Julius Kristjan Björnsson
Universitetet i Oslo

DOI: http://dx.doi.org/10.5617/adno.6273

Om lenkefeil og ekvivaleringsmetoder på nasjonale prøver: Evaluering av endring over tid[1]

Sammendrag
Nasjonale prøver i nåværende form, hvor Item Response Theory (IRT) benyttes for å bestemme oppgavenes egenskaper og hvor man måler utvikling over tid, har vært gjennomført siden 2014. Prøvene har vist seg å være stabile over tid, og en lenking og ekvivalering er blitt gjort siden 2014 for å gjøre sammenlikninger over tid mulige. For å kunne avgjøre om endringer over tid er signifikante, er det nødvendig å kvantifisere den usikkerheten som er knyttet til prosedyren for lenking fra år til år. Denne usikkerheten betegnes som lenkefeilen. Denne artikkelen gjør rede for ulike måter å gjøre dette på, og med bakgrunn i dette beregnes størrelsen av den lenkefeilen som er til stede i regning og engelsk for 5. og 8. trinn. I tillegg presenteres resultater fra en undersøkelse av mulig bias i lenkingen. Konklusjonen er at lenkefeilen er akseptabel, men likevel såpass stor at evaluering av endring over tid må ta hensyn til den. Det blir derfor viktig å ha et prøvedesign og bruke metoder som gir riktige (unbiased) estimater og som bidrar til å minimere lenkefeilen.

Nøkkelord: IRT, nasjonale prøver, ekvivalering, lenkefeil

Linking error and equating methods on the national tests: Estimating change over time

Abstract
The Norwegian national tests, utilizing Item Response Theory (IRT) to determine item characteristics and measure changes over time, have been administered since 2014. The tests have turned out to be stable over time, and linking and equating has been done each year to make comparisons over time possible. Central for these methods is to quantify the uncertainty in the linking from year to year, as this must be known to determine whether a change from year to year is significant or not. This article presents some often-used methods to estimate the linking error. Based on this, the size of the error due to linking is estimated

[1] Denne studien er gjennomført med tillatelse fra Utdanningsdirektoratet.

for English and Numeracy for the 5th and 8th grades. The article also presents an examination of possible bias in the linking. The main conclusion is that the linking error is acceptable, but nevertheless so large that a determination of changes over time must take it into account. It remains important to make use of a test design and methods that result in an appropriately small and unbiased estimate of the linking error.

Keywords: IRT, national tests, equating, linking error

Introduksjon

Nasjonale prøver skal måle endring av prestasjoner over tid for grunnskolen i Norge, i tillegg til å gi lærere og elever presise og nyttige opplysninger om elevenes ferdighetsnivå. For å muliggjøre måling av endring over tid brukes et NEAT (Non Equivalent groups with Anchor Test) ankerdesign, hvor anker-oppgaver gjenbrukes og gjennomføres i et tilfeldig og representativt utvalg elever hvert år. Ved å bruke Item Response Theory (IRT) i skaleringen av elevenes skårer på prøven oppnår man at resultatene fra forskjellige år plasseres på den samme underliggende skalaen. Dette betyr at en bestemt verdi på skalaen representerer den samme ferdighet hvert år. De nasjonale prøvene bruker en 2-parameter (2PL) IRT-modell for dikotome oppgaver og en «graded response» modell for polytome oppgaver. De grunnleggende metodene, IRT-kalibreringen og bruk av ankersett som er brukt i prøvene, er beskrevet detaljert i Björnsson (2016).

Ekvivalering er nødvendig hvis resultater fra to eller flere forskjellige prøver skal sammenliknes. To prøver med forskjellige oppgaver kan aldri sammenliknes direkte, og derfor er det vanlig praksis å endre resultatene over til såkalte standardiserte eller skalerte skårer. Nasjonale prøver bruker såkalte *skalapoeng* som er en standardisert skår med gjennomsnitt 50 og standardavvik 10. Men dette er ikke nok for å sikre at samme tall beskriver samme ferdigheter. I tillegg må man ha et design som gjør det mulig å studere sammenliknbarheten mellom skårene på to eller flere prøver. Dette kalles et lenkedesign, og har som mål å plassere skårer fra flere prøver på samme skala (Kolen & Brennan, 2014; Dorans, Moses & Eignor, 2010).

De metodene som presenteres i denne artikkelen, er alle basert på en IRT-kalibrering av data fra et prøvedesign med ankeroppgaver, men det finnes mange andre metoder for ekvivalering som ikke er basert på IRT eller et anker-oppgavedesign. Interesserte henvises til Kolen & Brennan (2014) og von Davier (2011).

Hensikten med denne artikkelen er å studere kvaliteten til den ekvivaleringsmetoden som nasjonale prøver anvender og å belyse hva en lenkefeil er når endringer over tid skal fanges opp. Hensikten er også å beskrive og forklare

noen av de grunnleggende metodene som har vært brukt i storskalaundersøkelser som PISA og TIMSS. Metoden som PISA brukte inntil 2015, er en av de metodene som blir omtalt. Lenkemetoden som de nasjonale prøvene bruker, blir sammenliknet med noen andre vanlige metoder, og lenkefeilen på prøvene estimeres ved å bruke de forskjellige metodene. I tillegg utforsker artikkelen om det finnes såkalt bias eller systematiske feil i ekvivaleringen.

Artikkelens forskningsspørsmål er derfor:

1. Hvordan fungerer den ekvivaleringsmetoden som nasjonale prøver bruker?
2. Hvilke ekvivaleringsmetoder fungerer best med nasjonale prøver?
3. Hvor stor er lenkefeilen fra år til år?
4. Gir dagens metoder for å beregne lenkingen systematiske feil som kan reduseres?

Hvorfor er det viktig å beregne lenkefeil?

I alle ankerdesign er det fare for at ankringen fra år til år ikke er feilfri. Dette skyldes først og fremst at ankeroppgavenes vanskegrad kan endres over tid. Et prøvedesign hvor ankeret ikke endrer seg i det hele tatt, finnes ikke, og derfor er det viktig å kunne kvantifisere usikkerheten i ekvivaleringen og lenkingen[2] fra år til år (van der Linden & Wiberg, 2010). Dette er spesielt viktig når gjennomsnittet for grupper eller hele populasjonen blir sammenliknet fra år til år. I tillegg til en viss usikkerhet (standardfeil) på gjennomsnittet for hvert år, er det også en usikkerhet knyttet til hvorvidt skalaen blir nøyaktig den samme hvert år. Dette kalles *lenkefeil* (linking error). Beregninger av slike lenkefeil gjennomføres blant annet rutinemessig i internasjonale storskalaundersøkelser som TIMSS og PISA, og metoder og resultater for dette er grundig diskutert i de tekniske rapportene til studiene og i separate artikler om lenkefeilen til selve studiene (Monseur & Berezner, 2007). Denne forskningen handler også om lenking mellom storskalaundersøkelsene og andre prøvesystemer (Hastedt & Desa, 2015) og understreker viktigheten av å estimere lenkefeilen.

TIMSS og PISA har historisk sett brukt litt forskjellige metoder, hvor PISA helt inntil 2015 brukte en såkalt «mean-alignment» metode, som er basert på en undersøkelse av oppgavevanskegraden på ankeroppgavene fra en IRT-analyse. TIMSS har derimot lenge benyttet en samkalibrering (concurrent calibration) av alle prøvens oppgaver (inkludert ankeroppgaver), en metode som nå også PISA har tatt i bruk fra og med PISA 2015 (OECD, 2017). Beregningene av

[2] Ekvivalering og lenking er begreper som ofte blir brukt om det samme. Helt presis ordbruk ville være å ikke snakke om ekvivalering når IRT-metodene er brukt, men bare bruke begrepet lenking. I realiteten er forholdet mellom disse to fenomenene slik at en lenking alltid er en ekvivalering, mens en ekvivalering ikke trenger å være en lenking. Andre metoder enn IRT kan også brukes for ekvivalering.

lenkefeilene for disse studiene viser at de er av en størrelsesorden som kan påvirke tolkningen av resultatene. Dersom man ikke hadde beregnet disse lenkefeilene (dvs. at man antar at feil knyttet til lenking ikke finnes), kunne man feilaktig ha konkludert at endringer over tid for noen land er signifikante. I den internasjonale leseundersøkelsen PIRLS (4./5. trinn) viser det seg at feilen er 1,1 poeng internasjonalt (på en skala med gjennomsnitt 500 og standardavvik 100), men noe varierende i forskjellige land (Martin, Mullis, Foy, Brossman & Stanco, 2012), og i TIMSS-studien er den i samme størrelsesorden, rundt 1 poeng. I disse to studiene er altså lenkefeilene veldig små. Lenkefeilene knyttet til endringen fra 2012 til 2015 i PISA er imidlertid betydelig større: 3,5 for matematikk, 5,2 for lesing og 3,9 for naturfag (OECD, 2017). Sannsynligvis kan disse forskjellene knyttes til at TIMSS og PIRLS har nesten tre ganger så mange ankeroppgaver som PISA innen hvert fagområde. Hensikten her er ikke å diskutere de internasjonale undersøkelsene i seg selv, men disse eksemplene viser godt det som er denne artikkelens anliggende: Lenkefeil kan være så store at det blir viktig å estimere dem, og lenkefeilen er i stor grad et resultat av prøvenes utforming, men også avhengig av lenkemetoden.

