Tor Midtbø
Kompetanse Norge
Arne Rossow
Kompetanse Norge
Brikt Sagbakken
Kompetanse Norge

DOI: http://dx.doi.org/10.5617/adno.6358

Måling av sensorreliabilitet ved vurdering av norskprøve i skriftlig framstilling

Sammendrag
Sensorer vurderer skriftlige tekster ulikt, og menneskelig sensur er en utfordring for prøvers reliabilitet. Dette er en utfordring som Kompetanse Norge må ta høyde for i arbeidet med å utvikle og kvalitetssikre Norskprøven for voksne innvandrere. Denne artikkelen redegjør for hvordan den statistiske modellen Many-Facets Rasch Measurement (MFRM) er brukt til å undersøke sensorkorpsets reliabilitet ved sensurering av Norskprøvens delprøve i skriftlig framstilling for desemberavviklingen 2017. MFRM-modellen gir oss informasjon om hvor streng og pålitelig hver sensor er i vurderingen av kandidatbesvarelser. Analysen viser at det er klare forskjeller i strenghet innad i sensorkorpset, og at kandidatens endelige resultat kan være påvirket av hvilke sensorer som vurderer besvarelsen. Samtidig finner vi at de fleste av de 77 sensorene sensurerer stabilt og pålitelig, som vil si at de har høy intra-sensorreliabilitet. Dette viser at sensorkorpset i stor grad oppfyller målsetningen om sensorer som uavhengige eksperter med konsekvent vurderingsadferd. Avslutningsvis diskuteres utfordringene knyttet til begrensninger ved prøvens utforming for analyse av sensorreliabilitet. I lys av diskusjonen vurderer vi MFRM sin rolle og egnethet, og peker på noen utviklingsområder.

Nøkkelord: norskprøve, skriftlig vurdering, reliabilitet, inter-sensorreliabilitet, intra-sensorreliabilitet, Many-Facet Rasch Measurement

Norwegian language test - Measuring rater reliability in the assessment of written presentation

Abstract
Raters assess written texts differently, and rater-mediated assessment is a challenge for test reliability. This is something Skills Norway has to take into consideration as test developer of the Norwegian test for adult immigrants. In this article, we demonstrate how the statistical model Many-Facets Rasch Measurement (MFRM) has been used to examine rater reliability in the written part of the test, using data from the December 2017 test. The MFRM model produces estimates on all raters in terms of severity and consistency. The results show large and significant variation in severity among the raters, and the candidates' final results can be affected by which raters have assessed the test. Nevertheless, we find that most of the 77 raters assess consistently, showing high intra-rater reliability. This finding suggests that the raters, to a large degree, fulfil their role as independent experts with consistent rating behaviour. Finally, we discuss the challenges associated with the limitations of the test's design, with respect to analysing rater reliability. We assess MFRM's role and suitability, and identify possible areas of future study.

Keywords: language testing, written assessment, rater-mediated assessment, inter-rater reliability, intra-rater reliability, Many-Facet Rasch Measurement

Introduksjon

Kompetanse Norge (tidligere Vox) har siden 2014 fått i oppdrag fra Kunnskapsdepartementet å ha ansvaret for å utvikle, administrere og legge til rette for gjennomføring av avsluttende prøver i norsk og samfunnskunnskap for voksne innvandrere, samt Statsborgerprøven siden 1. januar 2017. Kompetanse Norge har derfor administrert og sørget for sensur av over 60 000 delprøver i skriftlig framstilling siden 2014 (Kompetanse Norge 2018).

Siden resultatet på prøvene kan brukes til å oppfylle krav ved søknad om permanent opphold, statsborgerskap og høyere utdanning, blir prøvene ansett som «high-stakes», og Kompetanse Norge jobber kontinuerlig med å kvalitetssikre prøvene. Det er viktig at prøvene som Kompetanse Norge administrerer, er valide og pålitelige fordi de fungerer som instrument med sosiale og juridiske implikasjoner for kandidatene som tar prøvene (McNamara & Ryan, 2011). Den overordnede problemstillingen enhver testutvikler må kunne svare på, er derfor hvordan man kan være sikker på at resultatet en kandidat har oppnådd, er representativt for kandidatens faktiske evnenivå.

Formål og forskningsspørsmål

Med denne studien ønsker vi å gi leseren innblikk i reliabilitetsutfordringen ved sensurering av skriftlige langsvarsoppgaver. Videre er formålet å anvende den statistiske modellen Many-Facet Rasch Measurment (MRFM) til å diagnostisere reliabiliteten til sensorkorpset som vurderer besvarelser i Norskprøvens delprøve i skriftlig framstilling. Til slutt ønsker vi å diskutere MRFM sin egnethet og bruksområde sett i lys av Norskprøvens utforming og praktiske begrensinger.

Artikkelen har to forskningsspørsmål:

1. Hvor reliabelt vurderer sensorkorpset til Norskprøvens delprøve i skriftlig framstilling?
2. Hvor egnet er MFRM til å analysere sensorreliabilitet, og hvilken rolle kan modellen ha sett i lys av Norskprøvens utforming og praktiske begrensinger?

Dette bidraget handler om hvordan vi kan måle reliabilitet i sensorkorpset som vurderer besvarelser i skriftlig framstilling. Bidraget prøver ikke å gi et fullstendig bilde av Kompetanse Norge sitt arbeid med å utvikle og kvalitetssikre Norskprøven. Informasjon om hvordan sensoropplæringen foregår, rekruttering av sensorer og dialog tilknyttet vurderingsarbeidet er ikke omtalt. Bidraget vil heller ikke gå inn i dybden på de kognitive og språkfaglige aspektene bak vurderingsprosessen.

Før resultatet av analysen presenteres, beskriver vi hvordan Norskprøven er bygd opp, samt utfordringene med menneskelige sensorer for vurdering av skriftlige tekster. Deretter introduserer vi MFRM-modellen, som lar oss estimere graden av sensorstrenghet til hver sensor, og kan avdekke hvorvidt sensorene vurderer pålitelig og stabilt. Videre beskriver vi hvordan datamaterialet behandles før resultater blir presentert. Til slutt diskuterer vi utfordringer med valgt metode og datagrunnlag.

Norskprøven

Avsluttende prøve i norsk for voksne innvandrere, eller Norskprøven, er en obligatorisk avsluttende prøve for voksne innvandrere som er omfattet av Introduksjonsordningen, definert i Introduksjonsloven av 2005. I tillegg er prøven tilgjengelig for personer som ønsker å dokumentere norskferdighetene, også kalt privatister. Prøven har forskjellig funksjon for kandidater. Noen tar den som et mål på hva de har oppnådd i løpet av norskopplæringen, og andre tar den kun for å dokumentere språkferdighet. Selv om kandidater tar Norskprøven av forskjellige årsaker, anses den som «high-stakes», ettersom den blir brukt til å oppfylle krav til søknad om permanent opphold, og for å kunne søke opptak til høyere utdanning.

Norskprøven består av fire separate delprøver som måler ferdighetene lytteforståelse, leseforståelse, skriftlig framstilling og muntlig kommunikasjon. Resultatene på delprøvene slås ikke sammen, noe som gir mulighet for en differensiert språkprofil med ulik måloppnåelse i de fire ferdighetene. Det er per i dag seks mulige resultat på Norskprøven: B2, B1, A2, A1, Under A1 og Ikke grunnlag for vurdering. Karakterskalaen som blir brukt på prøven, er den samme som rammeverksnivåene i Det felles europeiske rammeverket for språk (Council of Europe 2011), og Læreplan i norsk og samfunnskunnskap for voksne innvandrere (Vox 2012). Hvert rammeverksnivå representerer et relativt stort ferdighetsområde. En kandidat som så vidt blir vurdert til å være på B1-nivå, og en som nesten er på B2, men har B1 på et av kriteriene, vil ha klart ulik norskferdighet, men begge vil få B1 som resultat på delprøven.

Delprøven i skriftlig framstilling

Delprøven i skriftlig framstilling er delt inn i tre prøvenivåer: A1–A2, A2–B1, og B1–B2, hvor A1–A2 er laveste nivå og B1–B2 er høyeste nivå. Når kandidatene melder seg opp, velger de hvilket nivå de går opp til. Prøven besvares digitalt på datamaskinen, hvor kandidatene skriver tekstene rett inn i prøvesystemet. Fordeling av kandidatbesvarelser til sensorene skjer etter en tilfeldig fordeling. Hver kandidatbesvarelse blir vurdert av to sensorer helt uavhengig av hverandre. Hvis de har ulik vurdering, går besvarelsen til en tredje sensor, også kalt oppmann, som tar den endelige avgjørelsen. Dette skjer i omtrent ett av fire tilfeller.

En kandidat kan oppnå alle nivåer opp til høyeste nivå på den oppmeldte prøven. Det vil si at en kandidat meldt opp til A2–B1, kan oppnå alt fra Under A1 til B1, men ikke B2, og en kandidat meldt opp til B1–B2, kan få Under A1 opp til B2. Kandidatbesvarelsene blir vurdert av et sensorkorps med norsklærere hovedsakelig fra voksenopplæringen. Det er et krav til alle sensorer å delta på en årlig sensorsamling for å kunne sensurere det kommende året. På sensorsamlingen får sensorene se analyser fra foregående prøveperiode med informasjon om hvor strengt og reliabelt de vurderer. I forkant av samlingene skal alle sensorene vurdere fellestekster som blir gjennomgått og diskutert på samlingen. Det er også lagt opp til fellestekster for prøveperiodene, som ikke er tilknyttet en sensorsamling hvor tilbakemeldinger skjer via mail. Fordi lærerne underviser på ulike norsknivåer, sensurerer de også på ulike nivåer. Kompetanse Norge har to separate sensorkorps delt etter hvilket nivå lærerne underviser på. Det ene sensorkorpset vurderer B1–B2-nivået, og det andre vurderer både A1–A2- og A2–B1-nivåene.

