om diverse forhold ved utdanningen på elevnivå. Studiene inkluderer dessuten også alltid et spørreskjema til skolen eller

skoleleder på skolenivå. Noen studier har i tillegg spørreskjemaer til lærere og foreldre. I tillegg har vi studier som TALIS («Teaching and Learning International Study») som er en spørreskjemaundersøkelse blant lærere og skoleledere om skolens og de ansattes arbeidsforhold, utdanning, videreutdanning og andre viktige forhold som påvirker elevers læring.

## 5.3. Matrix sampling-metodikken og utvalgene

Det er viktig å understreke at disse de internasjonale storskalaundersøkelsene er en helt annen type målinger enn for eksempel nasjonale prøver. For det første så gir de en vesentlig bredere måling og man bruker såkalt matrix-sampling-metodikk for å distribuere innholdet i undersøkelsene. Dette betyr at man organiserer oppgaver i prøvehefter eller blokker som elevene besvarer. I for eksempel TIMSS besvarer hver elev oppgaver i 90 minutter, mens hele testen ville tatt mer enn syv timer å gjennomføre hvis den samme eleven svarte på alle oppgaver i TIMMS. Dette betyr at det kanskje er hver tiende eller femtende elev som får de samme oppgavene i undersøkelsene, noe som også fører til at det er umulig å produsere individuelle resultater. Dette er heller ikke hensikten med disse målingene. Hensikten er å måle hele systemet. Siden konstruktene som måles er operasjonalisert i mange oppgaver, er denne typer målinger godt egnet til nettopp dette. Igjen henvises det til Nielsen et.al (2022), for eksempel kapittel 14, som gir en utførlig beskrivelse av designet i IEA-studiene TIMSS og PIRLS.

For å få en god, aggregert måling på systemnivå kreves det nøyaktige utvalgsprosedyrer, der man må forsikre seg om at utvalgene i undersøkelsene er representative for hele landet. Vanligvis benyttes en såkalt «proportional to size»-utvalgsprosedyre, som betyr at store skoler har en høyere sannsynlighet for å bli trukket ut enn små skoler. Prosedyren innebærer å først velge hvilke skoler som er representative for landet (i PISA, TIMSS og PIRLS er dette som regel rundt 250 skoler for hvert årstrinn i undersøkelsen), men antallet skoler er avhengig av hvor omfattende den faglige målingen skal være. den foreløpig siste runden av PISA- studien ble for eksempel såkalt «Financial Literacy» innlemmet, noe som betyr at det trengtes et større utvalg elever slik at det ble sikret at mange nok elever besvarte alle deler av prøven. Dette er gjort for å sikre at analysene har nok styrke og kan brukes for å konkludere om tilstanden til hele skolesystemet.

Metoden innebærer altså at man velger skoler først, og deretter velges elever fra disse skolene. TIMSS og PIRLS velger alle elever på visse klassetrinn (dvs. de som havnet i utvalget), mens PISA velger de som er 15 år gamle på uttrukne skolene, uansett klassetrinn.

## 5.4. Hvordan resultatene er produsert

Som tidligere nevnt, er det ikke mulig å produsere individuelle resultater i disse studiene, men en ganske kompleks metode for å regne ut sannsynligheten for svar ble vanligvis benyttet. Dette er en multidimensjonal IRT-modell som tar hensyn til ulike bakgrunnsvariabler, og den leverer sannsynligheter for et resultat for hver elev uavhengig om eleven har besvart oppgavene eller ikke. I disse beregningene brukes det såkalte sannsynlighetsvekter og hver elev får alt fra 5 til 10 såkalte «plausible verdier» som gir elevens resultat. Dette betyr i praksis at for å kunne få ut et gjennomsnitt fra PISA, må det i gjennomsnitt regnes ut 80 vekter og dette må man gjøre ti ganger, og sammenlagt blir dette 810 beregninger, for å få ut et enkelt gjennomsnitt, for eksempel for et land. Dette har vært nokså tunge prosesser tidligere, men med kraftige datamaskiner og nye og bedre analyseprogrammer, er dette noe som er relativt lett i dag. Alle analyser fra disse studiene krever en grunnforståelse av hva studiene måler, og hvordan de måler og hva resultatene kan brukes til. Dette er på mange måter fascinerende metoder, men her blir de ikke nærmere beskrevet. Interesserte henvises til Nilsen et. al (2022) og til teknisk dokumentasjon fra studiene, for eksempel «PISA Technical Report» (OECD, 2012) eller «TIMSS 2019 Methods and Procedures» (Martin et al., 2020).

## 5.5. Rammeverk

De internasjonale studiene publiserer et rammeverk i forbindelse med hver gjennomføring. I rammeverket finner vi teoriene som ligger til grunn for målingen og begrunnelser for innholdet man har valgt å teste. Rammeverkene inneholder også prøvespesifikasjoner, det vil si beskrivelser av hvordan testen skal operasjonaliseres, hva skal måles og hvordan. Her finner man nøyaktige beskrivelser av kompetanser, beskrivelser av hvor mange oppgaver testen inneholder i hver blokk eller del, hvor mange oppgaver det er totalt, hvilke kompetanser, fag eller ferdigheter som måles og begrunnelser for de valgene man har tatt. Disse rammeverkene er tilgjengelige og kan lastes ned fra enten IEA sine nettsider (www.iea.nl) eller fra OECD (www.oecd.org).

## 5.6. Endringer over tid

Alle de internasjonale storskalaundersøkelsene har til hensikt å følge utviklingen i hvert lands utdanningssystem over tid. For å få til dette, bruker de et visst antall ankeroppgaver som gjenbrukes (og derfor er hemmelige). Disse oppgavene erstattes gradvis, men siden det er viktig å bevare trenden gjøres dette nokså konservativt. Både TIMSS og PISA har et ganske stort antall ankeroppgaver som gjentas i hver gjennomføring. Dette kan være bortimot en tredjedel av oppgavene som er gjennomført hver gang.

