KAPITTEL 14

Rammeverk og metoder

Torgeir Onstad
Institutt for lærerutdanning og skoleforskning, UiO
Liv Sissel Grønmo
Institutt for lærerutdanning og skoleforskning, UiO

Dette kapitlet gir en kortfattet beskrivelse av bakgrunnen for TIMSS Advanced og en gjennomgang av hvordan studien ble planlagt og gjennomført. Framstillingen bygger på kapittel 12 i den norske matematikkrapporten fra TIMSS Advanced 2008 (Onstad, 2010b) og kapittel 7 i den norske rapporten fra TIMSS Advanced 2015 (Onstad & Grønmo, 2016).

14.1 Hva er TIMSS Advanced?

14.1.1 Historikk

TIMSS er en forkortelse for *Trends in International Mathematics and Science Study*. Det er først og fremst en stor internasjonal undersøkelse av matematikk og naturfag i grunnskolen. TIMSS beskriver og sammenlikner elevprestasjoner i disse fagene, så vel nasjonalt som internasjonalt, og søker å belyse og forstå forskjeller i prestasjoner ut fra andre data i undersøkelsen. Slik kan man si noe om hvilke faktorer som fremmer læring, og hvilke som hemmer læring.

Etter noen slike studier i grunnskolen på 1960-, 1970- og 1980-tallet utvidet man i 1995 omfanget til også å gjelde elever i videregående skole. Da definerte man følgende tre populasjoner på øverste trinn i videregående skole:

- **Generalistene**
  Denne populasjonen besto av alle elever i samtlige studieretninger på øverste trinn i videregående skole. Disse elevene ble testet i allmenne matematikk- og naturfagkunnskaper.

- **Fysikkspesialistene**
  Denne populasjonen besto av de elevene som tok høyeste spesialisering i fysikk; i Norge betydde det den gangen elevene på kurset 3FY.
- **Matematikkspesialistene**
  Denne populasjonen besto av de elevene som tok høyeste spesialisering i matematikk (på engelsk *advanced mathematics*); i Norge betydde det den gangen elevene på kurset 3MX.

Etter 1995 har TIMSS-undersøkelser i grunnskolen blitt gjennomført regelmessig hvert fjerde år, nå senest i 2015. *TIMSS Advanced* er en videreføring av undersøkelsene av fysikk- og matematikkspesialistene i videregående skole og har etter 1995 blitt gjennomført i 2008 og 2015.

Norge har deltatt i nesten samtlige TIMSS- og TIMSS Advanced-studier. I 1995 deltok vi imidlertid bare i de to første av de tre populasjonene på videregående nivå, altså generalistene og fysikkspesialistene. Myndighetene ønsket likevel en undersøkelse også av matematikkspesialistene, og i 1998 gjennomførte man den samme matematikkstudien i Norge som hadde vært gjennomført internasjonalt i 1995. Det ble utgitt en samlet norsk rapport for disse tre undersøkelsene (Angell, Kjærnsli & Lie, 1999).

Det at Norge gjennomførte matematikkundersøkelsen i etterkant av den internasjonale studien, hadde visse konsekvenser. De norske resultatene kom ikke med i den internasjonale databasen. De var ikke med i grunnlaget for den standardiserte skalaen og beregningen av det internasjonale skalerte gjennomsnittet. Det betyr at det er noe større usikkerhet forbundet med norske matematikkdata fra 1998 enn det ville ha vært dersom Norge hadde deltatt i 1995. Vi får likevel et godt inntrykk av hvordan Norge gjorde det i 1998 i forhold til andre land i 1995, slik dataene ble analysert i den norske rapporten den gang (Angell et al., 1999), og et godt grunnlag for å vurdere hvordan de norske prestasjonene i matematikk har forandret seg fra 1998 til 2008 og 2015.

Oppslutningen om TIMSS Advanced har vært betydelig lavere enn det vi er vant til i TIMSS. Tabell 14.1 viser de landene som deltok i henholdsvis 1995, 2008 og 2015.

Tabell 14.1 Deltakerland i TIMSS Advanced i 1995 2008 og 2015. Land som har deltatt flere ganger er gulfarget

Land	Deltok i 1995	Deltok i 2008	Deltok i 2015
Armenia		X	
Australia	(x)		
Canada	X		
Danmark	(x)		
Filippinene		M	
Frankrike	X		X
Hellas	X		
Iran		X	
Israel	(x)		
Italia	M	X	X
Kypros	X		
Latvia	F		
Libanon		X	X
Litauen	M		
Nederland		X	
Norge	Fm	X	X
Portugal			X
Russland	X	X	X
Slovenia	(x)	X	X
Sveits	X		
Sverige	X	X	X
Tsjekkia	X		
Tyskland	X		
USA	(x)		X
Østerrike	(x)		

X: Deltok på ordinær måte i begge fag
(x): Deltok, men med for små utvalg
M: Deltok bare i matematikk
F: Deltok bare i fysikk
Fm: Deltok ordinært i fysikk, men avholdt matematikkstudien i 1998

Totalt har altså 25 land deltatt minst én gang i TIMSS Advanced. Av de ni landene som deltok i 2015, har åtte deltatt én eller to ganger før.

14.1.2 Organisering

Det overordnede ansvaret for utviklingen og gjennomføringen av alle TIMSS-studiene, deriblant TIMSS Advanced, ligger hos den internasjonale organisasjonen IEA (*International Association for the Evaluation of Educational Achievement*). IEA er et internasjonalt nettverk for utdanningsforskning som ble etablert i 1959. Det internasjonale prosjektsenteret er lagt til Boston College i USA. Ansvar knyttet til statistisk design og databehandling er delegert til Data Processing and Research Center i Hamburg og Statistics Canada i Ottawa.

I Norge er det Utdanningsdirektoratet som på vegne av Kunnskapsdepartementet har ansvaret for norsk deltakelse og bevilgning av midler. Ansvaret for gjennomføringen av og rapporteringen fra studiene er delegert til Institutt for lærerutdanning og skoleforskning (ILS) ved Universitetet i Oslo. Prosjektet er der organisert med en prosjektleder og prosjektgruppe som har arbeidet med TIMSS Advanced i flere år. Det er en tilsvarende prosjektgruppe på ILS for TIMSS-undersøkelsene i grunnskolen. Disse prosjektgruppene er tilknyttet Enhet for kvantitative utdanningsanalyser (EKVA) ved ILS.

Den norske prosjektgruppen for TIMSS Advanced har samarbeidet med prosjektsenteret i Boston, IEAs sekretariat i Amsterdam, Data Processing and Research Center i Hamburg, Statistics Canada, og med de nasjonale prosjektgruppene i noen av de andre deltakerlandene. Den norske prosjektgruppen har hatt to medlemmer i SMIRC (*Science and Mathematics Item Review Committee* – en internasjonal gruppe oppnevnt av prosjektsenteret i Boston), som har hatt et overordnet ansvar for oppgavene som er blitt brukt i de faglige testene. Disse to medlemmene har også sittet i et mindre arbeidsutvalg (*Task Force*) for SMIRC.

Informasjon om ulike hovedaktører finnes på følgende nettsider:

- IEA: http://www.iea.nl/
- Prosjektsenteret i Boston: http://timssandpirls.bc.edu/
- ILS: http://www.ils.uio.no/
- TIMSS Advanced og TIMSS i Norge: http://www.timss.no/

14.1.3 Populasjoner

Når det gjelder hvilke populasjoner som blir undersøkt, er det viktige forskjeller mellom TIMSS i grunnskolen og TIMSS Advanced i videregående skole. I grunnskolen undersøker TIMSS et representativt utvalg av *hele årskullet* på 4. trinn og på 8. trinn (5. og 9. trinn i Norge i 2015). TIMSS Advanced undersøker betraktelig snevrere grupper, nemlig de elevene på øverste trinn i den videregående skolen som har valgt det eller de kurs som vedkommende land har definert som avansert matematikk eller fysikk. I Norge i 2015 gjaldt det kursene Matematikk R2 og Fysikk 2. Elever som tok begge disse kursene tilhørte begge populasjonene.