I utformingen av Norskprøven er målet at kandidatene kun skal bli vurdert basert på deres ferdigheter i norsk. Det teoretiske konstruktet for delprøven i skriftlig framstilling er kommunikativ skriveferdighet basert på modellen for kommunikativ kompetanse, definert av Bachmann (1990) og Bachman og

Palmer (1996). Tabell 1 viser hva prøven skal måle på de ulike prøvenivåene, og hvordan de blir operasjonalisert i oppgaver.

De tre prøvenivåene består av to eller tre oppgaver som er ment å måle på forskjellige språknivåer. Det er overlappende oppgaver mellom de forskjellige prøvenivåene som lenker prøveversjonene på tvers av nivå. Oppgave 2 og 3 fra A1–A2 går igjen i A2–B1 som oppgave 1 og 2, og oppgave 3 fra A2–B1 er oppgave 1 i B1–B2. Dette sikrer overlapp mellom alle de tre prøvenivåene. På grunn av at prøvegjennomføringen foregår over en lengre periode på 10–14 dager, opereres det med minst tre forskjellige prøveversjoner på hvert prøvenivå. Dette er for å redusere risikoen forbundet med at oppgavene blir kjent, og at kandidatene leverer tidligere produserte tekster.

Tabell 1: Oversikt over hva prøven skal måle og tilhørende oppgaver per prøvenivå

Prøvenivå	Prøven skal måle kandidatens evne til å	Oppgaver
A1–A2	Skrive en enkel melding (A1) Beskrive et bilde (A1, A2 og B1) Fortelle om et kjent tema (A1, A2 og B1)	Oppgave 1: Invitasjon Oppgave 2: Bilde (beskriv et bilde) Oppgave 3: Fortell om et kjent tema (minimum 80 ord)
A2–B1	Beskrive et bilde (A1, A2 og B1) Fortelle om et kjent tema (A1, A2 og B1) Få fram egne synspunkter (A2, B1 og B2)	Oppgave 1: Bilde (beskriv et bilde) Oppgave 2: Fortell om et kjent tema (mellom 80 og 200 ord) Oppgave 3: Alltid en klage skrevet som e-post (minimum 80 ord)
B1–B2	Få fram egne synspunkter (A2, B1 og B2) Argumentere (B2)	Oppgave 1: Alltid en klage skrevet som e-post (minimum 80 ord) Oppgave 2: Argumentere (valg mellom to oppgaver)

Vurdering skjer med hjelp av et felles vurderingsskjema som operasjonaliserer konstruktet som måles og nivåbeskrivelsene i Det felles europeiske rammeverket for språk. Besvarelsene blir vurdert etter to hovedkriterier: formidlingsevne og språklige evner. Til sammen er det fem kriterier som hver sensor skal vurdere besvarelsene etter. Kriteriene er tekstoppbygging, rettskriving og tegnsetting, ord og uttrykk, grammatikk, og formidlingsevne. Formidlingskriteriet skal vurderes for hver deloppgave, mens de språklige kriteriene skal vurderes for prøven samlet. For å få A1 eller høyere forventes det måloppnåelse på samtlige kriterier med unntak av tekstoppbygning som ikke er krav for å oppnå A1. Det kommunikative språksynet, som er beskrevet i

modellen for kommunikativ kompetanse av Bachmann (1990), gjenspeiles i vurderingsskjemaet ved et gjennomgående fokus på at budskapet skal være forståelig. Formelle feil ses i forhold til i hvilken grad de fører til misforståelser/hindrer forståeligheten. For eksempel i vurdering av grammatikk-kriteriet skal sensorene godta noen grammatiske feil, og gi B2 så lenge de ikke står i veien for formidlingen.

For hvert kriterium er det i vurderingsskjemaet beskrevet et kort minimumskrav for hva en kandidat må klare for å være på de ulike nivåene. Sensorene blir oppfordret til å bruke disse beskrivelsene aktivt i sensureringen. Instruksen til sensorene er at kandidatene må ha klart samtlige minimumskrav for å få endelig karakter på nivået. For eksempel skal en kandidat som er vurdert til B2 på alle kriteriene bortsett fra et der han/hun får B1, også få B1 som endelig resultat. Selv om vurderingen skal skje ved aktiv bruk av kriteriene, rapporterer sensorene per dags dato kun inn endelig karakter per kandidatbesvarelse. Vurderingsskjemaet er designet for å være korte og konsise slik at sensorene kan bruke de aktivt i vurdering av tekster. Ulempen med korte beskrivelser er at det gir større rom for tolkningsmuligheter og avvikende sensurering. Derfor er det utarbeidet en forklaring av kriteriene. Her beskrives hvert av kriteriene og hva som kreves for å bli plassert på de ulike rammeverksnivåene. Forklaringen inneholder eksempler og presiseringer for å gjøre vurderingsskjemaet klarere. Vurderingsskjemaet for hvert av de tre prøvenivåene inkludert forklaring av kriteriene er lagt ved artikkelen.

Formålet med vurderingsskjemaet er å standardisere vurderingsprosessen og minimere muligheten for subjektive tolkninger. Det er likevel vanskelig å unngå at sensorer tolker tekster forskjeller spesielt ved vurdering av langsvarsoppgaver. Neste avsnitt introduserer konseptet sensorvariasjon og en modell for å måle reliabiliteten til sensorkorpset som vurderer besvarelser i skriftlig framstilling.

Sensorvariasjon og reliabilitet i vurdering av skriftlig prøve

Det å vurdere en skriftlig tekst er en krevende og kompleks oppgave. Sensoren må tolke, evaluere og bedømme en besvarelse basert på et vurderingsskjema. Komplekse og subjektive vurderinger åpner opp for stor variasjon i sensorens nivåsetting. Sensorvariasjon er definert som variasjon i kandidatens oppnådde resultat som skyldes sensoren og er en del av konstrukt irrelevant variasjon ved måling av kandidatferdighet (Eckes, 2015). Sensorvariasjon vanskeliggjør nøyaktig måling av prøvens konstrukt, og er derfor en trussel for validiteten av kandidatens endelige resultat (Lane & Stone, 2006; Eckes, 2015). Forskning har vist at det generelt er betydelig sensorvariasjon i vurdering av skriftlige tekster (Eckes, 2015; McNamara, 1996; Weigle, 1998).

Det er flere grep prøveutviklere kan ta i bruk for å begrense sensorvariasjon. For eksempel er god sensoropplæring avgjørende. Det er også viktig at flere sensorer vurderer hver besvarelse, og at vurderingene skjer uten at sensorene samarbeider (Eckes, 2015). Uavhengige vurderinger er ønsket fordi når sensorer vurderer uavhengig, vil ikke mellommenneskelige faktorer påvirke sensureringen. Ved å tvinge sensorer til å bli enige om et resultat vil ofte egen erfaring og ekspertise bli tilsidesatt til fordel for meningene til sterke personligheter (Weigle, 1998). Derimot, hvis besvarelsene vurderes uavhengig, vil denne formen for konstrukt irrelevant variasjon forsvinne. Et annet ledd i kvalitetssikringsarbeidet er at prøveutviklere undersøker reliabiliteten til sensorkorpset, og gir tilbakemelding/avskjedige sensorer som skiller seg ut negativt.

I klassisk testteori har sensorvariasjon typisk blitt undersøkt ved hjelp av korrelasjons- og samsvarsanalyser. Her undersøkes korrelasjon mellom sensorvurderingene for å se om sensorene rangerer besvarelsene likt. I tillegg undersøkes den absolutte enigheten til sensorpar. Denne tradisjonelle formen for å undersøke sensorvariasjon har blitt kritisert fordi høy konsensus og konsistens mellom sensorpar ikke betyr at de nødvendigvis er reliable sensorer. Grunnen er at begge sensorene kan enten være blant de strengeste eller mildeste i sensorkorpset (Eckes, 2015).

Sensorer som rettemaskiner eller uavhengige eksperter

Hvordan sensorvariasjon er behandlet i litteraturen, kan grovt sett deles inn i to grupper. På den ene siden er det fokus på konsensus. Her er målet å ha sensorer som oppfører seg som rettmaskiner som alltid vurderer besvarelser likt (Linacre, 2017). Selv med utvidet bruk av sensoropplæring har forskning vist at målet om lik sensurering er langt fra nådd (Eckes, 2015; Weigle, 1998). Skar og Jølle (2017) finner at erfarne norske sensorer fra et flerårig sensorkorps til den nasjonale utvalgsprøven i skriving for 8. klasse har vedvarende forskjeller i hvor strengt de vurderer skriftlige tekster. Varierende praksis skyldes blant annet at sensorer systematisk bruker vurderingsskjemaet ulikt, og vektlegger deler av konstruktet forskjellig (Eckes, 2012). I språkvurdering kan for eksempel noen sensorer legge større vekt på vokabular og syntaks, mens andre velger å vektlegge struktur og flyt i språket (Eckes, 2008).

Et alternativ til å ha mål om at sensorer skal fungere som rettemaskiner, er å betrakte sensorene som uavhengige eksperter der noe sensorvariasjon er forventet og faktisk ønsket. Her kan sensorvariasjon deles opp i to kategorier. Den ene typen sensorvariasjon er forskjeller i hvor strengt sensorene vurderer, også kalt inter-sensorreliabilitet. Den andre typen sensorvariasjon ser på intern konsistens eller stabilitet hos den enkelte sensor, også kalt intra-sensorrelabilitet (Eckes, 2015).