Hvorfor endres ankeret?

Det kan være mange årsaker til at vanskegraden til ankeroppgaver endres, men den største effekten kommer sannsynligvis av at ankeroppgavene blir satt sammen med forskjellige andre oppgaver hvert år. Naboliggende oppgaver har en gjensidig påvirkning som ofte er vanskelig å unngå. Her er både rekkefølgen av oppgavene og det faglige innholdet i prøven som helhet og i de naboliggende oppgavene viktig (Meyers, Miller & Way, 2008). Nasjonale prøver gjennomføres ved å administrere to litt ulike prøver. De aller fleste elevene får en prøve som er helt ny hvert år (den såkalte kohortprøven), mens et utvalg på noen tusen elever (omtrent 6 %) får en prøve hvor en del av oppgavene i kohortprøven er erstattet med de såkalte ankeroppgavene (20–25 oppgaver). Nasjonale prøver blir offentliggjort etter hver gjennomføring, og en ny kohortprøve må derfor konstrueres for å bli brukt sammen med de samme ankeroppgavene neste år. Dette kan ha en effekt som viser seg som endret vanskegrad på ankeroppgavene.

Vanskegraden til ankeroppgaver kan også endres ved at oppgavene «drifter» over tid. Med oppgavedrift menes det fenomenet at innholdet og utformingen av oppgavene kan bli delvis kjent, eller at det faglige innholdet av en eller annen grunn vektlegges svakere eller sterkere over tid, f.eks. ved at konteksten som inngår i oppgavene blir mindre relevant over tid. Man kunne eksempelvis for noen år siden ha hatt en regneoppgave knyttet til en tekst om salg av CD-plater. Dersom denne hadde vært beholdt over mange år, ville man med stor sannsynlighet fått en oppgave som elevene i mindre grad er kjent med og/eller som de opplever som uinteressant og lite engasjerende. Konsekvensen er at det blir

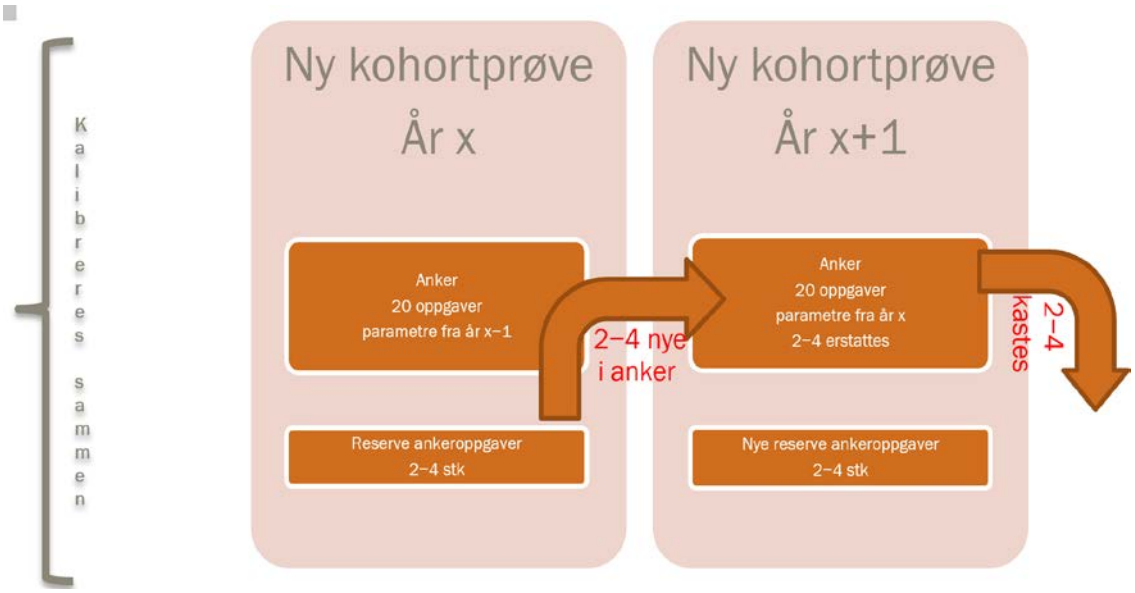
viktig å fornye og vedlikeholde ankeret på en systematisk måte (Antal, Proctor & Melican, 2014; Guo, Liu, Dorans & Feigenbaum, 2011).

Nasjonale prøver opererer derfor med et prinsipp hvor hele ankeret blir fornyet i løpet av omtrent fem år. Fornyingen skjer ved at omtrent 20 % av ankeret hvert år består av nye oppgaver som er blitt kalibrert sammen med anker og kohortprøve året før. Dette fører til at ankeret fornyes gradvis over tid, og gjør det også mulig å gjenspeile endringer i prøvenes innhold, og endringer i læreplaner og temaer som inngår i prøvene. På denne måten er det mulig å endre ankeret gradvis i takt med endringer i selve prøven og endringer i innhold og læreplaner, uten at trend over lengre tid går tapt.

Tidligere forskning har prøvd å svare på hvor stor en ankerprøve skal være (Cook & Eignor, 1989), og det finnes anbefalinger om at den skal være alt fra like stor som selve prøven som skal ankres, til at den kan være ned til 20 % av selve prøvestørrelsen. Noen snakker om at ankerprøven skal være en mini-prøve og måle samme konstrukt, men andre mener at det ikke er nødvendig. Ankerprøven i nasjonale prøver er ikke nødvendigvis tenkt som en «miniversjon» av kohortprøven, men både størrelsen av den (Cook & Eignor, 1989) og fordelingen i vanskegrad den inneholder (omtrent +/– et standardavvik rundt gjennomsnittet) er basert på tidligere forskning og erfaringer på området (Sinharay, Haberman, Holland & Lewis, 2012). I regning er ankerprøven rundt 38 % av størrelsen på selve prøven, i engelsk er ankerprøven rundt 45 %, mens i lesing er den like stor som selve prøven (Björnsson, 2016).

I praksis må de som utvikler prøvene, inkludere et så høyt antall ankeroppgaver som mulig. I designet med nasjonale prøver foregår det imidlertid to lenkinger samtidig, som det må tas hensyn til. For det første har man en ambisjon om å kalibrere prøvene til den samme skalaen over tid, og for dette blir de såkalte ankeroppgavene brukt. Imidlertid skal man samtidig også samkalibrere de to versjonene av prøven som brukes det samme året: kohortprøven og ankerprøven. I lenken mellom disse to er det ikke lenger de såkalte ankeroppgavene som er felles for de to prøvene. I stedet er det *de andre oppgavene* (de som er unike for hvert år) som utgjør ankeret. Hvis man øker antallet oppgaver som skal fungere som anker over tid for mye, vil antallet oppgaver som er felles for de to prøvene samme år bli mindre.

Figur 1 oppsummerer designet hvor man har kohortprøver, en versjon med ankeroppgaver og gradvis utskifting av ankeroppgavene over tid. Til sammen skal dette designet sikre at alle versjonene av prøvene, både kohortprøvene i påfølgende år og versjonene med ankeroppgaver, gir sammenliknbare resultater. Uavhengig av hvilken av prøvene elevene får, vil de få et resultat som refererer til den samme underliggende skalaen.