Læreplaner er forskjellige fra land til land. Man skal ikke lære nøyaktig det samme på samme trinn i alle land. Når det gjelder matematikkplanene for barnetrinnet, er likevel likhetene langt mer slående enn ulikhetene. Det er påfallende samstemmighet i de fleste land om det faglige innholdet i matematikken i barneskolen og ganske stor enighet om innholdet i naturfag. Forskjellene blir litt større når vi kommer til ungdomstrinnet, men fortsatt er det stor grad av samsvar. I videregående utdanning øker variasjonene. Det gjelder for eksempel hvor mye matematikk som er obligatorisk, hvilke kurs som tilbys, hvilket matematisk innhold disse kursene har, hvilke fagkombinasjoner kursene eventuelt inngår i og hvilke kurs som kreves for ulike typer høyere utdanning. Tilsvarende kan sies for fysikk.

Det er bare elevene som tar de kursene som er definert som avansert matematikk i det enkelte land, som utgjør landets matematikkpopulasjon i TIMSS Advanced. Tabell 14.2 øverst på neste side viser hvor stor prosentandel denne populasjonen er av årskullet i hvert deltakerland. Det dreier seg altså ikke om andelen av skoleelevene, men om *andelen av hele det aktuelle årskullet* i befolkningen. Denne prosentsatsen kalles *dekningsgraden* (*coverage index*) for hvert land.

Det er store variasjoner i dekningsgrad i matematikk, fra under 4 % til godt over 30 %. For Libanon avspeiler den lave prosentandelen trolig landets mangel på ressurser til videregående utdanning. I den andre enden av skalaen finner vi Slovenia; der tar ca. en tredel av årskullet avansert matematikk. Skulle vi overført Slovenias prosentsats til Norge, ville det betydd at de fleste av elevene på studieforberedende programmer skulle ha tatt Matematikk R2. I Norge lyder det helt utenkelig at en så stor andel av elevene skal ta avansert matematikk til topps, men i Slovenia er det tilfellet.

Tabell 14.2 Dekningsgrad: matematikkpopulasjonen i TIMSS Advanced i prosent av hele årskullet

Land	Dekningsgrad i matematikk i prosent av hele årskullet
Libanon	3,9
Russland*	10,1
Norge	10,9
USA	11,4
Sverige	14,1
Frankrike	21,5
Italia	24,5
Portugal	28,5
Slovenia	34,4

* I 2008 testet Russland bare elever som tok svært avanserte matematikkurs. Da var dekningsgraden bare 1,4 %. I 2015 valgte de å definere flere kurs som avanserte, og dermed ble dekningsgraden høyere. For å kunne gjøre fornuftige trendanalyser rapporterer de denne gangen resultatene både til hele gruppen, og til den delgruppen som svarer til populasjonen i 2008. I boka refererer vi stort sett bare russiske resultater fra hele gruppen.

Hvis vi vil sammenlikne prestasjonene i matematikk for flere land i TIMSS Advanced, er det viktig å ha dekningsgraden i mente.

14.1.4 Analysenivåer

I TIMSS Advanced og TIMSS analyseres data på tre nivåer:

Systemnivå — intendert læreplan

Dette nivået gjelder utdanningssystemet slik det legges til rette av nasjonale og regionale myndigheter i et land. Det dreier seg om organisering av skoletilbudet, rammefaktorer, ressurstilgang og elevenes muligheter til skole- og fagvalg. Ikke minst dreier det seg om læreplaner og vurderingsformer. Det er slike faktorer som forteller hva slags utdanningstilbud samfunnet og myndighetene ønsker og planlegger at elevene skal få. Opplysninger på dette nivået er primært hentet inn fra de nasjonale prosjektlederne i de enkelte deltakerlandene.

Det er utgitt en ensyklopedi med beskrivelser av skolesystemene i alle deltakerlandene i TIMSS 2015 (Mullis, Martin, Goh & Cotter, 2016). Samtlige deltakerland i TIMSS Advanced 2015 er med der. Selv om hovedvekten i ensyklopedien er på grunnskolen (*primary education* og *lower secondary*

education), kan den gi en viss støtte for å forstå ulikheter mellom landene på systemnivået. Dessuten inneholder den internasjonale rapporten for TIMSS Advanced 2015 (Mullis, Martin, Foy & Hooper, 2016b) ytterligere opplysninger om skolesystemene i deltakerlandene, med særlig vekt på videregående opplæring.

Klasseromsnivå — implementert læreplan

Dette nivået handler om hva som skjer i klasserommet, om undervisningen og læringsmiljøet. Hvordan blir intensjonene fra systemnivået omsatt i praksis? Hvordan blir den intenderte læreplanen iverksatt i skolen?

Både elevene, lærerne deres (i det faget elevene ble testet i) og skolelederne deres har svart på spørreskjemaer om situasjonen på skolen. Elevene ble blant annet spurt om hjemmebakgrunn, utdanningsplaner, trivsel på skolen, tidsbruk på skolearbeid og på arbeid utenom skolen, og om undervisningsmetoder i matematikk og fysikk. Lærerne ble blant annet spurt om alder, utdanning, erfaring som lærer, etter- og videreutdanning, faglige emner som er undervist, undervisningsmetoder, bruk av digitale verktøy, bruk og oppfølging av lekser, lærersamarbeid, trygghet og trivsel i jobben, og om eventuelle problemer i arbeidssituasjonen. Skolelederne ble blant annet spurt om utdanning og ledererfaring, skolens ressurser og begrensninger, elevenes bakgrunn, skolens vektlegging av matematikk og fysikk, eventuelle problemer med å rekruttere kvalifiserte lærere, og generelt om skolens miljø.

Elevnivå — resultert læreplan

Det siste nivået handler om hva som er oppnådd. Hvilke kunnskaper har disse elevene i matematikk og fysikk, og hvilke holdninger har de til fagene? Elevenes prestasjoner på den faglige testen ga informasjon om faglige kunnskaper og ferdigheter, mens elevspørreskjemaet ga informasjon om holdninger til fag og læring.

Med data på alle disse nivåene kan man beskrive og analysere situasjonen på en rekke måter. Vi kan studere forandringer i forhold til den forrige TIMSS Advanced-undersøkelsen. Vi kan sammenlikne elevprestasjoner i ulike land. Vi kan sammenlikne prestasjonene til jenter og gutter. Vi kan også analysere om det synes å være sammenheng mellom prestasjonene og noen av bakgrunnsvariablene, som for eksempel undervisningsmetoder, leksearbeid, lærernes utdanning eller elevenes hjemmebakgrunn.

14.2 Rammeverk og instrumenter

TIMSS Advanced baserer seg på et rammeverk som definerer hvilke kunnskaper og ferdigheter elevene skal testes i. Rammeverket er utviklet gjennom en drøftingsprosess mellom deltakerlandene som leder fram mot konsensus om hva som utgjør sentrale kunnskaper og ferdigheter i faget sett i forhold til de respektive landenes læreplaner. Det foregår en viss justering foran hver undersøkelse, noe som er naturlig ettersom skolesystemer utvikler seg og læreplaner revideres. Men det er samtidig et poeng å holde rammeverket relativt stabilt for å gi et solid fundament for pålitelige sammenlikninger over tid.

14.2.1 Rammeverk

Rammeverket for TIMSS Advanced 2015 (Mullis & Martin, 2014) bygger på rammeverket for TIMSS Advanced 2008 (Garden et al., 2006). Det er et mål at rammeverket skal ligge så tett som mulig opp til de aktuelle læreplanene i deltakerlandene. Det er selvsagt umulig å få det til fullt ut; til det er læreplanene for ulike, spesielt når man kommer til de høyere trinnene i skoleverket. Derfor blir målet i stedet at ikke noe land skal oppleve at det blir et urimelig stort avvik fra deres læreplan. Vi skal helst alle sammen kunne si at testen i hovedsak faller inn under vår læreplan. Samtidig aksepterer vi at noen av oppgavene ikke passer godt i vårt land og at noen deler av vår læreplan ikke dekkes av testen. For å oppnå dette er det viktig at alle deltakerlandene gis anledning til å påvirke prosessen med utvikling av rammeverket, slik at man oppnår konsensus om det.

Rammeverket definerer de *faglige innholdskategoriene* som testoppgavene skal hentes fra. Disse kategoriene er organisert i noen temaområder. Samtidig oppgis det hvor stor andel av oppgavene som bør høre inn under hver av disse innholdskategoriene.

I tillegg inneholder rammeverket en beskrivelse av *kognitive kategorier*. Det er et mål at oppgavene skal stille ulike kognitive krav til elevene. Derfor angir rammeverket også hvor stor andel av oppgavene som bør ligge i hver av de kognitive kategoriene.