## 5.7. Testsykluser

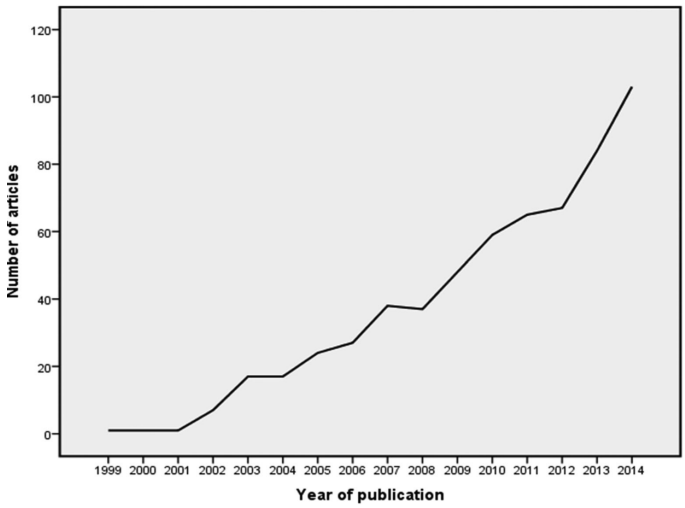
TIMSS gjentas hvert fjerde år og har blitt gjennomført i sin nåværende form siden 1995. PISA gjentas hvert tredje år, med unntak av at gjennomføringen i 2021 som ble utsatt med ett år på grunn av Corona-pandemien. PIRLS gjentas hvert femte år og, det samme gjelder for ICILS- og ICCS-studiene. I tillegg til studiene som måler elevprestasjoner i forskjellige fag og ferdigheter, har OECD en studie som heter TALIS (Teaching and Learning International Study) som inneholder spørreskjemaer til lærere og skoleledere. TALIS er med andre ord ikke en kunnskapsmåling slik de andre internasjonale undersøkelsene, selv om den på de fleste måter gjennomføres på samme måte som de andre studiene. Både utvalgsmetoder og analysemetoder er like, men en viktig forskjell er at TALIS gjentas hvert sjette år, det vil si ett år etter annen hver PISA-syklus. Intensjonen var opprinnelig at den skulle samtidig med annenhver PISA-omgang, men Corona-pandemien endret disse planene.

Det er viktig å minne om at testsyklusene er ikke fastsatte grunnet noen teori eller av hensyn til metoder, men mer ut ifra det som er praktisk mulig å få til. PISA-syklusene er ganske tett på hverandre, og der er et åpent spørsmål om endringer i utdanningsprestasjoner blir synlige i løpet av tre år.

## 5.8. Resultater fra ILSA-studiene

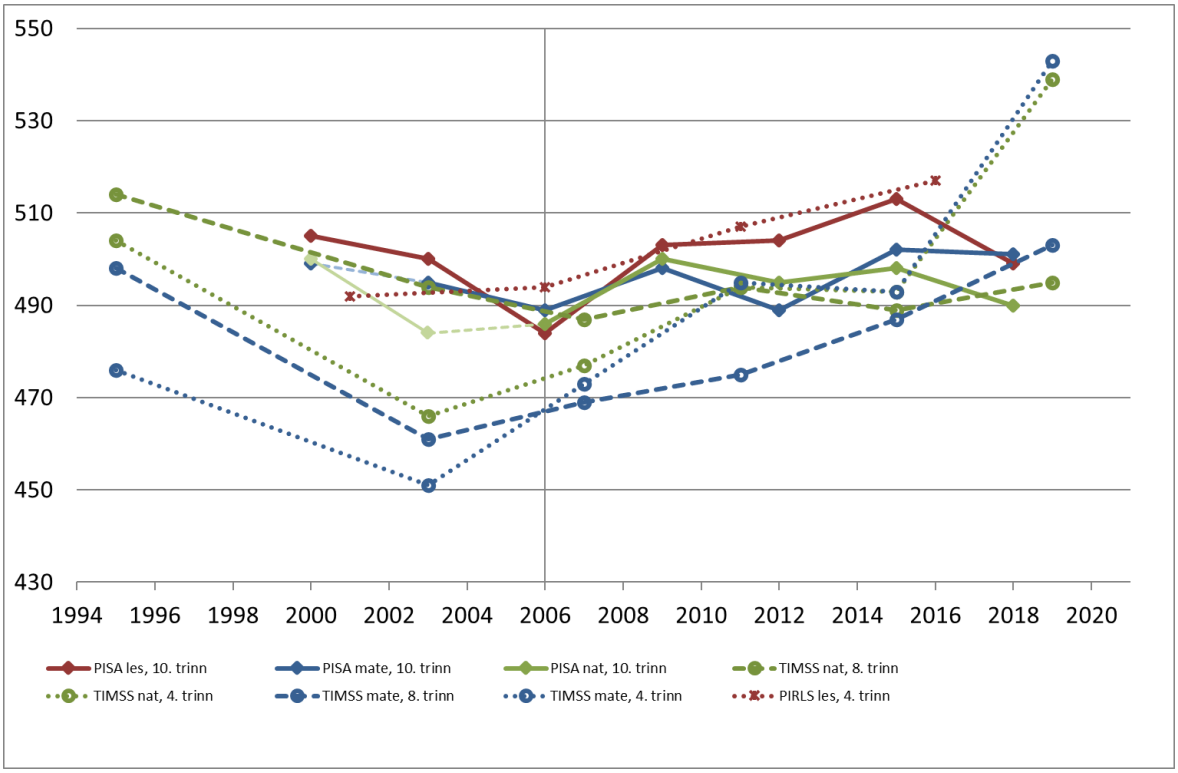
Tidligere var det ganske vanlig å kun publisere rangeringer av deltakerlandene i undersøkelsene, men i senere tid er det lagt mindre og mindre vekt på dette. Alle som jobber med storskalaundersøkelser og forsker på dette, er i dag enige om at rangeringer ikke lenger er like viktige. I større og større grad er grundige analyser av resultatene, og å kunne relatere resultatene til forhold ved utdanningssystemene kanskje de viktigste bidragene fra disse studiene i dag.

Det eksisterer pr. i dag et mangfold av resultater fra ILSA-studiene at det er nesten umulig å ha en god oversikt over dem. En nylig studie av Hopfenbeck et.al demonstrerer (Hopfenbeck et al., 2018) at det i perioden fra 1999 til 2015 var en eksplosiv økning i publikasjoner basert på PISA-studien. Figur 4 viser økningen. Det er interessant å merke seg at mesteparten av disse publikasjonene er såkalte sekundæranalyser med bruk av PISA-data. Dette understreker et forhold som er unikt i med ILSA-studiene: data fra hver gjennomføring blir gjort tilgjengelig i databaser for alle. Innen annen utdanningsforskning er ikke dette en like vanlig praksis.