Moderate faglige forskjeller i sensorstrenghet kan forsvares på to måter. På den ene siden vil vurderingsprosessen fortsatt være gjennomført etter de samme kriteriene, og det er kun faglig uenighet som ligger til grunn for uenigheten. På

den andre siden kan kandidater kompenseres for rene forskjeller i sensorstrenghet ved å matematisk tilpasse endelig resultat med hensyn til sensorens strenghet. Sagt på en annen måte så kan en besvarelse som blir vurdert av den strengeste sensoren i sensorkorpset, vektes likt som en gjennomsnittlig streng sensor hadde sensurert besvarelsen (Eckes, 2005; 2015; Skar og Jølle, 2017).

Istedenfor å oppnå absolutt enighet mellom sensorer, også kalt perfekt inter-sensor relabilitet, er målet med uavhengige eksperter at sensorene er konsekvente i sine vurderinger, og bedømmer besvarelser basert på like kriterier (Weigle, 1998). Her omtaler vi den interne reliabiliteten hos en sensor, også kalt intra-sensorreliabiliteten. En inkonsekvent sensor, som vurderer kandidater med samme ferdigheter ulikt, vil ha lav intra-sensorreliabilitet og er en trussel for prøvens reliabilitet. Mer konkret så vil det være noen aspekter ved en sensors vurderingsadferd som vil slå ut negativt på intra-sensorreliabiliteten. Et eksempel er sentraltendens i vurderingene. Sentraltendens er definert som at sensorene unngår å bruke de høyeste og laveste kategoriene på karakterskalaen, og heller foretrekker å gi vurderinger midt på skalaen. Konsekvensen vil være at sensorer overestimerer ferdighetene til kandidater med lav ferdighet, men underestimerer ferdighetene til kandidater med høy ferdighet. Det er også mulig at tendensen ikke gjelder midt på skalaen, men at vurderingene enten kun er blant de strengeste eller mildeste. Derimot kan det være vanskelig å skille disse effektene fra rene forskjeller i sensorstrenghet (Saal, Downey & Lahey,1980).

Forskjeller i sensorstrenghet er som nevnt ofte vedvarende og vanskelig å eliminere til tross for god opplæring. Det kan hevdes at det er mer fruktbart å eliminere ustabil retteadferd. I sin studie av et utvalg fra sensorkorpset for den nasjonale utvalgsprøven i skriving finner Skar og Jølle (2017) evidens for at opplæring over tid kan lære sensorer til å vurdere mer konsistent. Dette funnet viser viktigheten av gode metoder for å avdekke lav intra-sensorreliabilitet, og at denne informasjonen blir brukt riktig inn i sensoropplæringen.

I Kompetanse Norge ønsker vi at våre sensorer opptrer som uavhengige eksperter, og at hver besvarelse blir vurdert av minst to sensorer uavhengig av hverandre. Dette muliggjør å bruke Many-Facets Rasch Measurement (MFRM), presentert i neste avsnitt, til å tallfeste hvor strengt samtlige i sensorkorpset sensurerer, og til å avdekke i hvilken grad de vurderer stabilt og pålitelig.

Many-Facets Rasch Measurement

Many-Facets Rasch Measurement (MRFM) er en utvidet versjon av Raschmodellen, som ble utviklet av Georg Rasch (1960). Raschmodellen estimerer kandidatenes ferdigheter og oppgavens vanskelighetsgrader uttrykt på samme skala. For mer informasjon om Raschmodellen og andre psykometriske modeller, se for eksempel Embretson og Reise (2000).

MRFM er en utvidet versjon av Raschmodellen der sensorvariasjon er inkludert inn i modellen. Ved å bruke MFRM kan flere uavhengige

sensorvurderinger bli analysert sammen, og mål på sensorenes retteadferd kan gi et bilde av hvordan hver sensor vurderer i forhold til seg selv og resten av sensorkorpset. MRFM estimerer sannsynligheten for at en kandidat oppnår et resultat gitt kandidatens ferdighet og sensorens strenghet. Modellen vi har brukt, er som følger:

$$ln\left(\frac{p_{njk}}{p_{njk-1}}\right) = \theta_n + \alpha_j - \tau_k$$

$p_{njk} = sansynligheten\ for\ at\ kandidat\ n\ får\ en\ vurdering\ k\ av\ sensor\ j$
$p_{njk-1} = sansynligheten\ for\ at\ kandidat\ n\ får\ en\ vurdering\ k - 1\ av\ sensor\ j$
$\theta_n = norsk ferdighetsnivå\ til\ kandidat\ n$
$\alpha_j = sensormildhet\ til\ sensor\ j$
$\tau_k = vanskelighetsgraden\ av\ å\ få\ en\ karakter\ k\ relativt\ til\ k - 1$

Modellen er estimert med en joint maximum likelihood-estimator, som ved en gjentagende prosess estimerer kandidatens ferdighetsnivå og sensorens strenghet til de verdiene som er høyest sannsynlig basert på datamaterialet. Kandidatenes ferdighet blir overført fra sensorenes vurderinger til en lineær intervallskala kalt logits. Sensorenes grad av strenghet blir plassert langs samme skala. MRFM er kjørt med programvaren FACETS (Linacre, 2017). Oppgavene på prøven overlapper hverandre på tvers av de tre prøvenivåene. Sensorene vurderer derimot hele besvarelser, og det finnes ikke sensoroverlapp innad i en bevarelse. En konsekvens er at hvilken oppgave kandidaten har fått ikke kan inkluderes i modellen. Utfordringen med manglende sensoroverlapp er tatt opp i diskusjonen.

Å undersøke sensorenes reliabilitet i vurdering av skriftlige tekster ved hjelp av MFRM er ikke noe nytt. Flere internasjonale studier av sensorvariasjon baserer seg på MFRM-analyse, se blant annet Bahrouni (2016), Eckes (2015) og Myford og Wolfe (2003). I nordisk sammenheng er det også flere studier. Moe & Jones (2003) skriver i Acta Didactica om hvordan MRFM er brukt til å validere en tidligere versjon av Norskprøvens delprøve i skriftlig framstilling. Den nasjonale utvalgsprøven i skriving for norske 5.- og 8.-klassinger i 2016 har også blitt analysert med MFRM (Skar, 2017; Skar & Jølle, 2017).

Mål på intra-sensorreliabilitet

Målet om konsensus i sensorvurderingene har vist seg å være et vanskelig oppnåelig mål, gitt den komplekse oppgaven å vurdere skriftlige tekster. Et mer realistisk mål er at sensorene opptrer som uavhengige eksperter med pålitelig og stabil vurderingsadferd. En måte å undersøke intra-sensorreliabiliteten er å se hvorvidt sensorene passer Raschmodellen (Wolfe 2004). Sagt på en annen måte

undersøker vi hvilken grad de empiriske dataene, det vil si hvordan sensorene vurderer, passer med den teoretiske modellen. Overordnet er det nødvendig at de empiriske dataene passer modellen for at vi kan bruke den. Gitt at den totale tilpasningen er god nok, kan enkeltavvik blant sensorer avdekke inkonsekvent retteadferd. Graden av modelltilpasning kan undersøkes ved å se på Mean square statistics (MnSq) utregnet for hver sensor. Mean square statistics er det kvadrerte avviket mellom observert og forventet varians i vurderinger for hver sensor delt på gjennomsnittet av alle vurderingene. Litteraturen skiller også mellom Infit MnSq og Outfit MnSq. Infit MnSq er vektet slik at uteliggende/ekstremverdier får mindre vekt (Eckes, 2015).

Infit MnSq er alltid positiv og har en forventningsverdi lik 1. En sensor med Infit MnSq-verdi lik 1 har lik observert og forventet varians, og vurderer helt i tråd med Raschmodellen (Eckes, 2015). Verdier langt under 1 kalles overfit, som betyr at verdiene er for forutsigbare. Her brukes kun en begrenset del av karakterskalaen. Overfit indikerer ofte tilstedeværelse av sentraltendens i sensorvurderingene (Myford & Wolfe, 2004). Verdier høyt over 1 kalles underfit/misfit, og indikerer at det er mer variasjon i sensorvurderingene enn forventet av modellen, som betyr at vurderingene er uforutsigbare (Linacre, 2017). I tilfeller hvor sensorene ser bort ifra vurderingsskjema til fordel for egne vurderinger, kan det slå ut i høye verdier. Generelt er det et større problem med misfit, det vil si høye verdier, fordi de gir identifikasjon på inkonsekvent retteadferd løsrevet fra vurderingsskjemaet, og er derfor en direkte trussel for prøvens reliabilitet. I et slikt tilfelle kan like flinke kandidater få ulik vurdering av den samme sensoren (Myford & Wolfe, 2003). Tilfeller av misfit eller overfit kan ikke uten videre knyttes til inkonsekvent sensoradferd hos den enkelte sensor, ettersom det kan være uklarheter i vurderingsskjemaet som er årsaken til utslagene (Bahrouni, 2016).

Når det gjelder hva som er akseptable verdier av Infit MnSq så varierer det i litteraturen. Utvikleren av programvaren FACETS har foreslått verdier mellom 0,5 og 1,5 som akseptable verdier (Linacre, 2017). Innenfor vurdering av fremmedspråk har det blitt foreslått et smalere intervall, med verdier mellom 0,7 og 1,3 (McNamara, 1996). Derimot er det ikke anbefalt å bruke et bestemt intervall ukritisk blant annet fordi en sensors Infit MnSq-verdi er avhengig av det totale antallet vurderinger per sensor (Eckes, 2015). Det er vist at Infit MnSq er omvendt proporsjonal med antall vurderinger, slik at intervallet går mot 1 når antall vurderinger øker hvis alt annet er holdt konstant (Wu & Adams, 2013). I vårt datamateriale har noen sensorer betydelig færre besvarelser enn andre, og diagnostisering av intra-sensorreliabilitet må for sensorene med færrest vurderinger tas med forbehold ettersom disse verdiene er befestet med stor usikkerhet.