Innholdskategorier i matematikk

Innholdskategoriene i matematikk med anbefalt og faktisk fordeling av oppgavene er vist i tabell 14.3 Kategoriene og den anbefalte oppgavefordelingen er de samme som i 2008.

Tabell 14.3 Fordeling av matematikkoppgaver i TIMSS Advanced 2015 i innholdskategorier

Innholds-kategori	Anbefalt prosentandel av oppgavene	Faktisk prosentandel av oppgavene
Algebra	35 %	35 %
Kalkulus	35 %	36 %
Geometri	30 %	29 %

Tabell 14.4 viser hvilke temaområder som inngår i hver av innholdskategoriene.

Tabell 14.4 Temaområder i innholdskategoriene i matematikk i TIMSS Advanced 2015

Innholdskategori	Temaområder
Algebra	Uttrykk og operasjoner Likninger og ulikheter Funksjoner
Kalkulus	Grenseverdier Derivasjon Integrasjon
Geometri	Geometri uten og med koordinater Trigonometri

Fulle detaljer finnes i rammeverket (Mullis & Martin, 2014).

Kognitive kategorier i matematikk

TIMSS Advanced 2015 brukte de samme kognitive kategoriene og den samme anbefalte fordelingen av oppgaver som i 2008. Tabell 14.5 viser disse, samt den faktiske oppgavefordelingen i 2015.

Å *kunne* betyr blant annet å huske fakta, å gjenkjenne matematiske størrelser som er ekvivalente, å beherske algoritmer (som for eksempel løsning av enkle likninger og derivasjon av polynomfunksjoner), og å hente informasjon fra grafer

Tabell 14.5 Fordeling av matematikkoppgaver i TIMSS Advanced 2015 i kognitive kategorier

Kognitiv kategori	Anbefalt prosentandel av oppgavene	Faktisk prosentandel av oppgavene
Kunne	35 %	29 %
Anvende	35 %	41 %
Resonnere	30 %	30 %

og tabeller. Å *anvende* betyr blant annet å bruke kunnskapene og ferdighetene sine til å velge metoder og strategier, å representere matematisk informasjon på ulike måter, å modellere situasjoner, og å løse rutineoppgaver. Å *resonnere* betyr blant annet å tenke logisk, å analysere informasjon, å avgjøre hvilke framgangsmåter som trengs for å løse et problem, å kombinere ulike kunnskapselementer og representasjoner, å vurdere ulike strategier og løsninger, å trekke gyldige konklusjoner, å generalisere resultater, og å formulere matematiske argumenter og bevis.

Fulle detaljer finnes i rammeverket (Mullis & Martin, 2014).

Vi ser at testen inneholdt noen færre kunne-oppgaver og noen flere anvende-oppgaver enn planlagt. Det er vanskeligere å oppnå internasjonal enighet om den kognitive kategoriseringen enn den innholdsmessige. En oppgave som er klart rutinepreget i ett land – ut fra deres læreplan og undervisningstradisjoner – kan vurderes som en krevende problemløsingsoppgave med store utfordringer til resonnement i et annet land. Av den grunn har vi i denne boka valgt å legge liten vekt på å analysere resultatene i TIMSS Advanced basert på den internasjonale kategoriseringen av oppgavenes kognitive nivå.

Digitale hjelpemidler

Spørsmålet om å tillate bruk av kalkulator på testene har fulgt TIMSS og TIMSS Advanced i en årrekke. På 4. trinn har det aldri vært tillatt å bruke kalkulator. På ungdomstrinnet ble det imidlertid vanlig i mange land å introdusere kalkulatorbruk i undervisningen og til eksamen. For disse landene ville det være naturlig å tillate kalkulatorbruk også på tester. Andre land brukte ikke kalkulator i det hele tatt. Det kunne skyldes mangel på ressurser, men det kunne også være et pedagogisk begrunnet valg.

I TIMSS 2003 førte denne situasjonen til en ekstra studie knyttet til testen på 8. trinn. Elevene der fikk ikke bruke kalkulator på første del av testen, men fikk lov på siste del. Enhver oppgave var plassert i første del av noen oppgave-

hefter og i siste del av andre hefter. Dermed kunne man analysere effekten av å ha tilgang til kalkulator. Kalkulatorbruk ga signifikant utslag på prestasjonen på bare fem av i alt 194 oppgaver. Internasjonalt hadde 63 % av elevene tilgang til kalkulator under testen, men bare 15 % oppga at de hadde brukt den en del eller mye. I Norge hadde 80 % av elevene tilgang til kalkulator, men bare 8 % oppga at de hadde brukt den en del eller mye.

Med utgangspunkt i denne erfaringen ble det bestemt at kalkulatorbruk skulle være tillatt under hele testen på 8. årstrinn i TIMSS 2007. Dermed slapp elever som var vant til utstrakt kalkulatorbruk å føle at de ble satt i en uvant testsituasjon. Men mange oppgaver var konstruert slik at det ikke lå til rette for enkel kalkulatorbruk.

I TIMSS Advanced i 1995 var kalkulator tillatt. Det ble tidlig bestemt at kalkulator skulle være tillatt i 2008 også. En grunn til å tillate kalkulator var, som på 8. trinn, at elevene skulle kunne møte testen med de samme rammebetingelsene som de var vant til fra prøver og eksamener i egen skolegang. En annen begrunnelse var at for å kunne sammenlikne prestasjoner i 1995 og 2008, var det viktig å ha samme testbetingelser. De samme begrunnelsene har vært brukt for å tillate kalkulatorbruk også i 2015.

Den norske prosjektgruppen har problematisert denne argumentasjonen i forbindelse med TIMSS Advanced 2008 og 2015, spesielt når det gjelder matematikktesten. Vi pekte på den enorme teknologiske utviklingen på dette området fra 1995 til 2008. Kalkulatorer som var i vanlig bruk i undervisningen i en del land i 2008, kunne knapt sammenliknes med de som var tilgjengelige i 1995. Rammeverket for TIMSS Advanced 2008 erkjente denne problematikken: «it is noted that there have been tremendous changes in calculator technology since 1995» (Garden et al., 2006). Reglene for kalulatorbruk ble ikke endret, men ble presisert, og innspill fra norsk side førte til at en rekke oppgaver fikk spesielle koder for å avdekke bruk av kalkulator. I den norske matematikkrapporten for TIMSS Advanced 2008 ble norske elevers kalkulatorbruk drøftet (Grønmo, 2010; Onstad, 2010a; Pedersen, 2010). Et typisk trekk var at norske elever var gode til å bruke kalkulatoren i oppgavetyper som de umiddelbart gjenkjente. Derimot var evnen til kreativ kalkulatorbruk påfallende lav, selv i relativt enkle oppgaver. Dette ble ytterligere beskrevet og analysert i en masteroppgave (Sandstad, 2012).

Etter TIMSS Advanced 2008 har den teknologiske utviklingen fortsatt. I Norge er det mange elever som ikke lenger har kalkulator, men bruker

programvare på en bærbar datamaskin (f.eks. GeoGebra). Samtidig er vilkårene for bruk av digitale hjelpemidler til eksamen endret. Nå er matematikk-eksamener todelte; i del 1 er ingen hjelpemidler tillatt, mens det i del 2 forutsettes at elevene viser ferdigheter i bruk av relevant programvare. I utviklingen av matematikktesten til TIMSS Advanced 2015 ble det forsøkt å lage mange «kalkulatornøytrale» oppgaver, det vil si oppgaver der digitale hjelpemidler ville være til liten nytte.

Det planlegges nå en overgang til tester i TIMSS og TIMSS Advanced på digitale plattformer. I TIMSS for 4. og 8. trinn vil dette bli innført allerede i 2019. Bruken av eventuelle hjelpemidler i oppgaveløsingen vil da kunne styres på en helt annen måte enn hittil.

14.2.2 TIMSS Advanced og deltakerlandenes læreplaner

Et av målene med rammeverket for TIMSS Advanced er – som nevnt i det foregående – å sikre at elevene i ethvert deltakerland blir testet i oppgaver som i hovedsak faller innenfor landets læreplan. På grunn av de mange forskjellene mellom landene vil det alltid være en del oppgaver som ikke passer i enkelte land, men litt upresist kan man formulere det som et mål at testen skal være omtrent like «rettferdig» eller «urettferdig» i alle land.

