MASTEROPPGAVE

Samfunnsernæring

2013

Rasch-analyse av data fra et spørreskjema om hvordan helsesøstre oppfatter ulike brukergruppers funksjonelle og interaktive nutrition literacy

Erik Resaland

Fakultet for helsefag

Institutt for helse, ernæring og ledelse

HØGSKOLEN I OSLO
OG AKERSHUS

Forord

Det har vært en lærerik, spennende og tidvis krevende prosess å arbeide med masteroppgaven. Det er mange personer som fortjener en takk i forbindelse med arbeidet av denne oppgaven. Først vil jeg takke min hovedveileder førsteamanuensis dr. scient., Kjell Sverre Pettersen for all tilbakemelding og veiledning, og ikke minst for at jeg fikk muligheten til å skrive om SOMAH-prosjektet. Den største takken må imidlertid rettes til min biveileder førsteamanuensis dr. scient., Øystein Guttersrud. Takk for all hjelp og veiledning med Rasch-problematikk i forbindelse med både gjennomføring av Rasch-kurs, og oppgaveskriving. Din kunnskap har vært uvurderlig!

Jeg vil også gi en stor takk til min familie: min kjære Kjersti, du har vært en fantastisk støttespiller for meg gjennom hele masterprosessen. Og takk for din innsats med korrekturlesing. Til slutt men ikke minst vil jeg takke min herlige lille datter Erle for at hun viser meg at livet er så mye mer enn bare en masteroppgave! - det er noe man kan glemme i en slik skriveprosess. Jeg ser frem til å ha bedre tid med dere begge.

Høsten 2013

Erik Resaland

Sammendrag

Bakgrunn: Innen samfunnsernæring henter mange forskere inn data om holdninger til og kunnskaper om kosthold og helse gjennom spørreskjemaer og kunnskapstester. Dataene har i liten grad blitt validert ved bruk av moderne testteori eller «item response theory» (IRT).
Hensikt: Formålet med masteroppgaven har vært å validere data fra spørreskjemaet benyttet i det NFR-støttede forskningsprosjektet «Samtaler om mat på helsestasjon» (SOMAH). Dataene ble testet opp mot de to ulike parametriseringene (partial credit og rating scale) av den endimensjonale Rasch-modellen for polytome data (PRM). I tillegg ble faktoranalyse og reliabilitetsanalyser fra klassisk testteori (KTT) benyttet.
Metode: I 2010 besvarte 666 helsesøstre som arbeidet ved tilfeldig utvalgte helsestasjoner i ulike deler av Norge spørreskjemaet (62 % svarte på og returnerte skjemaet). Helsesøstrene ble bedt om å ta stilling til påstander vedrørende deres oppfattelse av den funksjonelle (FNL) og den interaktive (INL) ernæringsfremmende allmenndannelsen til de som bruker helsestasjonene. Holdningsnivået til helsesøstrene på disse to konstruktene ble målt ved bruk av fem-delte Likertskalaer (1 = helt uenig; 5 = helt enig).
Resultater: Dataene fra konstruktene hadde noe svak tilpasning til PRM på et overordnet nivå. Påstandenes affektive nivåer kunne vært bedre tilpasset holdningsnivået til helsesøstrene («targeting»), noe som påvirket standard målefeil til estimatene av påstandenes affektive nivåer. Antall personer med svak tilpasning til «Guttman-strukturen» varierte noe mellom de ulike analysene.

Enkelte påstander underdiskriminerte eller skilte for svakt mellom helsesøstre med høyt og lavt holdningsnivå. Siden de fleste påstandene hadde «ordnete svarkategorier», virket den fem-delte Likertskalaen relativt godt. Enkelte uordnete svarkategorier ble observert i forbindelse med skalaenes «nøytrale» midtkategori (svarkategori 3).

Videre ble det observert brudd på lokal uavhengighet gjennom svaravhengighet. Den kulturelle bakgrunnen til brukerne av helsestasjonene synes å påvirke hvordan helsesøstrene tolket og svarte på enkelte påstander («differential item functioning» DIF). Spørreskjemaet har følgelig enkelte utfordringer når det gjelder å måle invariant på tvers av brukernes kulturelle bakgrunn. Estimatene av den nedre grensen for reliabilitet (Cronbach's alpha) kan tyde på målinger med tilstrekkelig reliabilitet, men reliabilitetskoeffisientene kan fremstå som noe høye grunnet svaravhengighet.
Konklusjon: Underdiskriminerende påstander bør omskrives og prøves ut på nytt. Eventuelle revisjoner av spørreskjemaet bør bidra til å justere konstruktenes affektive nivåer for å bedre tilpasningen til helsesøstrenes holdningsnivå. Videre antas det at bruk av fire-delte Likertskalaer kan løse utfordringene knyttet til uordnete svarkategorier. FNL-konstruktet ser ikke ut til å måle invariant på tvers av brukernes kulturelle bakgrunn. De empiriske dataene er dermed ikke (direkte) tilstrekkelig valide til å teste hypoteser om ulike brukergruppers ernæringsfremmende allmenndannelse. INL-konstruktet ser derimot ut til, i modifisert form, å gi data som kan brukes til å beskrive slike forskjeller. Denne studien viser at Rasch-analyser av data fra utprøvinger av kvantitative måleinstrumenter har et stort potensiale hva angår kvalitetssikring, justering og eventuell videreutvikling av instrumentene *før* hovedinnsamling av data igangsettes.

Abstract

Background: The use of questionnaires is a common method to collect data about people's attitudes towards, and knowledge about health and nutrition, for researchers within the field of public health nutrition. The use of modern test theory or item response theory (IRM) as a validation method of quantitative data within this research field is limited.

Objective: The objective of this study is to validate data from a health related questionnaire used in The SOMAH-project – "Developing culture sensitive health and nutrition communication for Mother and Child Health Clinics"**.** The data were tested against two parameterizations of the unidimensional polytomous Rasch-model (PRM), the "rating scale" and the "partial credit". In addition, factor analysis and reliability analysis from KTT were done.

Method: 666 public health nurses from randomly selected health care centers in Norway (response rate 62%), answered in 2010 the questionnaire used in The SOMAH-project. The questionnaire collected data about public health nurses perception of the functional and interactive nutrition literacy level (FNL and INL) of pregnant women and mothers attending public health care centers. All items measuring FNL and INL had a Likert-five-point scale response format with the categories 1: strongly disagree … 5: strongly agree.

Results: Data from the constructs showed limited overall-fit to the PRM. The targeting of the constructs could have been better, which affected the standard error of the ability and difficulty estimates. The number of misfitting persons to the Gutman-structure did vary across different analysis.

Individual fit-analysis revealed both under-discriminating and over-discriminating items. Most items had ordered response categories indicating that the Likert-five-point response format was working as intended. Issues with the response categories were located to the "neutral" mid-category (neither agree nor disagree).

Some items might have response dependence indicating that they not are statistically independent of each other. The cultural background of minority users seemed to affect the answers from the public health nurses in some items ("differential item functioning" DIF). Hence, the instrument does not seem to measure invariant transversely of different cultural backgrounds. The internal consistence reliability of the FNL and INL constructs was adequate. However, response dependence in the data might enhance alpha in some constructs.

Conclusion: Underdiscriminating items should to be re-written and re-tested to a comparable sample, before being used in further data collection. Replacement of new items should contribute to improve the constructs overall targeting. The use of Likert-four-point response format would probably solve the issues with the disordered response categories.

There is no psychometric "evidence" that the FNL construct are collecting reliable and valid data to test hypotheses about, and to compare FNL of different groups. Hence, it needs further improvement if it is to measure invariantly across different cultural backgrounds.

However, the modified INL constructs seem to be able to collect data to compare the INL of different groups.

This study shows that Rasch-analysis of attitude data from pilot collection or psychometric instruments have a great potential for validating, adjusting and developing instruments before initiating main data-collection.

Begrepsavklaring

Dikotome data: Oppgaver med to svaralternativ, for eksempel ja og nei

Polytome data: Påstander med mer enn to svaralternativ, for eksempel Likert-skala

Latent trekk: De direkte observerbare variablene (svarene på påstandene, spørsmålene eller oppgavene) blir, gjennom en matematisk latent-trekk modell, «redusert» til én latent variabel kalt et underliggende «trekk» i dataene. Den matematiske latent-trekk modellen estimerer dermed sannsynligheten for ulike typer svar basert på personestimatet. Den enkleste matematisk latent-trekk modellen er den endimensjonale Rasch-modellen for bivariate/dikotome data kalt «simple logistic model» (SLM). Eksempler på latente trekk er engasjement i kosthold på et personlig, sosialt og globalt nivå (holdning) og «kompetanse til å sette sammen dietter med et variert kosthold basert på helsetilstand og energibehov» (kompetanse).

SLM: Simple Logistic Model – endimensjonal Rasch-modell for dikotome data

PRM: Polytom endimensjonal Rasch Modell, bygger på SLM og finnes i to parametriseringer, Rating Scale Model (RSM) og Partial Credit Model (PCM)

Dyktighet/holdningsnivå: Estimat av personers «plassering» langs skalaen som beskriver det underliggende trekket

Vanskegrad/affektivt nivå: Estimat av oppgavenes, spørsmålenes eller påstandenes «plassering» langs skalaen som beskriver det underliggende trekket

Targeting: Forholdet mellom fordelingen av personenes dyktigheter og oppgavenes, spørsmålenes eller påstandenes vanskegrader

Klasseintervall: Gruppering av personer basert på holdningsnivå eller dyktighet

Fit-residual: Avviket til en observert verdi fra Rasch-modellens forventningsverdi (teoretisk verdi)

ICC: Item characteristic curve grafisk representasjon av modellen. Denne fremstiller sannsynligheten for ulike typer svar som funksjon av personestimat (holdningsnivå eller dyktighet)

Diskriminering: Påstander, spørsmål eller oppgavers evne til å separere personer med lavt og høyt holdningsnivå eller lav og høy dyktighet

Overdiskriminering: Påstander, spørsmål eller oppgaver som, målt opp mot modellens forventninger, skiller «for godt» mellom grupper av personer med lavt og høyt holdningsnivå eller lav og høy dyktighet

Underdiskriminering: Påstander, spørsmål eller oppgaver som, målt opp mot modellens forventninger, ikke skiller tilstrekkelig mellom grupper av personer med lavt og høyt holdningsnivå eller lav og høy dyktighet

Lokal uavhengighet: Prinsippet om at påstander, spørsmål eller oppgaver skal gi relatert, men uavhengig informasjon, eller relevant, men ikke overflødig informasjon. Med uavhengighet menes at korrelasjonene mellom dataene fra de direkte observerbare variablene (påstandene, spørsmålene eller oppgavene) kan forklares av den ikke-observerbare «reduserte» variabelen (det underliggende trekket) «holdningsnivå» eller «dyktighet». Lokal uavhengighet i dataene kan bli brutt på hovedsakelig to måter, referert til som trekkavhengighet (flerdimensjonalitet) og svaravhengighet (statistisk avhengighet).

Trekkavhengighet: Det fenomenet at påstander, spørsmål eller oppgaver måler noe annet «utover» det underliggende trekket.

DIF: Påstander, spørsmål eller oppgaver som fungerer ulikt for personer med samme dyktighet eller holdning, men som tilhører ulike personfaktorkategorier. For eksempel at menn og kvinner med samme dyktighet eller holdningsnivå i gjennomsnitt oppfatter, tolker og svarer «signifikant» annerledes på en påstand, et spørsmål eller en oppgave. DIF er et måleteknisk problem. Hvis vi kan begrunne at en gruppe kan forventes å svare annerledes, kan imidlertid forskjellene i sannsynlighet for ulike typer svar være forårsaket av *faktiske* forskjeller mellom grupper av personer. I hvilken grad DIF representerer et problem må vurderes i hvert enkelt tilfelle.

Svaravhengighet: Når oppgaver ikke er statistisk uavhengige av hverandre, som for eksempel når svaret på en påstand, et spørsmål eller en oppgave avhenger av svaret på en/et tidligere gitt påstand, spørsmål eller oppgave.

Invarians:

Noe som ikke endrer seg under transformasjon, som fra en referanseramme til en annen. Hvis svaret på en påstand, et spørsmål eller en oppgave avhenger av andre personrelaterte faktorer enn det underliggende trekket (holdningsnivå eller dyktighet), er ikke målingen invariant. Dersom for eksempel kvinner og menn i gjennomsnitt svarer «signifikant» annerledes, virker påstanden, spørsmålet eller oppgaven på ulik måte i de to referanserammene. DIF er en viktig årsak til ikke-invariante målinger av kunnskap og holdning.

HL: Health literacy eller helsefremmende allmenndannelse

FHL: Functional health literacy eller funksjonell helsefremmende allmenndannelse

IHL: Interactive health literacy eller interaktiv helsefremmende allmenndannelse

CHL: Critical health literacy eller kritisk helsefremmende allmenndannelse

NL: Nutrition literacy

FNL: Functional nutrition literacy eller funksjonell ernæringsfremmende allmenndannelse

FNL: Interactive nutrition literacy eller interaktiv ernæringsfremmende allmenndannelse

CNL: Critical nutrition literacy eller kritisk ernæringsfremmende allmenndannelse

FNLmaj: Konstruktet som inneholder påstander om helsesøstrenes oppfattelse av majoritetbrukernes FNL-nivå. Data bygger på originalt datasett

FNLmin: Konstruktet som inneholder påstander om helsesøstrenes oppfattelse av minoritetsbrkernes FN- nivå. Data bygger på originalt datasett

INLmaj: Konstruktet som inneholder påstander om helsesøstrenes oppfattelse av majoritetbrukernes INL-nivå. Data bygger på originalt datasett

INLmin: Konstruktet som inneholder påstander om helsesøstrenes oppfattelse av minoritetsbrukernes INL-nivå. Data bygger på originalt datasett

FNL-stacked: Konstruktet som inneholder påstander om helsesøstrenes oppfattelse av både majoritet og minoritetsbrukerne sitt FNL-nivå. Data bygger på «stacked» datasett

INL-stacked: Konstruktet som inneholder påstander om helsesøstrenes oppfattelse av både majoritet og minoritetsbrukerne sitt INL-nivå. Data bygger på «stacked» datasett

CCA: Koeffisient Cronbach's alpha, som er et estimat for nedre grense for reliabilitet gitt endimensjonale data. Beregnet på grunnlag av råskår.

PSI: Person separasjonsindeks, som er CCA beregnet ved bruk av personestimater

Figuroversikt

Tabelloversikt

Innhold

1. Bakgrunn

Health literacy (HL) eller *helsefremmende allmenndannelse* (Pettersen, 2003) kan beskrives som den evne en person har til å søke etter, lese, tolke og vurdere informasjon i et helseperspektiv (Finbråten & Pettersen, 2009b; Nutbeam, 2000). Forskere operer gjerne med de tre dimensjonene funksjonell (FHL), interaktiv (IHL) og kritisk (CHL) HL.

I denne oppgaven tolkes imidlertid disse som tre ulike *kontekster* for å anvende i) grunnleggende ferdigheter, som lesing, skriving, regning, muntlige ferdigheter og digitale ferdigheter, ii) kunnskaper i og om helse og ernæring, som kunnskaper om hvilke faktorer som påvirker helsen, og iii) kunnskaper om naturvitenskapens egenart, som kunnskap om hva som kjennetegner vitenskapelig kunnskap, hvordan forskere jobber og hvordan vitenskapelig kunnskap oppstår og utvikler seg. Begrunnelsen for denne tolkningen er at det per i dag ikke foreligger evidens for at HL er et flerdimensjonalt begrep hvor FHL, IHL og CHL representerer ulike «literacies» som kan utvikles mer eller mindre uavhengig av hverandre. Det er heller ikke evidens for at FHL, IHL og CHL beskriver ulike kompetansenivåer av et endimensjonalt HL-begrep.

Vi finner igjen tilsvarende definisjon og underdimensjoner i nutrition literacy (NL) (Silk et al., 2008) eller *ernæringsfremmende allmenndannelse* (Pettersen, 2009). Her refererer funksjonell ernæringsfremmende allmenndannelse NL (FNL) til leseferdigheter som er nødvendige for blant annet å kunne lese og forstå matvaremerking, mens interaktiv ernæringsfremmende allmenndannelse NL (INL) handler om kompetanser til å søke opp informasjon og kommunisere med helsepersonell. Kritisk ernæringsfremmende allmenndannelse NL (CNL) innebærer evnen til å kritisk vurdere ernæringsinformasjon og avsendere av slik informasjon (Guttersrud, Dalane & Pettersen, 2013; Nutbeam, 2000).

Forskere innen samfunnsernæring henter inn data om ulike gruppers holdninger til og kunnskaper i og om kosthold og helse fra kvantitative holdningsundersøkelser og kunnskapstester (Azizi, Aghaee, Ebrahimi & Ranjbar, 2011; Gosse, 2012; Parmenter & Wardle, 1999). Det er derfor viktig å fremme strategier og metoder for utvikling av valide måleinstrumenter innenfor dette fagfeltet. Tradisjonelt har klassisk testteori (KTT), som faktoranalyse og Cronbach's alpha-koeffisient (CCA), blitt brukt for å beskrive dataenes validitet og reliabilitet (Belvedere & de Morton, 2010; A. Tennant & P. G. Conaghan, 2007; Wilson, Allen & Li, 2006). Moderne testteori, eller item response theory (IRT), er

et supplement til KTT for validering av kvantitative data (Masse, Wilson, Baranowski & Nebeling, 2006; A. Tennant & P. G. Conaghan, 2007).

