Eli Moe
Universitetet i Bergen
Hildegunn Lahlum Helness
Universitetet i Bergen
Craig Grocott
Universitetet i Bergen
Norman Verhelst
Eurometrics, Nederland

DOI: http://dx.doi.org/10.5617/adno.6281

Standardsetting av læringsstøttende prøver i engelsk for Vg1

Sammendrag
Formålet med denne artikkelen er å beskrive framgangsmåten som ble brukt for å bestemme kuttskårer (grenser) mellom tre nivåer i Det europeiske rammeverket for språk (A2, B1 og B2) på to læringsstøttende lytteprøver i engelsk for Vg1-elever. Målet har vært å undersøke om det er mulig å etablere enighet om kuttskårene, og om standardsetterne som deltok i arbeidet fikk tilstrekkelig opplæring på forhånd. Videre var det et mål å se på hvilke konsekvenser kuttskårene vil få for fordeling av elever på de ulike rammeverksnivåene.

Standardsettingen ble gjennomført med utgangspunkt i pilotdata fra 3199 elever på Vg1, Cito-metoden og 16 panelmedlemmer med god kjennskap til Rammeverkets nivåer. Flere av panelmedlemmene var eller hadde vært lærere i engelsk for elever på 10. trinn eller Vg1.

Cito-metoden fungerte bra for å etablere kuttskårer som standardsetterne var forholdsvis enige om. Sluttresultatene viser at målefeilen var relativt liten. Resultatene viser større enighet om kuttskåren mellom nivåene B1 og B2 enn mellom A2 og B1, og dette kan ha en sammenheng med at det ble brukt mer tid på forberedelsesarbeid for B1 og B2.

Lærere i panelet som kjenner elevgruppa godt, mener at konsekvensen kuttskåren har for fordeling av elever på de ulike rammeverksnivåene, stemmer med deres egen vurdering av elevenes lytteferdigheter.

Nøkkelord: standardsetting, testsentrert metode, Cito-metoden, standard, kuttskår, vippekandidat

Standard setting for English tests for 11th grade students in Norway

Abstract

This article presents the process used to determine the cut scores between three levels of the Common European Framework of Reference for languages (A2, B1 and B2) for two English listening tests, taken by Norwegian pupils at the 11th grade. The aim was to establish whether agreement can be reached on cut scores and whether the standard setters received enough preparation before the event. Another aim was to examine the potential consequences the cut scores would have for the distribution of pupils across the different levels.

The standard setting took place using pilot data from 3199 pupils, the Cito method and 16 panel members with a good knowledge of the framework levels. Some panel members were or had been 10th or 11th grade English teachers.

The Cito method worked well for establishing cut scores with which the panel members mostly agreed. The results indicated a small margin of error. The results showed a higher level of agreement for the cut score between B1 and B2 than between A2 and B1, possibly connected to the longer preparation time dedicated to B1 and B2.

Teachers on the panel with good knowledge of the pupil base believe that the consequences these cut scores have for the distribution of pupils, correlate with their own experiences of pupils' ability.

Keywords: standard setting, test-centered method, the Cito method, standard, cut score, borderline person / minimally competent user

Introduksjon

Standardsetting brukes for a) å definere ferdighets- eller kompetansenivåer på en skala, og b) å komme fram til kuttskårer som skiller de aktuelle nivåene.

> The term ‘standard setting’ in the field of educational measurement refers to a decision making process aiming to classify the results of examinations in a limited number of successive levels of achievement. (Kaftandjieva, 2004, s. 2)

Det finnes forskjellige metoder for standardsetting. Kaftandjieva (2004, s. 12) sier det er mer enn 50. Jaeger (1989, s. 493) klassifiserer to hovedgrupper av metoder: testsentrerte og elevsentrerte. Ved en testsentrert metode vil et panel med personer som kjenner en kompetanse- eller ferdighetsskala godt, vurdere testitem opp mot skalaen. Om en derimot bruker en elevsentrert metode, vil personer som kjenner testkandidater godt, vurdere deres kompetanse i forhold til skalaen. Denne hovedinndelingen er akseptert av de fleste i dag.

Denne artikkelen beskriver framgangsmåten som ble brukt for å bestemme grenser/kuttskårer mellom nivåene A2 (og under), B1, og B2 (og over) på skalaen til *Det felles europeiske rammeverket for språk: Læring, Undervisning, Vurdering* (Rammeverket) (Europarådet, 2001) for læringsstøttende lytteprøver i engelsk for Vg1. Vi tar for oss grensa mellom B1 og B2 først, fordi prøve-utviklere og lærere som deltok i arbeidet med prøvene, hadde en formening om at disse to nivåene antakeligvis er de som representerer lytteferdighetene til flest 16-åringer i Norge.

Mål

Målet med denne artikkelen er tredelt. Vi ønsker å undersøke følgende spørsmål:

- Er det mulig å etablere kuttskårer standardsetterne er relativt enige om ved hjelp av den standardsettingsmetoden (Cito-metoden) som studien bygger på?
- Har tiden avsatt til å forberede standardsetterne på det de skal gjøre, noe å si for hvor enige de er om kuttskåren, og hvor passende de synes kutt-skåren er?
- Hvilke konsekvenser har kuttskårene for fordeling av elever på ramme-verksnivåer?

Bakgrunn

Sentrale begreper

Følgende begreper er sentrale for denne studien: kuttskår, standard, vippe-kandidat og IRT-analyse.

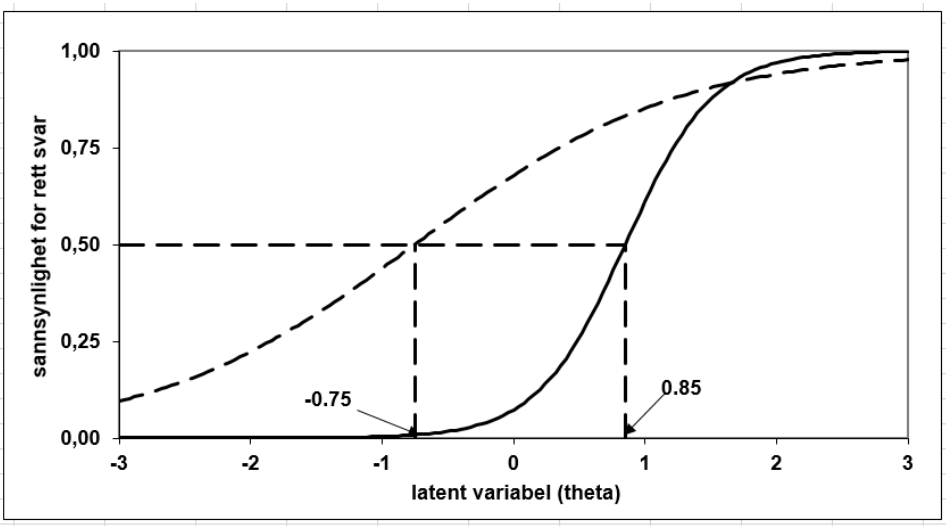
Kuttskår står for den lavest mulige skåren som gir et bestått resultat på en eksamen eller prøve. En kuttskår skiller altså mellom et bestått og ikke-bestått resultat. Noen prøver har flere kuttskårer som skiller mellom ulike ferdighets-nivåer, for eksempel en kuttskår som skiller mellom rammeverksnivåene A2 og B1, og en annen som skiller mellom nivåene B1 og B2.

Begrepet *standard* i utdanningssammenheng står for beskrivelser av hva elever skal kunne eller gjøre på bestemte trinn i et utdanningsløp. Med standard menes altså beskrevne kompetansenivåer. Mens begrepet kuttskår representerer et punkt på en skala, vil en standard stå for et intervall mellom to kuttskårer. Språkbeskrivelser som speiler B1-nivået vil være et eksempel på en standard som ligger mellom kuttskåren som skiller nivåene A2/B1 og kuttskåren som skiller nivåene B1/B2. Dette betyr at når man etablerer kuttskårer, for eksempel mellom rammeverksnivåene A2 og B1, bestemmer man samtidig hvor nedre grense for området for B1-standarden går.

En *vippekandidat* er en person som ligger på grensa til å bestå en prøve. Dersom en slik kandidat, hypotetisk sett, kunne gå opp til samme prøve 100 ganger, ville han klare seg i halvparten av tilfellene. I den engelskspråklige litteraturen blir en vippekandidat kalt *borderline person* eller *minimally competent user*.

Til grunn for *Item response theory* (*IRT-teori*) ligger oppfatningen av at en underliggende ferdighet (evne, kompetanse) kan representeres som en uavhengig kontinuerlig variabel på en skala fra $-\infty$ til $+\infty$ (fra minus uendelig til pluss uendelig). Denne variabelen, for eksempel lytteferdighet, kan ikke observeres direkte, derfor kalles den en latent variabel eller et latent trekk. Her refererer vi til variabelen som en ferdighet, og en antar at en persons ferdighet kan måles og plasseres på den latente variabelen. Vi kan også si at den latente variabelen korresponderer til en skala, og at individuelle personers lytteferdigheter kan plasseres på denne skalaen. Å måle lytteferdigheter blir da å finne ut, så nøyaktig som mulig, hvor på skalaen personer skal plasseres.

Ved hjelp av item og hvordan personer svarer på disse, kan en finne hvor både personer og item kan plasseres på skalaen. I en IRT-modell er forholdet mellom latente variabler og svar på item beskrevet på en formalisert måte ved hjelp av matematiske formler. Dette blir illustrert i figur 1. Den horisontale aksen viser den latente variabelen. I IRT-modeller er den greske bokstaven theta (θ) et uttrykk for ferdighet/evne/kompetanse (ability). Minusverdier viser til lavere ferdighet, mens plussverdier viser til høyere. Den vertikale aksen viser sannsynlighet for rett svar. Hvert item har en responsfunksjon som uttrykker sannsynligheten for rett svar som en funksjon av den latente variabelen. De to kurvene i figuren viser to slike funksjoner. Figuren viser at sannsynligheten for rett svar øker med stigende ferdigheter. Vi ser også at itemet som er representert ved en heltrukken linje, er mer til høyre og stiger brattere enn det andre. Dette betyr at itemet er vanskeligere.