Basert på disse målene, MFRM og programvaren FACETS gjennomfører Kompetanse Norge analyser etter prøvegjennomføringer for å se hvordan

sensorkorpset vurderer kandidatene i skriftlig framstilling. I neste avsnitt ser vi på analysen av desemberavviklingen 2017.

Desemberavviklingen i skriftlig framstilling 2017

Ved desemberavviklingen 2017 gjennomførte 7969 kandidater delprøven i skriftlig framstilling (Kompetanse Norge 2018). Det var et felles sensorkorps som vurderte besvarelsene for nivåene A1–A2 og A2–B1. Ettersom det ikke er mulig å oppnå B1 på en A1–A2-besvarelse, har vi valgt å kjøre to separate analyser slik at vurderingsskalaen er lik for alle besvarelsene som blir analysert sammen. Prøvenivået B1–B2 har et annet sensorkorps enn de lavere nivåene og har også blitt analysert separat. Tabell 2 viser karakterfordeling fordelt på de tre prøvenivåene. Tabellen viser at 88 prosent av kandidatene får A1 eller A2 på A1–A2-prøven, og at 89 prosent får A2 eller B1 på A2–B1-prøven, mens det er hele 24 prosent av kandidatene som får A2 på B1–B2 prøven.

Tabell 2: Deskriptiv statistikk per prøvenivå. Desember avviklingen 2017

Oppnådd nivå	A1–A2	A2–B1	B1–B2
IGFV	158 (7 %[1])	113 (3 %)	19 (1 %)
Under A1	131 (6 %)	4 (0 %)	0 (0 %)
A1	989 (42 %)	239 (7 %)	27 (1 %)
A2	1069 (46 %)	1849 (54 %)	529 (24 %)
B1		1196 (35 %)	1350 (61 %)
B2			296 (13 %)
Sum	2347 (29 %)	3401 (43 %)	2221 (28 %)
Sensorer	57	58	19

Vi har i den videre analysen valgt å ekskludere besvarelser med Ikke grunnlag for vurdering (IGFV) fra datagrunnlaget. Dette er fordi besvarelsene med dette resultatet enten er svært uferdige, eller fordi kandidatene ikke besvarer oppgaven selv om språknivået i seg selv kan være på et høyt nivå. Sagt på en annen måte passer ikke kategorien med den kontinuerlige stigende ferdighetskalaen fra Under A1 til B2. På prøvenivået A2–B1 har vi utelatt vurderinger med Under A1, mens på B1–B2 har både Under A1 og A1 blitt fjernet. Kandidater med oppnådd nivå A1 eller lavere på en B1–B2-prøve har en språkferdighet langt under det nivået prøven er ment for å måle. Valget med å ekskludere klart feiloppmeldte kandidater gir bedre modelltilpassing, og begrenser datamaterialet til tre mulige oppnådde nivå for hvert av de tre prøvenivåene.

I tilfellet hvor et sensorpar gir ulik vurdering vil besvarelsen sendes til en tredje sensor også kalt oppmann. Oppmannen gjør en uavhengig vurdering av

[1] Prosentene er avrundet til nærmeste heltall.

besvarelsen og fastsetter endelig resultat. Vi har valgt å ikke inkludere oppmennenes vurdering i analysen for at dataene bedre skal passe med forutsettingene for MFRM. Oppmannsvurderingene er utelatt fordi det ikke er tilfeldig hva slags typer besvarelser som går til oppmann, og fordelingen av oppmannsrollen heller ikke er tilfeldig fordelt i sensorkorpset.

For å få direkte sammenligning av sensorene har hvert prøvenivå 4–5 ankerbesvarelser. Dette er fellestekstene alle sensorene var oppfordret til å vurdere og sende inn resultatet på i forkant av prøveperioden. Dessverre manglet det ankervurderinger for 15 av 58 sensorer på A1–B1, og 3 av 19 på B1–B2. Sensorene uten ankertekster har blitt beholdt i analysen selv om de ikke vurderte fellesbesvarelsene. Der hvor kandidater står igjen med kun en sensorvurdering etter databehandling, vil også denne utelates fra analysen slik at hver kandidat står med to sensorvurderinger.

Resultater

Ved å kjøre modellen MFRM på det behandlede datasettet fra desember avviklingen 2017 med programvaren FACETS får vi en rekke verdier på de ulike målene presentert tidligere i artikkelen. Det gir oss mulighet til å undersøke forskjeller i sensorstrenghet mellom sensorene, også kalt inter-sensorreliabilitet. Videre kan MFRM-analysen være med på å avdekke hvorvidt hver sensor vurderer konsekvent og stabilt også kalt intra-sensorreliabilitet.

Inter-sensorreliabilitet

Etter databehandling inneholdt A1–A2-prøven totalt 2141 kandidatbesvarelser, mens det var 3231 kandidatbesvarelser i datasettet fra A2–B1 prøven. På B1–B2 vurderte de 19 sensorene 2183 besvarelser. 5,2 prosent av kandidatene forvant som følge av databehandlingen og av dem var om lag 4 prosent tilknyttet resultatet Ikke grunnlag for vurdering. Sensorparene hadde lik vurdering i omtrent 3 av 4 tilfeller både på A1–A2- og A2–B1-prøven. Til sammenligning er enigheten litt lavere på B1–B2, med 69,5 prosents enighet, se Tabell 3.

Et mål på om graden av sensorenighet er god, får vi ved å sammenligne faktisk observert enighet med forventet enighetsgrad basert på Raschmodellen. Siden modellen antar at sensorkorpset består av sensorer som vurderer med ulik strenghet vil ikke forventet enighet være 100%. Den forventede enigheten er satt sammen av sannsynligheten til hver sensor for å gi de ulike karakterene for en kandidat. Videre er sannsynligheten for at sensorene er enige om de ulike vurderingene summert opp for hvert rettepar. Dette gir forventet enighet for hvert rettepar. Den samlede forventede enigheten er gjennomsnittlig enighet basert på alle retteparene (Linacre, 2017). Vi ønsker at sensorene retter besvarelser som uavhengige eksperter istedenfor å være «rettemaskiner» som alltid retter likt. Hvis «rettemaskiner» var målet, burde den faktiske enigheten

vært over 90 prosent. Derimot ønsker vi at forventet og faktisk observert enighet er likest mulig. Hvis faktisk observert enighet overstiger forventet enighet, kan det være et signal om at sensureringen av en kandidat ikke har skjedd uavhengig fra hverandre, enten ved at sensorene har blitt enige om vurderingen, eller har blitt tvunget til å bli det (Linacre, 2017). Hvor høy den forventede enigheten er for hver sensor, avhenger av hvor streng sensoren er, og hvem han/hun er plassert i sensorpar med. Av Tabell 3 ser vi at observert enighet ligger under forventet enighet for de tre prøvenivåene. Dette gir en identifikasjon på at den samlede inter-sensorreliabiliteten er lavere enn målet med uavhengige eksperter (Linacre, 2017).

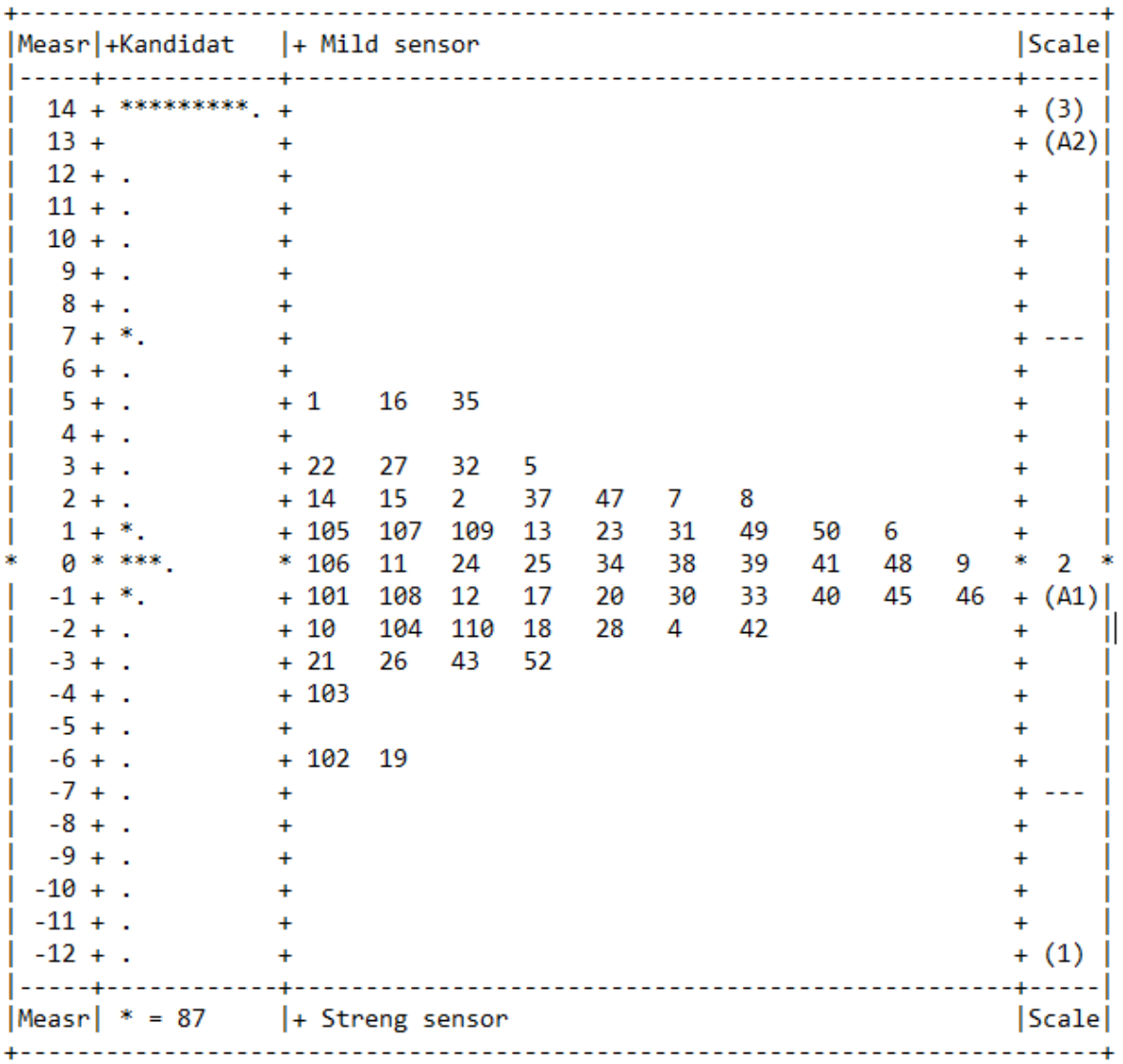
Tabell 3: Mål for inter-sensorreliabilitet. Observert og forventet enighet mellom sensorpar, Strata og H-Indeks.