Rasch-analyse har vært brukt av skoleforskere i mer enn 40 år (A. Tennant & P.G. Conaghan, 2007), der OECD-undersøkelsen "Programme for International Student Assessment" (PISA) (Kjærnsli, Lie, Olsen & Roe, 2007) er et eksempel fra nyere tid. I den senere tid er Rasch-analyse også blitt benyttet innen helseforskning (Masse et al., 2006; A. Tennant & P. G. Conaghan, 2007), og da særlig i forskning knyttet til rehabilitering (Belvedere & de Morton, 2010; A. Tennant & P.G. Conaghan, 2007; Tesio, Simone & Bernardinello, 2007). Søk og gjennomgang av publikasjoner i velkjente databaser, som PubMed, Cochrane og Google Scholar, tyder imidlertid på at Rasch-modeller i liten grad har vært benyttet ved validering av spørreskjemadata brukt i ernæringsstudier.

Forskningsprosjektet *"Samtaler om mat på helsestasjonen* (SOMAH)" har til hovedhensikt å beskrive og bedre den kommunikasjonen om kosthold og ernæring som finner sted mellom helsesøstrene og gravide og spedbarnsmødre med ulik kulturell bakgrunn på helsestasjoner i Norge. Prosjektet la særlig vekt på brukere med ikke-vestlig bakgrunn. Prosjektet hadde som mål å utvikle kultursensitive kommunikasjonsverktøy om ernæring til bruk på helsestasjonene (SOMAH, 2013). Gjennom den kvantitative delen av SOMAH-prosjektet (delprosjekt 2 kalt DP2), ble det hentet inn spørreskjemadata fra 666 helsesøstre som arbeidet ved helsestasjoner spredt rundt i landet (svarprosent lik 62 %). Spørreundersøkelsen samlet inn data om hvordan helsesøstre oppfatter kommunikasjonen om kosthold og ernæring med mødre med spedbarn og gravide kvinner med ulik kulturell bakgrunn. Spørreundersøkelsen dreide seg blant annet om hvordan helsesøstrene oppfatter disse brukergruppenes FNL- og INL-nivå.

1.1 Problemstilling

Formålet med masterprosjektet er å teste dataene fra DP2 opp mot den polytome endimensjonale Rasch-modellen (PRM) og undersøke hvor fri dataene er for systematiske feil (validitet) og tilfeldige feil (reliabilitet). Spesifikt søker masteroppgaven å gi svar på følgende forskningsspørsmål:

1. I hvilken grad gir spørreskjemaet i SOMAH-DP2 valide og reliable data om helsesøstres persepsjon av den ernæringsfremmende allmenndannelsen til gravide og spedbarnsmødre som bruker helsestasjoner?

2. I hvilken grad måler spørreskjemaet i SOMAH-DP2 invariant på tvers av brukernes kulturelle bakgrunn?

1.2 Avgrensninger

Beskrivelsene av Rasch-modellene er ment for lesere med begrenset forkunnskap om emnet. Inngående matematiske beskrivelser av modellenes oppbygning er derfor *ikke* prioritert i oppgaven. Ulike teknikker for å estimere personers holdningsnivå og påstanders affektive nivå ligger utenfor omfanget av masteroppgaven. Det samme gjelder analyse av dimensjonalitet i dataene (trekkavhengighet/multidimensjonalitet).

2. Bakgrunn og teori

I dette kapittelet blir først det NFR-støttede forskningsprosjektet "Samtaler om mat på helsestasjonen" (SOMAH) kort beskrevet. Videre følger en beskrivelse av det teoretiske grunnlaget for HL og NL og en gjennomgang av tidligere forsøk på å måle disse størrelsene. Hoveddelen av kapittelet handler imidlertid om Rasch-analyse.

Rasch-analyse blir av praktiske årsaker diskutert i lys av skoleforskning, og da med utgangspunkt i dikotome data knyttet til måling av kunnskaper og ferdigheter. Prinsippene kan generaliseres til validering av polytome data fra spørreskjemaundersøkelser, og prinsippene har også overføringseffekt til andre fagområder enn skoleforskning, som for eksempel ernæring og folkehelse.

2.1 SOMAH-prosjektet og kultursensitivitet

SOMAH-prosjektet er utviklet og ledet av forskere ved Institutt for samfunnsernæring, Avdeling for helse, ernæring og ledelse ved Høgskolen i Oslo og Akershus. Prosjektet har flere delprosjekter. Det er bare datamaterialet fra den kvantitative spørreskjemaundersøkelsen i delprosjekt 2 (DP2) som danner datagrunnlaget for denne masteroppgaven.

SOMAH-DP2 tar primært for seg å kartlegge hvordan helsesøstre og jordmødre kommuniserer om kosthold og ernæring med gravide og spedbarnsmødre med ulik kulturell bakgrunn (SOMAH, 2013). Hovedformålet med SOMAH-DP2 er å utvikle *kultursensitive* og «*health/nutrition literacy*»-tilpassede billedlige og skriftlige kommunikasjonsverktøy som helsesøstre og jordmødre kan bruke når de kommuniserer med brukergrupper med ulik kulturell bakgrunn.

Kultursensitivitet i ernæringskommunikasjon dreier seg om det å gi råd om et sunt kosthold på en slik måte at det appellerer til og når frem til personer med ulike matkulturer (Garnweidner-Holme, 2013). Kultursensitiv helse/ernæringskommunikasjon innebærer både å tilpasse materiell til observerbare karakteristikker ved målgruppene, og å innhente individuelle og kulturelle karakteristikker som kan påvirke valg av matvarer (Garnweidner-Holme, 2013; Resnicow, Baranowski, Ahluwalia & Braithwaite, 1999). Det anses derfor som nyttig å kartlegge FNL- og INL-nivåer hos brukerne av helsestasjonene, slik at kommunikasjonsverktøyene kan tilpasses en slik mulig variasjon av NL.

2.2 Kort om bakgrunnen for SOMAH-prosjektet

På verdensbasis har prevalensen av livsstilsykdommer som for eksempel diabetes type 2 økt de siste 30 årene (Helsedirektoratet, 2011). I Norge og Norden ser vi en lignende utvikling særlig blant personer med ikke-vestlig bakgrunn (Jenum A, Holme, Graff-Iversen & Birkeland, 2005). En svensk studie har for eksempel vist at innvandrere fra Tyrkia som er bosatt i Sverige har dobbelt så høy prevalens av diabetes type 2 som befolkningen i Tyrkia forøvrig (Hjörleifsdóttir, 2013). Utfordringer knyttet til endringer i kosthold ved migrasjon har vært brukt som forklaringsfaktor (Hjörleifsdóttir, 2013).

Videre er prevalensen av svangerskapsdiabetes i Norge økende (Larsen, 2000). Økningen er større hos gravide kvinner med ikke-vestlig bakgrunn enn blant gravide kvinner med vestlig bakgrunn (Abebe, 2010; Jenum et al., 2012; Larsen, 2000). I debatten om et likeverdig helsevesen trekkes det frem at det er viktig å forbedre kommunikasjonen mellom helsepersonell og brukergrupper med ulik kulturell bakgrunn og/eller personer med lav health literacy (HL) og lav nutrition literacy (NL) (beskrives nærmere i neste kapittel) (Easton, Entwistle & Williams, 2013; Naqv, 2012). Helsesøstre tilsatt ved helsestasjoner har en sentral rolle i formidlingen av helseinformasjon overfor ulike familier (Sosial- og helsedirektoratet, 2003). Siden nærmest 100 % av alle foreldre med små barn benytter seg av tilbudet om helseundersøkelser og vaksinasjonsprogram (Helsedirektoratet, 2010), har tjenesten en unik posisjon i forhold til å nå ut med helseinformasjon til alle barnefamilier uavhengig av sosial tilhørighet eller kulturell bakgrunn.

Helsesøstre ved helsestasjoner må imidlertid være oppmerksomme på at ulike brukergrupper kan ha forskjellige forutsetninger for å oppfatte og forstå helseinformasjonen som de mottar (Easton et al., 2013; Finbråten & Pettersen, 2012; Garnweidner-Holme, 2013). Dette kan skyldes brukernes HL-nivå (Easton et al., 2013; Paasche-Orlow, Parker, Gazmararian, Nielsen-Bohlman & Rudd, 2005), og det er påvist sammenhenger mellom HL og sosioøkonomisk status (Ishikawa, Takeuchi & Yano, 2008; Paasche-Orlow et al., 2005). Andre årsaker kan være at grupper med ulik kulturell bakgrunn har ulike utfordringer med å finne frem til, tolke og forstå informasjon om helse, og at de kan ha problemer med å orientere seg i helsevesenet i det landet de ankommer (Kreps & Sparks, 2008). Forskere har blant annet påpekt at innvandrere fra Pakistan til Norge har vansker med å forstå informasjon om diabetesforebyggende kosthold (Hjellset, 2010).

SOMAH-prosjektet har hentet inn kvalitative data gjennom intervjuer av gravide norske og gravide kvinner med ikke-vestlig innvandrerbakgrunn som har benyttet helsestasjonen under svangerskapet (Garnweidner, Pettersen & Mosdol, 2013). Kvantitative data er også innhentet gjennom fokusgruppeintervjuer med helsesøstre tilsatt ved helsestasjoner (foreløpig upublisert studie). Foreløpige analyser av datamaterialet peker mot at helsesøstre opplever at kommunikasjonen med personer av ikke-vestlig opprinnelse har særlige utfordringer (Helsesøster og PhD.student Bettina Fagerlund, HiOA, muntlig kommunikasjon, 17. oktober 2013). Videre kan kvinner med ikke-vestlig innvandrerbakgrunn bli forvirret av kostrådene som gis av helsesøstrene fordi det ikke er samsvar med deres egne mattradisjoner eller kostrådene de har fått i hjemlandet (Garnweidner-Holme, 2013; Garnweidner et al., 2013). Det er tiltro til at kommunikasjonsverktøy som er tilpasset personers HL og NL kan bedre kommunikasjonen mellom helsesøstre og brukere, og at dette kan føre til økt forståelse av muntlig og skriftlig informasjon (Finbråten & Pettersen, 2012b). En konsekvens av dette er at vi da trenger valide og reliable måleinstrumenter for å måle og sammenligne personers HL- og NL.

2.3 Health- og nutrition literacy

HL eller *"helsefremmende allmenndannelse"* er av Nutbeam (2000) og Pettersen (2003) definert som "personlige, kognitive og sosiale ferdigheter som er avgjørende for enkeltindividets evne til å få tilgang til, forstå, og anvende helseinformasjon som fremmer og ivaretar god helse". HL er foreslått delt inn i tre «dimensjoner» som viser til ulike «literacies» (Nutbeam, 2000). Som tidligere omtalt kan dette imidlertid dreie seg om ulike *kontekster* for å anvende grunnleggende ferdigheter, kunnskaper om naturvitenskapens egenart og kompetanse innen helse og ernæring.

Dimensjonen *funksjonell helsefremmende allmenndannelse* (FHL) handler om lese- og skriveferdigheter, og kunnskap om kroppen som er nødvendig for å kunne tolke og anvende helseråd. Dimensjonen *interaktiv helsefremmende allmenndannelse* (IHL) handler om evnen til å kunne hente inn kunnskap fra ulike kilder og aktivt anvende dem i interaksjon med for eksempel helsearbeidere i ulike situasjoner og kontekster. Evne til å orientere seg om, og finne frem i helsevesenet hører også til denne dimensjonen. Dimensjonen *kritisk helsefremmende allmenndannelse* (CHL) handler om å kunne forholde seg kritisk til informasjon og avsendere av informasjonen (Finbråten & Pettersen, 2012; Nutbeam, 2000).

Siden HL er et relativt bredt definert begrep, er det antatt at en persons HL-nivå kan variere mellom ulike kontekster (Sørensen, 2012). I ernæringssammenheng kan vi avgrense begrepet og snakke om *nutrition literacy* (NL) (Silk et al., 2008), eller *ernæringsfremmende allmenndannelse* (Pettersen, 2009).

Silk et al. (2008) har definert NL som "the capacity to obtain, process and understand nutrition information and the materials needed to make appropriate decisions regarding one's health". Definisjonen er nært forbundet med den definisjonen vi finner av HL hos Nutbeam (2000). Underdimensjonene av HL er antatt å ha høy overføringseffekt til NL (Dalane, 2011; Pettersen, 2009; Silk et al., 2008), hvor *funksjonell* (FNL), *interaktiv* (INL) og *kritisk* (CNL) ernæringsfremmende allmenndannelse nærmest kan oppfattes som enhetsvektorene i en tredimensjonal basis som spenner ut et underrom hvor forskning innen samfunnsernæring kan projiseres.

FNL handler om lese- og skriveferdigheter, og kunnskap om menneskekroppen til å tolke og anvende budskapet i ulike typer informasjon om ernæring. INL omfatter de kompetanser som er nødvendige for å kunne søke etter informasjon om ernæring med hensikt å bedre egen helse og kommunisere med helsepersonell om ernæring. Kritisk ernæringsfremmende allmenndannelse (CNL) handler om de kompetanser som er nødvendige for å kunne forholde seg kritisk til informasjon og råd om ernæring (Nutbeam, 2000), og engasjere seg i ernæringsutfordringer i personlige, sosiale og globale kontekster (Guttersrud et al., 2013; Nutbeam, 2000). Guttersrud et al. (2013) studerte mulige underdimensjoner av CNL som målte personers engasjement i kosthold og ernæringsutfordringer, og hvordan personer oppfatter sin egen evne til å kritisk vurdere informasjon om ernæring og kosthold. Siden CNL omfatter evne til å trekke evidensbaserte konklusjoner, har CNL sammenheng med begrepet *scientific literacy*, eller "*naturvitenskapelig allmenndannelse*" (Pettersen, 2007) som beskrevet i rammeverket til PISA (Kjærnsli et al., 2007; Kjærnsli & Roe, 2009).

2.4 Måling av health- og nutrition literacy

Begrepet HL dukket opp i forskningslitteraturen på 1970-tallet, og begrepet har siden vært forsøkt målt ved kunnskapstester og spørreskjemaundersøkelser (Sørensen et al., 2012) både nasjonalt og internasjonalt (Doyle, Cafferkey & Fullam, 2012; Finbråten & Pettersen, 2009a, 2012; Ishikawa, Takeuchi, et al., 2008; Steckelberg, Hülfenhaus, Kasper, Rost & Mühlhauser, 2009). FHL har særlig blitt målt gjennom kunnskapstester som består av regne-, lese- og ordgjenkjenningsoppgaver (Davis et al., 1991; Hanson-

Divers, 1997; Parker, Baker, Williams & Nurss, 1995; Weiss et al., 2005). IHL og CHL har i den senere tid blitt målt gjennom spørreundersøkelser både internasjonalt (Doyle et al., 2012; Ishikawa, Nomura, Sato & Yano, 2008; Ishikawa, Takeuchi, et al., 2008; Steckelberg et al., 2009) og nasjonalt (Finbråten & Pettersen, 2009a, 2012).

Basert på et måleinstrument for HL (Parker et al., 1995), utviklet Diamond (2007) et nytt instrument for måling av NL som ble oversatt til norsk (NLQ) og anvendt i ulike målgrupper (Aarnes, 2009; Kjøllesdal, 2009). Utprøving av instrumentet tydet på at NLQ kunne ha potensiale til å måle personers FNL, INL og CNL, men konstruktene ble imidlertid bare analysert ved bruk av KTT. I en tilsvarende studie målte (Dalane, 2011) INL og CNL hos norske sykepleierstudenter. Konstruktene som (Dalane, 2011)benyttet ble i ettertid validert ved Rasch-analyse av (Guttersrud et al., 2013). Studien avdekket mulige problemer med bruk av Likert-skalaer med nøytrale midtkategorier da dette ga uordnete svarkategorier, og at dataene fra spørreskjemaet hadde «svak tilpasning» til Rasch-modellen. Hva som menes med svak tilpasning vil bli omtalt senere.

2.5 Om analyse av kvantitative data fra holdningsundersøkelser og summative pedagogiske målinger

Følgende kapittel handler om analyse av kvantitative data hentet inn gjennom holdningsundersøkelser og summative pedagogiske målinger, der målet er å rangere personer basert på estimater av holdningsnivå og dyktighet. Summative pedagogiske målinger, heretter referert til som kunnskapstester, blir typisk brukt ved «sluttvurdering», men kan også ha formative egenskaper. Summative målinger kan måleteknisk og til dels innholdsmessig skille seg fra tester av mer diagnostisk art utviklet for formativ vurdering eller «underveisvurdering». Hensikten med analysene av summative målinger er å kunne gi begrunnede vurderinger av instrumentenes evne til å fremskaffe valide og reliable data fra de populasjonene som instrumentet er designet for.

Noen kunnskapstester er designet for å måle kompetanse til en gruppe ved et gitt tidspunkt, som for eksempel avsluttende eksamen i videregående skole. Andre følger utvikling av kompetanse for tilsvarende utvalg over tid, slik som PISA-undersøkelsen som blir gjennomført hvert tredje år. PISA-undersøkelsen måler 15-åringers kompetanse i lesing, matematikk og naturfag (Kjærnsli et al., 2007). Noen kunnskapstester er videre designet for å følge utvikling av kompetanse hos én og samme gruppe av personer over tid, som matematikkunnskapene til en bestemt skoleklasse fra 1. til 7. årstrinn (Looveer

& Mulligan, 2009). Den sistnevnte typen av tester vil typisk involvere linkeprosedyrer (Looveer & Mulligan, 2009) – ofte referert til som «test equating and linking» – mellom tilsvarende tester utviklet for ulike årstrinn. Testene består da av forskjellige oppgaver, og testene har ulik vanskegrad tilpasset de ulike alderstrinnene. Testene har likevel et felles sett oppgaver – «anchor test» – som gjør det mulig å måle endring over tid ved at elevene på hver prøve tilegnes et dyktighetsestimat langs den *samme* skalaen.