Figur 1. To itemresponsfunksjoner

I psykometri, og som oftest i standardsetting, fokuserer vi på en spesifikk verdi for sannsynlighet, nemlig 0,5. Når sannsynligheten for rett svar er 0,5, betyr det at det er like stor sjanse for et rett som et galt svar på itemet. Vi ser at itemet til venstre har en sannsynlighet på 0,5 når theta (en persons ferdighet) er –0,75. Itemet til høyre har en sannsynlighet på 0,5 når theta er 0,85. Mer om IRT-analyse under *Pilotering og prøvekonstruksjon / Analyse av pilotdata*.

Læringsstøttende prøver i engelsk for Vg1

Universitetet i Bergen har utviklet læringsstøttende prøver i engelsk for Vg1 på oppdrag fra Utdanningsdirektoratet i perioden 2016 til 2018. De læringsstøttende prøvene i engelsk kan brukes i underveisvurdering og kan gi læreren grunnlag for læringsfremmende tilbakemeldinger til elevene. Prøvene er frivillige (skolene bestemmer selv om og når de vil bruke prøvene), gjennomføres elektronisk og skåres automatisk. Det er også utviklet en lærerveiledning som beskriver hvordan elevenes resultater kan følges opp.

Kompetansemål i *Læreplan i engelsk for 10. trinn og Vg1* (Utdanningsdirektoratet, 2013) har vært utgangspunktet for utviklingen av prøvene. Disse kompetansemålene er svært generelt formulert, noe som gjør det vanskelig å bruke dem som utgangspunkt for vurdering. Derfor var det viktig at lærere som kjenner elevgruppa godt, var med i arbeidet med prøvene. Fire engelsklærere med lang fartstid fra ungdomsskole og videregående skole var med i gruppa som utviklet prøvene, nettopp for å sikre at oppgaver og tema skulle være så relevante som mulig.

Det er utviklet tre sett med læringsstøttende prøver i engelsk for Vg1: to lytteprøver, to leseprøver og to prøver i *Use of English* (koherens, vokabular, grammatikk, stil og register). *Leseprøve 1* og *2* måler det samme. Det samme gjør *Lytteprøve 1* og *2* og prøve i *Use of English 1* og *2*. Hvert sett med prøver er tilnærmet parallelle. De inneholder like mange oppgaver, de samme oppgaveformatene, og har tilnærmet samme vanskegrad. Prøvene diskriminerer mellom elevene på samme måte. Diskriminering i forbindelse med prøver er et mål på hvor godt resultatet skiller mellom sterke og svake elever. Hver prøve inneholder både lette og vanskelige oppgaver. Lærerne kan velge å la elevene ta én prøve i begynnelsen av skoleåret og den andre prøven mot slutten, og dermed få et mål på framgang.

Det har blitt arrangert to standardsettingsseminarer. I 2017 ble dett satt standarder for leseprøvene og prøvene i Use of English, mens seminaret i 2018 hadde fokus på lytteprøvene. Selv om det har blitt etablert kuttskårer for *Leseprøve 1* og *2* og prøve i *Use of English 1* og *2* også, er fokus for denne artikkelen standardsetting av *Lytteprøve 1* og *2*. Det er ingen spesiell grunn til at lytteprøvene er valgt i denne sammenhengen. Hovedresultatene for alle prøvene er rapportert i tabellene 10, 11 og 12 mot slutten av artikkelen.

Læreplanens kompetansemål som gjelder lytting under *Muntlig kommunikasjon* for 10. trinn og Vg1 (Utdanningsdirektoratet, 2013), har vært utgangs-

punktet for utvikling av *Lytteprøve 1* og *2*. Tabell 1 viser kompetansemålene som gjelder lytting.

Tabell 1. Relevante kompetansemål som omhandler lytteferdigheter

	Kompetansemål lytting etter 10. trinn	**Kompetansemål lytting etter Vg1/Vg2**
Lyttestrategier	• velge og bruke ulike lytte- og talestrategier tilpasset formålet	• vurdere og bruke egnede lytte- og talestrategier tilpasset formål og situasjon
Lytteforståelse (ord og uttrykk)	• forstå og bruke et generelt ordforråd knyttet til forskjellige emner • vise evne til å skille mellom positivt og negativt ladede uttrykk som refererer til enkeltindivider og grupper av mennesker	• forstå og bruke et bredt generelt ordforråd og et faglig ordforråd knyttet til eget utdanningsprogram
Lytteforståelse (muntlig tekst)	• forstå hovedinnhold og detaljer i ulike typer muntlige tekster om forskjellige emner	• forstå hovedinnhold og detaljer i ulike typer muntlige tekster om allmenne emner og faglige emner knyttet til eget utdanningsprogram
Forståelse av ulike varianter av engelsk	• lytte til og forstå varianter av engelsk fra forskjellige autentiske situasjoner	• lytte til og forstå sosiale og geografiske varianter av engelsk fra autentiske situasjoner

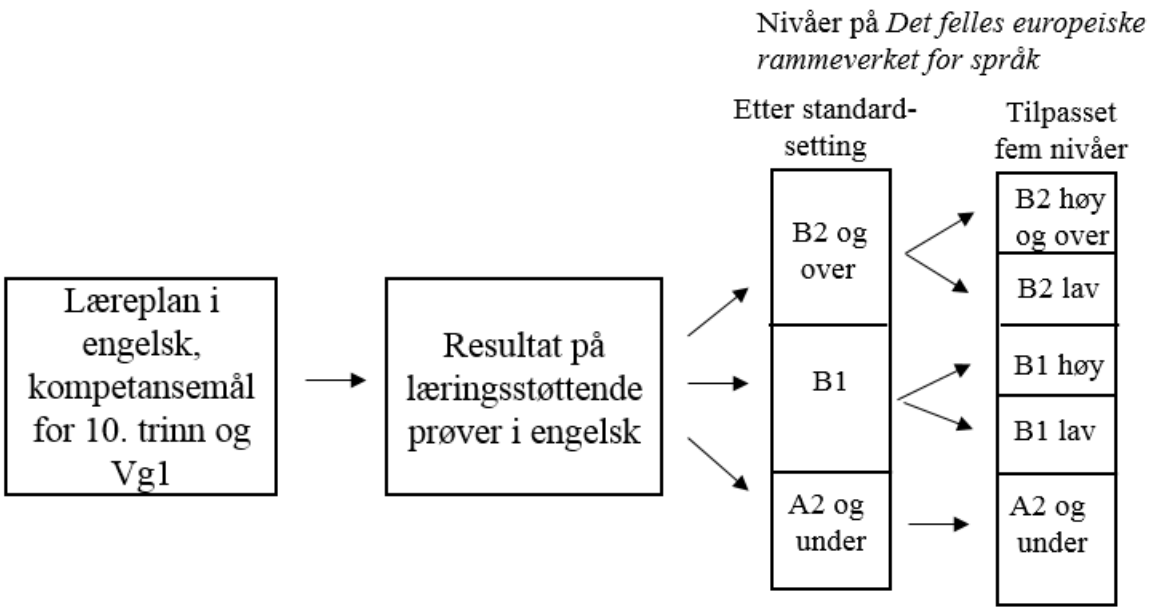
Kompetansemålene i lytting er svært generelt formulert. Det er vanskelig å bruke disse kompetansemålene som utgangspunkt for vurdering. Det samme gjelder andre kompetansemål for engelsk på 10. trinn og Vg1.

Prøvene i lytting måler:

- finne informasjon/forstå detaljer
- forstå hovedinnhold
- tolke og forstå
- forstå sammenheng i tekster

I tillegg blir elevene presentert for ulike uttalevarianter av engelsk, som britisk engelsk, australsk engelsk og amerikansk engelsk.

I avtalen med Utdanningsdirektoratet ble det slått fast at prøvene skal «som en regel utvikles etter standard fra nasjonale prøver» (Utdanningsdirektoratet, 2015). I samråd med Utdanningsdirektoratet ble det bestemt at resultatene fra de læringsstøttende prøvene i engelsk skulle rapporteres på fem mestringsnivåer, og at resultatene skulle knyttes til skalaen for Rammeverket, nærmere bestemt nivåene fra A2 (og under) til B2 (og over) slik figur 2 viser.



Figur 2. Sammenheng mellom læreplan i engelsk, resultat på Læringsstøttende prøver og Rammeverket

Ved å bruke metoder for såkalt standardsetting er det mulig å knytte resultatene på prøvene til skalaen for Rammeverket. Personer som kjenner Rammeverket godt, hjalp til med å etablere grensene mellom 1) A2 (og under) og B1, og 2) B1 og B2 (og over). Dette er det gjort greie for i delen som omhandler standard-setting senere i denne artikkelen. For å indikere hvem som ligger i lavere sjikt av B1 og B2, og hvem som har mer overbevisende nivåplassering, valgte prøve-utviklerne å dele disse nivåene i to. Vi tok utgangspunkt i spennet i poeng på hvert av disse nivåene. For eksempel: Elever med 23–34 poeng på lytteprøven havner på nivå B2 og over. Prøveutviklerne bestemte at 23–28 poeng skulle tilsvare B2 lav, mens 29–34 poeng skulle tilsvare B2 høy og over. Elever som havner på B2 lav kan en del av det som er beskrevet for B2, mens de som havner på B2 høy eller over, kan det meste av det som er beskrevet for nivået. Se omtalen av Det felles europeiske rammeverket for språk.

En viktig grunn til å dele nivåene B1 og B2 i to, er at det kan ta lang tid for en elev å utvikle ferdighetene sine fra ett nivå til et annet. Det kan derfor være nyttig å vise dem at de har hatt progresjon i utviklingen sin, selv om de ennå ikke har nådd neste rammeverksnivå.

Det felles europeiske rammeverket for språk

Rammeverket ble publisert i 2001, om lag 30 år etter at arbeidet med å beskrive et *terskelnivå* (B1) startet (Europarådet, 1975; Svanes, Hagen, Manne, Svindland & Husby, 1988). Den norske utgaven av Rammeverket ble publisert i 2011 (Utdanningsdirektoratet, 2011). Rammeverket deler språkinnlærere i tre hoved-grupper ut fra språkkompetansen de har: *Basisbrukere* (A1 og A2) har lært det aller mest nødvendige hverdagsspråket de trenger for å klare seg på et nytt språk, *Selvstendige brukere* (B1 og B2) kan klare seg språklig i sosiale settinger og i skolesammenheng, mens *Avanserte brukere* (C1 og C2) kan bruke språket

uanstrengt, sammenhengende, effektivt og presist i de aller fleste situasjoner, også profesjonelle.