Figur 1. Skjematisk bilde av system for ankerfornyelse i regning

Metoder

Her presenteres de metodene som er brukt for både utprøving av lenkingen mellom år og undersøkelsen av hvordan den fungerer. I tillegg til å regne ut lenkefeilen på prøvene ved hjelp av den metoden som brukes i nasjonale prøver, er det her gjort rede for fire andre metoder for ekvivalering. Dette gir en sammenlikning mellom ulike metoder som har til hensikt å gjøre det mulig å evaluere kvaliteten av den nåværende lenkemetoden (FCIP).

IRT-analyse

I de nasjonale prøvene brukes en såkalt 2-parameter IRT-modell (2PL). I tillegg til at oppgaver har ulik vanskegrad, tar denne modellen også hensyn til at oppgaver i ulik grad skiller mellom elever (vi sier da at oppgavene har ulik diskriminering). Utover dette presenteres ikke mer om IRT i denne artikkelen. Interesserte lesere henvises til Embretson og Reise (2000) eller de Ayala (2009) for detaljer. I tillegg er det viktig å nevne at både vanskegrad og diskriminering (som i det følgende blir betegnet som henholdsvis *b*- og *a*-parameterne) blir beregnet i en prosess hvor både ankerprøven og kohortprøven for det samme året inngår *samtidig* i beregningene. På denne måten får man verdier for *a*- og *b*-parameterne for ankeroppgavene som man oppfatter som kjente for senere års prøver. Det neste året beregnes derfor ikke disse på nytt for ankeroppgavene, men inngår som forhåndsbestemte verdier i analysen av neste års resultater. På denne måten beregnes verdiene (*a* og *b*) for de nye oppgavene året etter *relativt*

til denne kjernen av oppgaver med kjente egenskaper. Denne metoden kalles FCIP (Fixed Common Item Parameters) (Kolen & Brennan, 2014).

Ekvivaleringsmetoder

Før selve lenkefeilen kan regnes ut må ankeroppgavene fra to år settes på samme skala, eller *ekvivaleres*. I praksis betyr dette at uansett hvilken av prøvene elevene besvarer, så skal de få det samme resultatet. På nasjonale prøver brukes som kort presentert ovenfor FCIP (Fixed Common Item Parameters) som er metoden IRT-programmet Xcalibre (Guyer & Thompson, 2014) implementerer. Metoden bygger på en antakelse om at parameterne for ankeroppgavene ikke endrer seg fra én prøve til den neste. Men i den nye prøven (i år to) implementeres disse ankeroppgavene sammen med andre oppgaver. Det kan derfor tenkes at oppgavenes egenskaper endrer seg litt fra én gjennomføring til den neste. Før ekvivaleringen kan godtas, må man derfor undersøke om oppgaveparameterne til ankerprøvene blir vesentlig annerledes dersom de i stedet hadde blitt kalibrert sammen med oppgavene i den nye prøven. Dersom det er store endringer, vil lenkingen/ekvivaleringen være ustabil, og denne ustabiliteten er i så fall en kilde til målefeil. Denne metoden har vist seg egnet for komplekse ferdigheter, og slik den er implementert i Xcalibre er den ganske stabil over tid, selv om den ikke er perfekt, som litteraturen viser (Keller & Keller, 2011; Kim, 2006; Kim & Cohen, 1998; Strietholt & Rosén, 2016).

Det finnes imidlertid mange forskjellige metoder for å gjøre en ekvivalering. I tillegg til FCIP-metoden er derfor tre andre IRT-lenkemetoder som anvender ankeroppgave-parameterne, blitt undersøkt i denne studien: *Mean-Mean*-metoden (MM), *Mean-Sigma*-metoden (MS) og *Mean-Alignment*- (også noen ganger kalt «mean-geometric mean») metoden (MA). Disse metodene er valgt for å vise den underliggende tanken i ekvivalering av ankeroppgaver når kun parameterne til disse oppgavene er brukt. I tillegg evalueres her både FCIP-metoden som nasjonale prøver bruker og en *samkalibrering* (concurrent calibration) av alle oppgavene, som kanskje er den vanligste metoden i bruk i dag. Det bør understrekes at de to sistnevnte metodene ikke kun baseres på en vurdering av ankeroppgavene og hvordan de endrer seg, men bruker hele datasettet, dvs. alle anker- og kohortoppgaver som beskrevet senere.

Dette ble gjort for å evaluere forskjeller mellom metodene og hvorvidt det var samsvar mellom dem i det endelige resultatet. De første tre metodene (MA, MM og MS) implementerer lenking ved å beregne parameterne for ankeroppgavene i den ene prøven for deretter å bruke disse verdiene i en relativt enkel matematisk transformasjon av skårene i den nye prøven. Det finnes mange andre IRT-ekvivaleringsmetoder som opererer på ankerparameterne, for eksempel de som bygger på sammenlikninger av testkarakteristiske kurver (TCC) fra de to prøvene som skal ekvivaleres, for eksempel Stocking-Lord-metoden og

Haebara-metoden (Kolen & Brennan, 2014; von Davier, 2011). Disse metodene er matematisk noe mer kompliserte og blir ikke beskrevet her.

En mer krevende metode er en såkalt *samkalibrering* (concurrent calibration), som er den metoden TIMSS, PIRLS, NAEP og nå også PISA fra og med 2015 bruker for å følge med endringer over tid og sette resultater fra alle gjennomføringer på samme skala. Dette kan føre til litt forskjellige resultater fra de andre metodene (Kim & Cohen, 1998) men er likevel den foretrukne metoden i dag, spesielt for storskala-undersøkelser. Med denne metoden blir det gjort en kalibrering av alle elevsvar fra to eller flere år sammen, for eksempel alle svar i PISA fra 2012 og 2015, og den forskjellen som viser seg mellom de to årene, er endringen i ferdighet mellom år. Deretter kan man bestemme en lineær transformasjon fra det nåværende året til startåret av trenden, og alle skårer fra 2015 blir transformert med den. Dette har vist seg å være en bedre og mer presis metode for å evaluere forskjeller mellom år enn kun å bruke parameterne fra ankeroppgavene (Kim & Cohen, 1998). Metoden baserer seg på kalibrering av alle oppgaver i prøven fra alle elever, men den krever imidlertid kraftige datamaskiner og programvare som klarer å kalibrere meget store datasett med store andeler systematisk manglende data. En slik metode blir også prøvd ut her for prøven i regning på 8. trinn. På denne måten kan man få sammenliknet utfall av en slik samkalibrering med FCIP-metoden og andre metoder som baserer seg på å fiksere parameterne for ankeroppgavene. Alle de nevnte metodene er nært beslektet selv om de har forskjellig implementering. Man kan derfor ikke forvente at resultatene blir dramatisk ulike avhengig av metode. FCIP-metoden og samkalibreringen har spesielt mye til felles og burde derfor ha bedre lenkeresultater (von Davier & von Davier, 2007).

Beregning av lenkefeil

Hensikten med alle disse metodene er å finne den lineære matematiske funksjonen som gir en ekvivalent transformasjon fra den ene skalaen til den neste, og vice versa. Da må både stigningstallet og konstanten for den lineære transformasjonen (henholdsvis α og β) estimeres, og resultatet brukes til å transformere både oppgaveparametere og elevresultater fra den nye kalibreringen over til den originale skalaen. Disse er ofte kalt ekvivaleringskoeffisienter. Den lineære transformasjonen ser slik ut:

$$\theta_i = \alpha\theta_j + \beta \qquad (1)$$

Her er θ_i den transformerte skåren, hvor *i* og *j* står for henholdsvis transformert skala og den gamle skalaen. α er stigningstallet for linjen og β konstanten som beskriver hvor på *y*-aksen linjen starter.

De tre parametermetodene er i grunnen like, men mean-mean-metoden benytter et gjennomsnitt av vanskegraden til ankeroppgavene, mens mean-sigma-metoden bruker variansen i vanskegrad i hele ankersettet. Mean-alignment-metoden baseres på først å transformere alle parametere i to gjennomføringer av ankeroppgavene, slik at deres gjennomsnitt blir 0. I denne sistnevnte metoden trenger en ikke finne α og β spesifikt som for de to andre metodene, som beskrevet under.