Prøvenivå	Muligheter for enighet	Observert enighet	Forventet enighet	Strata (R-Indeks)	H-Indeks
A1–A2	6489	4951 (76,3 %)	5291 (80,0 %)	4,87	0,92
A2–B1	7657	5766 (75,3 %)	6064 (79,2 %)	6,54	0,96
B1–B2	2689	1870 (69,5 %)	2082 (77,4 %)	9,97	0,98

Tabell 3 angir Strata for hvert av prøvenivåene. Strata sier hvor mange distinkte statistiske nivå av sensorstrenghet det er mulig i identifisere i utvalget. I tilfellet med perfekt inter-sensorreliabilitet vil vi kunne forvente en Strata verdi nær 1. Tabell 3 viser at det ikke er tilfellet i våre data. Den største verdien er på B1–B2–prøven, der det er mulig å skille nesten 10 nivåer av sensorstrenghet mellom sensorene. H-indeksen i Tabell 3 kan ha verdi mellom 0 og 1, og angir graden av sikkerhet på at variasjonen i sensorstrenghet ikke skyldes målefeil. I tilfeller hvor alle sensorene er like strenge, vil all variasjon i sensorstrenghet skyldes målefeil, og verdien er lik 0. For alle prøvenivåene er H-ideksen nær 1, og vi kan med statistisk sikkerhet si at observert variasjon i sensorstrenghet ikke kommer av målefeil, og det vil si at sensorene ikke vurderer like strengt. For mer informasjon om reliabilitetsmålene presentert i Tabell 3, se Linacre (2017) og Eckes (2015). Resultatene viser klart at sensorene som vurderer delprøven i skriftlig framstilling, samlet ikke vurderer besvarelser like strengt. Resultatene kan likevel sammenlignes med tidligere studier. I en tilsvarende skriftlig prøve i tysk som andrespråk på høyere nivå fant Eckes (2005) Strata på 9,61 og en H-indeks på 0,98 sammenlignbart med funnene for B1–B2 prøven presentert i Tabell 3.

For å gi et bilde av hvor strengt hver sensor vurderer kandidatene, er sensorkorpset grafisk framstilt i Figur 1 for A1–A2, Figur 2 for A2–B1 og Figur 3 for B1–B2. Lengst til høyre i diagrammet er vurderingsskalaen. Her er 1 lik Under A1, 2 lik A1, 3 lik A2, 4 lik B1 og 5 lik B2. Graden av sensorstrenghet til

hver sensor er gitt i midten av diagrammet. De øverste sensorene (lengst opp i diagrammet) er de sensorene som vurderer mildest, mens de lengst ned vurderer strengest. Kandidatenes ferdigheter er gitt fra de sterkeste høyt opp i diagrammet til de svakeste lengst ned. Kandidatenes ferdigheter, sensorenes strenghet og karakterskalaen er overført til en lineær intervallskala med logits som enheter, her gitt som Measure. Logitsskalaen[2] er definert slik at den gjennomsnittlige strenge sensoren har en logit-verdi på 0.

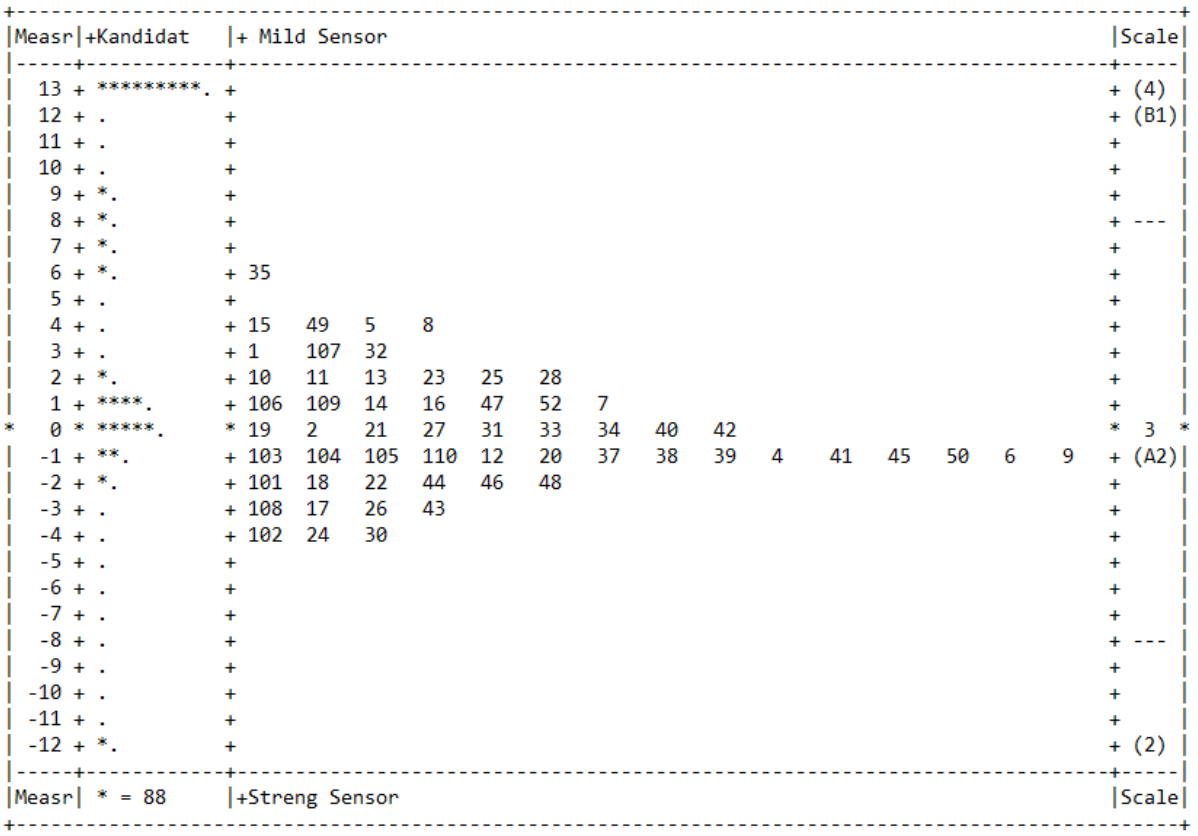


Figur 1: Graden av sensorstrenghet per sensor. Prøvenivå A1–A2

På A1–A2 (Figur 1) utgjør spennet i kandidatenes ferdigheter 26 logits, mens spennet i sensorenes strenghet er 11,6 logits mellom den mildeste (Sensor 1) og den strengeste sensoren (Sensor 19). Med andre ord kan forskjeller i

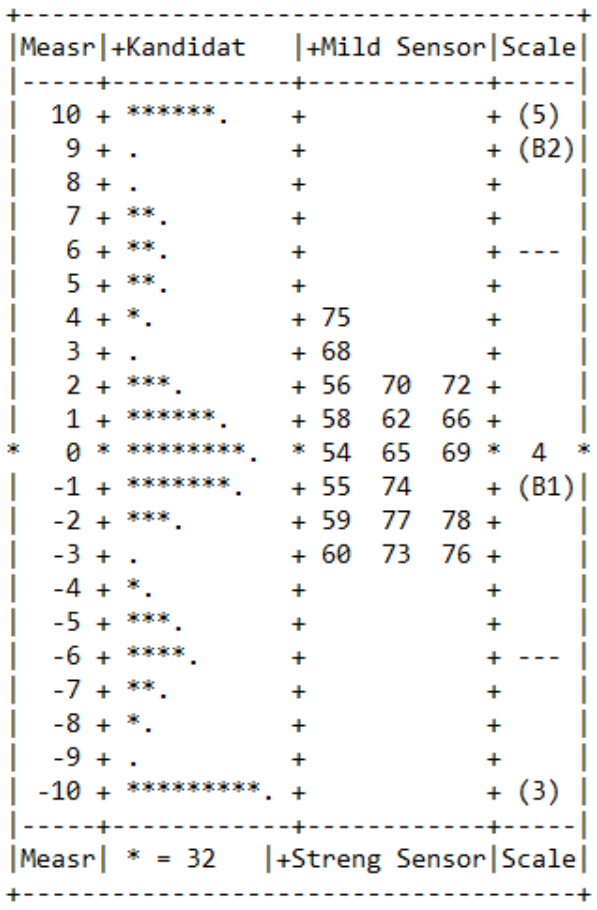
[2] Teoretisk kan logitskalaen gå fra minus uendelig til uendelig. Typisk har andrespråksprøver lang logitskala, siden det er stor forskjell i språkferdighet i kandidatmassen, se for eksempel Eckes (2005). Vi stiller likevel spørsmålstegn ved hvor lang skalaen er, spesielt for A1–A2- og A2–B1-prøven.

sensorstrenghet mellom den mildeste og strengeste sensoren sammenlignes med 45 prosent av spennet i kandidatenes ferdigheter. For A2–B1 (Figur 2) utgjør spredningen i sensorstrenghet 40 prosent av spennet i kandidatferdighet.