Figur 4. Økningen i PISA publikasjoner fra 1999 til 2015 (Hopfenbeck et al., 2018)

I Norge har man valgt å delta i veldig mange av de internasjonale undersøkelsene, og de er inkludert i det nasjonale kvalitetsvurderingssystemet. Dette er en ganske enestående situasjon: veldig få land deltar i like mange studier, og de fleste nøyer seg med å delta i en studie for hvert fag. I Norge har vi derfor en enestående situasjon hvor alle disse studiene kan brukes til å validere hverandre. Hvis alle studiene for eksempel viser de samme resultatene eller en tilsvarende trend over tid, kan vi være ganske sikre på at resultatene eller trendene er reelle. Dette er nettopp blitt gjort, og figur 5 viser utviklingen fra 1995 for PISA og TIMSS for fagområdene lesing, matematikk og naturfag.




Figur 5. Utviklingen i PISA, TIMSS og PIRLS fra 1995 til 2020. (Opprinnelig blide fra Kjærnsli og Olsen)

Figuren viser at trenden gikk skarpt ned i begynnelsen av tidsperioden, men har snudd, og viser en sakte oppadgående trend i siste del av perioden. Det viktigste er ikke at trenden viser framgang,

men at studiene viser *den samme trenden* over tid. Dette er en meget viktig indikator på at resultatene fra studiene er riktige.

## 6. Videre forskning på ILSA-data

Som tidligere nevnt, er det et mangfold i sekundære analyser av ILSA-data. I Norge har også denne forskningen økt, og et enkelt databasesøk viser flere titalls artikler og bøker om dette temaet. Hvis vi legger til andre artikler om disse studiene, ikke bare norske, så er det hundrevis av artikler publisert de siste 20 årene. Det er derfor umulig å gjengi alt det her, men noen få eksempler skal likevel nevnes.

I 2019 ble «Tjue år med TIMSS og PISA i Norge» (Olsen & Björnsson, 2018b) publisert. I boken presenterte en gruppe forskere sekundære analyser fra ulike ILSA-studier. I boken er også en god oversikt over hvilke studier Norge har deltatt i, og denne gjengis her i figur 6. Boken inkluderte resultater om: motivasjon for naturfag (Kaarstein & Nilsen, 2018), lærerkvalitet og undervisningskvalitet (Nilsen & Blömeke, 2018), effekten av fødselsmåned på skoleprestasjoner (Olsen & Björnsson, 2018a), en teoretisk sammenlikning mellom den norske lærerplanen i naturfag og PISA rammeverket (Jensen et al., 2018), utvikling i matematikkprestasjoner over tid (Olsen & Blömeke, 2018), utvikling i likeverd i norsk skole over tid (Nilsen et al., 2018) og skoleledereres syn på skoleklima (Hatlevik et al., 2018). Det finnes også mange andre. De nevnte artiklene er alle sammen skrevet på norsk, men det finnes også en god del publikasjoner om det norske skolesystemet på engelsk publisert i internasjonale tidsskrifter og bøker. I 2020 publiserte forskergruppen LEA ved Institutt for lærerutdanning og skoleforskning ved Universitetet i Oslo, en bok om likeverd i skolen. Denne antologien inneholder 16 kapitler om likeverd i skolen og er først og fremst basert data fra ILSA-studier (Frønes et al., 2020). Boken inneholder flere nordiske sammenlikninger, og er sånn sett et viktig bidrag for videre forskning og utvikling av likeverd i norsk skole. Vi kan derfor konstatere at det er fart på forskningen om norsk skole, og vi burde ha de beste forutsetningene for å lykkes med å forbedre skolen.

| Undersøkelse | Faglig område | Årstrinn/ alder | Når | Organi-sasjon |
|---|---|---|---|---|
| CIVED/ ICCS | Samfunnsfag: Demokratiforståelse og medborgerskap | 8./9. trinn | 1999, 2009 og 2016 | IEA |
| ICILS | Grunnleggende digital kompetanse | 9. trinn | 2013 og 2018 (Norge deltok ikke i 2018) | IEA |
| PIRLS | Lesing | 4./5. trinn | Hvert 5. år siden 2001 | IEA |
| TEDS-M | Matematikklærerstudenters pedagogiske, didaktiske og matematiske kompetanse | | 2008 | IEA |
| TIMSS | Matematikk og naturfag | 4./5. trinn og 8./9. trinn | Hvert 4. år siden 1995 (Norge deltok ikke i 1999) | IEA |
| TIMSS Advanced | Matematikk og fysikk | 13. trinn (vg3) | 1995, 2008 og 2015 | IEA |
| IALS/ALL/ PIAAC | Voksnes kompetanse (lesing, regning og IKT-basert problemløsing) | 16–65 år | 1998, 2003 og 2012 | OECD |
| PISA | Lesing, matematikk og naturfag | 15-åringer | Hvert 3. år siden 2000 | OECD |
| TALIS | Læreres og skoleledereres arbeidssituasjon | Ungdomstrinnet | Hvert 5. år siden 2008 | OECD |
| TALIS Starting Strong Survey | Barnehageansattes arbeidssituasjon | Barnehager | 2018 | OECD |

Figur 6. Norsk deltakelse i ILSA studier fra 1995 (Olsen & Björnsson, 2018b)

ILSA-studiene kan brukes til å belyse mange forhold i skolesystemet. De kan også brukes til mange ulike analyser av hvordan ting henger sammen og det er naturlig å bruke dem for å sammenlikne skolesystemer i forskjellige land.

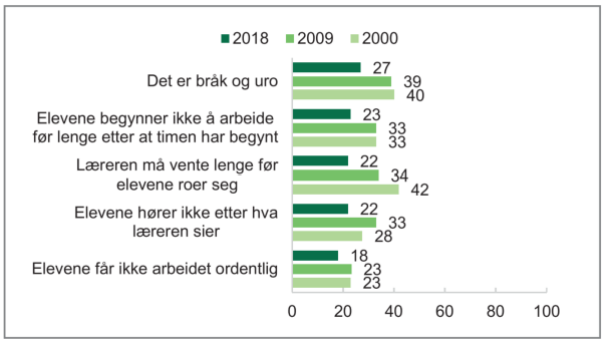
ILSA-studiene kan ikke brukes til å si noe om individuelle prestasjoner eller gjennomsnittlige prestasjoner for små kommune eller enkelteskoler, men de kan brukes for å si noe om trender og utvikling over tid i skolesystemet. Men vi må huske at ILSA-studiene er tverrsnittsmålinger: de tester nesten aldri de samme elevene på forskjellige klassetrinn, selv om det er mulig å få dette til, (for eksempel i TIMSS), så er ikke slike sammenligninger blitt gjort enda. Slike studier er likevel planlagt, det vil si longitudinelle studier som følger elever over tid.