Holdningsundersøkelser og internasjonale kunnskapstester har i enkelte tilfeller vært gjenstand for diskusjon og negativ kritikk (Bond & Fox, 2007; DeVellis, 2012). For eksempel har PISA-undersøkelsen blitt kritisert for ikke å måle hele naturfaget (Sjøberg, 2009), og at måledata ikke har tilstrekkelig tilpasning til valgt modell (Comins, Brodersen, Krogsgaard & Beyer, 2007; Forskning.no, 2011; Kreiner, 2011).

Siden testresultater kan ha store konsekvenser for både enkeltindivider (eksamen) og beslutninger på ulike nivåer, er det viktig å vurdere om instrumentene representerer valide og reliable målinger av de personene som en ønsker å måle. For SOMAH-prosjektet vil det for eksempel være viktig å ha tilstrekkelig valide og reliable data på brukernes ulike «grad» av FLN og INL. Dette fordi det vil danne grunnlag for utviklingen av HL og NL-differensierte *kultursensitive ernæringskommunikasjonsverktøy* til bruk ved helsestasjoner.

2.6 Validitet

Validitet er et uttrykk for hvor fri dataene er for systematiske feil (Frisbie, 1988). På et overordnet plan kan vi si at validitet handler om i hvilken grad empirisk evidens og rammeverk understøtter tolkning og bruk av resultater (Messick, 1989a, 1989b). For eksempel blir resultatene fra en avsluttende eksamen i videregående skole *tolket* som skoleflinkhet og *brukt* som seleksjonsmekanisme for høyere utdanning. Innholdsrelatert, kriterierelatert og konstruktrelatert evidens er ulike aspekter som inngår i begrepet validitet (Messick, 1989a). Inndeling i disse tre kategoriene synes å bli anvendt i stadig mindre grad, og at en oftere omtaler validitet på et overordnet plan.

Innholdsrelatert evidens er nært forbundet til det underliggende trekket instrumentet er designet for å måle (DeVellis, 2012). Oppgavene som inngår i instrumentet må innholdsmessig reflektere det underliggende trekket. I arbeidet med å utvikle relevante oppgaver er det ikke uvanlig at eksperter på området evaluerer oppgavene og at oppgavene blir knyttet til teoretiske rammeverk (DeVellis, 2012; Messick, 1989a). For eksempel kan eksamensoppgaver bli kategorisert i henhold til

kompetansemål i en lærerplan (rammeverk). På denne måten øker sannsynligheten for at eksamensoppgavene er relevante, og at oppgaven tester ulike nivåer av de kompetansene som rammeverket beskriver. På samme måte øker sannsynligheten for at holdningspåstandene er relevante når de knyttes til etablerte teorier om HL og NL.

Kvalitative forundersøkelser, som person- og fokusgruppeintervjuer, kan bidra til å i større grad utvikle relevante oppgaver eller holdningsutsagn (DeVellis, 2012; Messick, 1989a). Som forundersøkelse for utvikling av holdningspåstandene som operasjonaliserte FNL og INL i SOMAH-DP2, ble det gjennomført fem fokusgruppeintervjuer med totalt 26 helsesøstre som arbeidet ved helsestasjoner i det sentrale Østlandsområdet (Fagerlund & Pettersen, upublisert manuskript).

Konstruktrelatert evidens har tradisjonelt blitt definert som de empiriske holdepunktene vi har for å hevde at en test måler det den er utviklet for å måle (Brown, 2000). Begrepet kan defineres som: "[Construct validity is based on] an integration of any evidence that bears on the interpretation or meaning of the test scores" (Messick, 1989a s.7), slik at konstruktrelatert evidens dreier seg om en akkumulering av empirisk evidens hvor både den innholds- og kriterierelaterte evidensen inngår (Brown, 2000). I følge Messick (1989a) svekkes den konstruktrelaterte evidensen når oppgaver i en test ikke måler alle relevante underdimensjoner av trekket, eller når oppgavene måler noe annet i tillegg til det underliggende trekket (DeVellis, 2012; Messick, 1989a).

Kriterierelatert evidens handler om i hvilken grad resultatene fra to tester som er ment å måle det samme korrelerer med hverandre (DeVellis, 2012; Messick, 1989a). Hvis korrelasjonskoeffisienten er høy indikerer det at de to testene rangerer personer på tilsvarende måte (Scott & Mazhindu, 2009). Ved å beregne korrelasjonen mellom to tester gjort i nærliggende tidsrom kan det kalles «samtidig validitet», mens «prediktiv validitet» beregnes ved å korrelere for eksempel avsluttende eksamen med en inntaksprøve. Dette gir et mål på om opptaksprøven er egnet til å finne frem til de kandidatene som egner seg for et bestemt studium.

2.7 Reliabilitet

Reliabilitet er et mål på hvor fri dataene er for tilfeldige feil (DeVellis, 2012; Frisbie, 1988). Tilfeldige feil reduserer hvor konsistent eller pålitelig ulike tester måler samme størrelse, siden tilfeldige feil påvirker totalskår på ulike måter (Traub & Rowley, 1991).

Test og re-test er en metode for å etterprøve dataenes reliabilitet (Scott & Mazhindu, 2009). Test og re-test blir utført ved at samme utvalg gjennomfører to tester

som er ment å måle samme konstrukt eller underliggende trekk, for deretter å korrelere resultatene fra de to testene med hverandre (Scott & Mazhindu, 2009). Vi forventer da at personer med høy dyktighet oppnår høy skår både på test og på re-test (Traub & Rowley, 1991). Pearson korrelasjonskoeffisient er et mye brukt mål på slik «ytre konsistensreliabilitet» (Scott & Mazhindu, 2009). Reliabilitet er nært forbundet med kriterierelatert evidens, som handler om i hvilken grad resultatene fra to tester korrelerer med hverandre (DeVellis, 2012).

Et alternativ til test og re-test er å dele en test i to og korrelere resultatene fra de to deltestene med hverandre. Denne metoden er kjent som "*split-half metoden*" (Bond & Fox, 2007; DeVellis, 2012; Traub & Rowley, 1991). For å estimere reliabiliteten til hele testen basert på estimatet av reliabiliteten til en av de to deltestene, kan vi bruke Spearman-Brown formelen (DeVellis, 2012; Traub & Rowley, 1991):

$$P_{nn} = \frac{kP_{xx}}{1+(k-1)P_{xx}} \qquad (1)$$

Her betegner P_{nn} et estimat av reliabiliteten til den forlengede testen (hele testen), P_{xx} er et estimat av reliabiliteten til den originale testen og k er antall oppgaver i den forlengede testen (Traub & Rowley, 1991). Normalt vil estimatet av reliabiliteten øke når antall oppgaver i testen øker (Traub & Rowley, 1991).

Indre konsistens reliabilitet oppgis ofte gjennom estimatet av den nedre grensen for reliabiliteten til dataene kalt Cronbach's koeffisient alpha (CCA) (DeVellis, 2012). Denne koeffisienten er et mål på homogeniteten til dataene fra oppgavene i en test eller påstandene i et spørreskjemakonstrukt. Høy korrelasjon mellom data fra oppgavene i en test gir høy reliabilitet målt ved CCA. Høy CCA indikerer høy sann varians i testskår og at oppgavene i en test «diskriminerer» eller skiller mellom personer med lav og høy dyktighet på liknende måte (DeVellis, 2012). CCA kan utrykkes på følgende måte (Cronbach, 1951):

$$\alpha = \frac{k}{k-1}\left(\frac{s_x^2-\sum_{i=1}^{k} s_i^2}{s_x^2}\right) \qquad (2)$$

Her er k antall oppgaver, s_x^2 er variansen til fordelingen av personenes dyktighet, og s_i^2 er variansen til fordelingen av oppgavenes vanskegrad uttrykt ved antall riktige svar. Ved å erstatte P_{xx} i Spearman-Brown formelen med CCA, kan vi estimere hvordan reliabiliteten endrer seg når antall oppgaver (k) i testen øker og synker.

Ved å erstatte poengsum med dyktighetsestimater fra Rasch-modellering av dataene kan vi beregne «person separasjonsindeks» (PSI) og benytte denne som et mål på reliabilitet (Andrich & Marais, 2012; Bond & Fox, 2007). For datasett uten «missing» og

uten ekstremverdier (høyeste og laveste oppnåelige poengsum) vil PSI og CCA gi tilnærmet samme verdi. CCA er mindre følsom for ekstremverdier enn PSI, og dette skyldes at PSI er estimert fra dyktighetsestimater som er en non-linjær transformasjon av poengsum der variansen av målefeil øker ved økende antall ekstremverdier (Linacre, 1997). CCA er derimot beregnet på grunnlag av poengsum hvor den «klassiske standard målefeilen for enkeltmålinger» er lik for alle variablene (Demars, 2010).

2.8 Kort om item response theory, latent trait analysis, endimensjonalitet og lokal uavhengighet

Utgangspunktet for *item response theory* (IRT) eller moderne testteori er at vi ikke kan avgjøre om en person vil svare riktig på en testoppgave, men at vi kan estimere sannsynligheten for om personen vil svare riktig (Ryan, 1983).

Vi skiller mellom personers dyktighet og deres observerte skår. *Latent trait analysis* (LTT) handler om å beskrive hvordan dyktighet kan forklare observert skår (Ryan, 1983). Observert skår avhenger av personers dyktighet og oppgavenes vanskegrad, eventuelt personenes holdningsnivå og påstandenes affektive nivå (Rasch, 1960). Dyktighet kan tolkes som den dominante faktoren (for eksempel matematikkunnskap) som er avgjørende for å svare riktig på en oppgave (Ryan, 1983). Hvis dyktighet i tilstrekkelig grad dominerer hvordan en personen svarer, kan dataene ha tilstrekkelig *endimensjonalitet* (Andrich, 1988). Det betyr at det underliggende trekket forklarer «all» korrelasjon mellom oppgavene. En typisk misoppfatning er å betrakte endimensjonalitet som noe «snevert» (Goldstein, 1970). Spørsmålet om dimensjonalitet er et vurderingsspørsmål ut fra *graden* av endimensjonalitet snarere enn *om* dataene er endimensjonale eller ikke.

I Rasch-modellene er det et prinsipp om at variansen i fordelingen av poengsummer på en kunnskapstest utelukkende skal kunne forklares ved dyktighet (Marais & Andrich, 2008b). Oppgaver som i for stor grad måler andre personfaktorer enn dyktighet kan forårsake dimensjonsbrudd på «lokal uavhengighet» kalt trekkavhengighet (Andrich & Marais, 2012), men dette ligger utenfor denne masteroppgaven. Oppgaver som i for stor grad måler «det samme» kan gi svaravhengighet i dataene, og svaravhengighet representerer en annen kilde til brudd på lokal uavhengighet (Marais & Andrich, 2008b). Dette blir omtalt nærmere senere i oppgaven.

2.9 Invarians

Invarians betyr at det vi studerer *ikke* endrer seg fra et referansesystem til et annet. Invarians er dermed fundamentalt for alle målinger (DeVellis, 2012). Ved måling av fysiske størrelser som lengde, forventer vi at to målebånd er «enige» med hverandre om hvilken av to personer som er lengst. På tilsvarende måte må to testoppgaver som måler deler av samme underliggende trekk være «enige» om hvilken av to personer som er dyktigst.

Vi forventer at personer med høy dyktighet oppnår høyere skår på en test enn personer med lav dyktighet, uavhengig av det utvalget av valide oppgaver som inngår i testen (Masters, 2005). Dette forutsetter at forholdet mellom vanskegradene til to oppgaver er uavhengig av dyktighetene til personene som gjennomfører testen, og at dyktighetene til personene er uavhengige av vanskegraden til oppgavene (Rasch, 1960). Rasch (1960) uttrykte dette på følgende måte: «The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparisons. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for comparison» (Rasch, 1961 s.332, sitert i Andrich & Marais, 2012). Engelhard, (2013) trekker frem følgende krav for invariante målinger (fritt oversatt):

- Estimering av personers dyktighet må være uavhengig av oppgavenes vanskegrad
- Estimering av oppgavenes vanskegrad må være uavhengig av personenes dyktighet
- Alle personer skal ha høyere sannsynlighet for å svare riktig på oppgaver med lav vanskegrad enn på oppgaver med høy vanskegrad
- Personer med høy dyktighet skal ha høyere sannsynlighet for å svare riktig på enhver oppgave enn personer med lav dyktighet

Når data har tilstrekkelig tilpasning til endimensjonale Rasch-modeller, er målingen invariant, og dataene er tilstrekkelig endimensjonale og har tilstrekkelig reliabilitet (Andrich & Marais, 2012; Engelhard, 2013).

2.10 Rasch-modeller og andre IRT-modeller

Tre-parametermodellen kan utrykkes på følgende måte (DeAyala, 2009):

$$P\{X_{iv} = 1|\beta_v, \delta_i, \alpha_i, c_i\} = c_i + (1 - c_i)\frac{e^{\alpha_i(\beta_v - \delta_i)}}{1 + e^{\alpha_i(\beta_v - \delta_i)}} \quad (3)$$

Modellen angir at sannsynligheten (*P*) for riktig svar (*x* = 1) på en dikotom oppgave (*x* er 0 eller 1) er avhengig av (i) dyktigheten (β) til person *v*, (ii) vanskegraden (δ) til oppgave *i*, (iii) diskrimineringsparameteren (α) til oppgaven og (iv) gjetteparameteren (c) til oppgaven. Gjennom gjetteparameteren søker modellen å ta hensyn til at personer kan gjette på flervalgsoppgaver. Gjetteparameteren er en nedre asymptote for *P* hvor $c = 1/n$, der *n* er antall svaralternativer på oppgaven.

Selv om det er personene og ikke oppgaven som gjetter, er altså gjetting beskrevet som en egenskap ved oppgaven. Dette kan begrunnes ut fra at enkelte oppgaver er mer utsatt for gjetting enn andre. For eksempel har enhver person i *utgangspunktet* 25 % sannsynlighet for å gjette riktig svar på en vilkårlig flervalgsoppgave med fire svaralternativer, men vi kan anta at sannsynligheten for å gjette avtar med økende dyktighet. Siden hver oppgave har en konstant verdi for c, vil derfor ikke verdien av gjetteparameteren endre seg med personenes dyktighet. Ved å sette gjetteparameteren lik null reduseres tre-parametermodellen til to-parametermodellen (2PL) som er en logistisk funksjon (DeAyala, 2009):

$$P\{X_{iv} = 1|\beta_v, \delta_i\} = \frac{e^{\alpha_i(\beta_v - \delta_i)}}{1 + e^{\alpha_i(\beta_v - \delta_i)}} \quad (4)$$

Denne modellen «tillater» også bruk av oppgaver som diskriminerer på forskjellige måter. Dyktighetsestimatet til personer vil da (i likhet med tre-parametermodellen) være påvirket av *hvilke* oppgaver som personen har svart riktig på. Personer som svarer riktig på overdiskriminerende oppgaver vil få estimert en høyere dyktighet enn personer som svarer riktig på underdiskriminerende oppgaver. Oppgaver som gir data med god tilpasning til 2PL kan dermed bryte prinsippet om invarians (jamfør punkt 1 i ovenstående liste). Ved å sette $c = 0$ og $\alpha = 1$ reduseres 2PL til *Simple Logistic Model* (SLM) (Rasch, 1960) også kalt endimensjonal Rasch-modell for dikotome data:

$$\mathrm{P}\{\mathrm{X_{iv}} = 1|\beta_\mathrm{v}, \delta_\mathrm{i}\} = \frac{\mathrm{e}^{(\beta_v - \delta_i)}}{1 + \mathrm{e}^{(\beta_v - \delta_i)}} \quad (5)$$

Siden $c = 0$ og $\alpha = 1$ er sannsynligheten for riktig svar, gitt dyktigheten (β) til personen, bare bestemt av vanskegraden (δ) til oppgaven (Bond & Fox, 2007; Rasch, 1960). Modellen «tillater» dermed bare bruk av oppgaver som gir oss data som diskriminerer på én bestemt måte ($\alpha = 1$), men til gjengjeld vil data med tilstrekkelig tilpasning til modellen oppfylle kravene til invariante målinger (Engelhard, 2013). Vi kan dermed oppnå invariante målinger ved å prøve ut oppgaver og forkaste alle oppgaver som

ikke har tilstrekkelig tilpasning til SLM. Hva som menes med «tilstrekkelig tilpasning» er utdypet senere.

2.11 Odds og forholdet mellom dyktighet og vanskegrad

Når differansen mellom dyktighet og vanskegrad (β-δ) i SLM er større enn 0 er sannsynligheten for riktig svar større enn 50 %. Når differansen er mindre enn 0 er sannsynligheten for riktig svar mindre enn 50 %. Når dyktigheten tilsvarer vanskegraden er det 50 % sannsynlighet for riktig svar (Wright & Stone, 1979). Tabell 1 illustrerer hvordan sannsynligheten for riktig svar på en dikotom oppgave med vanskegrad 0,5 log units (logits) avhenger av differansen mellom dyktighet og vanskegrad for de ulike dyktighetene -2,0, -1,0, 0,5, 1,0 og 2,0 logits. Sannsynligheten er beregnet ved hjelp av SLM (5).

Tabell 1: Hvordan sannsynligheten for riktig svar (forventningsverdi) på en dikotom oppgave avhenger av differansen mellom dyktighet og vanskegrad.