Når denne ekvivaleringen av oppgaveparameterne er foretatt med en av de tre parametermetodene, kan selve lenkefeilen regnes ut, og et generelt uttrykk for dette er gitt i formel 2. Den matematiske formen for dette uttrykket er lik den som er brukt for å estimere utvalgsfeil i et populasjonsestimat (OECD, 2009a).

For eksempel gjør man følgende i mean-alignment-metoden, som er den enkleste ekvivaleringsmetoden: Først blir en gjennomsnittlig lineær transformasjon brukt på ankersettet fra to år, slik at gjennomsnittsvanskegraden på begge år blir 0. Så blir forskjellene (c_i) mellom de transformerte parameterne regnet ut, og kvadratene av alle disse avvikene summert. Når man så dividerer på antallet oppgaver som inngår (n) får man beregnet gjennomsnittlig kvadrert avvik på tvers av alle parameterne som inngår.

Selve lenkefeilen regnes ut med en av følgende formler:

$$Lenkefeil = \sqrt{\frac{1}{n}\sum_{i=1}^{L} c_i^2} \qquad (2)$$

En alternativ formulering anvender i stedet oppgavenes varians (σ^2):

$$Lenkefeil = \sqrt{\frac{\sigma^2}{n}} \qquad (3)$$

Formel 3 gir generelt lavere lenkefeil enn formel 2 selv om den bygger på samme forutsetninger og tar hensyn til antall ankeroppgaver (Monseur & Berezner, 2007).

Når mean-mean- eller mean-sigma-metodene er brukt, som nevnt tidligere, må man finne α og β først, og så foreta transformasjonen av parameterne, og etter det kan lenkefeilen regnes ut med formel 2 eller 3.

α og β er stigningstallet og konstanten i den lineære transformasjonen som blir anvendt på parameterne. Vanskegraden, eller b-parameteren, er i begge metodene transformert slik:

$$b_{y=}\alpha b_x + \beta \qquad (4)$$

Her er den transformerte vanskegraden b_y og den originale b_x. *a*-parameteren, eller diskrimineringen, transformeres slik:

$$a_y = \frac{a_x}{\alpha} \tag{5}$$

Siden både oppgaver og elever er på samme skala, kan også elevskåren transformeres lineært med samme α og β, som er det samme som formel (1):

$$\theta_y = \alpha\theta_x + \beta \tag{6}$$

Bestemmelsen av α og β

Med mean-mean-metoden estimerer man α og β ved å bruke gjennomsnittene av *a*- og *b*-parameterne på ankeroppgavene i begge prøvene. Mean-sigma-metoden estimerer α og β ved å bruke gjennomsnitt og standardavvik av *b*-parameteren på ankeroppgavene i begge prøver.

Med Mean-mean-metoden finner man α og β med følgende formler, hvor gjennomsnittene for diskrimineringen og vanskegraden (*a*- og *b*-parameterne) for begge ankersett er brukt. Metoden bruker to gjennomsnitt, derav navnet: Først må α bestemmes og så kan den verdien brukes til å bestemme β, sammen med gjennomsnittlig vanskegrad fra begge kalibreringer av ankersettet.

$$\hat{\alpha} = \frac{\bar{a}_y}{\bar{a}_x} \tag{7}$$

$$\hat{\beta} = \bar{b}_y - \bar{b}_x\hat{\alpha} \tag{8}$$

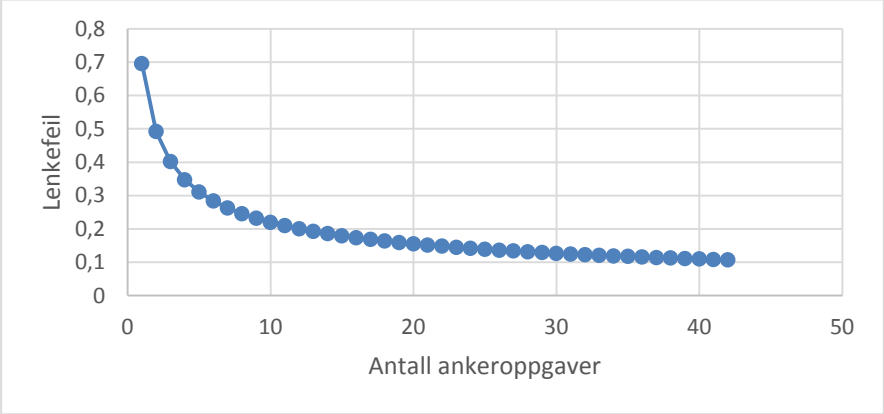
Mean-sigma-metoden gjør det samme, men finner α med gjennomsnittene av standardavviket av *b*-parameterne på de to prøvesettene (9) og β med samme metode som mean-mean metoden (8).

$$\hat{\alpha} = \frac{\bar{s}_y}{\bar{s}_x} \tag{9}$$

Størrelsen på lenkefeilen

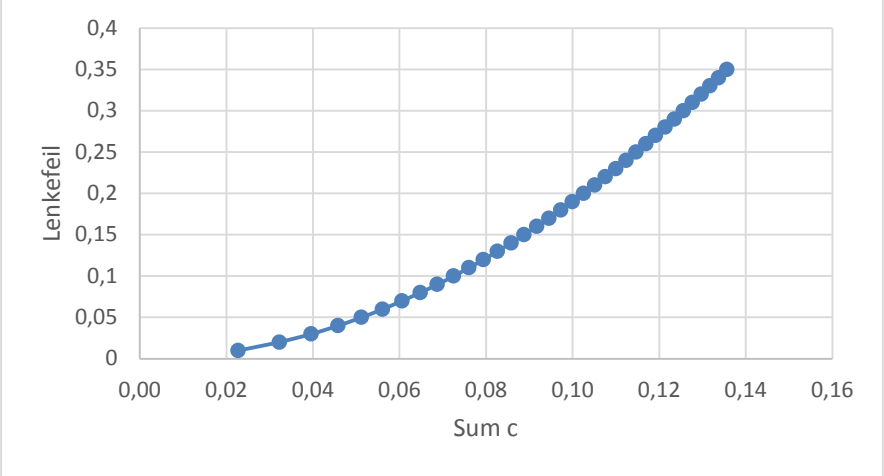
Det er klart fra formlene for lenkefeil (2 og 3) at den er avhengig av to forhold: For det første vil feilen minke når antallet oppgaver øker (større anker), og for det andre er den avhengig av hvor mye vanskegraden for ankeroppgavene endrer seg. Større forskjeller fører til større lenkefeil med begge metoder.

Hvis summen av forskjeller mellom alle ankeroppgaver var konstant, så ville antallet ankeroppgaver ha en gradvis mindre effekt på lenkefeilen som vist i figur 2, der antall oppgaver vises horisontalt og lenkefeil vertikalt. Det er verdt å merke seg at det å gå fra for eksempel 20 til 30 oppgaver, ser ut til å ha liten effekt på lenkefeilen.



Figur 2. Effekten av antall ankeroppgaver, gitt konstant forskjell mellom dem

Det viktigste er å holde ankeroppgavene stabile slik at *c* i formel 2 (summen av forskjellene mellom ankersettene) blir så lav som mulig eller at variansen i forskjellene blir så lav som mulig. Effekten på lenkefeilen av summen av forskjellene i ankeroppgaver vises i figur 3.



Figur 3. Effekt av summestørrelse på lenkefeilen

I realiteten er det alltid samvariasjon mellom antall oppgaver og summen /variansen av forskjeller.

Det er åpenbart at denne sammenhengen ikke er helt enkel, men er et samspill mellom to faktorer som begge må oppfylle minimumskrav, det vil si mange nok ankeroppgaver og så liten forskjell mellom gjennomføringer som overhodet mulig. For å oppnå lavere lenkefeil kan det derfor være fordelaktig i noen tilfeller å redusere antall ankeroppgaver, ved å ta bort de som har størst forskjell.

Når man bruker en samkalibrering for å ekvivalere prøver, kan lenkefeilen enkelt bestemmes ved å se på forskjellen i gjennomsnitt mellom to gjennomføringer og standardfeilen av den forskjellen, som er det samme som lenkefeilen. Ved å ta utgangspunkt i det året skalaen ble etablert og samkalibrere resultatene fra de to årene, kan derfor lenkefeilen bestemmes. Etter det gjør man det samme med alle år. Det samme gjelder for FCIP-metoden, der standardfeilen på forskjellen mellom år etter ekvivalering gir lenkefeilen.