Rammeverket inneholder 56 skalaer med språkbeskrivelser som dekker fem ulike språkferdigheter (lytting, lesing, muntlig produksjon, muntlig interaksjon og skriving) og seks nivåer A1–C2. På den ene siden må rammeverksnivåene stå fast. Om de ikke gjør det, mister de sin rolle som felles referansepunkt. På den andre siden skal ikke Rammeverket være dogmatisk, preskriptivt eller absolutt. Dokumentet er et referanserammeverk som oppfordrer ulike interessegrupper til å bruke de grunnleggende prinsippene og nivåene i dokumentet og deretter tilpasse disse til nasjonale og mer lokale behov.

Opprinnelig ble rammeverksbeskrivelsene utviklet med tanke på voksne fremmedspråksinnlærere, for eksempel turister eller unge/voksne studenter. Senere har nivåene og beskrivelsene blitt tilpasset andrespråksbrukere (Vox, 2012), barn (Hasselgreen, 2003, 2010) og til og med språket brukt i ulike skolefag som matematikk og historie (Moe, Härmälä, Kristmanson, Pascoal & Ramoniene, 2015. Tabell 2 er et eksempel på en av Rammeverkets opprinnelige skalaer for lytting.

Tabell 2. Eksempel på skala med nivåer i den norske utgaven av Rammeverket (Utdanningsdirektoratet, 2011, s. 85)

	Å LYTTE TIL AUDITIVE MEDIER OG INNSPILT MATERIALE
C2	Som C1
C1	Kan forstå et vidt spekter av innspilt og kringkastet auditivt materiale, til dels også når det ikke brukes standardspråk, og oppfatte finere detaljer, også underforståtte holdninger og forhold mellom dem som snakker.
B2	Kan forstå innspilt materiale på standard regionaltalemål om emner fra det sosiale, yrkesmessige eller akademiske liv, og i tillegg oppfatte holdningene og synspunktene til dem som snakker.
	Kan forstå de fleste radiodokumentarer og det meste av annet innspilt eller kringkastet auditivt materiale framført på standard regionaltalemål. Kan oppfatte sinnsstemningen, toneleiet osv. til dem som snakker.
B1	Kan forstå innholdet i det meste av innspilt eller kringkastet auditivt materiale om emner hun/han personlig er interessert i, når det er brukt klar standardtale.
	Kan forstå hovedpunktene i radioens nyhetssendinger og enklere innspilt materiale om kjente emner når stoffet er framført relativt langsomt og tydelig.
A2	Kan forstå og hente ut viktig informasjon fra korte, innspilte passasjer om forutsigbare, dagligdagse ting når stoffet er framført langsomt og tydelig.
A1	Ingen beskrivelse foreligger.

Rammeverket er per 2018 oversatt til 40 språk og blir brukt som grunnlag for læreplaner, undervisning og prøver i mange land i Europa. For å begynne å studere i utlandet, må studenter i mange land dokumentere språkferdigheter på B2-nivå. Utenlandske studenter som skal begynne å studere i Norge og ikke har eksamen fra norsk videregående skole, må dokumentere at de kan norsk godt, enten ved å oppnå B2 på *Norskprøven* (Kompetanse Norge, 2018) eller *Test i norsk, høyere nivå* (Folkeuniversitetet, 2018), alternativt ved å bestå eksamen fra universitetenes *Trinn 3 eksamen*. Norske studenter som vil studere i et engelskspråklig land, må dokumentere språkferdigheter på B2-nivå fra en

internasjonalt godkjent prøve, for eksempel *Toefl-testen* (ETS, 2018), *IELTS Academic* (IELTS, 2018) eller *First-testen* (FCE) (Cambridge Assessment, 2018).

Det at dokumentert rammeverksnivå B2 (ofte i både lytting, lesing, muntlig og skriftlig produksjon) blir brukt som inntakskriterium ved svært mange universiteter både i Europa og utenfor, er en grunn til å relatere resultatet på de læringsstøttende prøvene i engelsk til Rammeverket. På den måten kan elever få informasjon om de har nådd dette nivået eller ikke. En del av elevene i videregående skole ser for seg kortere eller lengre studieopphold i utlandet, og vil på et tidspunkt måtte dokumentere ferdighetene sine i engelsk. For disse kan det være viktig å vite noe om hvilke ferdigheter de har i lytting og lesing, og om de må gjøre en ekstra innsats før de går opp til en internasjonal prøve.

Pilotering og prøvekonstruksjon

Elevgruppe

Til sammen deltok 3199 Vg1-elever i pilotering av lytteoppgavene høsten 2017. De kom fra ulike fylker i Norge, fra store og små kommuner, og fra skoler med ulike størrelser. De fleste elevene gikk på studieforberedende utdanningsprogram, men også elever fra yrkesfaglige utdanningsprogram deltok. Hver skole ble bedt om et tilfeldig utvalg på 30 elever.

Piloteringsdesign og pilotering

I piloteringen ble det brukt et ufullstendig lenket design (incomplete linked design). Det vil si at hver pilotversjon er knyttet til alle andre pilotversjoner. Tabell 3 viser et eksempel på et slikt piloteringsdesign med seks pilotprøver.

Tabell 3. Eksempel på piloteringsdesign

Pilot 1	Pilot 2	Pilot 3	Pilot 4	Pilot 5	Pilot 6
X1	X1				
	X2	X2			
		X3	X3		
			X4	X4	
				X5	X5
X6					X6
	X7		X7		
		X8			X8
X9				X9	
X10		X10			
			X11		X11
	X12			X12	

X står for *en pakke* med item, for eksempel 8–10 item. Designet har til sammen 12 itempakker. X1 finnes i pilotprøve 1 og 2. X2 er i pilotprøve 2 og 3. Hver

prøvepakke går igjen i to pilotversjoner, og hver pilotversjon har fire itempakker. I et slikt design er hver pilotprøve lenket til alle andre pilotprøver, enten direkte eller indirekte. Pilotprøve 1 er lenket direkte til pilotprøvene 2, 3, 5 og 6 som den har felles itempakke med. Men pilotprøve 1 er også lenket indirekte til pilotprøve 4 via pilotprøve 3 som begge er direkte lenket til. Når alle pilotversjoner er lenket, kan data fra prøvene analyseres på samme skala.

Det ble pilotert 12 prøveversjoner for lytting. Hver versjon ble gjennomført av mellom 248 og 285 elever. Hvert item gikk som nevnt igjen i to versjoner, og det innebar at hvert item ble prøvd ut på minimum 499 og maksimum 569 elever. Lytteprøvene ble pilotert i løpet av tre uker i oktober 2017.

Analyse av pilotdata

I og med at itemene er pilotert i et delvis lenket design, kan vi ikke analysere data ved hjelp av klassisk testteori. Grunnen til dette er at elevene har svart på et subsett av oppgavene; ingen elev har svart på alle oppgavene som ble pilotert. Ved hjelp av IRT-analyse kan data fra alle pilotversjonene analyseres på en meningsfull måte. Her gir vi en kort oversikt over det viktigste som karakteriserer analysemåten, og som har noe å si for arbeidet med å konstruere endelige prøver.

En analyse av testdata må gjøre to ting: 1) *a*- og *b*-parametre må estimeres, og 2) en må undersøke om forholdet mellom den latente variabelen og responssannsynligheten kan beskrives relativt nøyaktig (ved hjelp av en funksjon). Dette betyr egentlig at man må undersøke om itemene har høy kvalitet, om de diskriminerer på en fornuftig måte mellom personer og har en vanskegrad en sikter mot. *a*-parametre uttrykker diskriminering og *b*-parametre vanskegrad. Det blir for teknisk og for omfattende å gå inn på hvordan parametre estimeres i denne sammenhengen. La oss bare si at det finnes mange gode programmer for IRT-analyse som kan gjøre den jobben. For å ta det andre punktet, kan vi nevne at i vårt tilfelle *overlevde* de aller fleste itemene IRT-analysen. Seks item ble kastet ut på grunn av lav diskriminering etter at 170 lytteitem var pilotert og analysert. Det er viktig å velge oppgaver som diskriminerer godt mellom elevene, fordi det medfører at prøven blir mer reliabel, noe som betyr at en kan stole på resultatene. Dersom prøven ikke er reliabel, har det ingen hensikt å gjennomføre standardsetting. Man kan ikke etablere stabile og reliable kuttskårer mellom ulike ferdighetsnivåer dersom prøven ikke er reliabel.

Prøvekonstruksjon

Basert på analyse av piloteringsdata ble oppgaver til hver prøve valgt ut med tanke på oppgaveformat, lytteaspekt (hva oppgavene måler), vanskegrad og hvor godt oppgavene diskriminerer mellom elevene. De to lytteprøvene som er konstruert, inneholder like mange oppgaver, og de inneholder oppgaver som måler de samme lytteaspektene. Oppgavene varierer fra relativt lette til vanskelige. Prøvene er samlet sett tilnærmet like vanskelige, og de diskriminerer mellom

elevene på tilnærmet samme måte. Prøvene inneholder ulike typer lyttetekster i ulike sjangre som for eksempel dialoger og intervju. I begge prøvene blir elevene eksponert for ulike uttalevarianter av engelsk. Videre har de to prøvene de samme oppgaveformatene og samme antall oppgaver innen hvert format. Emnene er varierte og sammenlignbare i begge prøvene, for eksempel med lyttetekster fra dagligliv, samfunn og yrkesliv. Tabell 4 viser oppsummerende statistikk for de ferdige lytteprøvene.

Tabell 4. Oppsummerende statistikk for læringsstøttende prøver i engelsk lytting for Vg1

	Lytteprøve 1	***Lytteprøve 2***
Antall item	25	25
Gjennomsnittlig *a*-verdi (diskriminering)	3,88	3,88
Gjennomsnittlig *b*-verdi (theta)	0,249	0,250

Vi har altså prøvd å gjøre de to prøvene så like som mulig. Se tabell 5 for mer detaljert informasjon om de to lytteprøvene.