## 6.1. Læringsmiljø, skolemiljø og internasjonale undersøkelser

Funn fra ILSA-studiene legger stor vekt på viktigheten av et godt læringsmiljø for at elevene skal prestere godt og trives på skolen. Internasjonale undersøkelser viser at de nordiske landene i gjennomsnitt kanskje har «verdens beste» skole- og læringsmiljøer, men slike funn må vil likevel ta med noen forbehold. Elevene i deltakerlandene blir spurt om det som ofte er kalt «Disciplinary Climate», som er en evaluering av blant annet bråk, uro og andre forstyrrelser av undervisningen og om elevene kommer for sent eller ikke. Det har vist seg, når skalaene er undersøkt, at det er vanskelig å sammenlikne slike forhold på tvers av land. Kulturelle faktorer spiller her en stor og viktig rolle, og gjør at spørreskjemaene ikke er helt sammenliknbare. Rangeringer av land ut ifra slike resultater er derfor ganske tvilsomme.

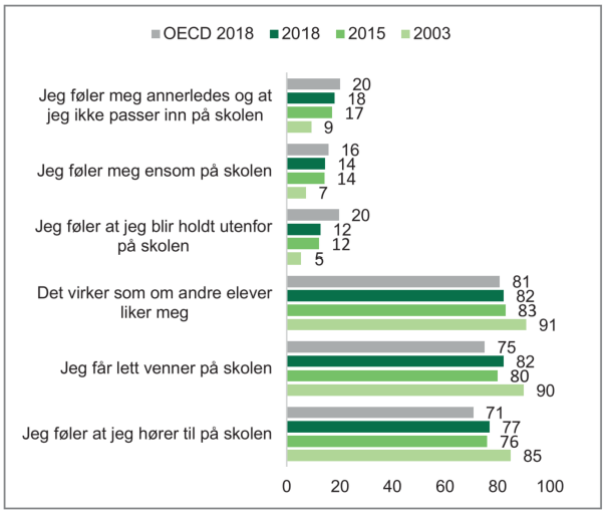
Det er likevel fullt mulig å undersøke sammenhengen mellom ulike variabler som beskriver skolemiljøet og elevprestasjoner innad i et land. I slike sammenlikninger havner ofte de nordiske landene langt nede på skalaen. Negative trekk ved skolemiljøet har mindre forklaringsverdi for prestasjoner i Norden enn i fleste andre land (OECD, 2013); (OECD, 2020).

I PISA 2018 (Jensen et al., 2019), viste det seg at situasjonen i norske klasserom ser ut til å ha bedret seg over tid. Figur 7, hentet fra kortrapporten fra PISA 2018, beskriver utviklingen gjennom de siste 18 årene, og viser at det har skjedd en sakte, men sikker forbedring i målingene av skolemiljø. I tillegg kan det nevnes at Jensen et. al (2019) finner en sterk positiv sammenheng mellom arbeidsro i timene og prestasjoner i lesing, og at dette ser ut til å gjelde for de andre deltakerlandene i PISA også. Det ser det ut til å være liten forskjell mellom gutter og jenter i disse variablene, og effekten av sosioøkonomisk status er liten i Norge i motsetning til fleste andre land. Dette gjelder også stort sett for de andre nordiske landene.




Figur 7. Prosentandel norske elever som svarer «alle timene» eller «de fleste timene) på spørsmålet «Hvor ofte skjer dette i norsktimene?». (Jensen et al., 2019)

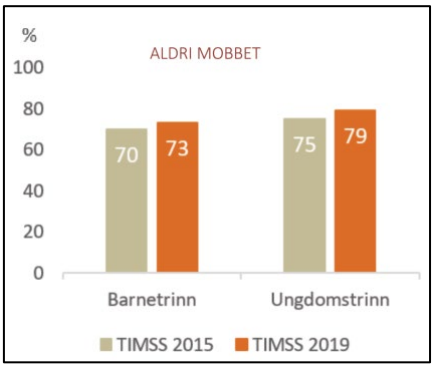
PISA 2018 viser også at de fleste norske elever opplever tilhørighet til skolen, men deres opplevelse er blitt litt dårligere over tid, noe figur 8. viser. Endringene er ikke store, men det ser likevel ut til å gå i negativ retning. Det kan se ut til at negativt formulerte utsagn har en sterkere tendens til å gi negative svar, mens når de blir spurt om positivt formulerte utsagn det mindre endringer. Norske elever er i gjennomsnitt litt under gjennomsnittet i OECD når formuleringene er negativt formulert, mens de positive formuleringene er litt over OECD-snittet. Her er det sikkert et utall kulturelle faktorer som påvirker, og det viktig at man overvåker og vurderer slike resultater nøye. Vi må huske på at man ikke nødvendigvis vil få samme resultat når man vinkler et spørsmål negativt som når det vinkles positivt. Syv prosent av norske elever rapporterer at de ensomme på skolen, men man kan undres om svaret ville blitt det samme hvis spørsmålet hadde handlet om «å ikke være ensom».




Figur 8. Prosentandel av de norske elevene i PISA 2018, 2015 og 2003 og for gjennomsnittet av OECD-landene som svarer «Svært enig» eller «Enig» på spørsmålet «Tenk deg skolen din: Hvor enig er du i disse utsagnene». (Jensen et al., 2019)

Det ser ut til å være mindre skulking i norske skoler sammenliknet med OECD, men at økt skulking har sammenheng med lavere prestasjoner i lesing.

Andelen elever som opplever mobbing er ganske uendret igjennom de siste 20 årene, men er litt lavere i Norge, enn i OECD generelt. Dette stemmer bra med resultatene fra TIMSS 2019, som viser lite mobbing i norske klasserom sammenliknet med andre land. Over 70% av elevene har aldri opplevd mobbing, ifølge TIMSS 2019 (Kaarstein et al., 2020), og TIMSS rapporterer om en økning i elever som rapporterer å aldri ha opplevd dette i 2019 sammenliknet med 2015.