Det matematiske innholdet i TIMSS Advanced er sammenliknet med de enkelte lands læreplaner på tre måter. For det første er innholdsbeskrivelsen i rammeverket holdt opp mot læreplanen (den intenderte). Som vi har beskrevet i delkapittel 14.2.1, er hver innholdskategori definert ved beskrivelse av en del faglige temaområder. Hvert temaområde er sammenliknet med læreplanen. Men siden det bare er 8 temaområder totalt i matematikk, vil hvert temaområde omfatte flere enkelttemaer, og det kan derfor være vanskelig å avgjøre om et område i hovedsak faller inn under landets læreplan eller ikke.

For det andre har lærerne blitt spurt om hvilke temaer de har undervist sine klasser i. Det gir oss informasjon om rammeverkets forhold til den implementerte læreplanen. Tabell 14.6 viser hvor stor andel av elevene som var blitt undervist i temaene i rammeverket før de tok testen — i snitt for hele testen, og for hver innholdskategori.

Tabell 14.6 Prosent av elevene som ifølge lærerne har blitt undervist i temaene i rammeverket for matematikk i TIMSS Advanced 2015 (gjennomsnittsverdier for samtlige temaer og for temaene i hver innholdskategori)

Land*	Alle temaer (19)	Algebra (8 temaer)	Kalkulus (7 temaer)	Geometri (4 temaer)
Frankrike	93	98	85	97
Italia	91	89	94	92
Libanon	98	98	98	99
Norge	92	84	98	99
Portugal	91	93	85	100
Slovenia	95	99	88	100
Sverige	94	90	96	97
USA	96	98	96	91

* Data for Russland var ikke tilgjengelige

Den laveste undervisningsdekningen har Norge med 84 % i algebra, tett fulgt av Frankrike og Portugal med 85 % i kalkulus. Det lave tallet til Norge skyldes i særlig grad temaet komplekse tall, og til en viss grad temaet følger og rekker. Samlet for hele rammeverket har samtlige land over 90 % undervisningsdekning.

Merk at tabell 14.6 antyder hvor godt rammeverket til TIMSS Advanced passer til et lands læreplan. Den viser derimot ikke det omvendte, nemlig hvor godt landets læreplan passer til rammeverket. Det vil si at dersom et faglig tema i rammeverket ikke er med i et lands læreplan, fanges det opp i tabellen. Men hvis det er temaer i landets læreplan som ikke er med i rammeverket, vises det ikke. Et eksempel på dette fra den norske læreplanen er emneområdet sannsynlighet.

I tillegg til disse sammenlikningene er hver enkelt testoppgave i TIMSS Advanced 2015 vurdert opp mot læreplanen i det enkelte land. Slik er det registrert hvilke av oppgavene i testen som er dekket av læreplanen og hvilke som må sies å ligge utenfor.

Tabell 14.7 viser sammenhengen mellom oppgavene og læreplanene, og hvilke utslag dette har gitt for prestasjonene.

Vi ser at testen passet ganske bra i de fleste landene, i den forstand at de aller fleste oppgavene ligger innenfor de respektive landenes læreplaner. Det eneste landet som skiller seg ut, er Russland med bare 91 av 120 testpoeng innenfor sin læreplan. Likevel gjorde de russiske elevene det godt på testen.

Tabell 14.7 Samsvar mellom matematikkoppgavene i TIMSS Advanced 2015 og landenes læreplaner. Prosent riktig på hele testen og på den delen av testen som faller innenfor det enkelte lands læreplan

Land	Antall poeng* innenfor læreplanen (av 120 poeng totalt)	Gjennomsnittlig prosent riktig på hele testen	Gjennomsnittlig prosent riktig på «egen del» av testen
Libanon	112	50	51
Russland	91	43	44
USA	119	43	43
Portugal	111	40	42
Norge	118	37	37
Slovenia	119	37	37
Frankrike	109	36	36
Sverige	111	33	32
Italia	117	31	32

** De aller fleste oppgavene i testen har ett oppnåelig poeng, mens noen få oppgaver har to poeng. Derfor er totalt antall oppnåelige poeng litt større enn antall oppgaver.*

Testen passet aller best i USA og Slovenia, der bare 1 testpoeng falt utenfor læreplanen. Tett etter følger Norge med 2 testpoeng utenfor vår læreplan; disse oppgavene handlet om komplekse tall.

Den midterste tallkolonnen viser hvor mange prosent av de 120 poengene elevene i hvert land skåret i gjennomsnitt. Best denne gangen var Libanon, der elevene i gjennomsnitt greide halvparten av oppgavene. Sist kom Italia og Sverige med bortimot en tredel av oppgavene korrekt. Det kan virke lite med 50 % korrekt på topp. Da må vi huske at dette er gjennomsnittet for alle elevene; de beste har selvsagt skåret langt høyere. I 2008 hadde Libanon 53 % korrekt (Onstad, 2010b, s. 251). Da var Russland best med 57 %. Men den gangen deltok Russland bare med en liten, svært elitepreget elevgruppe (dekningsgrad 1,4 %).

Dersom man likevel føler at dette må ha vært en vanskelig test, er det viktig å være klar over at det kompenseres for vanskelighetsgrad når testresultatene innpasses på den internasjonale trendskalaen med midtpunkt 500. (Mer om dette i delkapittel 14.3.6 om skalering.)

Viktigst er det kanskje å sammenlikne de to siste kolonnene. Mens den første viser hvor godt elevene i et land gjorde det på hele matematikktesten

i TIMSS Advanced 2015 (hvor mange prosent av de 120 poengene de greide), viser den siste kolonnen hvor mange prosent av poengene de greide på den delen av testen som lå innenfor dette landets læreplan. Det er slående hvor stor overensstemmelse det er mellom de to kolonnene. Størst forskjell har Portugal, som går opp 2 prosentpoeng fra hele testen til sin «egen del» av testen. Fire av landene, deriblant Norge, har ingen endring. Dermed blir det vanskelig å bruke argumenter om at testen er mer «urettferdig» for noen av deltakerlandene enn for andre.

14.2.3 Oppgaver

Når TIMSS utvikler oppgaver til undersøkelsene sine, tar de mange hensyn (Mullis et al., 2005):

- Oppgavene skal ligge innenfor læreplanen i de fleste deltakerlandene.
- Oppgavene skal kunne forsvare sin posisjon i en framtidig utvikling av matematikk og naturfag (fysikk i TIMSS Advanced) i skolen.
- Oppgavene skal være godt tilpasset de deltakende elevenes alderstrinn.
- Oppgavene skal fungere teknisk godt i en storskalaundersøkelse.
- Oppgavene skal fordele seg på innholdskategoriene og de kognitive kategoriene i samsvar med prosentangivelsene i rammeverket. (Se delkapittel 14.2.1.)

Oppgavene skal også fungere relativt godt i alle land, basert på resultatene fra piloteringen som gjennomføres året før hovedundersøkelsen. Videre er det et mål å få en balansert fordeling mellom flervalgsoppgaver og åpne oppgaver.

Punktet om å «fungere teknisk godt» betyr blant annet at en oppgave skal *diskriminere* godt, det vil si at den skal skille mellom sterke og svake elever. For å få høy reliabilitet på testen som helhet er det i tillegg viktig å ha oppgaver med ulik vanskegrad.

TIMSS Advanced er en *trendstudie*. Det betyr at den legger til rette for sammenlikning over tid. Et utvalg av oppgavene i TIMSS Advanced 1995 ble ikke offentliggjort, men lagt til side for gjenbruk i den neste TIMSS Advanced-studien i 2008. Dette er *trendoppgavene*, som knytter de to studiene sammen og gjør det mulig å sammenlikne prestasjonene. Tilsvarende skjedde i neste runde. Omtrent halvparten av oppgavene i 2008 ble hemmeligholdt og brukt som trendoppgaver i 2015.

Trendoppgavene fra TIMSS Advanced 2008 lå altså fastlagt som et utgangspunkt. Deretter var det behov for å utvikle mange nye oppgaver, slik at det samlede oppgavetilfanget fylte kriteriene ovenfor. Deltakerlandene ble invitert til å levere forslag til nye oppgaver. Oppgaveforslagene ble sendt til en internasjonal ekspertkomité hvor de ble vurdert mot rammeverket. Lå en oppgave utenfor rammeverket, ble den enten modifisert eller forkastet. Falt den innenfor, ble den plassert i en innholdskategori og en kognitiv kategori. Den internasjonale ekspertkomiteen hadde ansvaret for at det var tilstrekkelig med oppgaver innen de ulike faglige og kognitive områdene, at det var en akseptabel fordeling i oppgavenes vanskegrad, og at det var et passende forhold mellom flervalgsoppgaver og åpne oppgaver. Den hadde også ansvaret for beskrivelsene av de ulike *kompetansenivåene*. To norske forskere deltok aktivt i dette arbeidet med matematikkoppgavene.