Dyktighet	Vanskegrad	Differanse	Forventet sannsynlighet for riktig svar
- 2,0	0,5	- 2,5	8 %
- 1,0	0,5	- 1,5	18 %
0,5	0,5	0	50 %
1,0	0,5	0,5	62 %
2,0	0,5	1,5	82 %

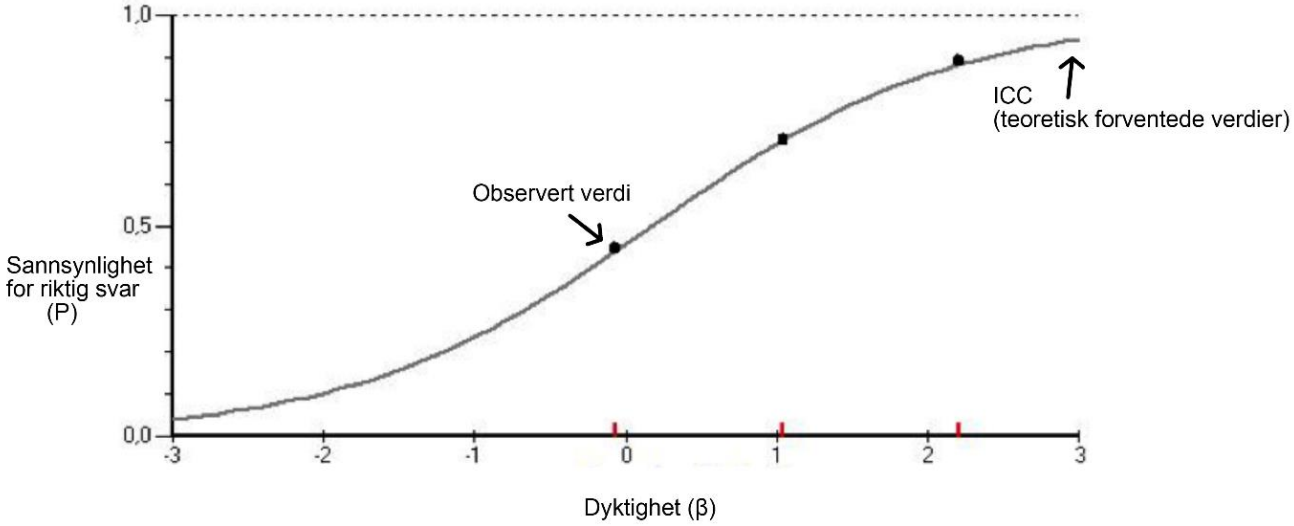
Odds er forholdet mellom sannsynligheten for riktig og feil svar. Den omtalte differansen mellom dyktighet og vanskegrad kan dermed uttrykkes som logaritmen til oddsen for at person *v* med dyktighet β_v svarer riktig på oppgave *i* med vanskegraden δ_i (Andrich & Marais, 2012):

$$Ln\frac{P\{x_{vi}=1\}}{P\{x_{vi}=0\}} = Ln\left(e^{(\beta_v-\delta_i)}\right) = \beta_v - \delta_i \qquad (6)$$

2.12 Oppgavekarakteristisk kurve – ICC

Den grafiske representasjonen av SLM blir referert til som *item characteristic curve* (ICC) (Baker, 2001). ICC viser sannsynligheten for riktig svar på en gitt oppgave som funksjon av dyktigheten (se figur 1). Ved Rasch-analyse blir data fra alle oppgaver testet opp mot én og samme ICC, men lokaliseringen av kurvens vendepunkt relativt til

førsteaksen avhenger av oppgavens vanskegrad. Data fra oppgaver med *samme vanskegrad* testes dermed opp mot *identiske* ICC. Figur 1 viser ICC til en dikotom oppgave, og kurven viser sannsynligheten for riktig svar som funksjon av dyktighet. Førsteaksen angir dyktigheter i logits med økende dyktighet mot høyre. Nullpunktet på førsteaksen angir gjennomsnittlig dyktighet. Andreaksen angir sannsynligheten for riktig svar. Den s-formede kurven har en horisontal asymptote ved P = 0 og en ved P = 1 (DeAyala, 2009).



Figur 1: ICC til en dikotom oppgave. Kurven viser sannsynligheten for riktig svar som funksjon av dyktighet. Førsteaksen viser dyktighet i logits og andreaksen angir sannsynlighet for riktig svar.

Personene som har svart på en test blir delt inn i grupper kalt «klasseintervaller» basert på dyktigheten (Andrich & Marais, 2012; Bond & Fox, 2007). Den gjennomsnittlige dyktigheten til personene i et gitt klasseintervall blir avsatt langs førsteaksen, og andel riktig svar blant personene i klasseintervallet blir avsatt langs andreaksen. Ut fra disse koordinatene avsettes de observerte verdiene i koordinatsystemet på figuren ovenfor, og vi kan få et bilde av oppgavens evne til å diskriminere.

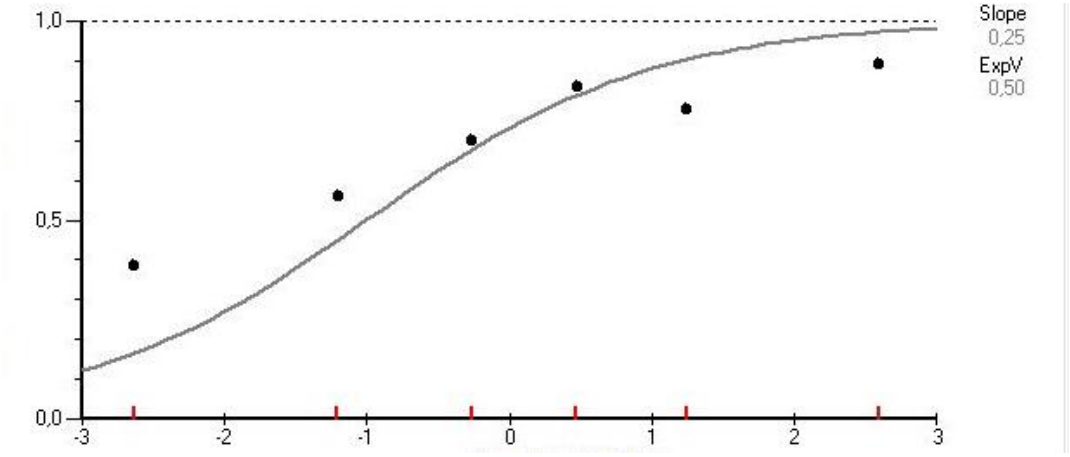
2.13 Under- og overdiskriminerende oppgaver

Med diskriminering mener vi den evnen en oppgave har til å skille mellom personer med høy og lav dyktighet (Andrich, 1988, 2005; Bond & Fox, 2007). Tatt oppgavens

vanskegrad i betraktning, kan en oppgave under- eller overdiskriminere langs hele eller langs deler av trekket.

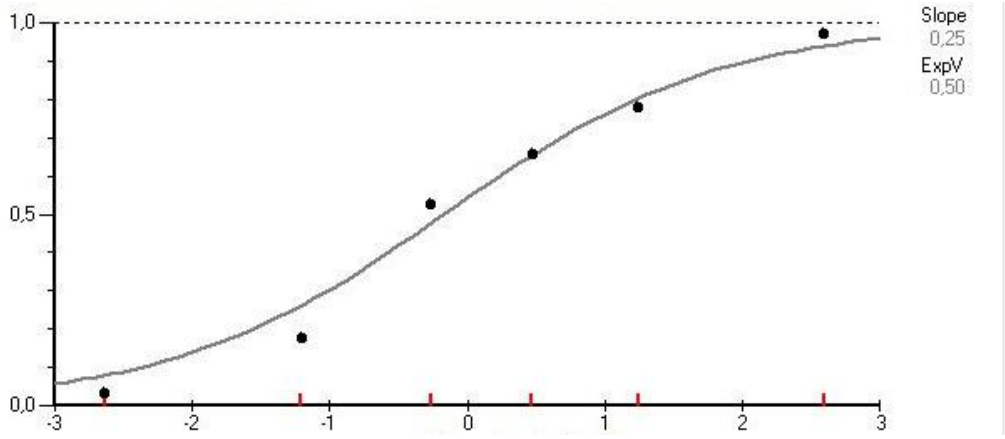
En oppgave med gjennomsnittlig vanskegrad underdiskriminerer langs hele trekket dersom personer med dyktighet under gjennomsnittet har høyere sannsynlighet for å svare riktig enn forventet, og hvis personer med dyktighet over gjennomsnittet har lavere sannsynlighet for å svare riktig enn forventet. Dataene fra en slik oppgave er vist i figur 2. En oppgave kan underdiskriminere dersom den måler «for mye» av noe annet enn den skal, og at det den måler for mye av diskriminerer negativt med det underliggende trekket (Masters, 1988). Dersom personer med lav dyktighet i stor grad gjetter på en oppgave, vil oppgaven kunne underdiskriminere bare i nedre del av trekket.

På tilsvarende måte vil en gjennomsnittlig vanskelig oppgave overdiskriminere langs hele trekket dersom personer med dyktighet under gjennomsnittet har lavere sannsynlighet for å svare riktig enn forventet, og hvis personer med dyktighet over gjennomsnittet har høyere sannsynlighet for å svare riktig enn forventet. Dataene fra en slik oppgave er vist i figur 3. En oppgave kan overdiskriminere dersom den måler «for mye» av noe annet enn den skal, og at det den måler for mye av diskriminerer positivt med det underliggende trekket (Masters, 1988).



Figur 2: ICC for en dikotom oppgave hvor klasseintervallenes observerte gjennomsnittsskår indikerer underdiskriminering (figuren er hentet fra RUMM2030).

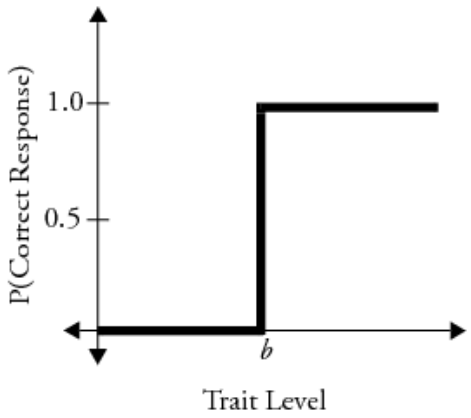
Det kan være ulike årsaker til at en oppgave overdiskriminerer (Masters, 1993). Oppgaver eller påstander hvor svaret avhenger av besvarelsen på en tidligere påstand er ofte (Marais & Andrich, 2008b). Det kan for eksempel være følgeoppgaver i kunnskapstester hvor det er forutsatt at for å svare riktig på en oppgave må man ha svart riktig på en foregående oppgave, eller hvis en oppgave gir «hint» om riktig svar til en påfølgende oppgave (Marais & Andrich, 2008b). Oppgaver eller holdningspåstander som oppsummerer tidligere oppgaver eller påstander kan også gi overdiskriminering, fordi svaret på oppgaven eller påstanden da avhenger av besvarelsen på en forutgående oppgave eller påstand (Masters, 1993). Oppgaver som måler andre faktorer som korrelerer med trekket,



Figur 3: ICC for en dikotom oppgave hvor klasseintervallenes observerte gjennomsnittskår indikerer overdiskriminering (figuren er hentet fra RUMM 2030).

for eksempel grunnleggende ferdigheter i lesing eller andre kunnskaper og ferdigheter som representerer «skoleflinkhet», kan føre til overdiskriminering (Araï, 2010; Masters, 1988).

En ekstremt overdiskriminerende oppgave, kalt Guttman-oppgave, fungerer som en «bryter» der alle personer med lavere dyktighet enn oppgavens vanskegrad svarer feil, mens alle personer med høyere dyktighet svarer riktig. En slik oppgave *måler* strengt tatt ikke, men stratifiserer personene over og under et visst nivå. Figur 4 illustrerer en slik oppgave. Oppgaver med høy diskriminering anses som svært gode oppgaver i KTT, men de gir måletekniske utfordringer i modeller basert på moderne testteori (Masters, 1988).



Figur 4: ICC til en ekstremt overdiskriminerende oppgave (Guttman-oppgave). Alle personer med dyktighet under oppgavens vanskegrad svarer feil, mens alle med høyere dyktighet enn oppgavens vanskegrad svarer riktig (figuren er hentet fra (www.jmp.com).

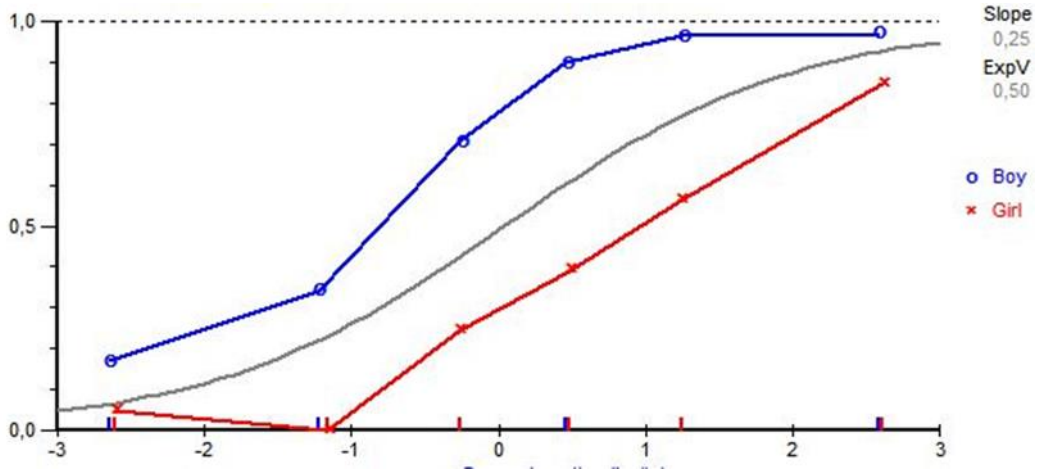
2.14 Trekkavhengighet og DIF

Trekkavhengighet indikerer at andre personfaktorer enn dyktighet påvirker svaret på en oppgave (Marais & Andrich, 2008b). Dette bryter prinsippet om endimensjonalitet og medfører ikke-invariante målinger (Andrich & Hagquist, 2004; DeAyala, 2009; Marais & Andrich, 2008b). Oppgaver med trekkavhengighet kan bidra til å minke variansen til fordelingen av dyktighetsestimatene slik at PSI og CCA koeffisientene synker (Marais & Andrich, 2008b). Analyser av trekkavhengighet/multidimensjonalitet ligger utenfor denne oppgavens omfang.

Som nevnt tidligere skal sannsynligheten for å svare riktig på en oppgave bare være bestemt av forholdet mellom oppgavens vanskegrad og personens dyktighet. I noen tilfeller kan likevel personfaktorer som for eksempel kjønn, kulturell bakgrunn, språklig bakgrunn og sosioøkonomisk bakgrunn påvirke hvordan personer svarer. Da tolker og oppfatter personer påstandene ulikt avhengig av andre faktorer enn dyktighet, og oppgaven «virker forskjellig» («differential item functioning» (DIF)) for ulike personfaktorkategorier som for eksempel mann og kvinne. Dette representerer et måleteknisk problem, for påstanden måler da for mye av noe «annet» enn det den utgir seg for å måle (Andrich & Hagquist, 2004). Dette skyldes at den andre faktoren som blir målt i tillegg til dyktighet (kjønn) korrelerer svakt med det underliggende trekket (Masters, 1988). DIF relatert til personfaktorer opptrer når personfaktorkategorienes observerte verdier er forskjellige for personer i samme klasseintervall. Det betyr at personfaktoren, i tillegg til dyktighet, påvirker sannsynligheten for riktig svar.

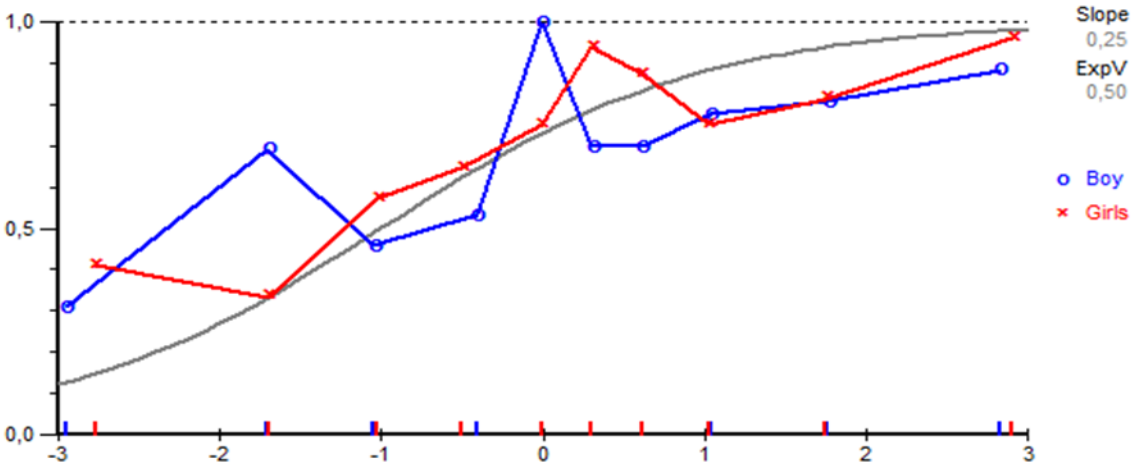
Litteraturen skiller mellom to hovedtyper DIF kalt «uniform DIF» og «non-uniform DIF» (Brodersen, Meads & Kreiner, 2007). Når en oppgave diskriminerer på

samme måte langs trekket for de ulike personfaktorkategoriene er kurvene som forbinder de observerte verdiene «parallelle», slik at kurvene ikke krysser hverandre. Denne formen for DIF kalles uniform (se figur 5). Figur 5 viser at kurvene som forbinder de observerte verdiene for gutter og jenter ikke krysser hverandre. Dermed diskriminerer oppgaven på samme måte langs hele trekket (langs førsteaksen) for begge personfaktorkategoriene, og DIF er uniform.