Figur 9. Andelen aldri mobbet – fra TIMSS 2015 og 2019. (Kaarstein et al., 2020)

TALIS-studien undersøker også lærings- og skolemiljøet. Der fremkommer det tydelig at skolemiljøet i Norge karakteriseres av samarbeid og samhold, gode relasjoner og høy trivsel (Throndsen et al., 2019). Disse resultatene var fra perioden før pandemi og lærerstreiker preget norsk skole i de siste to årene. Funnene viser likevel egenskaper ved det norske skolesystemet som har vedvart i mange år, og som bekreftes i mange forskjellige undersøkelser.

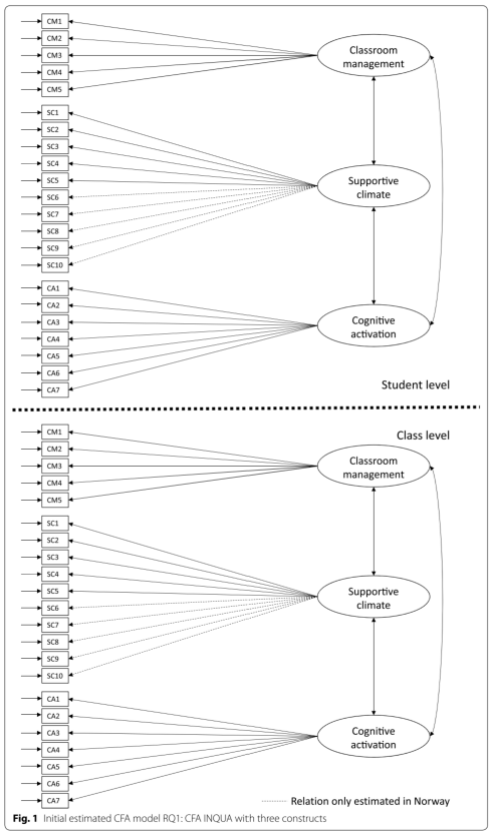
Elevundersøkelsen (Wendelborg et al., 2021) monitorerer noen av disse skolemiljø-variablene over tid, og i den nyeste rapporten er det ikke store endringer i elevenes trivsel. Den er vedvarende høy, selv om det er små indikasjoner på at noen av indeksene går i negativ retning. For 7. trinn rapporterer undersøkelsen at elevene opplever små negative endringer i skolemiljøet, selv om endringen er liten og trenden langsom. For 10. trinn er det liten endring over tid, og bortsett fra en liten reduksjon i mobbing på Vg1 er ingen eller veldig små endringer (ibid. s 29-31).

Selv om det er viktig å se på disse resultatene, er det enda viktigere er å se på hvordan variablene for lærings- og skolemiljøet er relatert til elevprestasjoner. Det finnes mange studier om disse sammenhengene, hvorav noen få blir nevnt her. Noen av disse studiene undersøker enkle sammenhenger, for eksempel hvordan prestasjoner er relatert til skolemiljøet, lærernes aktiviteter, klasseledelse og så videre. Andre studier undersøker liknende sammenhenger, men kontrollerer for flere variabler, for eksempel sosioøkonomisk status eller innvandring, for å nevne noen. Andre studier igjen tar dette enda et skritt videre, og undersøker om nivåene i skolen (elevnivået, klassenivået og til og med skolenivået) ser ut til å innvirkning på sammenhengen mellom skolemiljø og elevprestasjoner. Ofte benyttes modeller som inkluderer vesentlig flere variabler og nivåer samtidig i analysene. Det siste er uhyre viktig, ettersom alt som kan tenkes å påvirke elevprestasjoner virker samtidig, og da kan vi være ganske sikre på at hvis vi bruker *for* enkle sammenhenger eller relasjoner mellom variabler, gjør vi forenklinger, og vil muligens ende med feil konklusjoner.

Det er ikke mulig her å redegjøre for alle studier av den sistnevnte typen her, men noen må likevel nevnes. En studie som sammenlikner læringsmiljø i tre land, Belgia, Tyskland og Norge ser på læringsmiljø som et sammensatt fenomen, hvor klasseledelse, støttende klima og kognitiv aktivering inngår i ett felles konstrukt (Bellens et al., 2019). Denne studien brukte data fra 4. trinn i TIMSS 2015. Forskerne prøvde å svare på tre ulike spørsmål: 1) Virker konstruktet? 2) Hvilken effekt har det på læringsutbytte? 3) Hvordan er relasjonen mellom læringsresultater og bakgrunnsvariabler påvirket av sosioøkonomisk status og innvandrerstatus. Det siste dreier seg om likeverd i skolen og det andre om hva som forklarer prestasjoner på prøven.

Resultatene ble presentert på mange nivåer, og forskerne brukte en såkalt «multilevel structural equation modelling»-metodikk hvor det ble blant annet ble gjort analyser av hvorvidt konstruktet kunne sammenliknes på tvers av land.

Figur 10 viser grunnmodellen som ble prøvd ut.




Figur 10. Multilevel SEM modell

I figuren ser vi at alle relasjoner ble testet, og at alle variabler virker inn på alle nivåer. Det er viktig å understreke at også dette egentlig er en forenkling av realitetene, ettersom det sikkert er mange andre faktorer som påvirker skoler og elever som vi ikke fanger opp i statistiske analyser. Resultatene viser at klasseledelse har en reell innflytelse på matematikkprestasjoner i alle de tre landene. Støttende miljø påvirket matematikkprestasjoner bare i Belgia, men ikke i de to andre landene. Relasjonen mellom kognitiv aktivering og matematikkprestasjoner var ikke signifikant i studien. Forfatterne har lange diskusjoner om hvorfor det er slik, men for vårt formål er det viktigste å huske at *enkle relasjoner er nesten aldri riktige*! De er nesten alltid avhengige av mange andre faktorer. Relasjonene mellom disse variablene er ikke de samme på tvers av land, noe som understreker viktigheten av å gjøre analysene i hvert land, tilpasset kulturen i det aktuelle landet.

En annen nokså lik studie, der det ble brukt data fra de nordiske landene fra PISA 2015 (Rohatgi et al., 2022), viser at støttende klima, det vil se tilbakemeldinger fra læreren og at elevene opplever at læreren er rettferdig, har signifikant kobling til prestasjon i naturfag. Her var det igjen forskjeller mellom de fem landene, noe som kanskje skyldes nokså like kulturer og skolesystemer.