Denne gjennomgangen understreker at det ikke finnes noen «riktige» løsninger, men forskjellige metoder som brukt på en fornuftig måte kan gi nyttige opplysninger.

Bestemmelse av endring over tid og lenkefeilens rolle

Når man bestemmer om en endring fra ett tidspunkt til et annet er signifikant, er en vesentlig del av evalueringen å kunne regne ut standardfeil på *forskjellen* mellom to gjennomsnitt fra de to tidspunktene som skal sammenliknes:

$$SE = \sqrt{{SE_{Tid_1}}^2 + {SE_{Tid_2}}^2 + Lenkefeil^2} \qquad (10)$$

Vi ser her at standardfeilen av forskjellen i gjennomsnittet for de to måletidspunktene inkluderer lenkefeilen i tillegg til målefeil knyttet til hver av de to målingene. Dette er veldig viktig hvis man skal være sikker på om to gjennomsnitt i virkeligheten er forskjellige fra hverandre.

Resultater

Her blir resultatene fra sammenlikninger av de tre ekvivaleringsmetodene mean-alignment (MA), mean-sigma (MS) og mean-mean (MM) beskrevet. Transformasjoner av vanskegraden til de samme oppgavene blir vist, og effekten av metodene (MS, MM, samkalibrering og FCIP) på elevskår blir undersøkt, i tillegg til bias/skjevhet i transformasjonene og hvilken konsekvens dette har for tolkningene. Denne artikkelen viser kun resultatene for de nasjonale prøvene i regning for 8. trinn for årene 2014, 2015 og 2016. Analysene bruker data fra hele alderskohortene, dvs. omtrent 60 000 elever hvert år. Det var 19 felles ankeroppgaver som ble brukt i årene 2014, 2015 og 2016. Lenkefeilen mellom 2015 og 2016 blir beregnet, siden det er den, sammen med standardfeilene for hvert gjennomsnitt, som brukes til å bestemme om endring mellom disse årene er signifikant eller ikke.

Ekvivalering av oppgavenes vanskegrad (*b*-parameteren)

Figur 4 viser grafiske sammenlikninger av vanskegrad til ankeroppgavene for regning på 8. trinn, før og etter bruk av mean-alignment-, mean-mean- og mean-sigma-metodene for prøvene i 2014 og 2016. I den øverste grafen vises resultatene fra en separat kalibrering av oppgavene i 2014 og 2016. De tre neste figurene viser den samme kalibreringen for 2014-resultatene, men her er vanskegradene for oppgavene i 2016 beregnet ved å benytte henholdsvis mean-alignment-, mean-mean- og mean-sigma-metoden. Alle metodene gir samsvarende resultater. Gjennomsnittlig for hele settet finnes ingen signifikante forskjeller mellom 2014 og 2016, uansett hvilken metode som benyttes for å kalibrere 2016-parameterne. Et overordnet funn er likevel at mean-sigma-metoden gir de mest overlappende verdiene.

Figur 4. Separat kalibrering og kalibrering med tre lenkemetoder mellom 2014 og 2016 – regning 8. trinn

Når forskjellen mellom separat kalibrerte ankerparametere fra 2014 og 2016 er gransket, viser den seg også å være veldig liten gjennomsnittlig: –0,08. Tabell 1 har en oversikt over alle de 19 ankeroppgavene, gjennomsnittsforskjeller separat kalibrert og for hver av lenkemetodene.

Tabell 1. Forskjeller i vanskegrad (*b*-parameter) mellom 2014 og 2016, separat kalibrert og ekvivalert med tre parametermetoder

Oppgave-nummer	2016–2014	MA	MM	MS
R5041845	–0,1365	0,2185	–0,0550	0,0686
R5041705	–0,1151	–0,2172	–0,0419	0,0947
R5041701	–0,0692	–0,3172	0,0029	0,1007
R5041745	–0,3576	–0,5314	–0,2952	0,1135
R5041792	–0,1068	0,4524	–0,0200	0,0546
R5041693	0,1580	–0,2890	0,2350	0,0990
R5041835	–0,0191	0,5003	0,0703	0,0518
R5041736	–0,1160	0,0333	–0,0378	0,0797
R5041798	–0,2307	0,1542	–0,1522	0,0725
R5041809	–0,0736	–0,6603	–0,0084	0,1212
R5041704	–0,0785	1,0804	0,0213	0,0171
R5041826	–0,0685	–0,7398	–0,0048	0,1259
R5041687	–0,0293	0,4366	0,0586	0,0556
R5041729	–0,1710	–0,2614	–0,0997	0,0973
R5041664	–0,0295	–0,9708	0,0303	0,1397
R5041829	–0,0404	0,5862	0,0503	0,0467
R5041775	–0,3119	0,6660	–0,2247	0,0419
R5041797	0,2387	0,7164	0,3372	0,0389
R5041670	0,0325	1,0456	0,1337	0,0192
Snitt	–0,0802	0,1001	0,0000	0,0757
SD	0,1394	0,6054	0,1446	0,0362
Varians	0,0194	0,3665	0,0209	0,0013
SE	0,0320	0,1389	0,0332	0,0083

Andre metoder – «Concurrent» kalibrering

Parameterekvivaleringsmetodene som er nevnt og prøvd ut her, opererer alle med en separat kalibrering av hvert år. Som påpekt tidligere, er det nå mer vanlig å heller bruke samkalibreringsmetoder, dvs. metoder som ikke bruker én av prøveformene for å fiksere parameterne. Dette har også blitt gjort her ved å foreta en felles kalibrering av de tre prøvene i 2014, 2015 og 2016. Da ble alle datasettene satt sammen og transformert over på 2014-skalaen. Resultater for dette blir vist i sammenlikning med de andre metodene i det følgende.

Effekten av ekvivaleringsmetoden på elevskårer

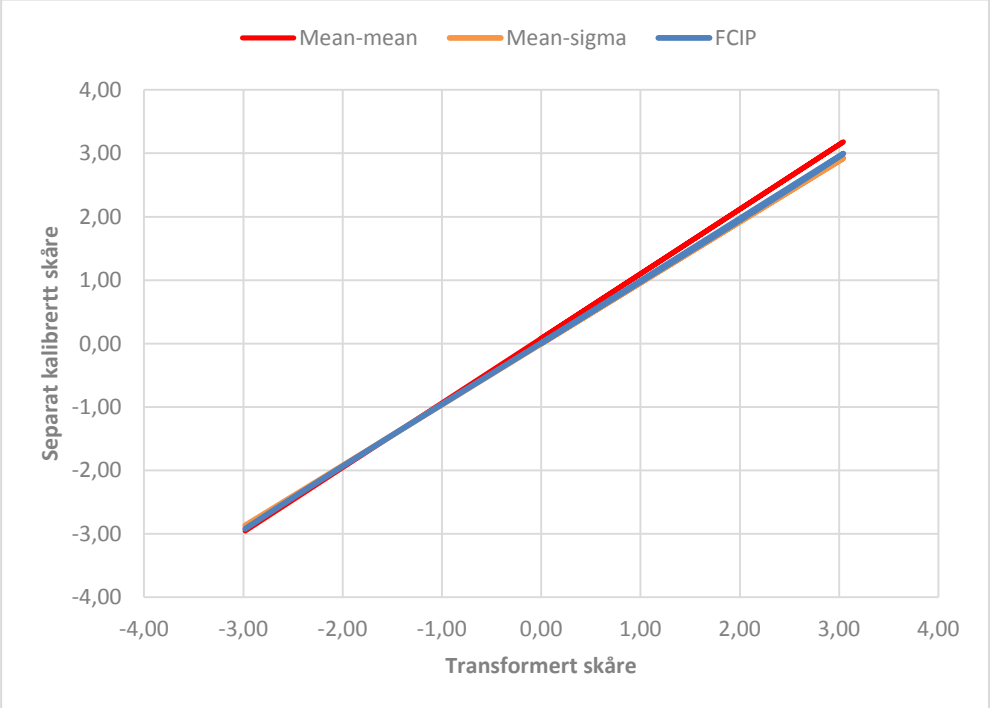
Alle ekvivaleringsmetoder har noen grad av feil, og i tillegg er det alltid en mulighet for at en form for bias eller skjevhet kan forekomme når skår fra én skala er ekvivalert med en annen. I praksis betyr dette ofte at en transformert skår ikke endres like mye alle steder på skalaen.