Figur 5: Grafisk representasjon av observerte verdier på en dikotom oppgave som funksjon av dyktighet, der personene i klasseintervallene er splittet basert på kjønn. Figuren illustrerer uniform DIF (figuren er hentet fra RUMM 2030).

Når en oppgave diskriminerer på forskjellig måte langs trekket for de ulike personfaktorkategoriene er kurvene som forbinder de observerte verdiene *ikke* «parallelle», men krysser hverandre (se figur 6). Denne formen for DIF kalles non-uniform DIF (Andrich & Hagquist, 2004). Som vist i figur 6 krysser kurvene som forbinder de observerte verdiene for de to personfaktorkategoriene hverandre. Dermed diskriminerer ikke oppgaven på samme måte langs hele trekket for de to personfaktorkategoriene gutt og jente.



Figur 6: Grafisk representasjon av observerte verdier på en dikotom oppgave som funksjon av dyktighet, der personene i dyktighetsgruppene er splittet basert på kjønn. Figuren illustrerer non-uniform DIF (figuren er hentet fra RUMM 2030).

Oppgaver som gir data med uniform DIF kan tolkes til å «favorisere» en av personfaktorkategoriene. Dette ser vi ved at den ene personfaktorkategorien uansett dyktighet har høyere observerte verdier enn forventningsverdiene fra Rasch-modellen. Mens den andre personfaktorkategorien har konsekvent lavere observerte verdier enn forventningsverdien (se figur 5). Dette betyr imidlertid ikke nødvendigvis at personer tilhørende den ene personfaktorkategorien er dyktigere på det temaet som oppgaven handler om, for personene i samme klasseintervall har i gjennomsnitt samme estimerte dyktighet. Forskjellene i de observerte verdier skyldes altså en «bias» forårsaket av en annen faktor eller dimensjon enn dyktighet og dataene er «flerdimensjonale» (Masters, 1988).

2.15 Svaravhengighet

Svaravhengighet opptrer når svaret på en oppgave er styrt av svaret på en forutgående oppgave (Marais & Andrich, 2008b). Den avhengige, etterfølgende oppgaven vil diskriminere sterkere enn om det ikke var avhengighet i dataene (Marais & Andrich, 2008b). Avhengighet betyr at det underliggende trekket ikke forklarer all korrelasjonen mellom dataene. Svaravhengighet fører til at variansen til fordelingen av dyktighetsestimater øker (Marais & Andrich, 2008a, 2008b). Som resultat vil estimatene av reliabilitetskoeffisientene PSI og CCA øke (Marais & Andrich, 2008b). Estimatene av PSI og CCA er gyldige mål for den nedre grensen for reliabilitet bare når data er tilstrekkelig uavhengige (tilstrekkelig lav svaravhengighet og trekkavhengighet i dataene) (Andrich & Marais, 2012). Mens trekkavhengighet typisk resulterer i

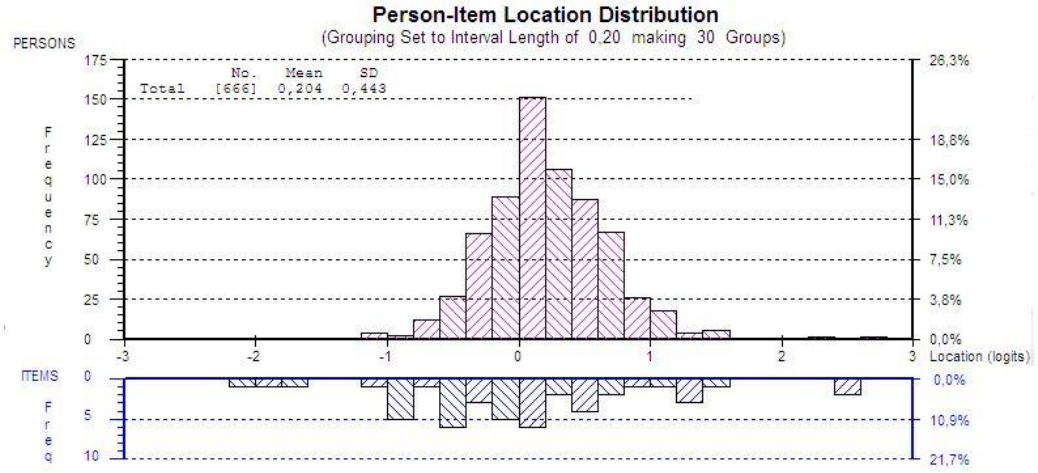
underdiskriminerende oppgaver, vil svaravhengighet gi overdiskriminerende oppgaver (Marais & Andrich, 2008b).

2.16 Targeting

Personenes poengsummer er tilstrekkelig (jamfør «sufficiency») for å beregne personenes dyktighetsestimat i Rasch-modeller (Andrich, 1988, 2005; Bond & Fox, 2007). Og alle personer med lik poengsum får estimert tilsvarende dyktighet (Andrich & Hagquist, 2012). Målefeilen til personenes dyktighetsestimater varierer langs trekket, og målefeilen er minst der «informasjonen er størst» (DeAyala, 2009). Mengden informasjon om personene er stor der «tettheten av terskler» er stor (Andrich & Marais, 2012; Andrich, Sheridan & Luo, 2012). For oppgaver som gir dikotome data, som flervalgsoppgaver, tilsvarer tersklene oppgavenes vanskegrader.

Når fordelingen av vanskegradene til oppgavene i en test har god tilpasning («targeting») til fordelingen av personenes dyktighet, vil målefeilen til dyktighetsestimatene avta (DeAyala, 2009), men svært høye og svært lave dyktighetsestimater vil normalt ha større målefeil da tester inneholder relativt få oppgaver med høy og lav vanskegrad. Personer med svært lav eller svært høy dyktighet kan oppnå «ekstremskår» («tak- og gulveffekt»). Personer som oppnår ekstremskår utgår følgelig fra Rasch-analyser, for de bidrar ikke med informasjon om oppgavenes vanskegradestimater.

For å begrense *standard målefeil* ved estimering av påstandenes affektive nivå, må vi som nevnt ha tilstrekkelig «informasjon» om påstandene (DeAyala, 2009). Da er det ikke nødvendigvis tilstrekkelig å ha stort nok utvalg. For å få tilstrekkelig informasjon om det affektive nivået til en påstand med høyt eller lavt affektivt nivå, vil det derfor være nødvendig å bruke «skjeve» utvalg hvor personer med svært høyt og personer med svært lavt holdningsnivå er «over-samplet». Vi må altså sørge for at påstandenes affektive nivå er godt tilpasset respondentenes holdningsnivå (god «targeting»). Det følgende eksempelet illustrerer sammenhengen mellom «targeting» og standard målefeil ved estimering av oppgavers vanskegrad: Hvis to personer med gjennomsnittlig dyktighet svarer på en oppgave med høy vanskegrad vil begge mest sannsynlig svare feil. Da har vi svært lite informasjon om oppgavens vanskegrad, og det vil være knyttet stor målefeil til estimatet av vanskegraden til oppgaven.



Figur 7: Fordelingen av personers dyktighetsestimater (over førsteaksen) og oppgavenes vanskegrad (under førsteaksen) i en test (figur hentet fra RUMM 2030).

Diagrammet i figur 7 illustrerer fordelingen av personers dyktighetsestimater (søylene over førsteaksen) og fordelingen av oppgavenes vanskegrad (søylene under førsteaksen) i en test. Slike diagrammer refereres gjerne til som «person-item location distribution» (Andrich, Sheridan, et al., 2012). Nullpunktet langs førsteaksen tilsvarer gjennomsnittlig vanskegrad. Når personenes gjennomsnittlige holdningsnivå er nært oppgavenes gjennomsnittlige vanskegrad er «targeting» god (Andrich & Marais, 2012; Andrich, Sheridan, et al., 2012). I figur 7 kan vi se at personenes gjennomsnittlige dyktighet er 0,2 logits noe som indikerer relativt god «targeting» mellom personenes gjennomsnittlige dyktighet og oppgavenes vanskegrad.

2.17 Tilpasningsstatistikk

Når data har tilstrekkelig tilpasning til endimensjonale Rasch modeller er målingene invariante og dataene er endimensjonale (Andrich & Marais, 2012), men i realiteten vil empiriske data aldri passe perfekt til modellen (Andrich & Marais, 2012; Bond & Fox, 2007). For å undersøke avviket fra modellen, blir data testet opp mot modellen ved at de observerte verdiene blir sammenlignet med de teoretisk forventete verdiene beregnet fra modellen (A. Tennant & P.G. Conaghan, 2007). Programmet RUMM2030 (Andrich, Sheridan, et al., 2012) rapporterer dataenes tilpasning til endimensjonale Rasch-modeller gjennom «fit residualer» og kji-kvadrat statistikk (X2).

2.17.1 Fit residualer for klasseintervaller og total kji-kvadratverdi (χ^2) for enkeltoppgaver (Bonferronijustert)

Et *fit residual* (*F*) angir differansen mellom antallet personer (T_{gi}) i et gitt klasseintervall (*g*) som svarte riktig på oppgave *i* (faktisk eller observert antall riktige svar) og teoretisk forventet antall riktige svar i klasseintervallet (Andrich & Marais, 2012). Det teoretisk forventede antall svar kan uttrykkes som produktet av totalt antall personer i klasseintervallet (N_g) og den teoretisk forventede sannsynligheten for riktig svar i klasseintervallet gitt ved den valgte Rasch-modellen (P_{gi}): $F_{gi} = T_{gi}\text{-}N_gP_{gi}$.

Alle klasseintervall hvor det observerte antallet riktige svar er høyere enn det teoretisk forventede har positive residualer. Alle klasseintervall hvor det observerte antallet riktige svar er lavere enn det teoretisk forventede har negative residualer (Smith, 2000). Jo større residualet for et klasseintervall er, jo svakere er tilpasningen til Rasch-modellen.

Ved å dividere på standardavviket til antall riktige svar blir residualet standardisert (z-fit-residual), og vi kan sammenligne med den standardiserte normalfordelingen som har gjennomsnitt lik 0 og standardavvik lik 1. Z-fit-residualer større enn 2,5 eller mindre enn -2,5 tolkes gjerne som store avvik fra modellen (Andrich & Marais, 2012; Smith, 2000).

$$z_{gi} = \frac{T_{gi} - N_g P_{gi}}{\sqrt{N_g P_{gi}(1 - P_{gi})}}$$

Summen av de *kvadrerte* z-fit-residualene over alle klasseintervall gir en «testobservator» for oppgaven (Smith, 2000). Ved høyt antall observasjoner kan vi forvente at denne observatoren følger en χ^2-fordeling (Bond & Fox, 2007; Smith, 2000). Verdien av observatoren kan sammenlignes opp mot χ^2-fordelingen for det tilsvarende antall frihetsgrader og si noe om hvordan data fra enkeltoppgaver passer til modellen (Wright & Mok, 2000).

Hvis en test består av *n* antall oppgaver utfører vi n signifikanstester. Når antall signifikanstester øker, øker sannsynligheten for at noen av testene gir signifikant utslag. For å «veie opp» for dette blir signifikansnivået (for eksempel 5 %) «Bonferroni»-justert (Bland & Altmann, 1995) ved å dividere på antall signifikanstester, som er lik antall oppgaver: 5/n.

2.17.2 Fit residualer for enkeltpersoner og totalt fit residual for en oppgave

Istedenfor å dele personer inn i klasseintervaller, kan formelen ovenfor uttrykkes ved skårverdien *x* til en enkeltperson *n* på enkeltoppgaven *i*:

$$z_{ni} = \frac{x_{ni} - E(x_{ni})}{\sqrt{V(x_{ni})}}$$

der $E(x_{ni})$ er forventningsverdien for skårverdien *x* gitt ved $E(x_{ni}) = \sum_{x=0}^{m_i} x\ P(x_{ni})$. $P(x_{ni})$ er sannsynligheten for skårverdien *x* gitt ved Rasch-modell, og uttrykket er summert opp over alle skårverdier fra $x = 0$ til $x = m_i$. Verdien m_i betegner høyeste tillatte skårverdi, for eksempel på en Likert-skala. Variansen $V(x_{ni})$ til x_{ni} er $V(x_{ni}) = E(x^2_{ni}) - E(x_{ni})^2$, der $E(x^2{}_{ni}) = \sum_{x=0}^{m_i} x^2\ P(x_{ni})$.

Siden fit-residualer har positive og negative fortegn er summen av standardiserte fit residualer alltid nær null (Bond & Fox, 2007). For å estimere «størrelsen» på avviket blir summen av de *kvadrerte* residualene brukt (Bond & Fox, 2007). Det første uttrykket nedenfor er summen av kvadrerte residualer over alle oppgaver *I* (residualet for person *n*), mens det andre uttrykket er summen av kvadrerte residualer over alle personer *N* (residualet for oppgave *i*):

$$y_n^2 = \sum_{i=1}^{I} z_{ni}^2$$

$$y_i^2 = \sum_{i=1}^{N} z_{ni}^2$$

Disse variablene kan igjen standardiseres (med forventningsverdiene $E(y^2)$ lik antall frihetsgrader til variabelen) og betegnes da Z_n og Z_i (dvs. standardiserte summer av de kvadrerte *z*-fit-residualene z_{ni}).

Underdiskriminerende oppgaver har positive Z_i verdier, mens overdiskriminerende oppgaver har negative negative Z_i verdier (Marais & Andrich, 2008a). Personer med «uventet god» tilpasning til Guttman-struktur (se nedenfor) har negative z_n verdier, mens personer med «uventet svak» tilpasning til Guttman-struktur har positive z_n verdier. Z-verdier med absoluttverdi større enn 2,5 tolkes som mulige avvik fra modellen (Andrich & Marais, 2012).

2.17.3 Tilpasning til SLM

Personers z-fit-residualer indikerer om svarmønsterne deres likner «Guttman-struktur» (Bond & Fox, 2007; Guttman, 1950). En perfekt Guttman-struktur oppnås når alle svar opp til en gitt vanskegrad er riktige og de resterende svarene er feil (Andrich & Marais, 2012; Engelhard Jr, 2008). Tabell 2 viser Guttman-strukturen til fire personer i stigende dyktighet som har svart på tre dikotome oppgaver (A-C) med stigende vanskegrad. Siden Rasch-modeller angir sannsynligheten for riktig svar, forventer vi svarmønstre som *ligner* Guttman-struktur (Bond & Fox, 2007; DeAyala, 2009). Store positive og negative person z-fit-residualer indikerer svarmønstre som har «for svak» og «for god» tilpasning til Guttman-strukturen (Andrich & Marais, 2012).

Tabell 2: Svarmønster med Guttman-struktur (1 = riktig svar og 0 = feil svar) for fire personer med poengsum 0, 1, 2 og 3 på tre dikotome oppgaver (A-C). Personene er rangert etter dyktighet (poengsum) og oppgavene er rangert etter vanskegrad (andel riktige svar).

Totalskår	A	B	C
0	0	0	0
1	1	0	0
2	1	1	0
3	1	1	1

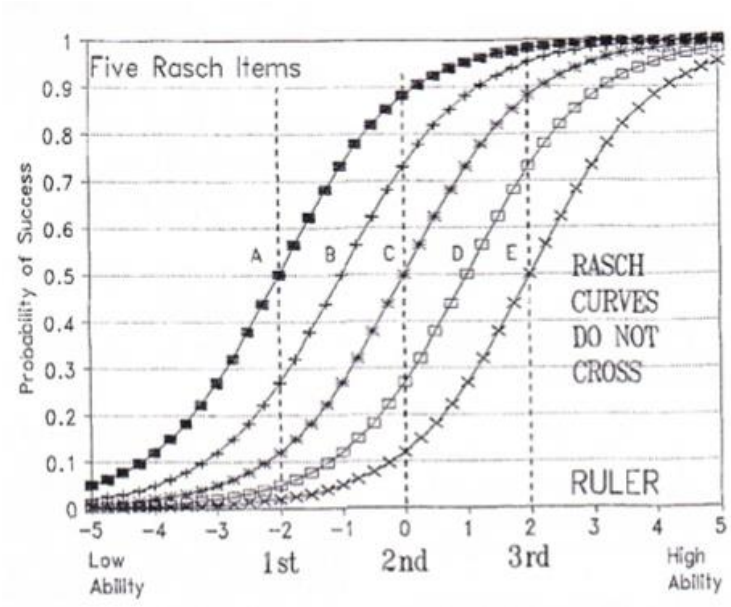
2.17.4 Undersøke brudd på invarians ved inspeksjon av tilpasning til ICC

Som nevnt tidligere kan vi studere DIF grafisk ved å inspisere de observerte verdienes tilpasning til ICC (Andrich & Hagquist, 2004). DIF bryter prinsippet om invarians fordi en persons svar på en oppgave er påvirket av andre faktorer enn dyktighet (for eksempel kjønn) slik det er vist i figur 5 og figur 6.