## 6.2. Til slutt om ILSA-studiene

ILSA-studiene har et solid teoretisk og metodologisk grunnlag, men de eksisterer på en måte i sin egen verden og de som ikke har tatt et skritt inn i den verdenen, oppfører seg ofte som om denne forskningen ikke eksisterer. Vi bør gjøre disse resultatene bedre tilgjengelige for skoler og lærere spesielt. Det er mange spennende resultater og interessante funn, og disse bør ikke være forbeholdt akademia eller grupper av forskere på utdanningsfeltet.

Det å utvikle skolepolitikk basert på ILSA-studiene er ikke en enkel oppgave, og det krever forståelse for hva studiene kan brukes til og hva de ikke er egnet for. De leverer først og fremst systemopplysninger, og man bør være varsom med å bruke dem på individ- eller skolenivå.

De siste årene har vist oss at sammenliknbarheten på tvers av land, på bakgrunn av resultatene fra disse studiene, kanskje ikke var så god som mange trodde. Kulturelle faktorer og andre ting som påvirker læringsresultater i forskjellige land er viktige, og påvirker ikke bare hvordan elever presterer på skolen, men også hvordan skolepolitikken utformes. Det går ikke an å kopiere løsninger fra et land til et annet, selv ikke fra land med en nokså lik kultur. Men det å bruke ILSA-resultatene på en fornuftig måte både til å iverksette forbedringer og reformer og, ikke minst, til å følge med på utviklingen av skolesystemet over tid, er fornuftig og god bruk av ILSA-resultater, og som ikke ville være mulig uten dem. Derfor kan man konkludere ganske sikkert at de internasjonale undersøkelsene er noe av det viktigste redskapene vi har for å kunne forbedre skolen. Det dreier seg både om direkte og indirekte effekter: direkte gjennom økt viten og kunnskap om tilstanden i skolen, og indirekte gjennom økt kompetanse i vurdering, og økt forståelse av skolesystemet på alle nivåer fra politikere til lærere.

# 7. Sammenfatning og konklusjoner

Hensikten med dette siste avsnittet er ikke å gjengi noe av det som er blitt beskrevet over, men å trekke noen konklusjoner som kan muligens benyttes i utvalgets videre arbeid. Utvalgets oppdrag er

***«Hvilke konsekvenser har innføringen av kvalitetsvurderingssystemet og de ulike elementene i systemet hatt på elevenes prestasjoner og lærernes pedagogiske praksis i skolen på 2000-tallet.»***

Det var nevnt her innledningsvis, at dette er sannsynligvis en forenklet problemstilling. Det er ikke en enkel oppgave å se og forstå hvordan alle elementene i kvalitetsvurderingssystemet samtidig kan føre til en forbedring i elevenes prestasjoner og/eller lærernes praksis. Her er det nesten ingen enkle årsakssammenhenger. Vi har sett her på noen eksempler på modeller som er blitt brukt for å forstå nettopp denne kompleksiteten og det er tydelig at det er slike modeller hvor alle faktorer og alle nivåer blir analysert samtidig, som vil forhåpentligvis føre oss videre til bedre forståelse av hvilke faktorer det er vi kan endre, til å forbedre både prestasjoner og trivsel i skolen. Å endre bare en ting, basert på enkle sammenhenger, kan ikke bare være galt, men kan føre til feil konklusjoner og ha helt andre effekter enn de tilsiktede.

Det er påfallende når vi sammenlikner empiriske studier av nasjonale prøver og internasjonale storskalaundersøkelser at det finnes få empiriske studier av de førstnevnte prøvene. Stort sett er dette begrenset til kvalitative undersøkelser om brukeropplevelser og effekter på lærere og skoler. Dette betyr slett ikke at disse undersøkelsene er uviktige, men understreker gode kvantitative undersøkelser av nasjonale prøver sjeldent publiseres. Dette problemet har sikkert mange ulike årsaker, og kanskje er mangel på data en av dem.

I ILSA-studiene samles det inn informasjon om elevenes kompetanse og prestasjoner og bakgrunnsdata på flere nivåer. Dette er naturligvis spørreskjemaene som er administrert sammen med selve prøvene. Uten disse bakgrunnsopplysningene ville det vært begrenset med muligheter til å forklare variasjoner i prestasjoner og hvilke faktorer som påvirker læring. Disse forholdene gjør ILSA-studiene velegnet til sekundæranalytisk forskning.

Dette gjelder naturligvis også for nasjonale prøver - prøvene eksisterer i relasjon til andre variabler i skolen, men siden vi ikke samler inn andre opplysninger under gjennomføringene, blir det krevende, for ikke si umulig, å fastslå hvilke faktorer som virker inn på resultatene. Det gjør det vanskelig å

iverksette for eksempel endringer på skolenivå basert på dem, spesielt på små skoler og i små kommuner. Likefullt er det et valg å la alle elever gjennomføre disse prøvene, og det er mange gode grunner for å gjøre det. Det doble formålet er, etter mitt syn, i seg selv problematisk, men det er forståelig at man har valgt å bruke det – det bidrar til å redusere antallet prøver og det demper de konkurransevridende effektene av å gjennomføre styringsprøver. Etter mitt syn er ikke prøvene egnet til å gi en systemoversikt, spesielt hvis de ikke fanger opp endringer over tid.

Trenger vi egentlig nasjonale prøver for å følge med hvordan systemet virker? Svaret er kanskje nei, og da står vi igjen med det formative formålet med prøvene. Dette formålet kan videreutvikles slik at både lærere og elever opplevde nasjonale prøver som mer relevante enn de sannsynligvis gjør i dag. Dette ville bidra til å øke og forbedre kompetansen hos skoler og lærere, og dermed ha en positiv effekt på elevprestasjoner og evalueringskulturen i norsk skole.