Tabell 5. Oppgavetyper og antall item per format i læringstøttende prøver i engelsk lytting

Teksttype	**Format**	***Lytteprøve 1***		***Lytteprøve 2***		**Begge prøver**	
		Antall tekster	Antall poeng	Antall tekster	Antall poeng	Antall av-spillinger per lyttetekst	Tekst-lengde
Conversations	Multiple choice	2	5	2	5	2	Lang
Move key words (conversation)	Sorting	1	4	1	4	2	Lang
Interview	Multiple choice	1	4	1	4	1	Lang
Announcements	Multiple choice	3	3	3	3	1	Kort
Short talks and monologues	Multiple choice	5	5	5	5	1	Kort
News items	Multiple choice	5	5	5	5	1	Kort
Pros and cons (argumentation)	Multiple choice	4	4	4	4	1	Kort
Radio/podcast (discussion)	Pairing	1	4	1	4	1	Lang
Totalt		**22**	**34**	**22**	**34**		

Rapportering av resultater

Når elevene har gjennomført en prøve, vil lærerne få tilbakemelding om elevenes resultater i form av oppnådd poengsum, et mestringsnivå fra 1 til 5 og nivå i Rammeverket, fra A2 og under til B2 og over. Tabell 6 viser sammenhengen mellom mestringsnivåer og rammeverksnivåer.

Tabell 6: Mestrings- og rammeverksnivåer

Mestrings-nivåer	**Nivåer i *Det europeiske rammeverket***
5	B2 høy og over
4	B2 lav
3	B1 høy
2	B1 lav
1	A2 og under

Standardsetting

Generelt om standardsetting

Vårt mål har vært at elever som tar de læringsstøttende prøvene i engelsk, skal få tilbakemelding om hvilket rammeverksnivå resultatet deres representerer. For å etablere valide og reliable kuttskårer som skiller mellom ulike kompetanse-nivåer, må standardsetting gjennomføres. Det innebærer å ta hensyn til både empiriske prøvedata og menneskelig vurdering av kompetanse. Når en prøve er satt sammen etter at itemene er pilotert, kjenner vi den empiriske vanskegraden som enkeltitem og prøven som helhet har. Vi vet også resultatene til de ulike elevene som har deltatt. Men for å bestemme hvor grensene mellom ulike kompetansenivåer går, må vi bruke kompetente mennesker som kan vurdere oppgavene i en prøve opp mot beskrivelser av ulike kompetansenivåer.

Kort om Angoff-, Bookmark- og Direct Consensus-metoden

Vi skal her kort referere det som kjennetegner de tre standardsettingsmetodene som er utgangspunktet for metoden vi har brukt, Cito-metoden eller The Data Driven Direct Consensus procedure. Disse metodene er Angoff-metoden, Book-mark-metoden og Direct Consensus-metoden. Alle tre metodene er testsentrerte og understreker betydningen av å ha et panel med standardsettere som kjenner godt prestasjonsstandarden som ligger til grunn for kuttskåren som skal settes. Standardsetterne må få anledning til å studere nøye den aktuelle prøven det skal settes standard for. I tillegg bør de fokusere på hva vippekandidater kan klare så vidt på en prøve. Alle metodene anbefaler at en gjennomfører flere runder med standardsetting for hver kuttskår, med innlagte gruppediskusjoner om vanske-grad mellom rundene for å oppklare misforståelser og ting som er uklare. Målet er selvsagt at panelmedlemmene skal være så enige som mulig.

I klassisk Angoff-metode (Angoff, 1971) skal panelmedlemmene vurdere hvert enkeltitem og anslå hvor stor sannsynligheten er for at vippekandidaten kan svare rett. I neste omgang kommer en fram til grensa mellom bestått og ikke-bestått, altså kuttskåren, ved å finne gjennomsnittet av standardsetternes gjennomsnittsvurderinger. I Ja/Nei-metoden, en modifisert versjon av Angoff-metoden (Impara & Plake, 1997), vurderer standardsetterne om vippekandidaten kan klare enkeltitem i en prøve ved å angi *ja* eller *nei* for hvert item. En regner deretter ut hvor mange prosent *ja* hvert medlem har svart. Deretter finner en

kuttskåren på prøven ved å finne gjennomsnittsprosenten for hele panelet. En gjennomsnittsprosent på 63 betyr at for å stå på prøven, må en prøvekandidat klare 63 % av oppgavene.

Bookmark-metoden (Mitzel, Lewis, Patz & Green, 2001; Cizek & Bunch, 2007, kapittel 10) er kanskje enklere for standardsettingspanelet å bruke enn Angoff-metoden. Metoden forutsetter at prøvedata er samlet inn og IRT-analyse er gjennomført. Panelmedlemmene får utdelt prøven som skal standardsettes, organisert etter itemenes vanskegrad (IRT), det letteste itemet først og det vanskeligste sist. Oppgaven er å sette et *bokmerke* i prøveheftet, der en mener kuttskåren bør være. Denne metoden fungerer når hvert item er binært, altså kan skåres som rett eller galt. En kan ikke uten videre bruke prøver med polytome item, for eksempel item som kan gi mer enn ett poeng.

Direct Consensus-metoden er en relativt ny metode (Sireci, Hambleton & Pitoniak, 2004; Cizek & Bunch, 2007, kapittel 7). Metoden ble foreslått for å

> improve upon some of the perceived shortcomings of the Angoff method and to give subject area experts more direct control in recommending where the passing score is set. (Sireci et al., 2004, s. 21)

Itemene i prøven blir organisert i grupper. Grunnlaget for inndelingen kan for eksempel være tema, det oppgavene måler, oppgaveformat m.m. I stedet for å vurdere enkeltitem, vurderer panelet grupper av item. Panelmedlemmene blir bedt om å indikere antall item vippekandidater trolig vil svare rett på i en delprøve. Deretter summeres antall foreslåtte rette svar hver standardsetter har angitt for hver gruppe. Kuttskåren blir gjennomsnittet av hver standardsetters summerte skår.

Standardsettere har en kompleks oppgave. De må ha en stabil oppfatning av en vippekandidats kompetanse, og en klar idé om sammenhengen mellom kompetansen/ferdighetene de vurderer, og spesifikke item. Sist, men ikke minst, må de også implisitt vurdere vanskegraden til itemene i prøven. Alle disse faktorene vil ha noe å bety for hvor stabil og reliabel standarden blir.

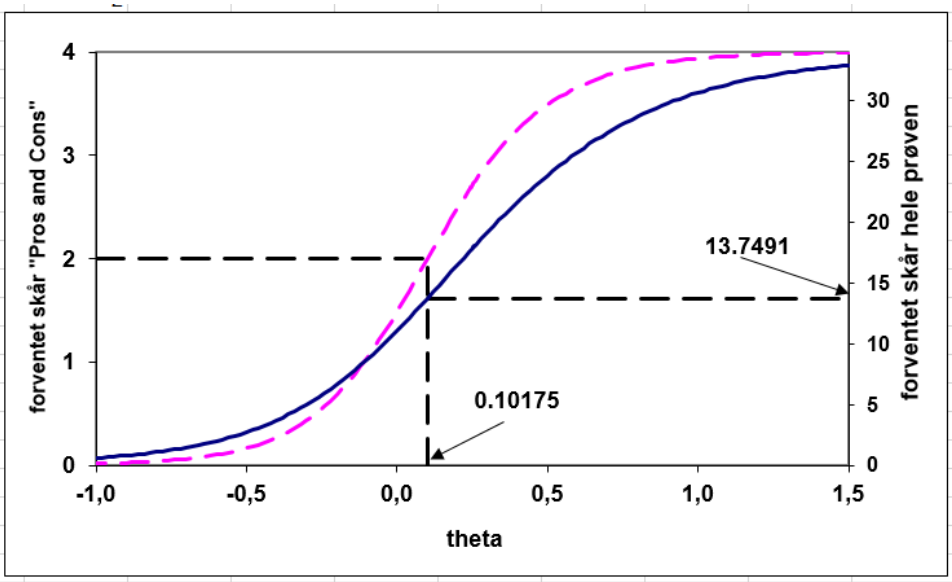
Cito-metoden (The Data Driven Direct Consensus (3DC) procedure)

Denne metoden er beskrevet av Europarådet (2009), Van der Schoot (2009) og Keuning, Straat og Feskens (2017). Metoden bygger på kjente prinsipper både fra klassisk Angoff-metode, fra Bookmark-metoden og fra Direct Consensus-metoden. I tillegg kan en si at grafisk framstilling av oppgavenes vanskegrad er helt sentral for metoden. I denne artikkelen bruker vi termen Cito-metoden når vi refererer til metoden vi har brukt. (CITO er navnet på det nederlandske instituttet som har ansvar for å utvikle prøver og måle skoleresultater.)

Angoff-metoden og Direct Consensus-metoden har blitt kritisert for ikke å informere standardsettingspanelet om hvor vanskelige itemene i en prøve er. I

begge tilfellene skal panelmedlemmene vurdere hvor sannsynlig det er at vippe-kandidater svarer rett på henholdsvis enkeltitem eller grupper av item uten at de vet vanskegraden på itemene.

I alle unidimensjonale IRT-modeller kan en konstruere en responsfunksjon for ethvert subsett av item som har blitt kalibrert sammen. Dersom ett subsett er hele testen, og en annen en delprøve med for eksempel 4–6 item, får vi et spesielt tilfelle, slik som vist i figur 3. Eksempelet er tatt fra *Lytteprøve 1* i de læringsstøttende prøvene. Den stiplete kurven viser responsfunksjonen for del-prøven *Pros and Cons* som hadde fire binære item. Den heltrukne kurven viser responsfunksjonen for hele prøven. Den forventete skåren for hele prøven vises på den vertikale aksen til høyre (full skår 34), mens forventet skår på *Pros and Cons* vises på den vertikale aksen til venstre (full skår 4). Denne spesifikke kombinasjonen av delprøve og hel prøve ble brukt til å sette kuttskår mellom B1 og B2 på *Lytteprøve 1*. Figuren viser at en forventet skår på 2 på delprøven *Pros and Cons* korresponderer til en forventet skåre på 13,7491 på hele prøven, med utgangspunkt i samme theta-verdi (0,10175).