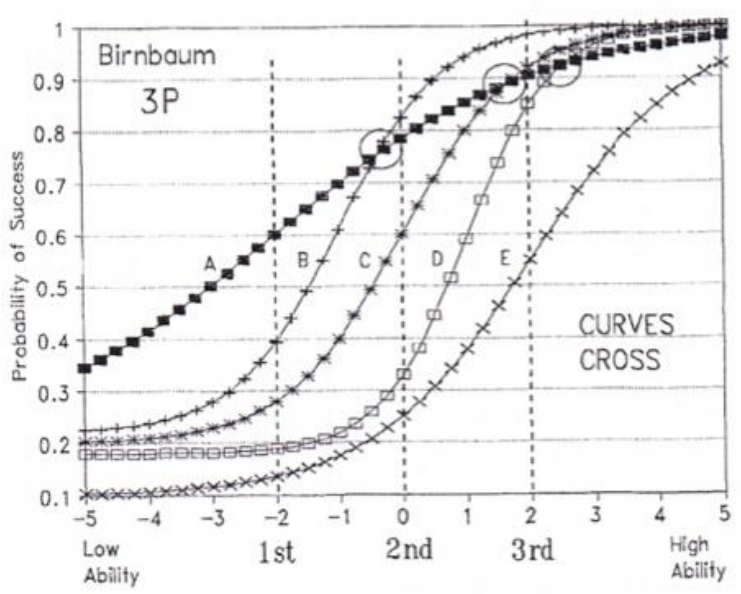
Videre kan inspeksjon av tilpasning til ICC for flere oppgaver samtidig også avdekke brudd på invarians (Wright, 1997). Siden diskrimineringsparameteren i endimensjonale Rasch-modeller er lik for alle oppgaver, skal teoretisk sett ikke de «tenkte» kurvene som forbinder de observerte verdiene krysse hverandre hvis måledataene har tilstrekkelig tilpasning til modellen (Engelhard, 2013; Wright, 1997). Dermed skal de tenkte kurvene som forbinder de observerte verdiene fremstå som forskyvninger langs førsteaksen av én og samme kurve (Lie, 2010) (se figur 8).

Figur 8 viser et eksempel med ICC for fem oppgaver A-E med økende vanskegrad. Kurvene viser at alle dyktigheter langs førsteaksen har størst sannsynlighet for å svare riktig på oppgave A og lavest sannsynlighet for å svare riktig på oppgave E.

Figur 9 viser tilsvarende kurver for de samme oppgavene ved bruk av treparametermodell. Kurvene har ulike diskrimineringsparametre og ulike nedre asymptoter grunnet modellering av gjetteparameter (DeAyala, 2009; Wright, 1997). Siden kurven til oppgave A krysser kurvene til oppgave B, C og D, er personer med ulik dyktighet ikke «enige» om oppgavenes vanskegrad. For eksempel har personer med dyktighet lik -2 logits som forventet større sannsynlighet for å svare riktig på oppgave A (lav vanskegrad) enn på oppgave B (høyere vanskegrad), mens for personer med dyktighet lik +2 logits gjelder det motsatte.



Figur 8: ICC for fem oppgaver (A-E). Siden kurvene er parallelle og ikke krysser hverandre er målingen invariante. Førsteaksen viser dyktighet i logits og andreaksen viser sannsynligheten for riktig svar. Figuren er hentet fra (Wright, 1997).



Figur 9: ICC for fem oppgaver (A-E). Siden kurvene krysser hverandre er målingen ikke invariant. Førsteaksen viser dyktighet i logits og andreaksen viser sannsynligheten for riktig svar. Figuren er hentet fra (Wright, 1997).

2.18 Polytome Rasch modeller: svarkategorier og terskelverdier

Polytome data blir testet opp mot polytome Rasch-modeller (A. Tennant & P.G. Conaghan, 2007; Van Wyke & Andrich, 2006; Wright & Mok, 2000). Påstander som søker å måle personers holdninger ved bruk av Likert-skalaer gir polytome data (Masters, 1982). Ved måling av holdningsnivå vil typisk *oppgave* erstattes med *spørsmål* eller *påstand*, *vanskegrad* erstattes med *affektivt nivå*, og *dyktighet* erstattes med *holdningsnivå* (Andrich, 1978). Sannsynligheten for at en gitt person krysser av i en viss svarkategori på Likert-skalaen for en bestemt påstand, er gitt ved forholdet mellom holdningsnivået til personen og det affektive nivået til påstanden. Polytom endimensjonal Rasch-modell (PRM) kan utrykkes ved:

$$P\{X_{vi} = x\} = \frac{e^{(x(\beta_v - \delta_i) - \sum_{k=1}^{x} t_k)}}{\sum_{x=0}^{m} e^{(x(\beta_v - \delta_i) - \sum_{k=1}^{x} t_k)}}$$

Uttrykket angir sannsynligheten (P) for at person v med holdningsnivå β krysser av i svarkategori x av totalt m+1 (ordnet) svarkategorier på påstand i med affektivt nivå δ (Van Wyke & Andrich, 2006).

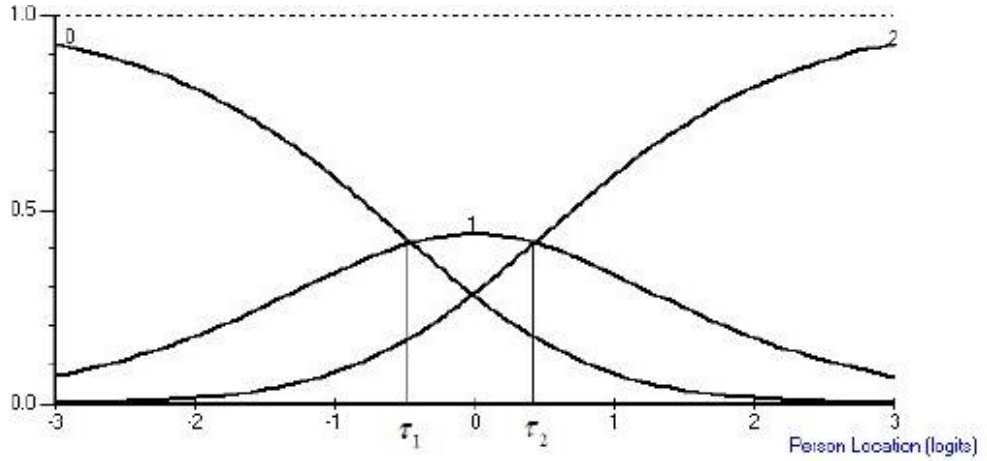
I motsetning til dikotome oppgaver som bare har én terskelverdi (oppgavens vanskegrad), gir påstander som benytter Likert-skalaer polytome data og følgelig minst

to terskelverdier (Araï, 2010). Terskelverdiene (τ) angir de holdningsnivåene hvor sannsynligheten er 50 % for å krysse av i en bestemt kategori eller i tilstøtende kategori nedenfor (Andrich & Marais, 2012; Van Wyke & Andrich, 2006). For at ulikesvarkategorier stadig skal reflektere mer av det underliggende trekket, må de suksessive terskelverdiene være «ordnet» slik at $\tau_{\text{mi-1}\ldots} > \tau_3 > \tau_2 > \tau_1$ (Andrich & Marais, 2012). Når høyeste svarkategori er *m* er antallet terskler *m*-1, slik at en femdelt Likert-skala vil ha fire terskelverdier.

Det finnes to parametriseringer av PRM, og disse refereres ofte til som Rating scale model (RSM) (Andrich, 1978) og Partial credit model (PCM) (Masters, 1982). RSM passer best til dataene når *avstanden* mellom terskelverdiene (τ) er konstante på tvers av spørsmålene (Andrich, 1978). Terskelverdiene kan imidlertid være lokalisert på ulike steder langs skalaen for hver av påstandene. PCM passer best til dataene når avstanden mellom terskelverdiene varierer på tvers av påstandene (Masters, 1982). RSM blir bare anvendt ved data fra Likert-skalaer, mens PCM blir brukt til å modellere data fra både Likert-skalaer og testoppgaver hvor det gis mer enn ett poeng.

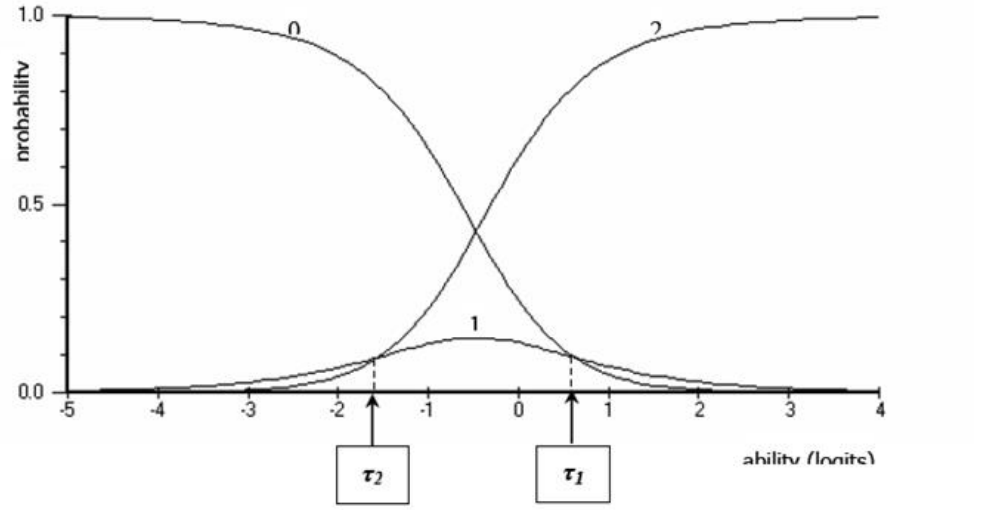
Selv om tilpasningsstatistikken viser at data har tilstrekkelig tilpasning til PRM er det ikke gitt at tersklene er ordnet (Hagquist & Andrich, 2004). I RUMM2030 indikerer *kategorisannsynlighetskurver* (category probability curves) om tersklene er ordnet. I noen tilfeller kan uordnete svarkategorier simpelthen være et resultat av lave andeler svar i enkelte kategorier. I andre tilfeller kan uordnete svarkategorier muligens indikere underliggende og mer alvorlige problemer i dataene.

Figur 10 viser kategorisannsynlighetskurver for en påstand med tre svarkategorier og to terskler. Kurvene viser sannsynligheten for å krysse av i kategori 1, 2 eller 3 (rekodet til 0, 1 eller 2) som funksjon av holdningsnivået. Personer med lavt holdningsnivå vil ha størst sannsynlighet for å krysse av i kategori 1 (verdien 0), mens sannsynligheten for å krysse av i kategoriene 2 og 3 (verdien 1 eller 2) øker med økt holdningsnivå. Terskelverdiene er ikke reverserte og svarkategoriene er «ordnete».



Figur 10: Kurvene viser sannsynligheten for å krysse av i kategori 1, 2 eller 3 (rekodet til verdiene 0, 1 eller 2) som funksjon av holdningsnivået. Terskelverdiene er ikke reverserte og svarkategoriene fungerer godt (Van Wyke & Andrich, 2006).

Figur 11 illustrerer data fra en tilsvarende påstand med uordnete svarkategorier og reverserte terskelverdier (Van Wyke & Andrich, 2006). De reverserte terskelverdiene gir et inntrykk av at skillet mellom den midtre og den øvre svarkategorien (τ_2) er lokalisert ved et lavere holdningsnivå enn skillet mellom den midtre og den laveste svarkategorien (τ_1). Personer med lavt holdningsnivå har også nå størst sannsynlighet for å krysse av i kategori 1 (rekodet til verdien 0), men midtkategorien (rekodet til verdien 1) er ikke det mest sannsynlige svaret for noe holdningsnivå. I et slikt tilfelle bør det bli vurdert om svarkategorier bør slås sammen.



Figur 11: Kurvene viser sannsynligheten for å krysse av i kategori 1, 2 eller 3 (rekodet til verdiene 0, 1 eller 2) som funksjon av holdningsnivået. Siden midtkategorien ikke er det mest sannsynlige svaret for noe holdningsnivå er terskelverdiene reverserte (Van Wyke & Andrich, 2006).

Når vi har påstander med reverserte terskelverdier bør andelen svar i de ulike kategoriene undersøkes og sammenlignes (Andrich & Marais, 2012; A. Tennant & P.G. Conaghan, 2007; Tennant & Lundgren-Nilsson, 2013; Van Wyke & Andrich, 2006). Reskåring ved å slå sammen kategorier kan i noen tilfeller være aktuelt. Det bør også vurderes om «problemene» kan løses ved for eksempel å omformulere påstanden. Vi kan sjelden med stor grad av sikkerhet si hva som forårsaker reverseringen, og det må derfor gjøres en subjektiv vurdering sett i sammenheng med påstandens kontekst (Andrich & Marais, 2012). Påstander med reverserte terskelnivåer bør imidlertid ikke forkastes før påstandene er forsøkt endret og testet på nytt (Andrich & Marais, 2012).

3. Metode

Første del av dette kapittelet dreier seg om utviklingen av spørreskjemaet brukt i SOMAHs DP2 og dets FNL og INL konstrukt. Dernest følger en beskrivelse av utvalget i SOMAH prosjektet. Videre i kapittelet er analysene fra KTT og Rasch-analysen beskrevet.

3.1 Utvikling av spørreskjemaet

Spørreskjemaet i SOMAH-DP2 ble utviklet høsten 2010 og består av totalt 69 påstander, hvorav 26 påstander samler inn data om helsesøstrenes oppfatning av FNL- og INL-nivået til mødre med spedbarn og gravide kvinner som besøker helsestasjonen. Spørreskjema dekker i tillegg andre sentrale aspekter ved kosthold- og ernæringskommunikasjon, som blant annet samtaler om overvektsproblematikk, opplevd tidspress og samarbeidsmiljø/kommunikasjon mellom helsepersonell ved de aktuelle helsestasjonene. Det er imidlertid kun påstandene i spørreskjemaet som tenderer å måle FNL og INL samt enkelte bakgrunnsvariabler, som inngår i analysene i masteroppgaven.

FNL og INL konstruktene er utviklet og operasjonalisert på bakgrunn av NL-teori og empiriske studier (Dalane, 2011; Guttersrud et al., 2013; Pettersen, 2009; Silk et al., 2008), samt analyser av kvalitative data fra fem separate fokusgruppeintervjuer med erfarne helsesøstre ved helsestasjoner i Oslo/Akershus/Buskerudregionen (N=26). Intervjuene ble gjennomført i løpet av studieåret 2010. Alle fokusgruppeintervjuene varte i om lag en time. Påstandene ble inndelt i konstruktene FNL og INL etter kvalitative vurderinger av fem forskere innen ernæring og helsekommunikasjon.

For hver påstand skal respondentene gi sin grad av tilslutning langs en femdelt Likert-skala med svarkategoriene: (1) «I veldig liten grad», (2) «I liten grad», (3) «Verken eller», (4) «I stor grad» og (5) «I veldig stor grad». Respondentenes svar er ment å reflektere deres holdningsnivå for den aktuelle påstanden.

Spørreskjemaet SOMAH-DP2 var designet slik at helsesøstrene først svarte på påstander som handlet om hvordan de opplevde FNL- og INL-nivået hos etnisk norske gravide og mødre med spedbarn (heretter kalt *majoritetsbrukere*). Deretter svarte helsesøstrene på de samme påstandene, men da med bakgrunn i hvordan de opplevde FNL- og INL-nivået hos mødre med spedbarn og gravide kvinner med ikke-vestlig innvandrerbakgrunn (heretter kalt *minoritetsbrukere*). Avslutningsvis inneholder

spørreskjemaet «åpne» og «lukkede» spørsmål om demografiske bakgrunnsvariabler for å beskrive helsesøstrene. Påstandene i konstruktene FNL og INL er gjengitt i tabell 3.

Tabell 3: Påstandene i spørreskjemaet SOMAH-DP2 kategorisert i henhold til konstruktene «funtional nutrition literacy» (FNL) og interactive nutrition literacy» (INL). Påstandene 1-13 refererer til helsesøstrenes svar på påstander om FNL- og INL-nivået til brukere av helsestasjonen med majoritetsbakgrunn. Påstandene 14-26 refererer til helsesøstrenes svar på påstander om FNL- og INL-nivået til brukere av helsestasjonen med flerkulturell bakgrunn. Påstandene 14-26 er identiske med påstandene 1-13.

Påstand	Ordlyd	Konstrukt
1 (14)	De leser godt	FNL
2 (15)	De forstår innholdet i helse- og kostholdsbrosjyrene som jeg gir dem	FNL
3 (16)	De gir inntrykk av å ha lest helse- og kostholdsbrosjyrene som jeg gir dem	FNL
4 (17)	De forstår faguttrykk som jeg bruker i samtale om helse og kosthold	FNL
5 (18)	De har gode nok kunnskaper om menneskekroppen til å forstå helse- og kostveiledningen	FNL
6 (19)	De stiller meg ofte spørsmål om helse og kosthold under konsultasjonen	INL
7 (20)	De klarer å fortelle meg om sine eventuelle helsebekymringer på en klar og forståelig måte	INL
8 (21)	De klarer å fortelle meg om sine eventuelle bekymringer om matvaner på en klar og forståelig måte	INL
9 (22)	De gir meg beskjed dersom det er noe de ikke forstår i den helse- og kostholdsinformasjonen jeg gir dem	INL
10 (23)	De deltar aktivt i våre samtaler om deres matvaner under konsultasjonen	INL
11 (24)	De vil gjerne diskutere med meg om informasjon de har fått via media om hva som er sunn mat for barn	INL
12 (25)	De vet hvilke instanser i helsevesenet de bør henvende seg til dersom det oppstår helseproblemer knyttet til mat og vekt hos barn	INL
13 (26)	De er interessert i å få vite hva som regnes for å være et sunt kosthold for deres barn	INL

Tabell 3 viser oversikt over alle påstandene med ordlyd, kategorisert i konstruktene FNL- og INL. Påstandene 1-13 referer til helsesøstrenes svar på påstander om FNL- og INL-nivået til majoritetsbrukere, mens påstandene 14-26 refererer til helsesøstrenes svar på påstander om FNL- og INL-nivået til minoritetsbrukere. Påstandene 1-13 og 14-26 er identiske.

3.2 Utvalg

I Norge er det nær 700 helsestasjoner (Kommuneforlaget, 2010) og totalt 4000 personer er tilsatt som helsesøster (Ersvik, 6.9, 2013, muntlig kommunikasjon). Basert på disse tallene ble det beregnet at et utvalg på ca. 350 helsesøstre fra nær 250 helsestasjoner ville være et representativt utvalg. Størrelsen på utvalget ble beregnet ved bruk av «Sample size calculation» (Creative Research System, 2012). Dette verktøyet beregner hvor mange respondenter som er nødvendig for at utvalget skal være generaliserbart til populasjonen det er trukket ut fra (Kadam & Bhalerao, 2010).