Hva kan man så gjøre? De følgende punktene er et forsøk på å dra noen konklusjoner etter denne gjennomgangen, og jeg skal forsøksvis prøve å belyse både fordeler og ulemper med de ulike løsningene:

- Nasjonale prøver måler for snevert til å fange opp systemopplysninger på en god måte, og spesielt endringer over tid. De måler upresist både øverst og nederst på ferdighetsskalaen, noe som gjør de mindre nyttige for å hjelpe svake/sterke elever i den videre opplæringen.
- Nasjonale prøver kunne videreutvikles til såkalte «Multi-stage»-adaptive prøver. Da ville de kunne måle bredere i den forstand at de ville fange opp kompetanser både øverst nederst på skalaen med bedre presisjon. Dette vil selvsagt få noen negative konsekvenser for det formative formålet og for mulighetene til å gi god oppfølging i klasserommet.
- Nasjonale prøver leverer sannsynligvis ikke gode systemmålinger, men dette kunne forbedres hvis man samlet inn bakgrunnsopplysninger om elevene, læringen deres, om skolene og muligvis om skolenes miljø. Dette kan gjøres uten å samle inn følsomme data, men kunne konsentreres om forhold ved skolene og skolemiljøet og selve opplæringen i klasserommet. Dette ville kunne, samtidig som man gjorde målingen mer presis over hele skalaen, føre til betydelig økt nytte av prøvene for systemutvikling.
- Nasjonale prøver blir kvalitetssikret og gjennomgått på mange måter før de blir brukt, men Utdanningsdirektoratet publiserer ikke rapporter regelmessig fra prøvekonstruksjonen. Dette ville gjøre det mer attraktivt for forskere å arbeide med prøvene og ville kunne fasilitere økt og forbedret forskning på hele prøvesystemet.
- ILSA-studiene leverer sannsynligvis meget god systeminformasjon, men denne kunne relateres bedre til både læreplaner og til nasjonale prøver. Disse studiene egner seg ikke til å rapportere på individnivå og sier heller ikke noe om mindre grupper, kommuner eller skoler, og blant annet derfor burde nasjonale prøver videreutvikles for å måle bredere og mer presist.
- Forskningen basert på ILSA-studiene har påvist viktigheten av bakgrunnsopplysninger for å forstå variasjonen i elevprestasjoner, og hvis noen slike opplysninger kunne lenkes til prøveresultatene fra nasjonale prøver ville deres forskningsverdi bli mangedoblet. Her kunne man kanskje tenke seg en kobling til de mindre følsomme delene av elevundersøkelsen.
- En stor mangel i det nåværende systemet er at det finnes ikke noen mulighet for å følge elever over tid, dvs. til å gjøre longitudinelle undersøkelser. ILSA-studiene gjør ikke dette, ettersom de er tverrsnittsundersøkelser. Det er også vanskelig å følge elevenes utvikling fra 5. trinn til 8./9. trinn på nasjonale prøver. For det første er tallene som prøvene leverer ikke sammenliknbare for de ulike trinnene og for det andre så er resultatene ikke oppbevart sentralt slik at en slik longitudinell oppfølgning ville være mulig. Det har likevel blitt gjort et

forsøk til å samkalibrere elevdata fra 5. og 8. trinn (Ræder & Olsen, 2020) og det er mye som tilsier at dette burde være en mulighet. Det ville kunne gjøre det enklere å følge elevers progresjon. Ellers kan det nevnes at det er igangsatt et par forskningsprosjekt som har planer om å følge opp elever som har tatt TIMSS, men det gjenstår å se hvordan det vil gå. Men en longitudinell komponent i prøvesystemet ville være en positiv utvikling.

# Referanser


Allerup, P., Kovač, V. B., Kvåle, G., Langfeldt, G., & Skov, P. (2009). *Evaluering av det Nasjonale kvalitetsvurderingssystemet for grunnopplæringen* (Vol. 8/2009). Agderforskning.

Arntzen, R., Andreassen, U. R., Karlsen, J., & Kvifte, B. H. (2019). Teststrategier og testmotivasjon: Fire niendeklassingers erfaringer med nasjonal prøve i lesing. *Nordic journal of literacy research*, *5*(2), 79-99. https://doi.org/10.23865/njlr.v5.1411

Bellens, K., Van Damme, J., Van Den Noortgate, W., Wendt, H., & Nilsen, T. (2019). Instructional quality: catalyst or pitfall in educational systems' aim for high achievement and equity? An answer based on multilevel SEM analyses of TIMSS 2015 data in Flanders (Belgium), Germany, and Norway. *Large-scale assessments in education*, *7*(1), 1-27. https://doi.org/10.1186/s40536-019-0069-2

Bjørnsson, J. K. (2015). *Metodegrunnlag for nasjonale prøver*. Utdanningsdirektoratet.

Björnsson, J. K. (2018). Om lenkefeil og ekvivaleringsmetoder på nasjonale prøver: Evaluering av endring over tid. *Acta didactica Norge [elektronisk ressurs]*, *12*(4), 24-24. https://doi.org/10.5617/adno.6273

Blömeke, S., & Olsen, R. V. (2018). På vei mot et sammenhengende nasjonalt kvalitetsvurderingssystem. *Acta didactica Norge*, *12*(4), 1. https://doi.org/10.5617/adno.6278

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In (pp. 143-230). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-2324-5_4

Embretson, S. E., & Reise, S. P. (2013). *Item Response Theory for Psychologists*. Taylor and Francis.

Forsbakk, B., & Nortvedt, G. A. (2021). Prøven er gjennomført, hva nå? En studie av praksiser i arbeid med nasjonale kartleggingsprøver i regning. In. Universitetsforlaget.

Frønes, T. S., Pettersen, A., Radišić, J., & Buchholtz, N. (2020). *Equity, Equality and Diversity in the Nordic Model of Education* (1st 2020. ed.). Springer International Publishing : Imprint: Springer.

Grandemo, L. I. (2017). Hva kjennetegner ledelsens tenkning ved skoler som over tid scorer høyt på nasjonale prøver? *Norsk pedagogisk tidskrift*, *101*(1), 19-30. https://doi.org/10.18261/issn.1504-2987-2017-01-03

Gunnulfsen, A. E. (2017). School leaders' and teachers' work with national test results: Lost in translation? *Journal of educational change*, *18*(4), 495-519. https://doi.org/10.1007/s10833-017-9307-y

Hatlevik, O. E., Rohatgi, A., & Björnsson, J. K. (2018). Skoleledernes syn på skoleklima: noen erfaringer fra PISA og TIMSS fra 2003 og 2015. In. Universitetsforlaget.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333-353. https://doi.org/10.1080/00313831.2016.1258726

Hovdhaugen, E., Vibe, N., & Seland, I. (2017). National test results: representation and misrepresentation. Challenges for municipal and local school administration in Norway. https://doi.org/https://doi.org/http://dx.doi.org/10.1080/20020317.2017.1316636

Høst, H. (2015). Kvalitet i fag- og yrkesopplæringen: Sluttrapport. In: NIFU.