Figur 3. Sammenheng mellom forventete skårer på hele *Lytteprøve 1* og delprøven *Pros and Cons*

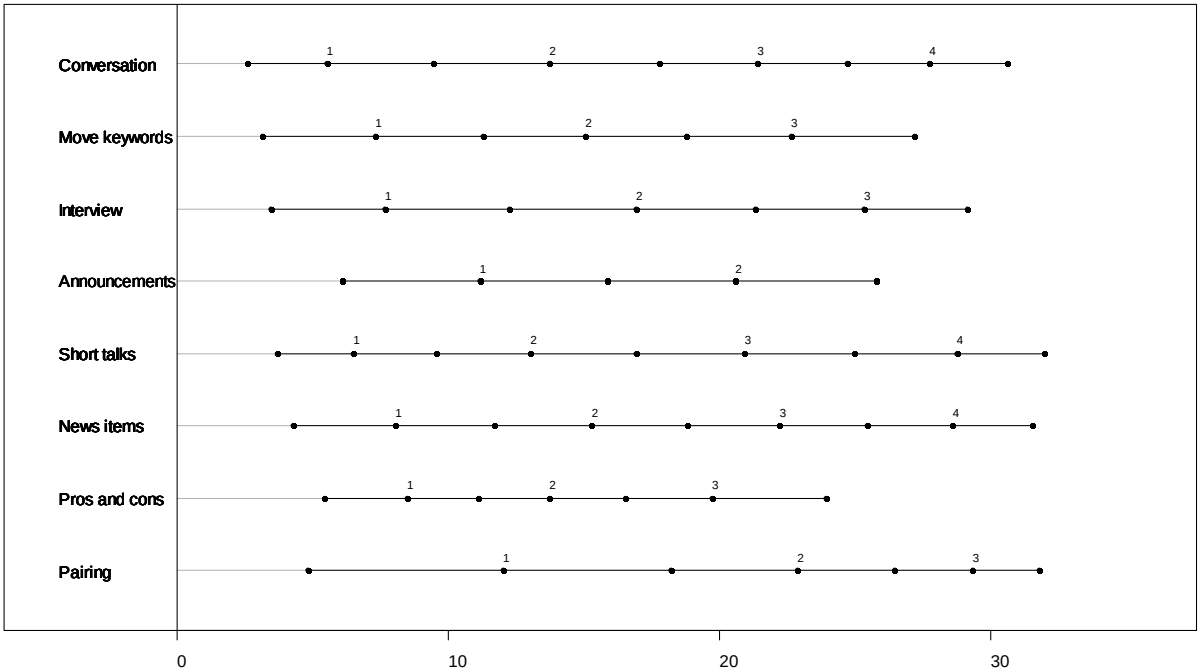
Det samme som er vist i figur 3, kan gjøres for alle subsett av en prøve. Denne sammenhengen ble brukt til å konstruere et vurderingsskjema til standard-settingspanelet, se figur 4. Navnene på delprøvene viser til sjanger eller opp-gaveformat. Se tabell 7. Det finnes ikke noen regel for hvordan slike item-grupper skal konstrueres.

Tabell 7. Delprøver i lytteprøven for Vg1

Delprøve	Maks poeng
Conversations	5
Move key words	4
Interview	4
Announcements	3
Short talks	5
News items	5
Pros and cons	4
Pairing	4

I figur 4 ser vi de samme delprøvene som i tabell 7. Skalaene viser skårer fra et halvt poeng til maks skår minus et halvt poeng. Den horisontale aksen viser skåren på hele prøven og hver av de åtte horisontale linjene representerer en delprøve. De åtte delprøvene til sammen utgjør hele prøven. Tallene langs de åtte linjene viser den nøyaktige sammenhengen mellom skår på en delprøve og på hele prøven.

Oppgaven til panelet er å markere vippekandidaters forventete skår på hver av de åtte delprøvene. Det var mulig å indikere skårer med et kvart poengs nøyaktighet. Summen av markerte forventete skårer på alle delprøvene er en individuell standard indikert av ett panelmedlem.



Figur 4. Vurderingsskjema for Lytteprøve 1 (brukt i arbeidet med å sette kuttskår B1/B2)

Det er ikke første gangen Cito-metoden blir brukt som utgangspunkt for standardsetting. Metoden har blitt brukt med hell tidligere, både i Nederland (Feskens, Keuning, van Til & Verheyen, 2014) og i Norge (Moe & Verhelst, 2017).

Standardsetting av læringsstøttende prøver for Vg1

Standardsetting ble gjennomført for prøvene i lytting, lesing og *Use of English*. Det ble satt to kuttskårer for hver prøve, en for å finne grensa mellom B1 og B2 og den andre for å finne grensa mellom A2 og B1.

Panel av standardsettere

Standardsettingen for lytteprøven hadde 16 panelmedlemmer med forskjellige bakgrunner og kunnskaper. Medlemmene kjenner Rammeverket godt. Tolv av panelmedlemmene jobber som testutviklere; fire av disse hadde vært med å utvikle lytteprøvene. De andre testutviklerne jobbet enten med å utvikle prøver i norsk for voksne innvandrere (som også er relatert til Rammeverkets nivåer) eller andre prøver i engelsk. Tolv av panelmedlemmene jobbet eller hadde jobbet som språklærere. Tabell 8 viser spredning i bakgrunn blant panelmedlemmene.

Tabell 8. Panelmedlemmenes bakgrunn

Standardsetter	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Testutvikler	x	x	x	x	x	x	x	x		x		x	x	x		
Lærer	x	x	x		x	x	x	x		x	x	x			x	x
Forsker							x	x		x					x	x
Annen bakgrunn									x		x					

Før standardsetting

Forberedende aktiviteter for standardsettere

Før selve standardsettingen begynte, var det viktig å sikre at standardsetterne kjente:

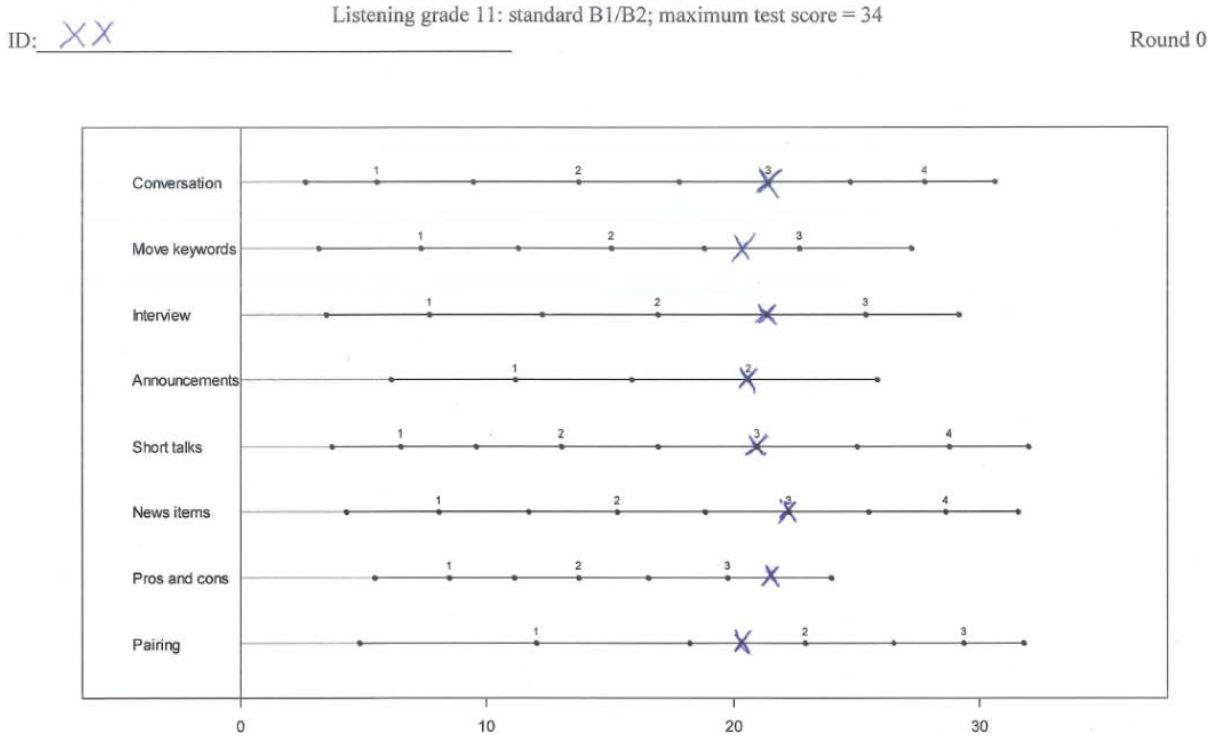
- prestasjonsskalaen som ligger til grunn for kuttskårene som skal settes, altså hva som karakteriserer rammeverksnivåene A2, B1 og B2
- prøvene det skal settes standard for
- standardsettingsprosedyrene

I utgangspunktet ble alle standardsetterne valgt ut fordi de kjente Rammeverket. Likevel var det viktig å forsikre seg om at de hadde en oppdatert oppfatning av skalaen. Seminaret startet med at panelmedlemmene måtte plassere ferdighetsbeskrivelser på ett av nivåene A2, B1 eller B2. Beskrivelsene var enten hentet fra den norske utgaven av Rammeverket (Utdanningsdirektoratet, 2011) eller fra skreddersydde mestringsnivåbeskrivelser som hadde blitt utviklet for prøvene. Til sammen måtte de plassere 18–20 beskrivelser for lytting på rammeverksnivåer. Deretter deltok de i et gruppearbeid der oppgaven var å formulere hva som kjennetegner lytteferdigheter på B1- og B2-nivå. Vi begynte med disse nivåene, siden vi først satte standard for grensa mellom B1 og B2. Deretter gikk alle panelmedlemmene gjennom prøven (på datamaskin) og svarte på spørsmålene slik at de kjente prøven de skulle sette kuttskår for.

Før standardsettingsprosedyrene startet, fikk deltakerne en introduksjon til standardsetting, inkludert en forklaring på hva som ligger i begrepene vippekandidat, gjennomsnitt, sannsynlighet på 0,5 og gjettefaktor. De fikk også en kort introduksjon til IRT-analyse og til konstruktet for lytteprøvene.

Vurderingsskjemaet

En viktig del av forberedelsen til standardsettingen var å gjøre panelmedlemmene kjent med vurderingsskjemaet som skulle brukes (se figur 4). Etter at alle hadde fått en kort forklaring på hvordan skjemaet skulle brukes, fylte de det raskt ut. Underveis kunne de få oppklart ting de ikke forstod. Det var hele tiden understreket at dette var en prøverunde som ikke talte. Figur 5 er et eksempel på hvordan skjemaet ble fylt ut av en av deltakerne.



Figur 5. Vurderingsskjema for Lytteprøve 2 (kryssvalidering B1/B2)

Gjennomføring av standardsetting

Kuttskår B1/B2

Først ble det satt kuttskår mellom B1 og B2 på *Lytteprøve 1*. Standardsetterne skulle ta for seg delprøve for delprøve og angi hvor mange oppgaver i gjennomsnitt vippekandidatene ville svare rett på.