Figur 2: Graden av sensorstrenghet per sensor. Prøvenivå A2–B1

For B1–B2-nivået (Figur 3) ser vi at sensorvariasjon i størrelse er lik 35 prosent av spredningen i kandidatenes ferdigheter. Det er ti prosentpoeng lavere enn A1–A2, og fem prosentpoeng lavere enn A2–B1. Ved hjelp av karakterskalaen lengst til høyre ser vi at kandidater med logits mellom -6 og 6 mest sannsynlig får B1 som endelig resultat. Sammenligner vi med avstanden mellom strengeste og mildeste sensor, ser vi at det utgjør 58 prosent av spredningen til en sannsynlig B1-kandidat. Med andre ord kan sensorvariasjon i sensorstrenghet sammenlignes med i overkant av halvparten av spredningen i kandidatferdighet innenfor B1-nivået.



Figur 3: Graden av sensorstrenghet per sensor. Prøvenivå B1–B2

Intra-sensorreliabilitet

Vi har til nå sett på inter-sensorreliabiliteten ved å undersøke hvordan sensorene varierer i sensorstrenghet i forhold til hverandre, og hvor stor denne spredningen er sammenlignet med forskjeller i kandidatferdighet. Resultatene viser betydelig og statistisk signifikant spredning i sensorstrenghet. Det er ideelt at kandidatenes resultat ikke skal avhenge av tildelt sensor, men det er vanskelig å unngå med menneskelige sensurering av langsvarsoppgaver. Et annet og vel så viktig aspekt av vurderingsprosessen er at sensorene bruker vurderingsskjemaet likt, og at hver sensor vurderer kandidatene konsekvent og pålitelig. At besvarelsene vurderes etter kriterier og konsekvent gjennom hele prøveperioden er et mer realistisk mål enn lik sensurering. Derfor er det viktige å studere intra-sensorreliabiliteten i sensorkorpset.

Tabell 4 viser informasjon om hver enkelt sensor som vurderer besvarelser på A1–A2 og A2–B1. Tabell 5 gir tilsvarende informasjon for B1–B2-sensorene. Tabellene inkluderer sensornummeret, antall vurderinger, gjennomsnittlig vurdering[3], målet på sensorstrenghet (Measure) uttrykt i logits og standardfeilen tilknyttet dette estimatet. Tabellene viser også Infit MnSq-verdi for hver sensor. Vi har valgt å kun se på Infit-verdier for å vurdere intra-

[3] A1=2, A2=3, B1=4 og B2=5. For eksempel har sensor 48 på A1–A2-prøven en gjennomsnittlig vurdering midt mellom A1 og A2, med verdi 2,50.

sensorreliabilitet, siden de har høyere presisjon i å vurdere sensorenes adferd (Myford & Wolfe, 2003), og for å minimere antall instrument.
Ser vi nærmere på Infit MnSq-verdiene til sensorkorpset som vurderer på A1–A2 og A2–B1, finner vi noen tilfeller av verdier utenfor foreslåtte intervall[4]. På prøvenivå A2–B1 har sensor 106 en verdi lik 1,85, mens på A1–A2 har sensor 104 og 4 Infit MnSq-verdi lik henholdsvis 2,11 og 2,71. Alle disse tre verdiene viser klar indikasjon av underfit/misfit, som vil si høyere variasjon i vurderingene enn det modellen forventer. Tilfeller av misfit kan tolkes som at sensorene ikke er konsekvente i sine vurderinger, eller at vurderingsskjemaet er tilsidesatt til fordel for egne vurderinger. I fellestekstene har sensor 4 vurdert de fleste besvarelsene et nivå strengere enn de fleste i sensorkorpset. Hvis vi derimot ser på desemberavviklingen separat, har den samme sensoren en gjennomsnittlig streng retteadferd med 78 prosents enighet, som ligger over snittet på 76,3 prosent. En mulig årsak til at den aktuelle sensoren kommer ut med høy verdi, er at han/hun vurderer fellestekstene strengere enn besvarelsene i desemberavviklingen.

Overfit eller lav Infit MnSq-verdi indikerer at sensoren har mindre variasjon i vurderingene enn forventet av modellen, og ikke bruker hele vurderingsskalaen til å skille mellom flinke og mindre flinke kandidater. På A1–A2 har sensor 102 den laveste Infit MnSq-verdien, lik 0,24, men er samtidig også den klart strengeste sensoren på det prøvenivået. Det betyr at en overvekt av de laveste karakterene er benyttet, og sensoren klarer ikke å skille tilstrekkelig mellom dyktige og mindre dyktige kandidater. På A1–A2-nivået samlet var omtrent halvparten av vurderingene A2. Sensor 102 gav til sammenligning A2 kun til 20 prosent av kandidatene. Tre av de fire sensorene med Infit MnSq-verdier utenfor kravet til Linacre (2017) har sensornummer over 100, som betyr at de for første gang var med og vurderte besvarelser i desemberavviklingen 2017. Ved riktige tilbakemeldinger, sensortrening og videre sensurering er det tenkelig at de vil forbedre seg.

Det er viktig å poengtere at Infit MnSq er sensitivt i forhold til antall vurderinger sensoren foretar seg. Ved flere vurderinger vil intervallet reduseres, det vil si at verdiene går mot 1, noe som kan ha innvirkning på hvorfor flere sensorer på A1–A2 og delvis A2–B1 har problematiske verdier sammenlignet med B1–B2, der hver sensor vurderer betydelig flere kandidater. Bruker vi det strengere kravet til McNamara vil ytterligere 11 sensorer ha for lave verdier hvor 9 av tilfellene er på A1–A2-prøven. Prøvenivå A1–A2 har lite variasjon i sensorvurderingene, ettersom de fleste kandidatene får enten A1 eller A2, og kun om lag 6 prosent av vurderingene er Under A1. Basert på disse funnene har vi valgt å evaluere intra–sensorreliabiliteten til sensorene som vurderte A1-A2-prøven ved å kun se på det mildere kravet til Linacre (2017).

[4] Målet for gode Infit MnSq-verdier satt av McNamara er mellom 0,7 og 1,3. Linacre sine verdier er mellom 0,5 og 1,5.

Tabell 4: Sensortabell for sensorer på prøvenivå A1–A2 og A2–B1

A1–A2						A2–B1					
Sensor	**Antall vurderinger**	**Gj.snittlig vurdering**	**Measure**	**Std. feil**	**Infit MnSq**	**Sensor**	**Antall vurderinger**	**Gj.snittlig vurdering**	**Measure**	**Std. feil**	**Infit MnSq**
102	77	2,04	-6,17	0,61	0,23	107	109	3,43	2,65	0,59	0,5
16	78	2,69	4,82	0,66	0,52	44	105	3,24	-1,72	0,46	0,68
2	69	2,55	2,29	0,62	0,53	9	128	3,27	-0,86	0,44	0,71
31	79	2,49	0,82	0,65	0,53	52	101	3,5	0,82	0,44	0,71
35	73	2,73	5,34	0,86	0,59	37	111	3,26	-1,24	0,46	0,71
32	82	2,55	2,58	0,56	0,63	48	80	3,2	-1,95	0,56	0,72
30	83	2,48	-1,1	0,62	0,63	6	116	3,29	-0,51	0,37	0,73
9	84	2,43	0,09	0,47	0,66	31	115	3,33	-0,37	0,46	0,73
10	90	2,33	-1,83	0,6	0,66	104	112	3,37	-0,61	0,41	0,75
21	80	2,24	-2,91	0,63	0,68	11	119	3,43	1,69	0,43	0,77
103	81	2,12	-3,73	0,51	0,72	25	133	3,42	1,75	0,4	0,78
23	85	2,48	0,86	0,52	0,73	38	104	3,23	-0,86	0,43	0,78
27	101	2,46	2,91	0,64	0,73	15	129	3,48	4,39	0,47	0,78
43	82	2,34	-3,17	0,65	0,74	24	136	3,1	-4,15	0,5	0,79
107	87	2,53	0,55	0,55	0,75	18	125	3,22	-2,42	0,38	0,81
26	88	2,26	-3,17	0,6	0,75	34	121	3,26	0,04	0,42	0,82
17	71	2,51	-0,77	0,58	0,76	47	59	3,42	0,69	0,55	0,82
12	87	2,41	-0,6	0,53	0,8	35	105	3,68	5,74	0,64	0,82
5	86	2,57	2,57	0,57	0,81	10	120	3,38	2,44	0,52	0,84
41	77	2,35	-0,35	0,43	0,86	22	133	3,29	-1,51	0,38	0,85
22	87	2,57	2,51	0,63	0,86	8	124	3,48	4,06	0,5	0,85
1	63	2,65	5,46	0,85	0,87	49	99	3,52	4,17	0,67	0,85
33	87	2,3	-0,68	0,46	0,87	105	132	3,3	-0,7	0,39	0,86
42	75	2,35	-1,51	0,52	0,9	32	106	3,41	2,84	0,57	0,86
108	78	2,42	-0,83	0,64	0,9	14	125	3,43	1,22	0,41	0,87
24	77	2,48	-0,2	0,61	0,92	103	105	3,29	-0,73	0,39	0,89
18	83	2,36	-1,58	0,49	0,92	26	122	3,12	-2,59	0,4	0,89
40	87	2,43	-0,57	0,4	0,93	7	119	3,38	0,91	0,42	0,89
48	56	2,5	-0,19	0,75	0,95	108	121	3,03	-3,34	0,46	0,9
34	117	2,44	-0,04	0,43	0,96	50	86	3,28	-1,05	0,49	0,9
52	48	2,29	-3,01	0,76	0,97	101	134	3,16	-1,73	0,36	0,91
14	77	2,55	1,95	0,55	0,98	12	127	3,2	-1,15	0,41	0,91
50	51	2,53	0,94	0,7	0,99	28	115	3,4	1,87	0,46	0,92
46	52	2,42	-0,69	0,58	1,01	46	101	3,18	-2,42	0,49	0,92
20	79	2,38	-0,87	0,57	1,02	20	132	3,27	-0,64	0,38	0,94
110	105	2,39	-2,01	0,52	1,02	1	131	3,46	3,16	0,4	0,94
7	69	2,54	1,92	0,63	1,03	16	123	3,21	0,61	0,35	0,95
25	87	2,43	0,16	0,54	1,03	45	87	3,14	-1,3	0,47	0,95
109	80	2,49	0,6	0,47	1,04	43	110	3,04	-3,1	0,38	0,96
101	88	2,38	-0,69	0,54	1,05	30	111	3,15	-3,64	0,44	0,97
39	102	2,38	0,01	0,44	1,05	5	107	3,55	4,35	0,51	0,97
38	120	2,33	-0,07	0,38	1,06	17	115	3,17	-2,51	0,41	0,98
6	85	2,41	0,74	0,57	1,07	21	123	3,32	0,11	0,4	0,99
105	73	2,49	0,72	0,7	1,08	40	129	3,38	0,4	0,46	1,01
11	90	2,49	0,34	0,53	1,08	33	107	3,26	-0,32	0,48	1,03
15	76	2,53	1,7	0,61	1,1	42	88	3,27	-0,25	0,39	1,04
8	83	2,61	1,76	0,6	1,11	102	133	3,1	-3,64	0,38	1,05
28	81	2,51	-1,57	0,72	1,19	39	106	3,32	-1,26	0,5	1,05
45	68	2,25	-0,5	0,56	1,2	27	73	3,29	0,3	0,49	1,07
37	81	2,37	2,43	0,64	1,2	13	121	3,48	1,57	0,47	1,11
49	66	2,52	1,21	0,61	1,25	23	137	3,39	1,58	0,41	1,13
106	69	2,54	0,01	0,6	1,25	109	123	3,35	0,53	0,43	1,17
19	80	2,26	-5,61	0,76	1,29	4	118	3,26	-1,02	0,44	1,2