Den store «oppgavebanken» som ble utviklet på denne måten, ble grundig gjennomgått. Fra denne valgte man ut omtrent dobbelt så mange oppgaver som man trengte til testen. Disse oppgavene ble utprøvd internasjonalt våren 2014. Resultatene i denne pilottesten ga grunnlag for å gjøre det endelige utvalget av oppgaver til selve TIMSS Advanced-undersøkelsen i 2015. Oppgaveutvalget ble diskutert internasjonalt med representanter fra alle deltakerlandene.

De utvalgte oppgavene er fordelt i såkalte *blokker*. En blokk består enten av trendoppgaver fra forrige runde eller av nye oppgaver som er prøvd ut i pilottesten. Blokkene er relativt like i arbeidsmengde og vanskegrad. Hver blokk inneholder omtrent 10 oppgaver og er anslått til å kreve 30 minutters arbeid for elevene.

I 2008 var det 7 blokker med matematikkoppgaver og 7 blokker med fysikkoppgaver. Av disse var 3 stykker i hvert fag trendblokker fra 1995. Antallet ble økt til 9 blokker i hvert fag i 2015. Begrunnelsen var at med flere oppgaver fikk man dekket rammeverket bedre. I hvert fag var det slik at 3 av blokkene inneholdt trendoppgaver fra 2008, mens de andre 6 blokkene inneholdt nye oppgaver. Fem av blokkene i hvert fag i 2015 blir nå hemmeligholdt slik at de kan brukes som trendblokker i neste runde av TIMSS Advanced.

14.2.4 Koder

Omtrent halvparten av oppgavene i TIMSS Advanced er *flervalgsoppgaver*. I slike oppgaver får elevene fire svaralternativer å velge mellom: A, B, C eller D. (I 1995 var det fem svaralternativer.) Eleven skal markere hvilket av svarene hun eller han mener eller tror er det riktige.

Det ligger et grundig arbeid bak konstruksjon av flervalgsoppgaver. Det er viktig at ett av svaralternativene er riktig, og at ingen av de andre er det. De gale alternativene kalles *distraktorer*. Gode distraktorer bør avspeile typiske misoppfatninger, regnefeil eller liknende. En distraktor som knapt velges av noen av elevene, er ikke ønskelig. Det er heller ikke ønskelig at en distraktor skal «lokke» eller «lure» elevene til å gi galt svar. For å finne gode distraktorer prøver man ofte ut oppgavene som åpne oppgaver først. De elevsvarene man da får, danner utgangspunkt for konstruksjon av distraktorer.

En flervalgsoppgave er enkel å kode etterpå. Det er én tallkode for hvert av svarene A, B, C og D (og eventuelt E). Det er også spesielle koder for elever som har svart på en gal måte – for eksempel markert to svar – eller ikke har svart i det hele tatt. Disse kodene registreres i en database som deretter kan behandles med statistisk programvare.

For de *åpne oppgavene* er kodingen mye mer krevende. Åpne oppgaver kalles *constructed response items* på engelsk. Det er altså oppgaver hvor eleven ikke skal velge mellom ferdigformulerte svarforslag, men må formulere svaret selv. Svaret som kreves, kan være av ulikt format. Det kan for eksempel dreie seg om å skrive ned bare et tall eller et ord, eller oppgaven kan kreve at eleven viser utregningen, redegjør for framgangsmåten, gir en begrunnelse eller forklarer et resonnement.

Gjennom utprøving av oppgavene danner man seg et inntrykk av hvordan de blir besvart. Dersom det viser seg at det er noen karakteristiske forskjeller mellom svarene, kan det ha diagnostisk interesse å gi visse svarkategorier særskilte koder. Det kan gi mulighet til å analysere både *hva* og *hvordan* elevene har svart på oppgaven. I TIMSS har man utviklet et tosifret kodesystem for å ta vare på slik informasjon. Norske forskere sto sentralt i denne utviklingen (se Lie, Angell & Rohatgi, 2010, s. 42).

Hvis en oppgave bare har ett riktig svar, gis det kode 10 for dette svaret. Dersom det er flere svar som anses som korrekte, eller dersom det er ulike måter å komme fram til svaret på, er det mulig å kode med for eksempel 10, 11 og 12. Hver kode er definert gjennom en beskrivelse (og eventuelt eksemplifisering) av hvilke typer elevsvar som skal falle inn under denne koden. Feilsvar kodes konsekvent på 70-tallet. Dersom det er interessant å skille mellom ulike feilsvar, kan de gis kode 70, 71 osv. Alle andre feilsvar får kode 79. Helt blanke oppgaver får kode 99.

Riktig svar på en slik oppgave (kode 10, 11, . . .) gir *ett poeng*.

Noen oppgaver er mer komplekse, og det er naturlig å kunne skille mellom helt eller delvis riktig svar. Da vil kodene 20, 21, 22 osv. betegne ulike typer korrekte svar, mens 10, 11, 12 osv. betegner ulike typer delvis korrekte svar.

Helt riktig svar på en slik oppgave gir *to poeng*, mens delvis riktig svar gir *ett poeng*.

Poengene gir grunnlag for å beregne prestasjonene, mens kodene for øvrig muliggjør nærmere studier av elevenes kunnskaper og strategivalg.

14.2.5 Spørreskjemaer

Hver elev som deltok i TIMSS Advanced, svarte på et *elevspørreskjema* i tillegg til den faglige testen. Lærerne til disse elevene (i det faget de ble testet i) fikk dessuten et eget *lærerspørreskjema*, og skolens ledelse fikk et *skolespørreskjema*. Gjennom skjemaene ble det samlet inn en rekke opplysninger om holdninger, hjemmebakgrunn, undervisningsmetoder, skolens ressurser med mer.

Spørreskjemaene i TIMSS Advanced 2015 gikk også gjennom en ekspertvurdering og en grundig internasjonal debatt før de ble ferdigstilt. Alle deltakerlandene hadde en demokratisk mulighet til å foreslå endringer og tillegg.

Det var mulig for land å sløyfe enkelte spørsmål som ble ansett som irrelevante for deres utdanningssystem, eller å legge til spørsmål som utdanningsmyndighetene eller den nasjonale prosjektgruppen fant interessante. Svarene på slike spørsmål blir ikke tatt med i den internasjonale rapporten.

14.2.6 Oversetting

Det internasjonale arbeids- og samarbeidsspråket i TIMSS er engelsk. Alle offisielle dokumenter, instruksjoner, oppgaver og spørreskjemaer foreligger på engelsk. Men når undersøkelsen gjennomføres, må oppgavene og spørreskjemaene foreligge på de språkene som brukes i skolene i de respektive landene. Elevene, lærerne og skolelederne skal møte oppgavene og spørsmålene på et språk de er vant til, ellers vil internasjonale sammenlikninger gi liten mening.

Oversetting er imidlertid vanskelig. For det første må det sikres at det spørres om nøyaktig det samme på alle språk. Videre bør oppgavene være like vanskelige, noe som ikke er opplagt når de reformuleres på et nytt språk. Noen enkle eksempler kan illustrere dette. Spørsmålet «What does a carnivore eat?» oversettes naturlig til «Hva spiser en kjøtteter?» på norsk. Vi ser at mens engelsk

bruker et vanskelig fremmedord, bruker norsk et selvforklarende ord. Det norske spørsmålet blir dermed lettere enn det engelske. Andre ganger ligger engelsk fagterminologi nærmere allmennspråket enn tilsvarende norske faguttrykk gjør. «Multiply» inngår i engelsk hverdagsspråk og kan bety «mangfoldiggjøre» eller «formere seg», mens «multiplisere» knapt brukes utenfor matematikken på norsk. En regulær trekant vil på enkelte språk kalles «likesidet» og på andre språk «likevinklet». Å være likesidet eller likevinklet er ekvivalent for trekanter (men ikke for firkanter!). Det er like fullt to ulike opplysninger om trekanten som formidles gjennom disse betegnelsene, og i noen oppgaver kan det spille en rolle for løsningen.