Vi brukte to runder for å sette kuttskår på *Lytteprøve 1*. Standardsetterne hadde fått utdelt papirkopier av oppgavene, slik at de kunne gå fram og tilbake i prøven mens de vurderte. I første runde fylte de ut vurderingsskjemaet individuelt. Deretter ble de individuelle vurderingene tastet inn og presentert grafisk (anonymt) i plenum.

I andre runde ble panelmedlemmene plassert i grupper for å diskutere delprøver og hvilke spørsmål det var sannsynlig at vippekandidater kunne klare å svare rett på. Målet med dette arbeidet var ikke å bli enige om en felles standard, men å oppklare ting medlemmene kunne ha misforstått eller oversett. Etter diskusjonen i gruppa satte hver og én en individuell skår. Prosedyren gjentok seg, individuelle skårer ble tastet inn og presentert i plenum og eventuelt kommentert.

Kryssvalidering

For å sjekke hvor pålitelig kuttskåren vi satte på *Lytteprøve 1* var, brukte vi *Lytteprøve 2* (som var tilnærmet like vanskelig og diskriminerte tilnærmet likt mellom elevene) til kryssvalidering. Standardsetterne gikk gjennom *Lytteprøve 2* og svarte på oppgavene, og vi gjennomførte en runde med standardsetting.

Kuttskår A2/B1

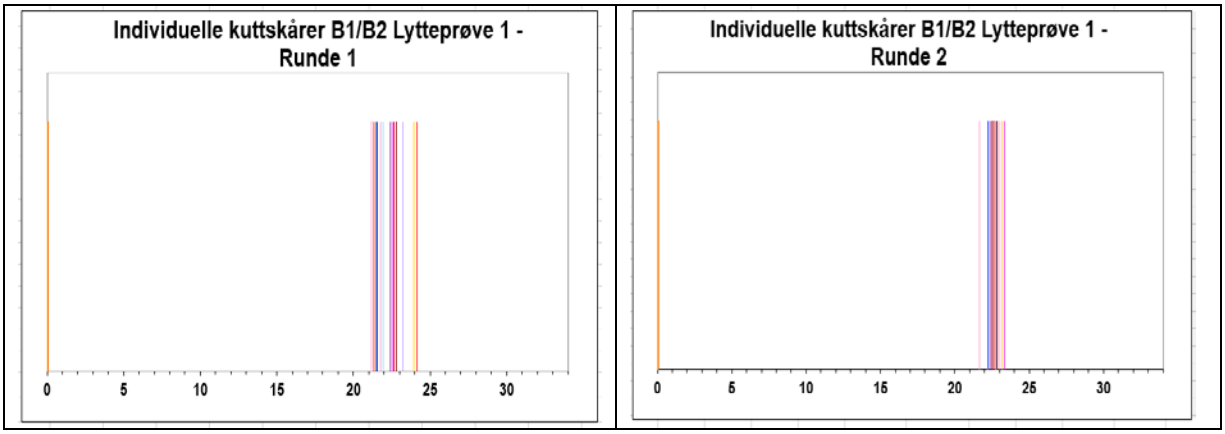
Mens standarden for B1/B2 ble satt på de reelle læringsstøttende prøvene som skolene får tilgang til, ble standarden for A2/B1 satt på en fiktiv prøve som var spesifikt konstruert for å sette denne kuttskåren. De reelle prøvene inneholdt for få item som speilet A2-nivå, derfor satte vi sammen en fiktiv, lettere lytteprøve. Itemene i den fiktive prøven var pilotert og kalibrert på samme skala som itemene i den reelle prøven. Målet var å følge de samme prosedyrene som for B1/B2-standarden. I gruppearbeidet før selve standardsettingen formulerte panelmedlemmene trekk som kjennetegner A2-ferdigheter og A2-oppgaver i lytting. På grunn av mangel på tid ble A2-nivået ikke sammenlignet med B1-nivået. Gruppene hadde jobbet med B1-nivået i forbindelse med kuttskåren for B1/B2 dagen før. Vurderingsskjemaet var selvsagt tilpasset den lette, fiktive prøven.

Selv om panelmedlemmene på seminaret sa seg fornøyd med A2/B1-standarden som ble satt på den fiktive prøven i løpet av runde 1 og 2, sa flere i ettertid at de var bekymret for at denne kuttskåren var for lav. Grunnen til dette var at arbeidet foregikk sent på dagen, og at de hadde konsentrert seg om A2-item og glemt at en vippekandidat også måtte kunne svare rett på en del B1-item. Siden det var umulig å samle hele standardsettingspanelet igjen, valgte prøveutviklerne å gjenta standardsettingen for A2/B1 den påfølgende uka med en mindre gruppe (åtte personer). Alle personene som deltok i andre omgang i runde 3 og 4, hadde også vært med i første. Endelig kuttskår ble basert på resultatet av runde 4 fordi prøveutviklerne oppfattet at panelmedlemmene i større grad enn i runde 1 og 2 vurderte både A1- og A2-ferdigheter da de foreslo kuttskår.

Resultater

Kuttskår for B1/B2 – B2-standard

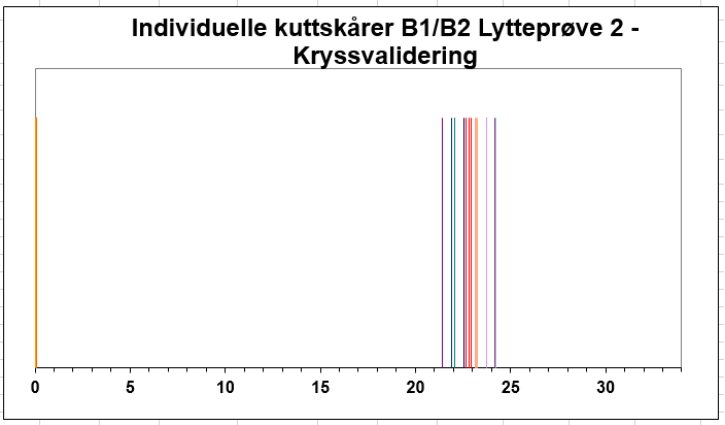
Figur 6 viser de individuelle vurderingene fra standardsettingsrunde 1 og 2. Vi ser at panelmedlemmene var mer enige i runde 2 enn i runde 1. Etter runde 2 ble panelmedlemmene enige om at resultatet fra runde 2 var tilfredsstillende, og derfor ble det ikke gjennomført en tredje runde med individuell vurdering.



Figur 6. Runde 1 og 2 for kuttskår mellom B1 og B2

Standardavviket gikk ned fra 0,94 til 0,37 og målefeilen fra 0,24 til 0,10.

For å sjekke hvor stabilt standardsetterne vurderte når de foreslo kuttskår, gikk de gjennom *Lytteprøve 2* og foreslo kuttskår mellom B1 og B2 for denne prøven også. Figur 7 viser de individuelle forslagene til kuttskår på *Lytteprøve 2*.



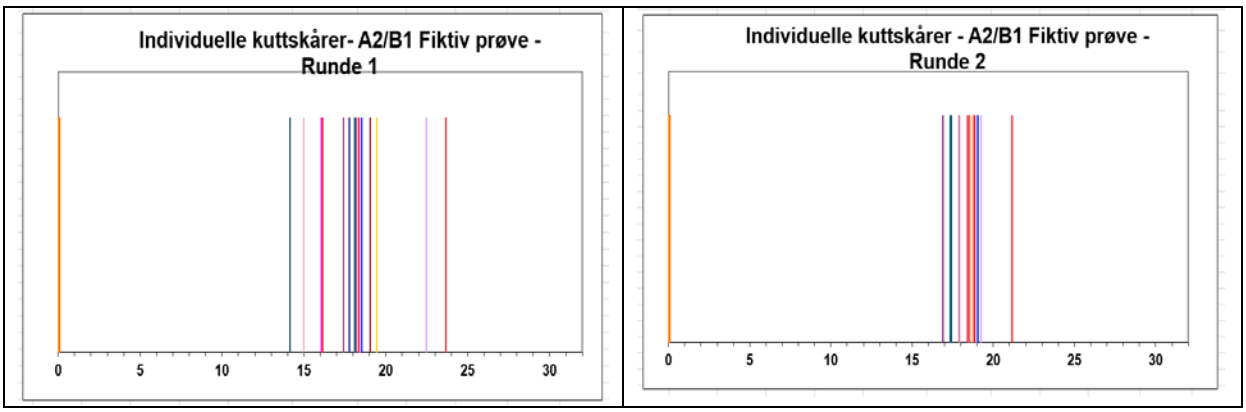
Figur 7. Kryssvalidering for kuttskår B1/B2 med *Lytteprøve 2*

Kuttskåren vi satte for *Lytteprøve 1*, var 22,65, mens kuttskåren for *Lytteprøve 2* ble 22,89. I praksis vil dette bety at elever som får 22 poeng (av 34) på *Lytteprøve 1* og *2*, blir plassert på nivå B1, mens 23 rette svar gir B2. Samlet oversikt over resultatene finnes i tabell 10.

Kuttskår for A2/B1 – B1-standard

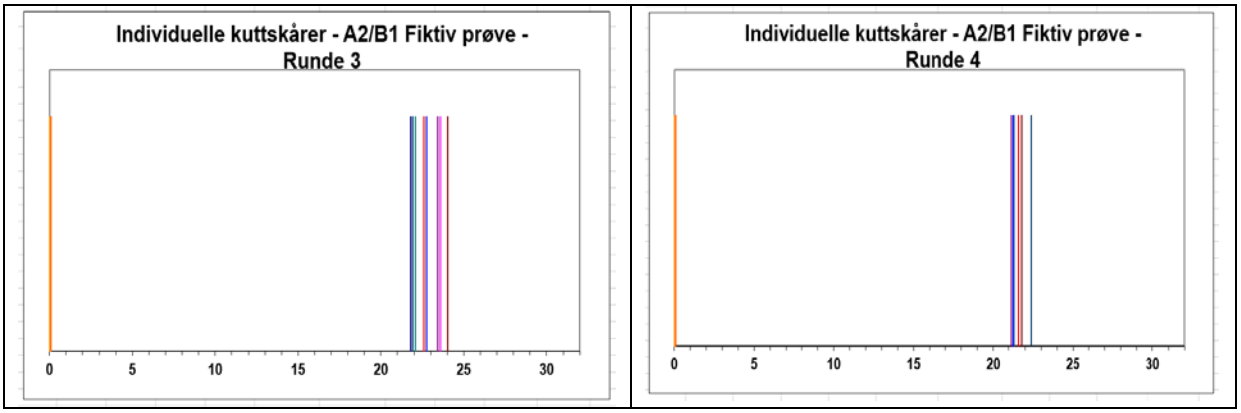
Som nevnt i *Gjennomføring av standardsetting / Kuttskår A2/B1*, ble kuttskåren mellom A2 og B1 satt på en lettere, fiktiv prøve som var spesielt satt sammen

for standardsettingen. Standardsettingen ble gjennomført i fire runder, to runder på selve seminaret og to uka etter. Figur 8 viser resultatene fra de to første rundene.