Lenkefeilen er av to typer: feil knyttet til tilfeldigheter i utvalget av elever som tar prøvene, og systematisk feil som følger av alle de grunnene som er

nevnt tidligere. Den første feilkilden er først og fremst knyttet til utvalgsfeil eller skjevheter i utvalgsmetoder, og den er ofte beskrevet med en såkalt «standard error of equating» (SEE) (Ogasawara, 2001). Denne feilen forventes å være meget liten på nasjonale prøver siden nesten hele populasjonen deltar.

Hvis den systematiske lenkefeilen er stor, vil den ha effekt langs hele skalaen. Hvis den for eksempel er 1 poeng, så vil alle elever på den ekvivalerte skalaen få 1 poeng høyere enn de ellers ville fått. Dette har ikke konsekvenser for sammenlikninger innenfor et år, men er viktig når resultater fra ett år blir sammenliknet med et annet år. I tillegg er det ikke helt sikkert at denne feilen er den samme alle steder på skalaen, det kan eksistere skjevheter (bias) i hvordan lenkefeilen viser seg.

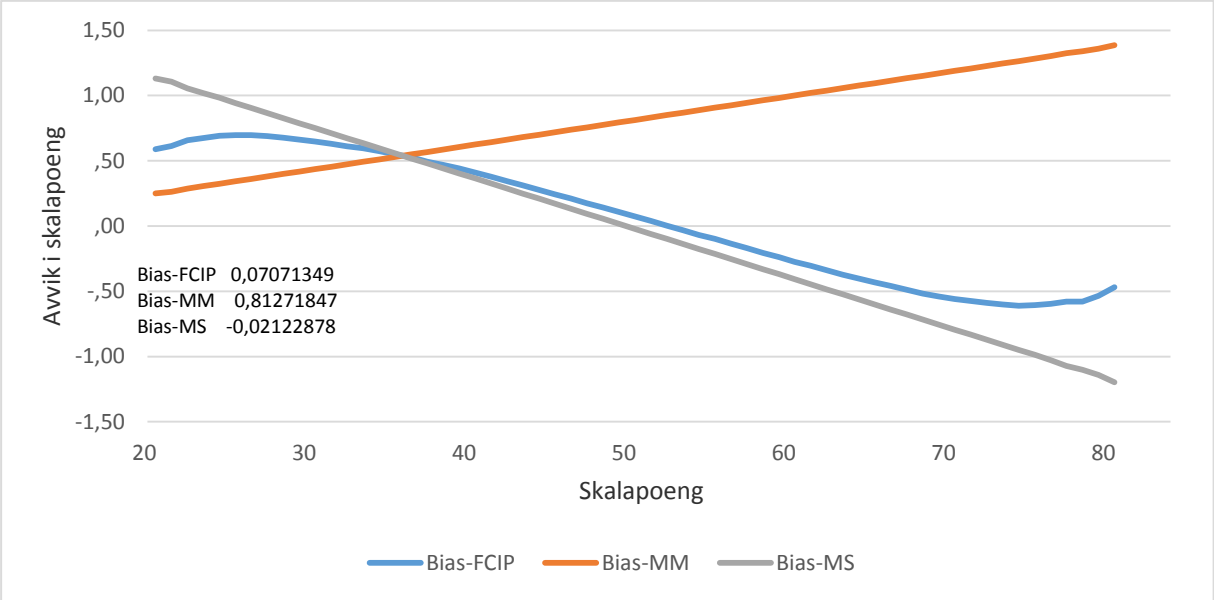
Derfor er det også viktig når en vurderer selve lenkefeilen, å sjekke om den er den samme alle steder på skalaen. For å studere dette ble alle skårer fra prøven i regning for 8. trinn transformert med metodene presentert ovenfor, og forskjellen på ulike steder på skalaen ble undersøkt[3]. Det bør understrekes at metodene leverer nesten identiske skårer, og en korrelasjon mellom elevskårer fra dem alle sammen er rundt 1.0, altså omtrent perfekt samsvar. Men de skjevhetene som er til stede, kan undersøkes ved å se på sammenhengen mellom det en separat kalibrering av alle oppgavene ville gitt og det som ekvivaleringen gir. Figur 5 viser tre ekvivalerte theta-skårer sammenliknet med en separat kalibrering av de samme oppgavene, fra samme oppgavesett hos samme elever. Merk at her ble det operert med theta-skår direkte fra analysene, som er rapportert på en skala fra –4 til 4 med gjennomsnitt på 0.



Figur 5. Effekten av tre forskjellige ekvivaleringsmetoder på elevskåren

[3] Dette ble imidlertid ikke gjort for mean-alignment-metoden, siden det ville ha krevd en helt ny skåring av alle elever med de transformerte parameterne.

Figuren viser små forskjeller mellom de tre metodene, og det er mean-mean-metoden som avviker mest fra de to andre, og avviket er størst øverst på skalaen. Hvis alle metodene fungerte like godt, ville alle linjene i figuren hatt samme stigningstall, selv om de kunne vært forskjøvet litt i forhold til hverandre. Men de tre metodene ser ut til å ha omtrent samme plassering på *x*-aksen, men litt forskjellig stigning, og FCIP-metoden er den som har stigning nærmest 45 grader, som er resultatet hvis en separat og en transformert skår var akkurat like. Dette tyder på at det kan finnes noen skjevheter (bias) i ekvivaleringene. Dette undersøkes nærmere i figur 6, som viser hvor forskjellene fra en separat kalibrering av alle oppgavene og den transformerte skåren ligger på skalaen. Her er resultatene vist med skalapoeng som har et gjennomsnitt på 50 og standardavvik på 10.



Figur 6. Forskjeller mellom separat kalibrering og tre ekvivaleringsmetoder forskjellige steder på skalaen

Sammenliknet med en separat kalibrering får vi det resultatet at MM-metoden overestimerer over hele skalaen og som sagt mest for høye skårer. En metode med ingen bias ville ha en linje på, over eller under 0 på *y*-aksen, men den ville vært vannrett. Merk igjen at selv om dette bildet viser forskjeller som ikke er større enn +/– 1,5 skalapoeng, så er bildet basert på gjennomsnitt, og noen av de egentlige tallene er høyere. Men disse forskjellene er uansett veldig små, og i gjennomsnitt har FCIP-metoden en bias på 0,07 og MS på –0,02 skalapoeng. I tillegg inkorporerer FCIP en feilretting øverst og nederst på skalaen slik at over- og underestimering i denne ekvivaleringen blir korrigert i noen grad. Denne feilrettingen er også forklaringen på hvorfor FCIP-metoden ikke er helt lineær som de andre. Dessverre finnes det ikke eksplisitt dokumentasjonen av hvordan Xcalibre-programmet har implementert denne korreksjonen i enden av skalaene. Bias for både FCIP- og MS-metodene er omtrent 0 på midten av skalaen (50), og selv om bias i begge tilfeller er liten, viser figur 6 at begge overvurderer

lavtpresterende elever og undervurderer elever med høye skårer. MM-metoden derimot overvurderer alle elever og mest øverst på skalaen. Merk at dette likevel er veldig små forskjeller.

Konsekvenser av bias i ekvivaleringen

Det er klart at konsekvensene av den bias som er i lenkingen, ikke er de samme alle steder på skalaen. Figur 6 viser at FCIP-ekvivaleringen medfører en liten overestimering av ferdighet nederst på skalaen, og en underestimering øverst på skalaen. Selv om bias i FCIP-metoden er liten i gjennomsnitt, 0,07 i den standardiserte skåren som brukes for å rapportere resultater (gjennomsnitt 50, SD 10), er det likevel klart at dette er forskjellig for forskjellige elever. For å belyse dette viser tabell 2 gjennomsnittsbias i skalapoeng for hvert av de fem mestringsnivåene i prøven.