I tillegg vil skifte av språk mange ganger gå sammen med skifte av kultur, tradisjoner, miljø og erfaringsverden. Slike ting kan også spille en rolle for hvordan situasjoner og spørsmål oppfattes. Sammen med oversettingen bør man derfor være oppmerksom på om dette kan skape ulikheter mellom elevene i forskjellige land. En matematisk modell for isbjørnpopulasjoner kan virke mer fremmedartet i Afrika enn i Norge.

TIMSS har omfattende rutiner for oversetting og språkkontroll. I Norge utarbeides alle instrumentene (testene og spørreskjemaene) for grunnskolen på både bokmål og nynorsk. Det brukes moderate språkformer, slik at instrumentene blir nokså like i de to målformene. I TIMSS Advanced 2015 brukte vi en annen tilnærming. Der lot vi noen av oppgaveblokkene være på bokmål og de andre på nynorsk. Det betydde at hver elev fikk oppgaver på begge målformer. Vi hadde myndighetenes støtte for dette, og det var nesten ingen elever som reagerte. Vi gjorde det samme med spørreskjemaene; noen av dem ble formulert på bokmål, andre på nynorsk. Oversettelsesforslagene våre ble sendt til IEA, som sendte dem videre til en norsk språkekspert som var ukjent for prosjektgruppen i Norge. Kommentarene og forslagene fra eksperten ble sendt via IEA tilbake til Norge, der prosjektgruppen gjennomgikk dem, vurderte dem fra en faglig og språklig synsvinkel og foretok nødvendige forbedringer av tekstene.

Det er også viktig at layout på oppgaver og hefter er så lik som mulig i alle land. Alle heftene sendes derfor til internasjonal godkjenning av layout før de trykkes.

14.2.7 Hefter

Oppgavene ble, som nevnt ovenfor, fordelt i 9 blokker i hvert fag. Blokkene hadde omtrent like mange oppgaver og like stor vanskegrad og arbeidsmengde. Blokkene ble kalt M1, M2, . . ., M9. Blokkene M1, M3 og M5 var trendblokker fra 2008. De øvrige seks blokkene inneholdt nye oppgaver som var utarbeidet til studien i 2015.

Den totale arbeidsmengden for alle blokkene ville bli altfor stor for en enkelt elev, anslagsvis $4\frac{1}{2}$ time (pluss nok en halvtime til spørreskjemaet). Det er behov for å bruke mange oppgaver for å gi en bred dekning av innholdskategoriene i rammeverket. Hver enkelte elev får imidlertid bare et utvalg av alle oppgavene som er med i testen. Blokkene er fordelt på seks forskjellige hefter. Hvert hefte inneholder tre blokker, som tilsvarer en estimert arbeidsmengde på halvannen time. Tabell 14.8 viser hvordan blokkene ble fordelt på heftene.

Tabell 14.8 Fordeling av blokker i hefter. Trendblokkene er rødmerket.

Hefter	Blokker		
Hefte 1	M1	M2	M4
Hefte 2	M4	M3	M6
Hefte 3	M6	M7	M5
Hefte 4	M3	M8	M7
Hefte 5	M8	M5	M9
Hefte 6	M2	M9	M1

Vi ser at hvert hefte inneholder én trendblokk og to nye blokker. Vi ser videre at hver blokk forekommer i to hefter, og på ulike plasser i de to heftene (først/midt/sist). Elevenes prestasjoner kan nemlig påvirkes av om en oppgave ligger tidlig eller sent i heftet. Mot slutten av en test er elevene ofte mer slitne og mindre konsentrerte.

Hver elev fikk altså ett hefte. Den enkelte elev fikk dermed prøve seg på en tredel av oppgavene i studien. TIMSS Advanced er derfor lite egnet til å si noe om den enkelte elev; studien er designet for å kunne trekke relativt sikre konklusjoner om hele den nasjonale populasjonen eller store deler av denne.

Alle oppgaveheftene i TIMSS inneholdt en kortfattet instruksjon til elevene om hvordan de ulike oppgavetypene – det vil si flervalgsoppgaver og åpne oppgaver – skulle besvares. Det var en kort formelsamling i begynnelsen av hvert hefte. Denne er gjengitt i et appendiks bakerst i boka.

14.3 Gjennomføring

TIMSS har utviklet grundige prosedyrer for å sikre en ensartet gjennomføring av undersøkelsen i alle deltakerlandene. Prosedyrene er nøye beskrevet i manualer for gjennomføringen av ulike deler av studien. En teknisk rapport er publisert av det internasjonale prosjektsenteret (Martin, Mullis & Hooper, 2016).

14.3.1 Tidspunkt

TIMSS Advanced-undersøkelsen skulle gjennomføres i slutten av det siste året i videregående skole. Det betydde våren 2015 innenfor tidsrammer som var fastsatt sentralt.

14.3.2 Utvalg

Bare et utvalg av elevene i hvert deltakerland blir testet. Dette utvalget trekkes ut etter bestemte statistiske regler og prosedyrer. For å kunne gjøre generaliseringer fra utvalget til hele populasjonen med liten usikkerhet (små feilmarginer), ble det satt som mål at utvalgene burde omfatte 3600 elever i hvert fag. Dette målet gjaldt i utgangspunktet alle land. At kravet til utvalgsstørrelsen er uavhengig av størrelsen på populasjonen, kan begrunnes statistisk, men vi går ikke inn på det her. For små land kunne ikke disse målene nås, og prosedyrer og mål måtte modifiseres. Av de 264 aktuelle videregående skolene i Norge ble 134 trukket ut til å delta i matematikk, og de andre 130 til å delta i fysikk. Den norske prosjektgruppen fant det ikke ønskelig at skoler skulle bes om å delta i begge studiene. Det ville lett føre til at samme elev måtte delta i begge studiene, siden svært mange av fysikkelevene også tar matematikk. Det ville være en urimelig belastning relativt kort tid før avsluttende eksamen. På skoler som ble trukket ut i matematikk, var alle elevene i Matematikk R2 med i utvalget, og på skoler som ble trukket ut i fysikk, var alle elevene i Fysikk 2 med i utvalget.

Den nasjonale prosjektgruppen kontaktet alle de uttrukne skolene med en oppfordring om å delta i undersøkelsen. Av de 134 skolene som ble bedt om å delta i matematikk, svarte 133 ja. Av de aktuelle elevene på disse skolene deltok 93 %. Av de 130 skolene som ble bedt om å delta i fysikk, var det 127 som svarte ja. Av de aktuelle elevene på disse skolene deltok 94 %. Det gir en samlet deltakelsesprosent på 93 % i både matematikk og fysikk. Til sammen deltok 2537 norske elever i matematikkundersøkelsen og 2472 i fysikkundersøkelsen.

TIMSS hadde detaljerte regler for hvordan disse utvalgene skulle trekkes. I tillegg var det strenge krav til deltakelsesprosentene for å anerkjenne utvalgene som *representative*. Norge tilfredsstilte disse kravene med god margin.

Dersom et utvalg er trukket *tilfeldig* og har en viss størrelse, regnes det som *representativt*, det vil si at det avspeiler situasjonen i hele populasjonen. I vårt tilfelle ville tilfeldig utvalg bety at enhver R2-elev i landet hadde samme sannsynlighet for å bli med i utvalget (og tilsvarende i fysikk). Dette var ikke tilfellet i TIMSS Advanced. Skolene hadde ikke samme sannsynlighet for å bli trukket ut, siden skoler som ikke hadde Fysikk 2 nødvendigvis måtte være med i matematikkutvalget. Dessuten var det ulikt antall elever fra skole til skole. Men i etterkant var det mulig å beregne hvor stor sannsynligheten for å bli trukket ut hadde vært for hver enkelt elev i utvalget. Disse sannsynlighetene ble brukt til å beregne hvor mange elever i populasjonen den enkelte elev i utvalget kunne sies å representere. Dermed kunne elevene tildeles *vekter* som tilsvarte denne representativiteten. På tilsvarende måte ble det beregnet vekter for skolene i utvalgene. Dataanalysene benytter disse vektene.