Figur 8. Runde 1 og 2 for kuttskår mellom A2 og B1 (fiktiv prøve)

Figur 9 viser hvordan standardsetterne fordelte seg på poengskalaen i runde 3 og runde 4.



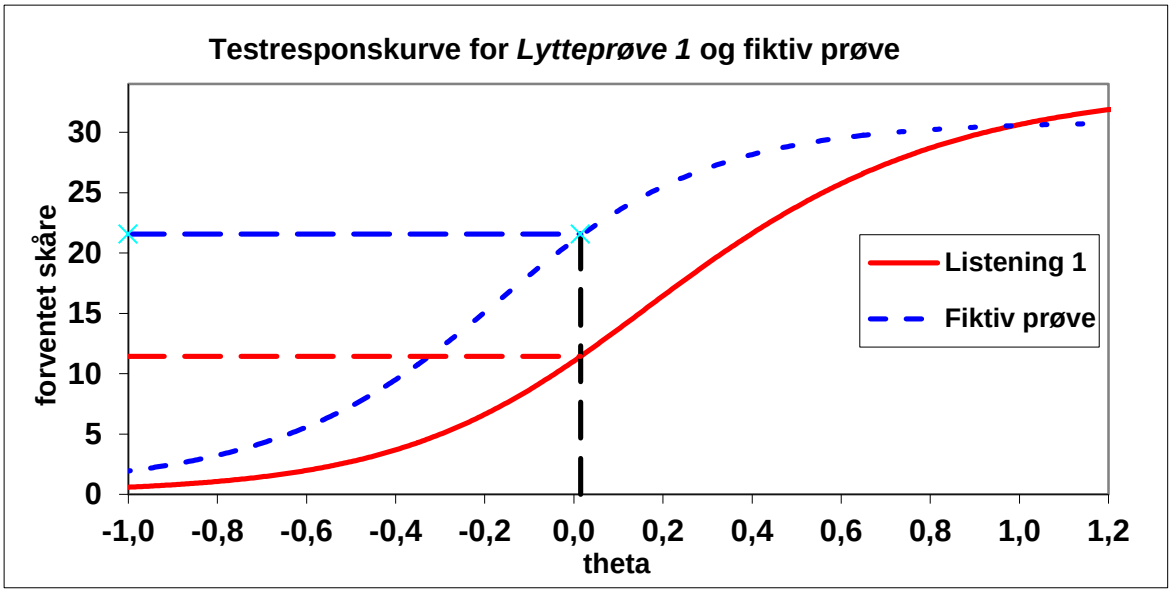
Figur 9. Runde 3 og 4 for kuttskår mellom A2 og B1 (fiktiv prøve)

Tabell 9 viser resultatene fra de fire omgangene med standardsetting som ble gjennomført for A2/B1-standarden for lytting.

Tabell 9. Standardsetting lytting A2/B1

Runde	Antall standardsettere	Gjennomsnittlig kuttskår	Standardavvik	Målefeil
1	16	18,28	2,35	0,59
2	16	18,56	0,95	0,24
3	8	22,75	0,78	0,28
4	8	21,56	0,41	0,14

Standarden for A2/B1 ble satt basert på den fiktive prøven og deretter overført til den reelle prøven. Figur 10 viser *Lytteprøve 1* (den heltrukne linja) og den fiktive prøven (den stiplete linja).



Figur 10. Responskurve for *Lytteprøve 1* (Listening 1) og fiktiv prøve

Gjennomsnittlig theta for kuttskåren A2/B1 var 0,020. På den fiktive prøven tilsvarte dette 21,56 poeng, mens når standarden ble overført til *Lytteprøve 1*, svarte theta lik 0,020 til 11,56 poeng. Oversikt over alle kuttskårene finnes i tabell 11.

Resultater for prøver i lesing, lytting og Use of English

I denne artikkelen har vi beskrevet gjennomføring og resultater av standard-setting for læringsstøttende prøver i lytting for Vg1. Det ble også utviklet læringsstøttende prøver i lesing og *Use of English*. Disse prøvene ble det gjennomført standardsetting for i 2017, på samme måte som for lytteprøvene. Det ble imidlertid ikke gjennomført kryssvalidering for prøvene i lesing og *Use of English*. Tabell 11 viser resultatene av standardsettingsprosedyrene som ble gjennomført for alle tre ferdigheter: lytting, lesing og *Use of English*.

Tabell 10. Kuttskårer for lesing, *Use of English* og lytting

Test	Kuttskår	Standard fiktiv prøve	Standard-avvik	Målefeil	Theta (θ)
Lesing B1/B2	21,07	–	1,51	0,34	0,322
Lesing A2/B1	8,76*	18,92	0,83	0,31	
***Use of English* B1/B2**	26,58	–	1,28	0,31	0,167
***Use of English* A2/B1**	11,21*	17,93	0,39	0,15	–0,252
***Lytteprøve 1* – B1/B2**	22,65	–	0,37	0,10	0,445
***Lytteprøve 2* – B1/B2 kryssvalidering**	22,89	–	0,82	0,21	0,450
Lytting A2/B1	11,56*	21,56	0,41	0,14	0,020

* Kuttskåren er overført fra den fiktive prøven.

Tabell 10 viser kuttskårene standardsettingspanelet kom fram til på seminarene i 2017 og 2018. Mange av de samme standardsetterne deltok i både 2017 og 2018. Når det gjelder kuttskårene for B1/B2, kan det se ut som 2018-seminaret der vi satte kuttskår for lytteprøvene, var mest vellykket i og med at både

standardavviket og målefeilen er betydelig lavere enn for leseprøvene og prøvene i *Use of English*. En grunn til dette kan være at standardsetterne var mer forberedt på hva de skulle gjøre i 2018 siden de hadde vært med på det samme før. Tabell 10 viser at standardavviket og målefeilen for kuttskåren for B1/B2 på leseprøven er høyere enn for de to andre ferdighetene. Det vil være viktig å følge resultatene ved de neste avviklingene av læringsstøttende prøver. Kanskje burde vi gjenta standardsetting for disse prøvene når det har gått en tid?

Tabell 11 viser standarder overført til poeng på prøvene. Tabellen viser at en elev som får 11 poeng på en av lytteprøvene, blir plassert på A2, mens en som får 12 poeng blir plassert på B1. Vi ser også at en elev som får 22 poeng på en lytteprøve, blir plassert på nivå B1, mens en som får 23 poeng, blir plassert på nivå B2.

Tabell 11. B1- og B2-standarder for prøvene i lytting, lesing og *Use of English*

Nivåer i *Det europeiske rammeverket*	***Lytteprøve 1* og *2*** (max skår: 34)	***Leseprøve 1* og *2*** (max skår: 34)	***Use of English 1* og *2*** (max skår: 41)
B1/B2	22/23	21/22	26/27
A2/B1	11/12	8/9	11/12

Med utgangspunkt i tallene i tabell 11 kan elevgruppa deles i tre nivåer: 1) A2 og under, 2) B1, og 3) B2 og over. Vi ender dermed opp med fem nivåer, se tabell 12.

Tabell 12. Poeng for hvert mestringsnivå

Mestrings-nivå	**Nivå i *Det europeiske rammeverket***	***Lytteprøve 1* og *2* Antall rette svar (poeng)**	***Leseprøve 1* og *2* Antall rette svar (poeng)**	***Use of English 1* og *2* Antall rette svar (poeng)**
5	B2 og over	29 – 34	29 – 34	35 – 41
4	B2 lav	23 – 28	22 – 28	27 – 34
3	B1	18 – 22	16 – 21	20 – 26
2	B1 lav	12 – 17	9 – 15	12 – 19
1	A2 og under	0 – 11	0 – 8	0 – 11

Det siste punktet vi ville undersøke, var hvilke konsekvenser kuttskårene ville ha på fordelingen av elever på rammeverksnivåer. Vi sa tidligere i artikkelen at både personers kompetanse og item kan plasseres på theta-skalaen. Hvis vi tar utgangspunkt i 1) pilotdata og pilotkandidatenes plassering på thetaskalaen, og 2) kuttskårer, kan vi sortere pilotkandidatene i tre grupper: A2 og under, B1, og B2 og over. Da får vi en prediksjon på hvor mange prosent av elevene som havner i de ulike gruppene. Se tabell 13.

Tabell 13. Predikert fordeling av elever på rammeverksnivåer

Rammeverksnivå	**Prosent elever**
B2 og over	31,2
B1	51,9
A2 og under	16,9

Sammenfatning og diskusjon

I denne studien har oppgaven vært å lenke resultatene på de læringsstøttende prøvene i engelsk for Vg1 til Rammeverket ved å etablere kuttskårer mellom rammeverksnivåene A2/B1 og mellom B1/B2 ved hjelp av Cito-metoden. I utgangspunktet formulerte vi tre mål for studien, og i punktene under diskuterer vi resultatene kort.

Er det mulig å etablere reliable kuttskårer ved hjelp av Cito-metoden?
Svaret på dette spørsmålet er *Ja*. Selv om panelmedlemmene var plukket ut nettopp fordi de hadde god kjennskap til Rammeverket, så hadde de en vanskelig oppgave. De skulle gjøre seg opp en mening om grupper av item eller delprøver i forhold til rammeverksnivåer. Dette skulle sikre sammenhengen mellom rammeverksnivåene og delprøvene. Det virket som om panelmedlemmene forstod hva de skulle gjøre, og at de fokuserte på det de mente vippekandidater kunne klare. Etter første individuelle runde med standardsetting ble standardsetterne plassert i heterogene grupper basert på vurderingene de hadde gjort. Antakeligvis la dette grunnlaget for diskusjon i gruppene fordi panelmedlemmene hadde gode innspill i diskusjonene, som foregikk både i grupper og i plenum. Figurene 6, 8 og 9 viser at gruppediskusjonene gjorde at panelmedlemmene ble mer enige om kuttskårene.