Tabell 2. Gjennomsnittsbias på hvert mestringsnivå med FCIP-lenking

Nivå	Snitt	SE
1	0,5849	0,00106
2	0,3722	0,00093
3	0,0787	0,00096
4	–0,2306	0,00125
5	–0,4944	0,00136
Total	0,0707	0,00136

Her er det tydelig at selv om gjennomsnittet for hele gruppen viser liten bias, er den mange ganger større på laveste og øverste nivå enn på midtnivået. Det er bare på nivå 1 at den er større enn 0,5, og avrundning av skåren til hele tall uten desimaler i rapporteringen vil stort sett fjerne dette. En liten gruppe elever som har skalapoeng over omtrent 67, vil tape et skalapoeng av skåren sin, og en mindre gruppe lavt presterende elever med skalapoeng under 36, vil få et skalapoeng ekstra. Biasen er altså ikke helt symmetrisk, og det er sannsynligvis fordi vanskegradsfordelingen på ankeroppgavene heller ikke er det. Dette har små eller ingen konsekvenser for størsteparten av elevene, unntatt de som har en skår som ligger på grensen mellom nivå 1 og 2 eller på grensen mellom nivå 4 og 5.

Tabell 3 viser en nærmere granskning av dette forholdet etter at avrundede skårer er plassert på nivåene. Det er 579 elever som havner på nivå 2 istedenfor på nivå 1 (siden metoden overestimerer de som er nederst på skalaen), og det er på samme måte 388 elever som havner på nivå 4 istedenfor på nivå 5 (underestimering). Det er forventet at noen elever flytter seg, og helst burde flyttingen være den samme på alle nivåer, men på grunn av den bias som er i ankringen og FCIP-justeringen, blir ikke dette helt jevnt over alle nivåer.

Tabell 3: Fordeling mellom nivåer basert på ankrede og separat kalibrerte data.

Nivåer før ankring	Nivåer etter ankring						
	Nivå	1	2	3	4	5	**Total**
	1	4 233	579	0	0	0	4 812
	2	0	10 340	599	0	0	10 939
	3	0	0	25 912	29	0	25 941
	4	0	0	210	9 690	2	9 902
	5	0	0	0	388	6 270	6 658
	Total	4 233	10 919	26 721	10 107	6 272	58 252

Det må understrekes at inndelingen på nivåer først skjer etter avrunding av elevskåren. Her må man også huske at i rapporteringen fra prøvene er det lagt vekt på at hver elevskår har en usikkerhet rundt seg og at alle elever som ligger på eller nær grensene mellom nivåer, bør sees nærmere på, siden måleusikkerheten på prøvene for den enkelte elev er mye større enn de lenkefeilene for gjennomsnittet og den bias som er rapportert her.

Lenkefeil for alle nasjonale prøver

Tabell 4 viser lenkefeilen for regning og engelsk på både 5. og 8. trinn basert på formel 2 og ved å bruke mean-alignment-metoden. I tillegg har lenkefeilen for regning på 8. trinn også blitt beregnet ved hjelp av formel 3 og mean-alignment. Disse verdiene er gjengitt i tabell 5. Gitt at mean alignment er den metoden som fungerer dårligst, vil dette si at de lenkefeilene som rapporteres her utgjør en øvre grense for lenkefeilene for disse prøvene.

Tabell 4. Lenkefeil på nasjonale prøver fra 2014 til 2016 regnet med formel 2 basert på MA

	2014–2016	**2014–2015**	**2015–2016**
Regning 8. trinn	1,35	1,62	0,92
Engelsk 8. trinn	1,07	0,85	1,09
Regning 5. trinn	1,35	0,98	0,93
Engelsk 5. trinn	1,61	1,14	0,85

Tabell 5. Lenkefeil på nasjonale prøver i regning på 8. trinn fra 2014 til 2016 regnet med formel 3 basert på MA

	2014–2016	**2014–2015**	**2015–2016**
Regning 8. trinn	0,36	0,84	0,21

De to metodene gir store forskjeller, og vi ser at bruk av formel 3 gir langt lavere estimat for lenkefeilene i alle de tre tilfellene. Imidlertid ser de ulike metodene ut til å gi relativt stabile estimater for lenkefeil, uavhengig av hvilke år resultatene lenkes mellom.

Den største lenkefeilen er for regning på 8. trinn mellom årene 2014 og 2015. Allerede i 2015 viste det seg at ankeret som ble benyttet så ut til å være systematisk lettere i 2015 enn i 2014 (lavere *b*-parametere). En sannsynlig grunn til at

dette skjedde, var at antallet oppgaver totalt på hele prøven ble redusert fra 58 i 2014 til 50 i 2015. Dette førte til at flere elever besvarte alle oppgavene (som nettopp var hensikten med endringen), og spesielt påvirket dette andelen elever som besvarte oppgaver sist i oppgaveheftet. Det er derfor grunn til å tro at den noe større lenkefeilen mellom disse to årene reflekterer denne endringen i hvordan prøvene ble utformet. For regning på 8. trinn gir mean-mean-metoden og mean-sigma-metoden svært like lenkefeil for alle år.

Men det er likevel klart at de fire ekvivaleringsmetodene virker på forskjellig måte; mean-sigma-metoden virker best av parametertransformasjonene, selv om transformasjonen av elevskårer er bedre i FCIP nettopp på grunn av rettelsen øverst og nederst på skalaen. Alle metodene har noen forskjeller når vi tar fram forstørrelsesglasset som vist i figur 6, og i tillegg er lenkefeilen avhengig av hvilken beregningsmetode som er brukt for å estimere den.

FCIP og samkalibrering

Av denne grunn var det viktig å prøve nyere og mer presise metoder, og derfor ble metoden for samkalibrering også prøvd ut, og den viser en lenkefeil i regning på 8. trinn som er 0,58 mellom 2014 og 2015, og 0,57 mellom 2015 og 2016. Verdiene ligger altså midt imellom de to beregningsmetodene for å estimere lenkefeilen. Lenkefeilen med samkalibrering er også mye mer stabil over tid, noe som skyldes at alle oppgavene inngår, ikke bare ankeroppgavene.

Tabell 6 viser lenkefeilen med FCIP-metoden for alle prøvene og mellom alle tre år, og den indikerer at FCIP-metoden virker tilsvarende som samkalibreringen når tre år inkluderes i analysen. I tillegg er en konsekvens av FCIP-metoden at parameterne for alle oppgavene i prøven endres relativt til ankeroppgavene, noe metoden også har til felles med samkalibrering. Det er derfor å forvente at samkalibrering og FCIP begge gir mer stabile og presise resultater enn de andre metodene som anvender lineære transformasjoner av oppgaveparameterne til kun ankeroppgavene. I regneprøven for 8. trinn gir disse metodene det samme resultatet.

Tabell 6. Lenkefeil i skalapoeng med FCIP-metoden

	2014–2015	**2015–2016**
Regning 8. trinn	0,58	0,57
Engelsk 8. trinn	0,59	0,57
Regning 5. trinn	0,58	0,56
Engelsk 5. trinn	0,58	0,57

Det gjenstår å se hvordan FCIP vil oppføre seg når flere enn tre år ligger bak lenkingen.

Bruk av lenkefeilen

I 2014 ble det nasjonale gjennomsnittet for regning og engelsk på 5. og 8. trinn satt til 50 poeng, og med et standardavvik lik 10 poeng.

Tabell 7 viser endringene i det nasjonale gjennomsnittet for alle de fire prøvene over tre år, etter FCIP-ekvivalering. De to kolonnene merket forskjell, angir standardfeilen for gjennomsnittet fra formel 10 ovenfor. Tallene i kolonnene merket z, gir resultater fra signifikanstestingen hvor verdier på +/– 1,96 angir statistisk signifikante forskjeller, og vi kan konkludere at ingen av forskjellene fra ett år til det neste er i nærheten av å være statistisk signifikante.