På denne måten fikk vi et representativt utvalg av skoler og et representativt utvalg av elever. Utvalget av lærere ble derimot *ikke* trukket tilfeldig. Lærerne fulgte med som et «attributt» til elevutvalget – det var de utvalgte klassenes lærere som deltok i undersøkelsen. Strengt tatt betyr det at lærerutvalget ikke med sikkerhet kan anses som representativt for hele lærerpopulasjonen; det er derfor litt mer usikkert å generalisere fra det. Men siden lærerutvalget omfatter så mange av de aktuelle lærerne – og det er et biprodukt av en tilfeldig utvalgsprosess – kan det vanskelig tenkes betydelige feilutslag om man antar at de på en god måte representerer samtlige lærere i dette faget. Vi kan anse lærerutvalget som «tilstrekkelig tilfeldig» til at vi kan generalisere fra det. Derfor tillater vi oss å bruke uttrykk av typen «23 % av de norske R2-lærerne» og liknende uttrykksmåter når vi strengt tatt burde ha skrevet «lærerne til 23 % av R2-elevene i Norge».

Vektingen av dataene ble beregnet av datasenteret til IEA. Dette blir beskrevet i den internasjonale tekniske rapporten til TIMSS Advanced 2015 (Martin et al., 2016).

Skolene som hadde sagt seg villige til å delta, sendte inn anonymiserte lister over de uttrukne elevene. Prosjektgruppen brukte et dataprogram spesiallaget for TIMSS Advanced til å trekke ut hvilken elev som skulle ha hvilket oppgavehefte.

14.3.3 Gjennomføring på skolene

Det internasjonale prosjektsenteret hadde utarbeidet detaljerte instrukser for hvordan testen skulle gjennomføres i klasserommet. Det var gjort for å sikre like testvilkår for alle elever, både nasjonalt og internasjonalt.

Alt elevmateriell ble sendt til skolene litt før undersøkelsen skulle gjennomføres. Materiellet besto av oppgavehefter og spørreskjemaer til elevene, samt instrukser for gjennomføringen. En av de tilsatte på skolen var ansvarlig for å sette seg inn i instruksene på forhånd og å påse at de ble fulgt nøye.

På den avtalte testdagen ble elevene samlet i klasserommet eller et annet egnet rom. Elevene fikk hvert sitt oppgavehefte. Hvem som skulle ha hvilket hefte, var angitt med en kodet klistrelapp foran på heftet. Dersom en elev ikke møtte, ble vedkommendes hefte inndratt. Dersom en frammøtt elev burde ha tilhørt utvalget, men ikke var registrert, ble vedkommende registrert og fikk et ekstrahefte som var klargjort for slik bruk. Elevene fikk ikke lov til å åpne heftene før de fikk beskjed om å gjøre det.

Elevene fikk opplest informasjon om testen og om gjennomføringen, og eksemplene forrest i heftene ble gjennomgått. Deretter fikk de nøyaktig 90 minutter til å løse oppgavene. Etterpå besvarte elevene spørreskjemaet.

Den internasjonale TIMSS-ledelsen hadde knyttet til seg én person i hvert land som kontrollerte gjennomføringen på en del tilfeldig valgte skoler. Vedkommende var uavhengig av den nasjonale prosjektgruppen og rapporterte direkte til den internasjonale ledelsen ved hjelp av et grundig rapporteringsskjema.

Den ansvarlige personen for gjennomføringen på den enkelte skole sendte alt materiellet tilbake til den nasjonale prosjektgruppen. Det ble kontrollert at ingen oppgavehefter forsvant i prosessen.

Spørreskjemaene til lærerne og skolelederne ble distribuert og utfylt på nett.

14.3.4 Koding

All informasjon fra oppgaveheftene og de ulike spørreskjemaene ble registrert i en databank. I prinsippet er det enkelt å kode svarene på flervalgsoppgavene og på spørsmålene i spørreskjemaene. Da skal det bare registreres hvilket svaralternativ vedkommende har valgt. På den annen side er det selvsagt mulig å gjøre tastefeil ved innskrivingen. I Norge ble dette lest og registrert elektronisk fra skannede versjoner av elevenes oppgavehefter.

Når det gjelder de åpne oppgavene er situasjonen mer krevende, noe som går fram av redegjørelsen for kodesystemet i delkapittel 14.2.4 ovenfor. Koden settes altså etter en subjektiv vurdering av elevens svar. Skal analyser av elevprestasjonene være pålitelige (reliable), må denne kodingen av åpne oppgaver utføres så likt som mulig av alle kodere i samtlige deltakerland. Det nedlegges et stort arbeid for å sikre dette best mulig. De tillatte kodene på en oppgave er utførlig beskrevet i de internasjonale kodemanualene Dette materiellet var grundig gjennomgått på en internasjonal samling. I det enkelte land ble kodedefinisjonene nøye gjennomgått i fellesskap før kodingen startet. Eventuelle uklarheter ble drøftet og avklart, i noen tilfeller i samråd med den internasjonale TIMSS-ledelsen. For mange av oppgavene var det utarbeidet et eksempelmateriell som illustrerte hvordan kodene skulle brukes. Dette ble gjennomgått og kommentert i fellesskap. I tillegg var det ofte øvingsoppgaver som alle koderne skulle vurdere hver for seg. Etterpå sammenliknet man de kodene man hadde valgt, drøftet vurderingene og holdt disse opp mot en internasjonal «fasit» som fastslo hvordan kodene skulle brukes på øvingsoppgavene. I noen land, blant annet Norge, var elevbesvarelsene skannet inn, og kodingen ble utført ved skjerm og tastatur.

Som en ytterligere kontroll ble det gjort tre typer ekstra koding:

- Omtrent en tredel av heftene var trukket ut til *reliabilitetskoding*, det vil si at to personer kodet disse heftene uavhengig av hverandre. På denne måten kunne man statistisk måle den nasjonale *sensorreliabiliteten*, det vil si graden av samsvar mellom koderne (sensorene) i et land.
- En del engelskspråklige elevbesvarelser var plukket ut til å bli kodet av to kodere fra hvert eneste deltakerland. På denne måten kunne man statistisk måle sensorreliabiliteten mellom land.
- En del besvarelser fra TIMSS Advanced 2008 på trendoppgaver som ble brukt på nytt i 2015, var plukket ut til å bli kodet av to kodere fra hvert land. På denne måten kunne man statistisk måle sensorreliabiliteten over tid.

14.3.5 Databehandling

De innlagte dataene ble kontrollert i flere omganger, først i Norge og deretter i det internasjonale datasenteret til TIMSS. Dataene ble «vasket», det vil si at man lette etter inkonsistente og overraskende data. Disse ble så kontrollert mot oppgaveheftene og spørreskjemaene. Prosedyrene skal sikre høy grad av samsvar mellom det elevene, lærerne og skolelederne faktisk hadde svart, og de dataene som ble lagret elektronisk.

Da datavaskingen var avsluttet, ble alle forbindelser mellom de elektroniske dataene og deltakerne i undersøkelsen slettet. Dermed lar det seg ikke gjøre å spore enkeltresultater tilbake til elever eller skoler. Prosedyrene var i Norge godkjent av Datatilsynet.

14.3.6 Skalering

Avanserte statistiske metoder er brukt for å behandle dataene på en måte som muliggjør sammenlikninger. Dette blir grundig beskrevet i den internasjonale tekniske rapporten (Martin et al., 2016).

Som nevnt ovenfor svarte hver enkelt elev bare på en tredel av det samlede oppgavetilfanget. Prestasjonene til to elever som hadde samme hefte, kan sammenliknes. To elever som fikk forskjellige hefter, fikk derimot helt eller delvis forskjellige oppgaver, og da kan ikke prestasjonene uten videre sammenliknes. Tilsvarende kan prestasjoner i 2015 ikke uten videre sammenliknes med prestasjoner i 2008.

Disse problemene løses ved hjelp av blokker som er felles mellom hefter og mellom de to undersøkelsene. Disse blokkene fungerer som «broer» som knytter de enkelte delene sammen.

La oss eksempelvis se på en elev, vi kaller henne Helga, som fikk hefte 2 i matematikktesten. Hefte 2 inneholdt blokkene M4, M3 og M6 (se tabell 14.8 på side 290). Blokk M4 fantes også i hefte 1. Med kunnskap om hvordan Helga presterte på blokk M4, og ut fra det statistiske materialet om hvordan elevene som fikk hefte 1, presterte, kan vi anslå hvordan Helga ville ha gjort det på blokkene M1 og M2 dersom hun i stedet hadde fått hefte 1. På samme måte kan vi ved hjelp av blokk M6 anslå hvordan hun ville ha gjort det i hefte 3, og ved hjelp av blokk M3 anslå hvordan hun ville ha gjort det i hefte 4. Og med denne kunnskapen og disse anslagene kan vi videre anslå hvordan hun ville ha gjort det i heftene 5 og 6.