Tabell 10 viser at standardavviket og målefeilen i de fleste tilfeller er mindre i 2018 enn i 2017. Det kan bety at panelmedlemmene var mer forberedt på hva de skulle gjøre i 2018, og at det kan være en grunn til større enighet.

Ble det brukt nok tid til å forberede standardsetterne?
I omtalen av mange standardsettingsmetoder blir det understreket at det er viktig at de som deltar i standardsetting, får en opplæring på forhånd som forbereder dem på oppgaven de skal gjøre, og at det er viktig å bruke tid på slik opplæring. Det kan være vanskelig å vite hvilke aktiviteter som er mest effektive som forberedelse til standardsetting. På seminaret begynte vi med å be panelmedlemmene plassere ferdighetsbeskrivelser på rammeverksnivåene A2, B1 og B2. Deretter var det et gruppearbeid der gruppene skulle formulere kjennetegn ved a) lytting på B1- og B2-nivåene, og b) lytteoppgaver som speilet ferdigheter på disse nivåene. Standardsettingsprosedyrene vi senere gjennomførte for å etablere kuttskår for B1/B2, resulterte i kuttskår med relativt lavt standardavvik og lav målefeil. Da vi dagen etter skulle sette kuttskår for A2/B1, var det seint på dagen og mindre tid til gruppearbeid. Panelmedlemmene fikk litt tid til å formulere kjennetegn ved lytting på A2-nivå og lytteoppgaver som representerer ferdigheter på A2. Dette ble imidlertid ikke sett i forhold til ferdigheter/oppgaver på B1-nivå fordi vi dagen før hadde fokusert på B1-nivået, og fordi tiden var knapp.

Da flere av standardsetterne i ettertid uttrykte bekymring for om A2/B1-kuttskåren som ble satt for lytting var riktig, kan det være et resultat av at det rett og slett ble brukt for lite tid til å fokusere på A2- og B1-nivå og oppgaver på seminaret. Avdøde Felianka Kaftandjieva pleide å si om det å forberede standardsettere: «Train them until it hurts» (personlig kommunikasjon). Akkurat dette er både et tids- og et pengespørsmål. Målet er i alle fall å gjøre standard-settingsprosedyrene så effektive som mulig, og det er all grunn til å reflektere over om noe kunne ha blitt gjort annerledes i forberedelsesfasen.

Hvilke konsekvenser har kuttskårene for fordeling av elever på rammeverksnivåer?

Piloteringsdata predikerer at 16,9 % av elevene vil bli plassert på nivå A2 eller under på Rammeverkets skala, 51,9 % på B1, og 31,2 % på B2 eller over (se tabell 13). Det vil være interessant å sjekke om denne prediksjonen holder stikk i framtiden når elever har tatt de ferdige prøvene. Ofte viser det seg at oppgaver er lettere i reelle prøver enn ved pilotering. Dette er selvsagt fordi elever anstrenger seg mer når de tar en reell prøve enn det de gjør ved pilotering. Selv om lærere som kjenner elevgruppa godt, sier at en slik fordeling stemmer godt med deres vurdering av elevgruppas ferdigheter i lytting, er prediksjonen avhengig av flere faktorer. En ting som kan påvirke fordelingen av elevresultater på rammeverkskalaen, er tidspunktet for prøvene. Oppgavene ble pilotert høsten 2017, mens prøvene kan tas hele skoleåret. Vi vet ikke om data vil komme til å vise et høyere gjennomsnittsresultat på prøver tatt om våren enn de som blir tatt om høsten.

I tiden som kommer, er det viktig at vi følger nøye med på om prøve-resultater og kuttskårer fungerer slik de skal. En viktig årsak til dette er at standardsettingen var basert på piloteringsdata. For å sikre reliable kuttskårer kan det være lurt å gjenta standardsettingsprosedyrene etter en tid, og da med reelle prøvedata som utgangspunkt.

Om forfatterne

Eli Moe er universitetslektor i engelsk ved Universitet i Bergen. Hennes forsk-ningsinteresser omfatter blant annet språktesting av barn og voksne innvandrere, validering av språktester, standardsetting og skolespråk.
Institusjonstilknytning: Institutt for fremmedspråk, Universitetet i Bergen, Post-boks 7805, 5020 Bergen.
E-post: eli.moe@uib.no

Hildegunn Lahlum Helness er universitetslektor i engelsk ved Universitet i Bergen. Hennes forskningsinteresser omfatter blant annet språktesting og digitale prøver.

Institusjonstilknytning: Institutt for fremmedspråk, Universitetet i Bergen, Postboks 7805, 5020 Bergen.
E-post: hildegunn.helness@uib.no

Craig Grocott er universitetslektor i engelsk ved Universitet i Bergen. Hans forskningsinteresser omfatter blant annet språktesting, sosiolingvistikk og historisk lingvistikk.
Institusjonstilknytning: Institutt for fremmedspråk, Universitetet i Bergen, Postboks 7805, 5020 Bergen.
E-post: craig.grocott@uib.no

Norman Verhelst er ansatt ved Eurometrics. Hans forskningsinteresser omfatter blant annet psykometri, kvantitativ metode og dataanalyse, flernivåmodellering, skalautvikling og -konstruksjon, IRT-teori, standardsetting.
Institusjonstilknytning: Eurometrics, Stationsstraat 31, 4001 CD, Tiel, Nederland.
E-post: norman.verhelst@gmail.com

Referanser

Angoff, W. H. (1971). Scales, norms and equivalent scales. I R. L. Thorndike (red.), *Educational measurement* (2. utg.) (s. 508–600). Washington DC: American Council on Education.

Cambridge Assessment (2018). *B2 First*. Lastet ned 22.05.2018: http://www.cambridgeenglish.org/exams-and-tests/first/

Cizek, G. J. & Bunch, M. B. (2007). *Standard Setting*. Thousand Oaks: Sage.

ETS (2018). *The Toefl test*. Lastet ned 22.05.2018: https://www.ets.org/toefl/

Europarådet (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR): a manual*. Strasbourg: Europarådet.

Europarådet (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press. Lastet ned 20.05.2018: www.coe.int/t/dg4/linguistic/source/framework_en.pdf

Europarådet (1975). *Threshold Level*. Council of Europe Publishing.

Feskens, R., Keuning, J., van Til, A. & Verheyen, R. (2014). *Performance Standards for the CEFR in Dutch secondary education. An international standard setting study*. Arnhem, Nederland: CITO.

Folkeuniversitetet (2018). *Test i norsk – høyere nivå*. Lastet ned 22.05.2018: https://www.folkeuniversitetet.no/Artikler/Spraaktester/Test-i-norsk-hoeyere-nivaa

Hasselgreen, A. (2003). *The Bergen 'Can Do' Project*. ECML, Graz. Lastet ned 20.05.2018: https://www.ecml.at/Resources/ECMLPublications/tabid/277/ID/67/language/en-GB/Default.aspx og http://archive.ecml.at/cando/

Hasselgreen, A. (2010). *The Ayllit project*. ECML,Graz. Lastet ned 20.05.18: http://ayllit.ecml.at/

IELTS (2018). *IELTS Academic*. Lastet ned 22.05.2018: https://www.ielts.org/about-the-test/two-types-of-ielts-test

Impara, J. C. & Plake, B. S. (1997). Standard setting. An alternative approach. *Journal of Educational Measurement*, *34*, 353–366.

Jaeger, R. M. (1989). Certification of student competence. I R. L. Linn (red.), *Educational Measurement* (Third Edition) (s. 485–511). Washington, DC: American Council on Education.

Kaftandjieva, F. (2004). Standard setting. I *Reference supplement to the Manual for relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment* (Section B). Strasbourg: Council of Europe, Language Policy Division . Lastet ned 21.05.2018: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a1d

Keuning, J., Straat, J. H. R. & Feskens, C. W. (2017). The Data-Driven Direct Consensus (3DC) Procedure: A New Approach to Standard Setting. I S.Blömeke & J. E. Gustafsson (red.), *Standard Setting in Education. The Nordic Countries in an International Perspective* (s. 263–278). Cham, Switzerland: Springer International Publishing.

Kompetanse Norge (2018). *Norskprøve*. https://www.kompetansenorge.no/norskprove/

Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001). The bookmark procedure: psychological perspectives. I G. J. Cizek (red.), *Setting performance standards: concepts, methods and perspectives* (s.249–281). Mahwah, NJ: Lawrence Erlbaum.

Moe, E. & Verhelst, N. (2017). Setting standards for Multistage Tests of Norwegian for Adult Immigrants. I S. Blömeke & J. E. Gustafsson (red.), *Standard Setting in Education. The Nordic Countries in an International Perspective* (s. 181–204). Cham, Sveits: Springer.

Moe, E., Härmälä, M., Kristmanson, P. L., Pascoal, J. & Ramoniene, M. (2015). *Language skills for successful subject learning. CEFR linked descriptors for mathematics and history/civics*. Strasbourg: Council of Europe publishing. Lastet ned 20.05.2018: https://www.ecml.at/ECML-Programme/Programme2012-2015/LanguageDescriptors/tabid/1800/Default.aspx

Sireci, S. G., Hambleton, R. K. & Pitoniak. M. J. (2004). Setting passing scores on licensure examinations using direct consensus. *CLEAR Exam Revue*, *15*(1), 21–25.

Svanes, B., Hagen, J. E., Manne, G., Svindland, A. S. & Husby, O. (1988). *Et terskelnivå for norsk*. J. W. Cappelen Forlag.

Utdanningsdirektoratet (2015). *Bilag til: Avtale med Universitetet i Bergen om å utvikle læringsstøttende prøver i engelsk for Vg1*.

Utdanningsdirektoratet (2013). *Læreplan i engelsk* (ENG1-03). Lastet ned 14.09.2018: https://www.udir.no/kl06/ENG1-03/

Utdanningsdirektoratet (2011). *Det felles europeiske rammeverket for språk: Læring. Undervisning. Vurdering*. Oslo: Utdanningsdirektoratet.

Van der Schoot, F. (2009). Cito variation of the bookmark method. I Council of Europe, *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Section I). Strasbourg: Council of Europe. Lastet ned 27.11.2018: https://rm.coe.int/1680667a24

Vox (2012). *Læreplan i norsk og samfunnskunnskap for voksne innvandrere*. Oslo: VOX, Nasjonalt fagorgan for kompetansepolitikk. Lastet ned 27.11.2018: https://www.kompetansenorge.no/contentassets/f6594d5dde814b7bb5e9d2f4564ac134/laereplan_norsk_samfunnskunnskap_bm_web.pdf