Tabell 7. Signifikanstest av forskjeller mellom alle år

	2014	**SE**	**2015**	**SE**	**2016**	**SE**	**Forskjell 14–15**	**z**	**Forskjell 15–16**	**z**
REG08	49,993	0,041	50,165	0,041	50,062	0,040	0,171	0,293	–0,103	–0,179
REG05	49,984	0,042	50,019	0,040	50,020	0,039	0,035	0,060	0,002	0,004
ENG08	49,990	0,042	50,044	0,041	50,045	0,040	0,054	0,092	0,002	0,003
ENG05	49,968	0,041	49,733	0,041	49,859	0,040	–0,236	–0,405	0,201	0,349

Det er for øvrig verdt å merke seg at når standardfeilen mellom gjennomsnitt for to år beregnes (formel 10 foran), så er denne i all hovedsak dominert av størrelsen på lenkefeilen. Som en tommelfingerregel viser beregninger at en endring må være på nesten 2 poeng på denne skalaen for at forskjellene mellom to år skal kunne slås fast med tilfredsstillende sikkerhet. Det er derfor klart at det er av overordnet betydning å holde lenkefeilen så lav som mulig. For mindre grupper (skolenivå) vil generelt standardfeil for gjennomsnittet for de to separate prøvene (SE_{Tid1} og SE_{Tid2} i formel 10) ha langt større betydning fordi elevtallene er mindre.

Diskusjon og sammenfatning

Konklusjonen etter denne øvelsen er at lenkefeilen på nasjonale prøver bør inkluderes i alle evalueringer av trendsignifikans. Hvis denne feilkilden ikke inkluderes, vil man i de fleste tilfeller trekke feil konklusjoner om endring over tid. Lenkefeilene for de nasjonale prøvene er eksempelvis større enn den som rapporteres for de fleste internasjonale undersøkelser – hvor tilsvarende metoder anvendes. Dette skyldes primært at prøvene i de internasjonale undersøkelsene har langt flere oppgaver (inkludert et høyere antall ankeroppgaver), og en annen årsak er at man i disse undersøkelsene også kan foreta samkalibrering gjennom å bruke data fra flere undersøkelsesår. Imidlertid er lenkefeilen for de nasjonale prøvene i engelsk og regning relativt lik den lenkefeilen som rapporteres for lesing i PISA (OECD, 2009b).

Gjennom resultatene som er vist i denne artikkelen, og ved å vise til størrelsen for lenkefeil også i de internasjonale undersøkelsene, er det grunn til å være skeptiske dersom man ser rapporteringer av signifikante endringer i elevprestasjoner på systemnivå fra ett år til det neste. Lenkefeilen alene tilsier at endringen fra ett år til det neste må være 0,2 standardavvik eller høyere for å være statistisk signifikant – og på systemnivå er dette å betrakte som en svært stor endring. For 8. trinn svarer en slik endring eksempelvis til den samlede effekten av om lag ett års undervisning i skolen (se Olsen & Björnsson, 2018).

Denne gjennomgangen har også forhåpentligvis belyst at det å bruke ekvivaleringsmetoder som benytter hele datasettet og inkluderer alle oppgavene og alle elevene (ikke bare ankeroppgavene), er av betydning for presisjonen av resultatet. De eldre parameterbaserte metodene overestimerer sannsynligvis lenkefeilen i noen grad. En samkalibrering ser ut til å være den metoden som er mest anvendbar og muligens mest transparent og best dokumentert. I tillegg til å regne ut endringer mellom år, er det enda viktigere å se på utviklingen over flere år for å fange opp de trender og tendenser som er til stede, og disse metodene er i stand til å fange opp dette på en sikker og metodologisk forsvarlig måte. For å se på endringer fra ett år til det neste ser imidlertid også FCIP ut til å virke tilfredsstillende og leverer lenking med tilsvarende stor sikkerhet. Det gjenstår imidlertid å se hvordan FCIP vil stå seg i en sammenlikning med samkalibreringsmetoden over lengre tid.

Om forfatteren

Julius Kristjan Björnsson er forsker ved Universitet i Oslo og leder av EKVA – Enhet for kvantitative utdanningsanalyser. Hans forskningsinteresser omfatter blant annet psykometri og prøveutvikling, storskala utdanningsanalyser og internasjonale komparative studier.
Institusjonstilknytning: Institutt for lærerutdaning og skoleforskning, Universitetet i Oslo, Postboks 1099 Blindern, 0317 Oslo.
E-post: j.k.bjornsson@ils.uio.no

Referanser

Antal, J., Proctor, T. P. & Melican, G. J. (2014). The Effect of Anchor Test Construction on Scale Drift. *Applied Measurement in Education, 27*(3), 159–172. doi: https://doi.org/10.1080/08957347.2014.905785

Björnsson, J. K. (2016). Metodegrunnlag for nasjonale prøver. Oslo: Utdanningsdirektoratet. https://www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-for-nasjonale-prover-august-2018.pdf

Cook, L. L. & Eignor, D. R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*(2), 161–173.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford press.
Dorans, N. J., Moses T. P. & Eignor, D. R. (2010). Principles and Practice of Test Score Equating. *ETS Research Report Series*, (2), i–41. doi: https://onlinelibrary.wiley.com/doi/full/10.1002/j.2333-8504.2010.tb02236.x
Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associates, Inc.
Guo, H., Liu, J., Dorans, N. & Feigenbaum, M. (2011). Multiple Linking in Equating and Random Scale Drift. *ETS Research Report Series*, (2), i–27. doi: https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2011.tb02282.x
Guyer, R. & Thompson, N. A. (2014). *User's Manual for Xcalibre item response theory calibration software, version 4.2.2 and later*. Woodbury, MN: Assessment Systems Corporation.
Hastedt, D. & Desa, D. (2015). Linking Errors between Two Populations and Tests: A Case Study in International Surveys in Education. *Practical Assessment, Research & Evaluation, 20*(14), 1–12.
Keller, L. A. & Keller, R. R. (2011). The Long-Term Sustainability of Different Item Response Theory Scaling Methods. *Educational and Psychological Measurement, 71*(2), 362–379. doi: https://doi.org/10.1177/0013164410375111
Kim, S. (2006). A Comparative Study of IRT Fixed Parameter Calibration Methods. *Journal of Educational Measurement, 43*(4), 355–381. doi: https://doi.org/10.1111/j.1745-3984.2006.00021.x
Kim, S. H. & Cohen, A. S. (1998). A Comparison of Linking and Concurrent Calibration under Item Response Theory. *Applied Psychological Measurement, 22*(2), 131–143.
Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*. New York, NY: Springer New York.
Martin, M. O., Mullis, I. V. S., Foy, P., Brossman, B. & Stanco, G. M. (2012). Estimating linking error in PIRLS. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, Volume 5.
Meyers, J. L., Miller, G. E. & Way, W. D. (2008). Item Position and Item Difficulty Change in an IRT-Based Common Item Equating Design. *Applied Measurement in Education, 22*(1), 38–60. doi: https://doi.org/10.1080/08957340802558342
Monseur, C. & Berezner, A. (2007). The Computation of Equating Errors in International Surveys in Education. *Journal of Applied Measurement, 8*(3), 323–335.
OECD (2009a). *PISA Data Analysis Manual SAS* (2nd ed.). Paris: Organisation for Economic Co-operation and Development.
OECD (2009b). *PISA Data Analysis Manual SPSS* (2nd ed.). Paris: Organisation for Economic Co-operation and Development.
OECD (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.
Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25*(1), 53–67.
Olsen, R. V. & Björnsson, J. K. (2018). Fødselsmåned og skoleprestasjoner. I J. K. Björnsson & R.V. Olsen (red.), *Tjue år med TIMSS og PISA i Norge* (s. 76–93). Oslo, Universitetsforlaget.
Sinharay, S., Haberman, S., Holland, P. & Lewis, C. (2012). A note on the choice of an anchor test in equating. *ETS Research Report Series*, (2), i–9. doi: https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2012.tb02296.x
Strietholt, R. & Rosén, M. (2016). Linking Large-Scale Reading Assessments: Measuring International Trends over 40 Years. *Measurement: Interdisciplinary Research and Perspectives, 14*(1), 1–26. doi: https://doi.org/10.1080/15366367.2015.1112711

van der Linden, W. J. & Wiberg, M. (2010). Local Observed-Score Equating with Anchor-Test Designs. *Applied Psychological Measurement*, *34*(8), 620–640. doi: https://doi.org/10.1177/0146621609349803

von Davier, A. A. (2011). *Statistical Models for Test Equating, Scaling, and Linking*. Dordrecht: Springer.

von Davier, M. & von Davier, A. A. (2007). A Unified Approach to IRT Scale Linking and Scale Transformations. *Methodology*, *3*(3),115–124. doi: https://doi.org/10.1027/1614-2241.3.3.115