13	79	2,35	1,18	0,55	1,49	2	102	3,41	-0,06	0,45	1,21
104	47	2,47	-1,7	0,79	2,11	110	167	3,23	-0,86	0,35	1,25
4	69	2,3	-2,23	0,73	2,71	41	90	3,31	-0,56	0,45	1,27
47	26	2,39				19	136	3,27	-0,01	0,33	1,28
						106	119	3,41	1,2	0,47	1,85

Tabellen er sortert fra lave til høye Infit MnSq-verdier. Measure og Infit MnSq er skjult for Sensor 47 på grunn av for få vurderinger.

Det er ingen av B1–B2-sensorene med Infit MnSq-verdier utenfor det strengeste kravet satt av McNamara (1996). Dette forteller oss at selv om sensorstrengheten varierer klart signifikant mellom B1–B2-sensorene, er det ikke påvist noen tilfeller av inkonsekvent retteadferd eller begrenset bruk av karakterskalaen blant de 19 B1–B2-sensorene.

Tabell 5: Sensortabell for sensorer på prøvenivå B1–B2

Sensor	Antall vurderinger	Gj.snittlig vurdering	Measure	Standardfeil	Infit MnSq
60	236	3,74	-2,69	0,29	0,73
65	251	3,96	0,35	0,28	0,79
62	250	4,04	1,3	0,26	0,83
66	241	3,91	0,8	0,25	0,87
70	242	4	1,59	0,27	0,88
54	241	3,96	0,17	0,28	0,89
58	234	4,05	1,41	0,27	0,9
69	245	3,86	-0,13	0,26	0,9
73	223	3,74	-3,17	0,3	0,9
59	232	3,81	-2,48	0,28	0,91
75	234	4,18	3,56	0,29	0,93
77	270	3,75	-2,24	0,26	0,93
76	243	3,74	-2,76	0,3	0,97
68	239	4,13	3,01	0,28	1
74	229	3,87	-0,51	0,27	1,04
78	230	3,84	-1,55	0,25	1,08
55	157	3,89	-1,22	0,35	1,13
72	242	4,07	2,17	0,28	1,14
56	191	4,12	2,39	0,3	1,25

Overordet sett ser vi at de fleste sensorene har akseptable verdier og oppfyller sin rolle som uavhengig ekspert. Spesielt er det viktig at høye Infit MnSq-verdier er lite utbredt siden det er den største trusselen for prøvens reliabilitet.

Diskusjon

Norskprøvens delprøve i skriftlig framstilling skal vurderes med en analytisk vurderingsskala. Med analytisk vurderingsskala vurderer sensorene besvarelsen ved å gi karakter på flere kriterier. Selv om besvarelsene skal vurderes analytisk, rapporterer sensorene per i dag kun inn en holistisk/samlet vurdering per kandidat. Hvis sensorene hadde rapportert inn nivå på alle kriteriene i vurderingsskjemaet de skal vurdere en besvarelse etter, ville vi fått mer informasjon om kandidatens måloppnåelse enn de to holistiske vurderingene vi får i dag. En MRFM-analyse basert på en analytisk vurderingsskala ville gitt oss informasjon om hvor vanskelige kriteriene er i forhold til hverandre. I tillegg vil en analyse kunne belyse hvordan sensorene vurderer de ulike kriteriene og hvor godt vurderingsskjemaet fungerer. Det kan tenkes at fordi sensorene kun trenger å skrive inn en holistisk vurdering vil det ha innvirkning på hvordan sensorer vurderer kandidatbesvarelser, og at det derfor ikke er utenkelig at noen sensorer velger å se bort ifra kriteriene til fordel for fokus på overordnet måloppnåelse. Selv om en holistisk vurderingskala ofte er fortrukket av praktiske og økonomiske årsaker, har flere studier har vist at sensorreliabiliteten øker når sensureringen blir gjennomført med analytisk vurderingsskala (Chi, 2001; Weigle, 2002). Denne studien har påvist noen sensorer med verdier for intra-sensorreliabilitet, som kan indikere begrenset bruk av vurderingsskalaen eller sensurering løsrevet fra vurderingsskjemaet. Med en operasjonalisert analytisk vurderingsskala kunne vi fått vite om det var noe med vurderingsskjemaet eller spesifikke kriterier som var årsak til de problematiske verdiene.

Delprøven i skriftlig framstilling består som tidligere nevnt av to eller tre separate oppgaver med overlapp mellom prøvenivåene. Hvordan de ulike deloppgavene fungerer og hvor vanskelige de er, får vi derimot lite informasjon om. Et lenket design der sensorene vurderte forskjellige deloppgaver i stedet for kun en samlet prøve, ville muliggjort en utvidet MFRM-modell, der oppgavene som kandidatene får er inkludert i modellen. I vurderingsskjemaet er det spesifisert at det endelige resultatet aldri skal være høyere enn den laveste kriterievurderingen. En konsekvens er at kandidater med høy skriveferdighet kan bli straffet sterkt av et manglende aspekt av konstruktet. Sagt på en annen måte er ikke skillet mellom to ferdighetsnivåer definert som et fast punkt midt mellom de to nivåene. MFRM antar at prøven måler en endimensjonal ferdighet, og gir en kontinuerlig ferdighetsskala fra lav til høy norskferdighet. Denne antakelsen passer ikke helt med instruksjonen om oppnådde minstekrav.

For å få en direkte lenking mellom sensorene er vurderingene av fellestekstene blitt inkludert i datasettet. Disse fellestekstene er gitt ut i forkant av prøveavviklingen med mål om å forbedre sensorene til prøveavviklingen. Vurderingene blir nøye gjennomgått, og sensorene får tilbakemelding på hva som er riktig nivå på hver tekst. Derfor er det mulig at sensorene bruker mer tid på fellestekstene enn det de gjør med en vanlig besvarelse. Det beste ville vært

hvis fellestekstene ble gitt som en del av prøveperioden uten at sensorene fikk vite at disse oppgavene blir spesielt vurdert. En annen uheldig side ved fellestekstene er at de ikke ble vurdert av alle sensorene. Resultatet er at 3 av B1–B2-sensorene og 12 av A1–B1-sensorene ikke er direkte lenket sammen gjennom fellesvurderte tekster. Sensorene er fortsatt indirekte lenket sammen ved at de fleste sensorene har sensurert minst én kandidat sammen, eller ved at sensorer er lenket sammen gjennom en tredje sensor. Likevel er det uheldig at en direkte link mangler for noen sensorer. Et alternativ ville vært å utelate alle sensorene uten felles vurderte tekster, men det ville ha ført til at nesten halvparten av sensorvurderingene hadde blitt ekskludert fra analysen, ettersom alle kandidater knyttet til sensorene uten fellesvurderinger måtte fjernes.

MFRM blir i dag brukt av Kompetanse Norge som et diagnostisk verktøy for å kvalitetssikre sensureringen av delprøven i skriftlig framstilling. Resultatene fra analysen blir presentert for sensorene i forbindelse med den årlige sensorsamlingen. Målet med dette er å bevisstgjøre sensorene i forhold til hvordan de vurderer, og at de mest avvikende sensorene tilpasser seg. Resultatene blir også brukt internt i Kompetanse Norge som kunnskapsgrunnlag, og til å følge opp sensorkorpset.