I alt 390 helsestasjoner ble tilfeldig trukket etter stratifisering for 20 fylker og antall enheter (mange helsestasjoner i de største byene). Ved disse utvalgte helsestasjonene var det registrert til sammen 1107 tilsatte helsesøstre. Alle disse fikk tilsendt invitasjons- og informasjonsbrev om frivillig deltakelse i studien. Til sammen 280 helsestasjoner og N = 666 helsesøstre besvarte det web-linkete spørreskjemaet (Quest back, 2013) eller i papirversjon etter ønske fra enkelte helsestasjoner. Etter én e-postpurring var svarresponsen henholdsvis 72 % og 62 % for helsestasjonene og helsesøstrene. Deskriptive data for deltagende helsesøstre er rapportert i tabell 4.

Tabell 4: Deskriptive data for de deltagende helsesøstrene (N=666).

	Mode	*Median*	*Mean ±SD*	*Range*
Alder	47	47	48 ± 9	28 - 67
Antall år tilsatt som helsesøster	5	11	13 ± 9	1 - 40
Antall år tilsatt ved nåværende helsestasjon	5	8	9 ± 7	1 - 35
		Majoritet	Minoritet	
Kulturell bakgrunn (helsesøstre)		94 %	6 %	

Tabell 4 viser deskriptive data som alder, antall år tilsatt som helsesøster og antall år tilsatt ved nåværende helsestasjon for de deltagende helsesøstrene. Dataene beskrives ved bruk mode, median, gjennomsnitt (*mean*) med standardavvik (*SD*) og *range*. Tabellen viser videre at kun 6 % av helsesøstrene hadde minoritetsbakgrunn.

Alle norske fylker og de største byene var representert med helsestasjoner og helsesøstre i studien. Tabell 5 viser antallet helsesøstre og svarprosent per fylke. Dataene i tabell 5

viser at de folkerike fylkene på Østlandet og Sørlandet var godt representert, mens kystfylkene i nord og i vest i mindre grad var representert. Manglende svar fra de utvalgte helsestasjonene i Bergensregionen førte til at Hordaland var svakest representert i undersøkelsen (se tabell 5).

Tabell 5: Antallet helsesøstre og svarprosent per fylke.

Fylke	Antall deltakende helsesøstre i studien (n)	Totalt antall kontaktete helsesøstre i utvalget (N)	Svarprosent n/N(%)
Svalbard	1	1	100
Finnmark	15	17	88
Buskerud	41	55	75
Sør-Trøndelag	37	50	74
Oslo	67	95	71
Akershus	64	91	70
Østfold	49	72	68
Vestfold	37	55	67
Vest-Agder	23	36	64
Møre og Romsdal	39	62	63
Aust-Agder	16	26	62
Telemark	24	41	59
Oppland	31	54	57
Hedmark	19	34	56
Rogaland	58	111	52
Troms	24	51	47
Nordland	28	61	46
Nord-Trøndelag	17	39	44
Sogn- og Fjordane	19	49	39
Hordaland	57	115	39
SUM	**666**	**1107**	
GJENNOMSNITT			**62**

3.3 Statistiske analyser fra KTT

All databehandling med KTT ble gjennomført med programmet Statistical Package for the Social Sciences (SPSS) versjon 20,0 (IBM Corp, 2011).

3.3.1 Koding av bakgrunnsvariabler med åpent format

Svar avgitt på bakgrunnsvariabler med åpent format ble kodet og slått sammen til kategorier for å bruke variabelen i DIF-analyser. For eksempel ble variabelen Q10 «*Hva er din alder?*» kodet ved bruk av kategoriene «lav» for svarene 28 år til 47 år (50 % av helsesøstrene) og «høy» for svarene 48 år til 67 år (50 %). Variabelen Q18 *«Omtrent hvor mange innbyggere (avrund til nærmeste 1.000 eller 10.000) er det i kommunen du arbeider i? (Innbyggertallet i hele kommunen - ikke bydel/kommunedel)*» ble kodet ved å

dele kommunene inn i kategorier basert på innbyggerantall. Helsesøstrene ble gruppert etter kategoriene «liten» for kommuner med færre enn 5000 innbyggere (17,1 % av helsesøstrene), «mellomstor» 5000 – 20000 innbyggere (31,5 %), «stor» 20000 – 100000 innbyggere (39,0 %) og «svært stor» for kommuner med mer enn 100000 innbyggere (11,3 %). Denne kategoriseringen svarer til kategorier brukt av SSB (NOU 2007: 12, 2007). Kodene for innbyggertall ble deretter rekodet til en bivariat/dikotom kode med kategoriene «landsbygd» for kommuner med færre enn 75000 innbyggere (87 % av helsesøstrene) og «by» for kommuner med flere enn 75000 (13 %). Innbyggertallet i Tromsø som var minst av byene i undersøkelsen og bestemte dermed kriteriet for kategorien «by» (Statistisk sentralbyrå, 2013).

3.3.2 «Stacking» av data

Siden helsesøstrene svarte på det samme skjemaet to ganger, først med utgangspunkt i brukere med majoritetsbakgrunn og deretter med utgangspunkt i brukere med minoritetsbakgrunn, frembringer spørreskjemaet SOMAH-DP2 to sett med data der personene (helsesøstrene) *ikke* er uavhengige. Disse to datasettene gir tilsvarende data som for eksempel pre- og posttesting av elever ved bruk av identisk kunnskapstest før og etter undervisning. Designet benyttet ved spørreskjemaundersøkelsen SOMAH-DP2 kan dermed potensielt sett medføre brudd på lokal uavhengighet i dataene ved at samme påstand i de to datasettene, hvor helsesøstrene svarte med utgangspunkt i brukere med majoritetsbakgrunn (sett 1) og minoritetsbakgrunn (sett 2), er avhengige.

Datasettene gir videre mulighet til å «følge helsesøstrene» og studere deres relative gjennomsnittlige holdningsnivå til de to brukergruppene ved å «stacke data (Wright, 1996).

Datasettene ble «stacked» ved at datasettene ble slått sammen slik at datasett 1 representerte person 1-666 (linje 1-666 i en SPSS eller Excelfil) og datasett 2 representerte person 667-1332 (linje 667-1332 i en SPSS eller Excelfil). Dette betyr at hver helsesøster var representert to ganger i det totale datasettet. En ekstra variabel (en kolonne i datasettet) ble opprettet for å betegne om dataene tilhørte datasett 1, eller datasett 2, altså om helsesøstrene har svart med tanke på majoritet eller minoritetsbrukere.

3.3.3 Definering av FNL og INL konstruktene

Påstandene er kategorisert i ulike FNL- og INL-konstrukter avhengig av om helsesøstrene har svart på påstander med hensyn til majoritet- eller minoritetsbrukere sitt FNL- og INL-nivå. Påstandene som omhandler helsesøstrenes svar med tanke på majoritetsbrukerne sitt FNL- og INL-nivå er kategorisert i konstruktene *FNLmaj* (påstandene 1 til 5) og *INLmaj* (påstandene 6 til 13). Tilsvarende omhandler konstruktene *FNLmin* (påstandene 14 – 18) og *INLmin* (påstandene 19 til 26) påstandene hvor helsesøstrene har svart med tanke på minoritetsbrukerne sitt FNL- og INL-nivå. Disse konstruktene bygger på det originale datasettet. I det «stackede» datasettet er konstruktene *FNLmaj* og *FNLmin* slått sammen til ett felles konstrukt og kalles *FNL-stacked*, mens konstruktene *INLmaj* og *INLmin* er slått sammen til konstruktet *INL-stacked*.

3.3.4 Faktoranalyse

Faktoranalyse en korrelasjonsmodell som vurderer korrelasjonen (ladningen) mellom påstander og én eller flere komponenter/faktorer (Pallant, 2007). Den kan brukes for å påvise mønstre i korrelasjonene for å undersøke om indikatorer måler en eller flere underdimensjoner (Johannessen, 2009).

Faktoranalyse kan deles inn i hovedgruppene *eksplorerende* metode og *konfirmerende* metode (Lie, 2010). Førstnevnte er en induktiv metode hvor målet er å finne frem til en faktormodell som reproduserer testskårenes kovarians best mulig (Lie, 2010). Konfirmerende faktoranalyse er en deduktiv metode som gjerne brukes når vi vil teste en forhåndsspesifisert faktormodell mot data (Johannessen, 2009; Lie, 2010). Semi-konfirmerende faktoranalyse, som kan ses som en mellomvariant mellom eksplorerende og konfirmerende faktoranalyse (Holbert & LaMarre, 2011), er benyttet i denne masteroppgaven.

I hovedsak er det to forutsetninger som må ligge til grunn for at et datasettet egner seg for faktoranalyse (Pallant, 2007). Den første forutsetningen dreier seg om antall enheter (personer) i datamaterialet (Johannessen, 2009). Tabachnick & Fidell (2007) foreslår at det bør minst være 300 enheter for å kunne utføre faktoranalyse, men et mindre antall kan aksepteres dersom det er høye faktorladninger (gjerne over 0,80) på nøkkelvariabler (Johannessen, 2009; Pallant, 2007). Andre mener man kan se bort fra antall enheter så lenge det er en ratio på mellom 5 til 1 og 10 til 1; det vil si fra 5 til 10 enheter per faktor som skal inngå i analysen (Pallant, 2007).

Den andre forutsetningen dreier seg om det forholdet mellom variablene, og det er foreslått at korrelasjonen mellom variablene bør være minst 0,30 (Pallant, 2007). For å oppnå dette kan det være nødvendig å rekode enkelte variabler (Johannessen, 2009). Videre krever faktoranalyse at dataene er kontinuerlige med minimum fire verdier (Johannessen, 2009). I tillegg er det en fordel for styrken til analysene om dataene er normalfordelt (Johannessen, 2009), men dette anses ikke for å være en absolutt nødvendighet (Pallant, 2007).

Kaizer-Meyer-Olkin (KMO) og Bartletts sfær-test er målemetoder for å stadfeste om variablene egner seg for faktoranalyse (Johannessen, 2009; Pallant, 2007). KMO undersøker partielle korrelasjoner og gir et mål på utvalgstilstrekkelighet (Johannessen, 2009). KMO-indeksen er fra 0 til 1, hvor 0,60 er minimumsverdien for å gjennomføre faktoranalyse (Pallant, 2007). Bartletts sfær-test tester nullhypotesen om at alle korrelasjonene i korrelasjonsmatrisen kan komme fra utvalg hvor alle korrelasjonene er lik null (Johannessen, 2009). Man ønsker å forkaste nullhypotesen ved å påvise korrelasjoner i korrelasjonsmatrisen som viser at faktorene i konstruktet faktisk har en sammenheng. Bartletts sfær-test må være statistisk signifikant for å kunne gå videre med faktoranalysen (Johannessen, 2009; Pallant, 2007).

3.3.5 Indre konsistens reliabilitet målt ved Cronbach's koeffisient alpha

Høy indre konsistens reliabilitet betyr at en høy andel av variansen i personenes holdningsestimater er sann varians (Lie, 2010). Høy indre konsistens reliabilitet bidrar dermed til å separere personene basert på deres holdningsnivå. Når data har høy indre konsistens reliabilitet rangerer ulike deler av instrumentet personene på tilsvarende måte (Lie, 2010). Ulike delsett av påstander gir dermed data som korrelerer relativt sterkt (Scott & Mazhindu, 2009). Vi kan tolke CCA som den gjennomsnittlige korrelasjonen mellom dataene fra ulike delsett av datamaterialet (Lie, 2010). En CCA på 0,70 regnes ofte som nedre grenseverdi dersom et konstrukt skal kunne sammenligne grupper av personer (Kjærnsli et al., 2007; Lie, 2010). I tilfeller hvor vi ønsker å sammenligne enkeltpersoners summerte skår på et konstrukt bør nedre grenseverdi være høyere, gjerne over 0,80 (Frisbie, 1988; Kjærnsli et al., 2007; Scott & Mazhindu, 2009).

3.4 Rasch analyser

Rasch-analyse av dataene i denne oppgaven er utført med programmet RUMM2030 (Andrich, 2012).

3.4.1 Valg av parametrisering

Fisher's «*Likelihood-ratio Test*» i RUMM2030 sammenligner kji-kvadrat statistikk for dataenes tilpasning til PCM og RSM (RUMM, 2012). Et signifikant testresultat ($p < 0,05$) indikerer at PCM gir «signifikant» mer informasjon om dataene og bør brukes fremfor RSM (Tennant & Lundgren-Nilsson, 2013).

3.4.2 Vurdering av konstruktenes, påstandenes og personenes tilpasning til Rasch-modellen

Tilpasning til Rasch-modellen på et overordnet nivå ble vurdert ut fra konstruktenes totale X^2-verdi og X^2-sannsynlighetsverdi («*item trait interaction*» i RUMM2030). En X^2-sannsynlighetsverdi signifikant forskjellig fra null ($p > 0,05$) indikerer tilstrekkelig tilpasning til Rasch-modellen (til 5 % nivå). Tilpasningen til Rasch-modellen for data fra hver enkelt påstand ble vurdert ut fra z-fit-residualer og X^2-statistikk. Tilsvarende ble tilpasningen til data fra personer vurdert ut fra z-fit-residualer. Personenes overordnete tilpasning til Guttman-strukturen ble vurdert ut fra deres gjennomsnittlige z-fit-residualer.

3.4.3 Personseparasjonsindeks (PSI)

PSI for hvert konstrukt blir beregnet i RUMM2030 og angir dataenes indre konsistens reliabilitet (Lundgren-Nilsson, Jonsdottir, Ahlborg & Tennant, 2013). På samme måte som CCA gir PSI informasjon om konstruktets evne til å diskriminere mellom personer eller grupper med ulike holdningsnivåer (Hendriks, Fyfe, Styles, Skinner & Merriman, 2012).

3.4.4 Targeting

Analyser av påstandenes affektive nivå relativt til personenes holdningsnivå ble analysert grafisk ved bruk av «Person-Item Location Distribution» i RUMM2030. Personenes gjennomsnittlige holdningsnivå ble sammenlignet med påstandenes gjennomsnittlige affektive nivå.

3.4.5 Svarkategorier og terskelverdier

Påstander med uordnete svarkategorier ble påvist ved bruk av «*threshold map*» i RUMM2030. Uordnede svarkategorier er et tegn på at tilstøtende svarkategorier bør vurderes slått sammen.

3.4.6 Analyse av DIF – differential item functioning

DIF kan påvises grafisk ved analyse av ICC og statistisk ved analyse av varians (ANOVA) (Andrich & Marais, 2012; Smith, 2000). Variansanalysen kan avdekke om personfaktorer i vesentlig grad påvirker svarmønstre (Scott & Mazhindu, 2009).

3.4.7 Svaravhengighet

RUMM2030 påviser svaravhengighet gjennom analyse av residualkorrelasjoner (RUMM, 2012). Residualene er forventet å representere «tilfeldig støy» i målingene, slik at residualkorrelasjoner har forventningsverdier lik null. Positive residualkorrelasjoner er tegn til at dataene fra påstander har noe mer eller noe annet til felles utover det underliggende trekket (Andrich & Marais, 2012). Høye positive residualkorrelasjoner > 0,30 kan indikere svaravhengighet mellom to påstander (Andrich, Humpry & Marais, 2012). Andre har foreslått at residualkorrelasjoner som er større enn 0,2 i forhold til den gjennomsnittlige residualkorrelasjonen kan tolkes som svaravhengighet mellom påstander (Lundgren-Nilsson et al., 2013). Det er utviklet metoder for å estimere styrken på svaravhengigheten (Andrich, Humpry, et al., 2012; Andrich & Kreiner, 2010), men slike analyser av polytome data ligger utenfor denne oppgaven.

3.4.8 Hvordan håndtere mulige brudd på lokal uavhengighet

Subtest-funksjonen i RUMM2030 «slår sammen» dataene fra flere påstander til ulike delskalaer eller underdimensjoner (Andrich & Marais, 2012). Subtestanalyser basert på mistanke om trekkavhengighet/flerdimensjonalitet i dataene «tar hensyn til og justerer for» trekkavhengighet, mens subtestanalyser basert på høye residualkorrelasjoner «tar hensyn til og justerer for» svaravhengighet.

Split-funksjonen i RUMM2030 kan brukes for å splitte påstander med mulig svaravhengighet (Andrich & Marais, 2012). Den avhengige påstanden splittes da opp i like mange «virtuelle» påstander som det er svarkategorier i den «uavhengige» etterfølgende påstanden (Andrich, Humpry, et al., 2012). Splitfunksjonen kan også brukes for å splitte påstander med uniform-DIF opp i «virtuelle» påstander – én for hver

personfaktorkategori for de personfaktorene som påvirker hvordan personer svarer (RUMM, 2012).

4. Resultat

I dette kapittelet blir resultater fra KTT kort presentert før resultatene fra Rasch-analysene blir presentert i påfølgende underkapitler.

4.1 Resultater fra KTT

Det ble gjennomført semi-konfirmerende faktoranalyse, hvor antall ekstraherte faktorer ble satt til én, og minste faktorladning satt til minimum 0,300. Ortogonal Variamax-rotasjon ble brukt. Korrelasjonsmatrix av dataene fra påstandene viste at de fleste korrelasjonskoeffisientene var > 0,30. Alle konstruktene hadde KMO verdier > 0,77 (større enn minimumsverdien 0,60). Videre var Bartlett's Test of Sphericity signifikant < 0,05 for konstruktene *FNLmaj, FNLmin, INLmaj* og *INLmin* og støttet de foregående resultatene om at dataene var egnet for faktoranalyse.