Jensen, F., Mork, S. M., & Kjærnsli, M. (2018). En sammenligning av naturfagkompetanser i PISA-rammeverket 2015 og den norske læreplanen. In. Universitetsforlaget.

Jensen, F., Pettersen, A., Frønes, T. S., Eriksen, A., & Narvhus, E. (2019). *pisa 2019: Norske elevers kompetanse i lesing, matematikk og naturfag*. Universitetsforlaget.

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355-381. https://doi.org/DOI 10.1111/j.1745-3984.2006.00021.x
Kaarstein, H., & Nilsen, T. (2018). Norske elevers motivasjon for naturfag gjennom 20 år. In J.K.Bjørnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i Norge*. Universitetsforlaget.
Kaarstein, H., Radisic, J., Lehre, A.-C. W. G., Nilsen, T., & Bergem, O. K. (2020). *TIMSS 2019 Kortrapport*. Universitetsforlaget.
Landmark, E. (2018). Nasjonale prøver som kvalitativt styringsverktøy for alle elever - et systembidrag til kvalitet i opplæringen. In.
Lie, S., Hopfenbeck, T. N., Ibsen, E., & Turmo, A. (2005). Nasjonale prøver på ny prøve : rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005. *Acta didactica Norge*, *1/2005*.
Martin, M. O. E., von Davier, M. E., & Mullis, I. V. S. E. (2020). *Methods and Procedures: TIMSS 2019 Technical Report* (9781889938530,188993853X).
Mausethagen, S., Prøitz, T. S., Skedsmo, G., Practices of Data Use in, M., & Schools. (2018). *Elevresultater : mellom kontroll og utvikling*. Fagbokforl.
Nilsen, T., Björnsson, J. K., & Olsen, R. V. (2018). Hvordan har likeverd i norsk skole endret seg de siste 20 årene? In J. K. Bjørnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i Norge*. Universitetsforlaget.
Nilsen, T., & Blömeke, S. (2018). Lærerkvalitet, undervisningskvalitet, -kvantitet og prestasjon: Analyser av TIMSS 2015 data i naturfag på barnetrinnet. In J. K. Bjørnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i Norge*. Universitetsforlaget.
Nilsen, T., Stancel-Piątak, A., & Gustafsson, J.-E. (2022). *International Handbook of Comparative Large-Scale Studies in Education : Perspectives, Methods and Findings*. Springer International Publishing AG.
OECD. (2012). *PISA 2009 Technical Report*. OECD Publishing.
OECD. (2013). How the quality of the learning environment is shaped. In (pp. 165-188). Paris: OECD Publishing. https://doi.org/10.1787/9789264201156-9-en
OECD. (2020). *PISA 2018 Results (Volume V)*. OECD Publishing.
Olsen, R. V., & Björnsson, J. K. (2018a). Fødselsmåned og skoleprestasjoner. In. Universitetsforlaget.
Olsen, R. V., & Björnsson, J. K. (2018b). Tjue år med internasjonale skoleundersøkelser i Norge: Bakgrunn, læringspunkter og veien videre. In J.K.Bjørnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i Norge*. Universitetsforlaget.
Olsen, R. V., & Blömeke, S. (2018). Hva forklarer endringer i elevenes matematikkprestasjoner over tid? In J.K.Bjørnsson & R. V. Olsen (Eds.), *Tjue år med TIMSS og PISA i NOrge*. Universitetsforlaget.
Olsen, R. V., Tveit, S., & Björnsson, J. K. (2018). Nasjonale prøver og eksamener i norsk og svensk grunnopplæring. *Acta didactica Norge*, *12*(4). https://doi.org/10.5617/adno.6647
Roe, A., Ryen, J. A., & Weyergang, C. (2018). *God leseopplæring med nasjonale prøver : om elevers leseutfordringer i et mangfold av tekster*. Universitetsforl.
Rohatgi, A., Hatlevik, O. E., & Björnsson, J. K. (2022). Supportive climates and science achievement in the Nordic countries: lessons learned from the 2015 PISA study. *Large-scale assessments in education*, *10*(1), 1-28. https://doi.org/10.1186/s40536-022-00123-x
Ræder, H. G., & Olsen, R. V. (2020). *Utvikling av nasjonale prøver - rapport 2b. Endelig rapport av vertikalt lenkedesign for de nasjonale prøvene i regning 5.-8. trinn*. U. i. Oslo.
Scheerens, J. (1990). School Effectiveness Research and the Development of Process Indicators of School Functioning. *School effectiveness and school improvement*, *1*(1), 61-80. https://doi.org/10.1080/0924345900010106
Seland, I., Vibe, N., & Hovdhaugen, E. (2013). Evaluering av nasjonale prøver som system. In: NIFU.
Sinharay, S., & Holland, P. (2006). CHOICE OF ANCHOR TEST IN EQUATING. *ETS Research Report Series*, *2006*(2), i-43. https://doi.org/10.1002/j.2333-8504.2006.tb02040.x

Tan, X., & Michel, R. (2011). Why Do Standardized Testing Programs Report Scaled Scores?: Why Not Just Report the Raw or Percent-Correct Scores? *ETS R & D Connections*, *16*.

Throndsen, I., Carlsten, T. C., & Bjørnsson, J. K. (2019). *TALIS 2018:Første hovedfunn fra ungdomstrinnet*.

Utdanningsdirektoratet. (2022). *Rammeverk for nasjonale prøver*. Utdanningsdirektoratet Oslo. Retrieved 04.11.22 from https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover2/

Vestheim, O. P. (2018). Nasjonale prøver – hemmende styringsverktøy eller lokale redskap for praksisutvikling? *Acta didactica Norge*, *12*(4), 3. https://doi.org/10.5617/adno.6249

Walgermo, B. R., Uppstad, P. H., Lundetræ, K., Tønnessen, F. E., & Solheim, O. J. (2018). Kartleggingsprøver i lesing - tid for nytenking? *Acta didactica Norge*, *12*(4), 7. https://doi.org/10.5617/adno.6499

Wendelborg, C., Wendelborg, C., & inkludering, N. s. M. o. (2021). *Elevundersøkelsen 2020 : analyse av Utdanningsdirektoratets brukerundersøkelser*. NTNU Samfunnsforskning.

Werler, T., & Færevaag, M. K. (2017). National testing data in Norwegian classrooms: a tool to improve pupil performance? *Nordic journal of studies in educational policy*, *3*(1), 67-81. https://doi.org/10.1080/20020317.2017.1320188