Et av forskningsspørsmålene til denne studien er å avdekke om MFRM er en egnet metode for avdekking av inter-sensorreliabilitet og intra-sensorreliabilitet, og å se på hvilken rolle modellen kan ha sett i lys av Norskprøvens utforming og praktiske begrensinger. Den MFRM-modellen er presentert i denne artikkelen, tar hverken hensyn til de underliggende kriterievurderingene til det endelige resultatet, eller hvordan deloppgavene fungerer separat. Resultatet er en enkel modell som er kritisert for å oversimplifisere prosessen med å identifisere komponenter som påvirker vurderingen av skriftlige tekster (Eckes, 2015). Vi har også problemet med at manglende direkte lenking av sensorer gjør datamaterialet mindre egnet til MFRM-analyse. Fordi MFRM kvantifiserer sensorstrenghet, kan metoden brukes til å trekke ut sensorvariasjonen fra det endelige resultatet, og vekte karakteren slik som en gjennomsnittlig streng sensor vurderer besvarelsen. På denne måten kan rene forskjeller i sensorstrenghet kompenseres for. For å implementere en slik matematisk vekting av kandidatens endelige resultat må Kompetanse Norge adressere de begrensingene som er beskrevet her for å sikre gyldigheten av modellen. Vi oppfatter likevel at MRFM allerede i dag er et viktig bidrag i kvalitetssikringsprosessen, men at metoden må kombineres med andre analyser for å gi et helhetlig bilde av sensorreliabiliteten. Her kan korrelasjonsstudier være et nyttig supplement. Det er viktig å presisere at fokus her har vært på måling av sensorreliabilitet. God opplæring og kontinuerlig kursing av sensorer med mål om en samlet forståelse av vurderingsskjema og vurderingsprosessen er nøkkelen til reliabel vurdering.

Konklusjon

Denne artikkelen har prøvd å beskrive reliabilitetsutfordringen med menneskelig sensurering av skriftlige tekster og en statistisk metode Kompetanse Norge bruker for å kvalitetssikre sensorkorpset som sensurerer Norskprøvens delprøve i skriftlig framstilling. Selv erfarende sensorer varierer betydelig i hvor strengt de vurderer skriftlige tekster. Forskjeller i det endelige resultatet som kan knyttes direkte til sensoren, også kalt sensorvariasjon, er en utfordring for prøvens validitet og reliabilitet.

Ved hjelp av modellen Many-Facet Rasch Measurement (MFRM) kan graden av sensorstrenghet kvantifiseres for hver enkelt sensor. Inter-sensorreliabiliteten, det vil si enigheten i sensorkorpset, er undersøkt ved å se på mål for spredning av sensorstrenghet, og ved å sammenligne den faktisk prosentvise enigheten mellom sensorene med forventet enighet basert på modellen. Om sensorene vurderer stabilt og konsekvent igjennom hele prøveperioden, kan undersøkes ved å se hvor godt sensorervurderingene passer modellen. Retningen på avvik fra modellen kan gi indikasjon på om den aktuelle sensoren enten i for liten grad bruker vurderingskalaen til å skille mellom kandidatene, eller om det er identifikasjon på tilfeldig sensurering løsrevet fra vurderingsskjemaet. Samlet sett gir MRFM-modellen prøveutviklerne verdifull informasjon om hvordan sensorkorpset fungerer, og kan være en hjelp til å identifisere tiltak som bidrar til mer rettferdig sensurering.

Overordnet viser resultatene at sensorene ved desemberavviklingen 2017 ikke vurderer besvarelser like strengt. Det er flere signifikant skillbare nivåer av sensorstrenghet mellom den mildeste og strengeste sensoren for alle de tre prøvenivåene. Intra-sensorreliabiliteten er på den andre siden for de fleste sensorene høy. De 19 sensorene på B1–B2-sensorene skiller seg ut positivt, med ingen verdier utenfor anbefalte intervall. Av de 58 sensorene som vurderte besvarelser på A1–A2 og A2–B1, fant vi noen tilfeller av uønskede verdier. Det lavere antallet vurderinger per sensor og at kun et fåtall kandidater oppnår laveste vurdering, kan være med å forklare hvorfor det er flere problematiske verdier på de lavere nivåene. Om vi bruker det anbefalte intervallet til Linacre (2017) er det kun 4 av 77 sensorer med verdier som indikerer lav intra-sensorreliabilitet. Flere av sensorene med problematiske verdier vurderte for første gang besvarelser ved denne gjennomføringen, og de vil forhåpentligvis forbedre seg med mer sensortrening og videre sensurering. Fordi resultatet på Norskprøven kan stor betydning for kandidatene er det viktig at Kompetanse Norge klarer å bruke informasjonen fra MFRM til å følge opp sensorkorpset og øke sensorreliabiliteten. For eksempel må det adresseres tiltak for å redusere gapet mellom de strengeste og mildeste sensorene. Videre må tilfellene av lav intra-sensorreliabilitet undersøkes nærmere, gjerne med et kritisk blikk på vurderingsskjemaet. Hvordan Kompetanse Norge på best mulig kan dra nytte av

innsiktene fra MFRM inn i sensoropplæringen, er en viktig oppgave som fortjener en studie i seg selv.

Denne studien har vist at delprøven i skriftlig framstilling har begrensinger knyttet til hvor godt MFRM-modellen passer datamaterialet. Dette gjelder for eksempel ved at oppgavens vanskelighetsgrad ikke kan inkluderes i modellen på grunn av manglende sensoroverlapp innad i en besvarelse. Et annet problem er at ikke alle sensorene er direkte lenket gjennom vurdering av fellestekster. Manglende vurderinger på kriterienivå for hver kandidat gjør at vi mister informasjon om hvordan sensorene tenker når de vurderer besvarelser, og derav også om vurderingsskjemaet fungerer. Disse manglene bør Kompetanse Norge adressere for å få styrket modellens troverdighet, og for å unytte potensialet med MFRM-metoden til det fulle.

Om forfatterne

Tor Midtbø er rådgiver ved Kompetanse Norge.
Institusjonstilknytning: Kompetanse Norge, Postboks 236 Sentrum, 0103 Oslo.
E-post: tor.midtbo@kompetansenorge.no

Arne Rossow er rådgiver ved Kompetanse Norge.
Institusjonstilknytning: Kompetanse Norge, Postboks 236 Sentrum, 0103 Oslo.
E-post: arne.rossow@kompetansenorge.no

Brikt Sagbakken er fungerende seksjonsleder ved Kompetanse Norge
Institusjonstilknytning: Kompetanse Norge, Postboks 236 Sentrum, 0103 Oslo.
E-post: brikt.sagbakken@kompetansenorge.no

Referanser

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Bahrouni, F. (2016). Using multi-facet Rasch model (MFRM) in rater-mediated assessment. *Journal of Teaching English for Specific and Academic Purposes, 4*(1), 195-212.

Council of Europe (2011) *Common European Framework of Reference for Languages: learning, Teaching, Assessment*. Hentet fra: https://rm.coe.int/16802fc1bf.

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch model. *Journal of Applied Measurement, 2*(4), 379-388.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal, 2*(3), 197-221.

Eckes, T. (2008). *Rater types in writing performance assessments: A classification approach to rater variability. Language Testing, 25*(2), 155-185.
Eckes, T. (2012). O*perational rater types in writing assessment: Linking rater cognition to rater behavior. Language Assessment Quarterly, 9*(3), 270-292.
Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement – Analyzing and Evaluating Rater-Mediated Assessments"* 2nd edition. Peter Lang GmbH: Internationaler Verlag der Wissenschaften.
Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. *Psychology Press.*
Kompetanse Norge (2018) Norskprøveresultater, fordelt på fylke og kommune – norsk for innvandrere, 2014-2018, Hentet fra: http://status.vox.no/webview/?language=no
Lane, S. Stone, C.A. (2006). *Performance Assessment. In R. L. Brennan (Ed.): Educational Measurement* (s. 387-431). Wesport, CT: ACE/Praeger.
Linacre, J. M. (2017). *A user's guide to FACETS. Rasch-model computer programs. Program manua*l 3.80.1. Hentet fra: http://www.winsteps.com/a/Facets-ManualPDF.zip
McNamara, T. F. (1996). Measuring second language performance. Addison Wesley Longman.
McNamara, T. & Ryan, K. (2011). Fairness Versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. *Language Assessment Quarterly, 8*(2), 161-178.
Moe, E. & Jones, N. (2003). Using multi-faceted Rasch analysis to validate test of writing. *Acta Didactica. 1.* 110-127.
Myford, C. & Wolfe, E. W. (2003) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of applied measurement, 4*(4), 386-422.
Myford, C. & Wolfe, E. W. (2004) Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. Journal of applied measurement, 5(2), 189-227.
Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Oxford, England: Nielsen & Lydiche.
Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.
Skar, G. B. U. (2017). *The Norwegian National Sample-Based Writing Test 2016: Technical Report*. Skrivesenterets skriftserie 2.
Skar, G. B. & Jølle, L. J. (2017). Teachers as raters: Investigation of a long term writing assessment program. *L1 Educational Studies in Language and Literature, 17* (Open Issue).
Vox (2012) Læreplan i norsk og samfunnskunnskap for voksne innvandrere. Hentet fra: http://www.kompetansenorge.no/contentassets/f6594d5dde814b7bb5e9d2f4564ac134/laereplan_norsk_samfunnskunnskap_bm_web.pdf
Weigle, S. C. (1998) Using FACETS to model rater training effects. *Language Testing 15* (2), 263-287.
Weigle, S. C. (2002). *Assessing writing.* Cambridge, UK: Cambridge University Press.
Wolfe, E. W. (2004) Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35-51.
Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*, 339–355.

Vedlegg

Vurderingsskjema for Norskprøven, delprøve i skriftlig framstilling, inkludert forklaring til kriteriene.
Lenke**:** **https://www.kompetansenorge.no/contentassets/3e8bccee0dad40a3ab69a8b122f89d46/vurderingsskjema_skriftlig_a1_a2_b1_b2.2_bm.pdf**