Et slikt resonnement er ganske usikkert for én enkelt elev. Men tilknytningen til de virkelige elevene er kuttet – det finnes ingen «Helga». Dataene kan ikke brukes til å si noe om enkeltelever. De blir bare anonyme representanter som kan hjelpe oss til å si noe om den nasjonale populasjonen eller deler av denne.

På grunn av usikkerheten blir det kjørt flere simuleringer ut fra prestasjonene til Helga i hefte 2 og anslagene for hvordan hun kunne ha gjort det på oppgavene i resten av blokkene. Disse simuleringene produserer fem verdier som representerer hva Helga kunne ha skåret totalt dersom hun hadde tatt hele testen. Disse verdiene kalles *plausible verdier*. Variasjonen mellom de plausible verdiene er et uttrykk for usikkerheten i anslagene. Det regnes ut fem plausible verdier for hver eneste elev som deltok i testen. De fleste statistiske analysene som tar for seg elevenes skår på testen (for eksempel i forhold til en bakgrunnsvariabel) benytter alle de plausible variablene. Det minsker usikkerheten i resultatene.

Når samtlige elever på denne måten har fått plausible verdier for sine prestasjoner, kan man regne ut gjennomsnittsskår og standardavvik for utvalget og bruke det til å generalisere til hele populasjonen eller til deler av denne. For alle slike generaliserte verdier er det beregnet *standardfeil*, som brukes til å avgjøre om forskjeller er *signifikante*.

Alle enkeltskårene ligger spredt omkring gjennomsnittet på en skåringsakse. Da er det mulig å justere selve måleaksen. På samme vis som vi kan regne om temperaturer mellom celsius-verdier og fahrenheit-verdier, kan vi regne om skårene til nye verdier langs en ny skala. Vi får andre tall og et annet nullpunkt, men det er fortsatt den samme statistiske fordelingen.

En slik *skalering* ble gjort med dataene i TIMSS Advanced 1995. Elevskårene i alle deltakerlandene ble regnet om til en ny skala slik at det internasjonale gjennomsnittet ble 500 «poeng» og standardavviket ble 100 «poeng». Disse tallene er ikke poeng oppnådd på selve testen, men de er likevel mål for hvor godt elevene presterte. En slik skalering ble utført for matematikk og fysikk hver for seg.

Elevene som ble testet i TIMSS Advanced 2008 hadde tre trendblokker fra 1995. Med liknende teknikker som vi nettopp har antydet kunne elevens prestasjoner på trendoppgavene i 2008 brukes til å anslå hvordan disse elevene ville ha prestert dersom de hadde tatt hele testen fra 1995. Resultatene deres kunne derfor innpasses på den skalaen som ble fastlagt i 1995.

Teknikkene som brukes for slik «brobygging» mellom undersøkelser baserer

seg på *Item Response Theory* og er statistisk avanserte – atskillig mer avanserte enn beskrivelsen ovenfor kan gi inntrykk av. De blir beskrevet i den internasjonale tekniske rapporten til TIMSS Advanced (Martin et al., 2016). «Brobyggingen» ble foretatt for alle de landene som deltok i både 1995 og 2008. Slik ble altså skalaen i 2008 definert i samsvar med skalaen fra 1995. De nye deltakerlandene i 2008 ble innpasset på denne skalaen. En tilsvarende «brobygging» ble foretatt med resultatene i 2015 ved hjelp av trendoppgavene fra 2008.

Denne prosessen ga en skala (for hvert av fagene) som kan brukes som fast målestokk for prestasjoner i den første undersøkelsen i 1995, for TIMSS Advanced 2008, for TIMSS Advanced 2015 og for eventuelle nye TIMSS Advanced-studier. Dette muliggjør trendanalyser.

Den internasjonale gjennomsnittsskåren var 500 per definisjon i 1995. I 2008 var den ikke lenger 500. Det kunne heller ikke forventes. For det første må vi forvente at de landene som hadde deltatt i 1995, ikke presterte akkurat likt i 2008. Viktigere er det likevel at det ikke var samme gruppe land som deltok i begge undersøkelsene. Noen land som deltok i 1995, uteble i 2008, og nye land kom til, se tabell 14.1 på side 273. På samme måte var det en viss utskifting av deltakerland fra 2008 til 2015. Det er ingen grunn til å forvente at én gruppe land skal prestere nøyaktig like godt i gjennomsnitt som en (delvis) annen gruppe land.

Å relatere prestasjoner til det internasjonale gjennomsnittet på den enkelte studien kan gi liten mening, siden et slikt gjennomsnitt naturlig varierer fra studie til studie. Kommer det for eksempel inn et fattig land som presterer svakt – som Filippinene i 2008 – vil det kunne trekke gjennomsnittet ned i forhold til den foregående studien. Det ville være sterkt misvisende om en bedring i norske prestasjoner i forhold til det internasjonale gjennomsnittet på én studie ble framstilt som en framgang i forhold til en tidligere studie, mens det i virkeligheten skyldtes at gjennomsnittet hadde endret seg fordi nye land med svakere prestasjoner deltok. Dersom vi tenker oss at Singapore hadde deltatt i TIMSS Advanced 2008 i stedet for Filippinene, hadde utvilsomt totalbildet vært ganske annerledes. Men vurderingen av den norske utviklingen skal ikke avhenge av hvilke andre land som valgte å delta.

De prestasjonsdataene som foreligger, gir god anledning til å studere et enkelt lands utvikling over tid. Da sammenliknes landet med seg selv på den faste skalaen fra undersøkelse til undersøkelse. Sammenlikninger mellom land i samme undersøkelse er også meningsfulle. Dersom to eller flere land har

deltatt i flere av undersøkelsene, kan landenes utvikling over tid også sammenliknes. Det som derimot gir liten mening, er å sammenlikne prestasjoner for et land med de internasjonale gjennomsnittene fra undersøkelse til undersøkelse, siden disse altså varierer og er avhengige av hvilke land som deltar. I de internasjonale rapportene for TIMSS og TIMSS Advanced unnlater prosjektsenteret i Boston å gjøre dette. I tabellene over deltakerlandenes gjennomsnittsskår er skalamidtpunktet på 500 oppgitt, men ikke årets internasjonale gjennomsnitt. Samme valg er gjort i denne boka.

14.3.7 Analyser og rapportering

Det internasjonale prosjektsenteret for TIMSS Advanced har ansvaret for en første grundig gjennomgang og analyse av dataene fra samtlige deltakerland. Det er de som beregner vekter for dataene i alle land, som beregner plausible verdier for alle elevene, og som foretar den internasjonale skaleringen av skårene. De utgir en teknisk rapport om gjennomføringen av studien og om hvordan dataene er behandlet (Martin et al., 2016). De utgir også en rapport om de internasjonale resultatene (Mullis, Martin, Foy et al., 2016b). Det enkelte land har ansvar for å kontrollere at landets data som brukes i disse analysene er korrekte.

Til hjelp i analysene er det utviklet en del *samlevariabler*. Eksempler på slike er *faglig selvtillit*, *indre motivasjon* og *ytre motivasjon*. En samlevariabel er en slags sammenfatning av flere variabler. Etablering av en samlevariabel er en omfattende prosess som baserer seg både på faglig innsikt og på statistiske metoder. Med bakgrunn i erfaring og tidligere forskning vil man ofte anta at flere variabler måler aspekter av samme fenomen, det vil si at man antar at de sammen danner et naturlig og interessant *konstrukt*. Denne antakelsen blir testet med *korrelasjonsundersøkelser*, *regresjonsanalyser* og *eksplorerende faktoranalyse*, og i etterkant med *konfirmerende faktoranalyse*. På denne måten søker man å etablere et solid faglig og statistisk grunnlag for bruken av samlevariablene.

For den som er interessert i statistiske resonnementer og metoder som brukes i slike store studier, finnes det mye teori man kan sette seg inn i. Eksempler er bøkene *Introduction to classical and modern test theory* (Crocker & Algina, 1986), *Statistics for social data analysis* (Knoke, Bohrnstedt & Mee, 2002) og *Structural Equations with Latent Variables* (Bollen, 1989).

Denne boka er en oppfølger til den norske rapporten fra TIMSS Advanced 2015 (Grønmo, Hole & Onstad, 2016). Nye analyser av data fra TIMSS Advanced, TIMSS og PISA presenteres og drøftes i et fagdidaktisk og utdanningspolitisk perspektiv.