Tabell 6 viser de største og minste faktorladningene, og de høyeste og laveste gjennomsnittspoengene (mean) med standardavvik (SD) til hvert av konstruktene *FNLmaj, FNLmin, INLmaj* og *INLmin.* Hvert av konstruktenes CCA-verdi er oppgitt sammen med hvilke påstander som har størst «innflytelse» på konstruktenes CCA-verdi. Disse tallene bygger på «*alpha if item deleted*» analysen i SPSS som forteller hva CCA-verdien for konstruktet ville vært hvis påstanden var fjernet (Pallant, 2007).

Vi ser at helsesøstrenes gjennomsnittspoeng på FNL og INL konstruktene var høyere når de svarte med tanke på majoritetsbrukere, enn når de svarte for minoritetsbrukere. Alle konstruktene hadde akseptable CCA-verdier som vare nære 0,80 (Scott & Mazhindu, 2009).

Tabell 6: Resultater fra KTT for *FNLmaj, FNLmin, INLmaj* og *INLmin* konstruktene beregnet ved N=666.

.

Konstrukt	*Største og minste faktorladning*	*Mean (SD)*	*CCA*	CCA uten påstandene	
FNLmaj	0,826 og 0,622	3,74(0,50)	0,77	Påstand 1 Påstand 2	0,71 0,69
FNLmin	0,819 og 0,674	2,46(0,61)	0,81	Påstand 15(2) Påstand 17(4)	0,74 0,76
INLmaj	0,755 og 0,474	3,65(0,46)	0,79	Påstand 8 Påstand 10	0,76 0,76
INLmin	0,759 og 0,435	2,83(0,58)	0,81	Påstand 21(8) Påstand 23(10)	0,77 0,77

4.2 Resultater fra Rasch-analysen

4.2.1 Valg av parametrisering

Konstruktene *FNLmaj, FNLmin, INLmaj* og *INLmin* ga data med tilnærmet like god tilpasning til PCM og RSM (se tabell 7). Valget av parametrisering ble derfor avgjort ved bruk av i Fishers «*likelihood-ratio test*» som indikerte at PCM ga «signifikant» mer informasjon om dataene i alle konstruktene enn det RSM gjorde.

4.2.2 Tilpasning til Rasch-modellen og Guttman-strukturen på et overordnet nivå («summary statistics»)

Tabell 7 viser de observerte verdiene for de originale konstruktene *FNLmaj, FNLmin, INLmaj* og *INLmin* ved bruk av de to parametriseringene av PRM («rating scale» og «partial credit»). Analysene nummerert 1-11 gjengir tilsvarende overordnete data for modifiserte konstrukter etter at underdiskriminerende påstander ble fjernet, påstander med non-uniform DIF ble fjernet og avhengige påstander ble slått sammen i sub-tester eller splittet. Disse analysene vil bli referert til i forbindelse med at modifiseringene blir utført.

Forventningsverdiene til *påstandenes* gjennomsnittlige fit-residual og standardavvik er henholdsvis 0 og 1 siden de er standardiserte z-verdier (Bond & Fox, 2007; Lundgren-Nilsson et al., 2013). Siden alle med unntak av én av de gjennomsnittlige z-fit residualene er negative (se tabell 7), kan vi tolke det slik at konstruktene totalt sett overdiskriminerer noe. Da standardavvikene er mye større enn 1, er det tegn til at variasjonen i påstandenes evne til å diskriminere er langt større enn forventet.

Forventningsverdiene til *personenes* gjennomsnittlige fit-residual og standardavvik er også henholdsvis 0 og 1 siden de er standardiserte z-verdier (Bond & Fox, 2007; Lundgren-Nilsson et al., 2013). Siden alle med unntak av én av de gjennomsnittlige z-fit residualene er negative (tabell 7), kan vi tolke det slik at svarmønstrene til helsesøstrene totalt sett har noe bedre tilpasning til Guttman-strukturen enn forventet. Da standardavvikene er i nær 1, er det tegn til at variasjonen i helsesøstrenes tilpasning til Guttman-strukturen er omtrent som forventet.

Konstruktenes totale X^2-verdier varierer mellom 77 og 177 (se tabell 7). De oppgitte X^2-sannsynlighetene (se tabell 7) viser at det er lite sannsynlig at X^2-verdier av denne størrelsesorden opptrer tilfeldig gitt god tilpasning til endimensjonal polytom Rasch-modell.

Koeffisientene CCA og PSI, som indikerer nedre grense for *reliabilitet*, varierer mellom 0,75 og 0,81. Disse verdiene er tilstrekkelig høye for å gjøre analyser på gruppenivå av helsesøstres svar på spørreskjemaet i SOMAH-DP2.

Ingen påstander viste DIF for personfaktorene «alder» eller «by og landsbygd» som beskrev helsesøstrenes alder og arbeidssted. I konstruktene *INLmaj* og *INLmin* viste påstand 12 og 25 non-uniform DIF for personfaktoren «kommunestørrelse» (se tabell 8). Begge påstandene hadde høye positive z-fit-residualer (se tabell 8). Ved å fjerne påstandene ble *INLmaj* og *INLmin* konstruktenes overordnede tilpasning til Rasch-modellen bedret (analyse 7 og 11, se tabell 7).

DIF-analyse av påstander i konstruktene FNL-stacked og INL-stacked

I konstruktet *FNL-stacked* viste påstandene 1, 2, 3, og 5 non-uniform DIF for variabelen som betegnet helsesøstrenes svar basert på minoritet- og majoritetsbrukere (dvs. datasett 1 og 2) (se tabell 9). I konstruktet *INL*-stacked viste påstandene 7, 8 og 12 non-uniform DIF og påstandene 6, 11 og 13 uniform-DIF for variabelen som betegnet helsesøstrenes svar basert på minoritet- og majoritetsbrukere (dvs. datasett 1 og 2) (se tabell 9). Denne variabelen beskriver *ikke* en personfaktor knyttet til helsesøstrene, men indikerer om

dataene beskriver helsesøstrenes vurdering av ernæringsfremmende allmenndannelse hos majoritet- og minoritetsbrukere av helsestasjonen. Analysene er ikke gjort på «original-dataene», men på «stacked data» slik det er beskrevet i metodekapittelet.

Tabell 7: Tilpasningsstatistikk for konstruktene *FNLmaj, FNLmin, INLmaj* og *INLmin*

	Holdningsnivå		Påstand z-fit		Person z-fit		X^2-verdier			Reliabilitet		Modell
Konstrukt (analyse nr)	Mean	SD	Mean	SD	Mean	SD	*Verdi*	*DF*	*P*	*PSI*	*CCA*	
FNLmaj	1,86	1,73	-0,65	2,72	0,45	0,94	97	30	0,000	0,76	0,77	PCM
FNLmaj	2,22	1,67	-0,67	2,43	-0,48	1,02	156	30	0,000	0,75	0,77	RSM
FNLmin	-1,68	1,64	-0,26	1,95	-0,61	1,49	77	45	0,002	0,79	0,80	PCM
FNLmin	-1,46	1,60	-0,34	2,10	-0,63	1,51	120	45	0,000	0,78	0,80	RSM
INLmaj	1,31	1,28	-1,12	2,60	-0,58	1,28	127	72	0,000	0,78	0,79	PCM
INLmaj	1,19	1,29	-1,40	3,53	-0,58	1,24	132	72	0,000	0,78	0,79	RSM
INLmin	-0,54	1,15	0,58	3,13	-0,37	1,34	177	72	0,000	0,81	0,80	PCM
INLmin	-0,60	1,12	-0,25	4,32	-0,42	1,36	126	72	0,000	0,80	0,80	RSM
FNLmaj (analyse 1)	2,24	2,00	-0,46	0,80	-0,98	2,27	147	20	0,000	0,75	0,76	PCM
FNLmaj (analyse 2)	1,51	1,51	-0,18	2,66	-0,39	0,87	75	24	0,000	0,72	0,77	PCM
FNLmaj (analyse 3)	1,06	1,60	-0,52	0,69	-0,50	0,92	49	41	**0,182**	0,65	-	PCM
FNLmin (analyse 4)	-1,62	1,50	0,12	2,02	-0,55	1,32	59	36	0,008	0,76	0,74	PCM
FNLmin (analyse 5)	-1,87	1,67	-0,03	1,17	-0,59	1,29	72	55	**0,059**	0,72	-	PCM
INLmaj (analyse 6)	1,15	1,19	-0,60	2,18	-0,47	1,18	119	63	0,000	0,75	0,75	PCM
INLmaj (analyse 7)	1,35	1,31	-1,16	1,26	-0,54	1,17	104	42	0,000	0,74	0,75	PCM
INLmaj (analyse 8)	1,13	1,22	-0,54	1,72	-0,55	1,19	107	81	0,016	0,71	-	PCM
INLmaj (analyse 9)	1,34	1,35	-0,99	1,45	-0,69	1,28	83	68	**0,104**	0,69	-	PCM
INLmin (analyse 10)	-0,57	1,07	0,78	2,78	-0,34	1,19	159	63	0,000	0,78	0,75	PCM
INLmin (analyse 11)	-0,29	1,32	0,26	1,86	-0,45	1,16	61	45	**0,052**	0,77	0,74	PCM
Ønsket verdi	**0**	**1**	**0**	**< 1,4**	**0**	**< 1,4**			**> 0,05**	**>0,85**	**>0,85**	

Tabell 7 viser tilpasningsstatistikk for konstruktene *FNLmaj*, *FNLmin*, *INLmaj* og *INLmin*, for de opprinnelige dataene og de «modifiserte» (analyse 1-11). Tilpasningsstatistikken er beskrevet ved bruk av helsesøstrenes gjennomsnittlige holdningsnivå med standardavvik (SD), gjennomsnittlig z-fit-residual (z-fit) med SD for data fra påstander og personer. X^2-verdier, frihetsgrader (DF) og X^2-sannsynlighet (*P*) for alle konstrukt. Indre konsistens reliabilitet er oppgitt som person separasjonsindeks (*PSI*) og Cronbach`s alpha (*CCA*) (CCA er ikke oppgitt for konstrukter med ufullstendige datasett («missing»).

Targeting

Konstruktet *INLmin* hadde relativ god *targeting* da helsesøstrenes gjennomsnittlige holdningsnivå lå nær påstandenes affektive gjennomsnittsnivå (satt til 0 logit). Konstruktene *FNLmaj*, *FNLmin* og *INLmaj* hadde svakere targeting da helsesøstrenes gjennomsnittlige holdningsnivåer avvek noe fra påstandenes gjennomsnittlige affektive nivå (satt til 0 logit). De gjennomsnittlige holdningsnivåene med SD var 1,86 (1,73), -1,68 (1,64), 1,31 (1,28) og -0,54 (1,15) for henholdsvis *FNLmaj*, *FNLmin*, *INLmaj* og *INLmin* (se tabell 7).

For å kunne sammenligne helsesøstrenes holdningsnivå når det gjelder allmenndannelse innen kosthold og ernæring blant spedbarnsmødre og gravide med minoritet- og majoritetsbakgrunn, må vi «linke» datasettene 1 og 2 (helsesøstrene svarte ut fra brukere med majoritet- og minoritetsbakgrunn) til hverandre og danne et felles sammenligningsgrunnlag.

Ved bruk av *«stack»-analyse* ble datasettene 1 og 2 slått sammen for å «linke» datasettene til hverandre, slik det er beskrevet i metodekapittelet (Looveer & Mulligan, 2009). Denne analysen gjør det mulig å sammenligne helsesøstrenes holdningsnivå med hensyn på de to brukergruppene.

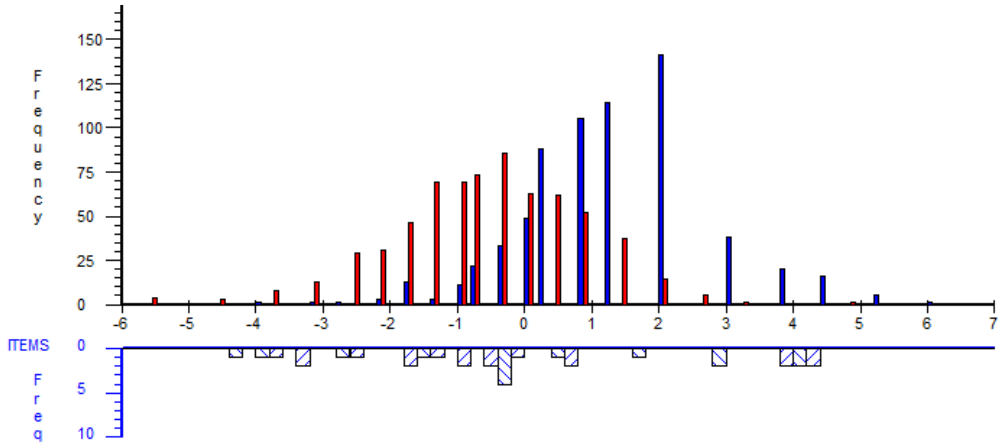
Helsesøstrenes gjennomsnittlige holdningsnivå var signifikant forskjellig ($p < 0{,}001$ beregnet ved ANOVA) avhengig av om de hadde svart på påstandene med tanke på majoritet eller minoritetsbrukere (data ikke vist). Helsesøstrenes gjennomsnittlige holdningsnivå med SD på konstruktet *FNL-stacked* var 1,74 (1,63) og -1,57 (1,56) for henholdsvis majoritet- og minoritetsbrukere (estimert på bakgrunn av «stacked» data). Tilsvarende tall for konstruktet *INL.stacked* var 1,20 (1,20) og -0,55 (1,17).

Etter å ha slettet påstander som ga data med non-uniform DIF og splittet påstander som ga data med uniform DIF for variabelen som betegnet helsesøstrenes svar basert på minoritet- og majoritetsbrukere (dvs. datasett 1 og 2), ble helsesøstrenes gjennomsnittlige holdningsnivå med SD på konstruktet *INL-stacked* estimert til 1,21 (1,37) og -0,55 (1,31) for henholdsvis majoritet- og minoritetsbrukere (fortsatt signifikant forskjell $p < 0{,}001$ beregnet ved ANOVA). Det betyr at «bias» i dataene som skyldes DIF ikke markant endret de gjennomsnittlige estimerte holdningsnivåene på konstruktet *INL-stacked*.

Det var imidlertid ikke mulig å estimere helsesøstrenes gjennomsnittlige holdningsnivå på konstruktet *FNL-stacked*, fordi alle med unntak av én påstand ga data med non-uniform DIF for variabelen som betegnet helsesøstrenes svar basert på minoritet- og majoritetsbrukere (dvs. datasett 1 og 2). Påstandene diskriminerer dermed

ikke på samme måte langs trekket for de to brukergruppene. Påstander eller oppgaver som benyttes som «ankere» i en linkeprosedyre bør være frie for «bias» i form av DIF (Looveer & Mulligan, 2009).

Diagrammet i figur 12 viser fordelingen av helsesøstrenes holdningsnivåer (søyler over førsteaksen) i konstruktet *INL-stacked* etter at påstander med uniform DIF var splittet og påstander med non-uniform DIF var slettet. De blå søylene viser fordelingen av helsesøstrenes holdningsnivåer når de har svart med tanke på majoritetsbrukere og de røde søylene viser fordelingen av holdningsnivåene når de har svart med tanke på minoritetsbrukere. Søylene under førsteaksen viser fordelingen av påstandenes terskelverdier.



Figur 12: Fordeling av helsesøstrenes holdningsnivåer (over førsteaksen) for konstruktet *INL-stacked*. Helsesøstrenes holdningsnivå avhenger av om de har svart på påstander med tanke på majoritetsbrukere (blå søyler) eller minoritetsbrukere (røde søyler). Søylene under førsteaksen viser fordelingen av påstandenes terskelverdier.

Figur 12 viser grafisk forskjellen mellom helsesøstrenes oppfattete INL-nivå hos majoritet og minoritetsbrukerne. Når helsesøstrene har svart på påstandene med tanke på majoritetsbrukere (blå søyler) ligger hoveddelen av holdningsnivåene til høyre for nullpunktet på førsteaksen, mens når de har svart med tanke på minoritetsbrukerne (røde søyler) ligger hoveddelen av holdningsnivåene til venstre for nullpunktet.

4.2.3 Tilpasning til Rasch-modellen og Guttman-strukturen på «individuelt nivå» (data fra enkeltpåstander og personer)

Ifølge påstandenes X^2-statistikk og z-fit-residualer hadde påstandene 1, 2, 12, 23 og 25 svakest tilpasning til Rasch-modellen, men av disse var bare påstandene 12 og 25

underdiskriminerende (se tabell 8 for tilpasningsstatistikk og tabell 3 for påstandenes ordlyd). Svarene på de to påstandene var i tillegg påvirket av personfaktoren «kommunestørrelse», slik at påstandene viste non-uniform DIF (se tabell 8). Påstand 3 i konstruktet *FNLmaj* var underdiskriminerende (se tabell 8), men fjerning av påstanden førte ikke til en betydningsfull forbedring av *FNLmaj* konstruktet (analyse 1) og påstanden ble beholdt i videre analyser.

I konstruktet *FNLmaj* hadde én person z-fit-residual utenfor intervallet ±2,5. I konstruktene *FNLmin*, *INLmaj* og *INLmin* hadde henholdsvis 79, 64 og 35 personer z-fit-residualer utenfor intervallet ±2,5. Hovedandelen av disse helsesøstrene hadde z-fit-residualer mindre enn -2,5 (data ikke vist) – altså statistisk sett «uventet god» tilpasning til Guttman-strukturen. Helsesøstre med z-fit-residualer større enn 2,5 har statistisk sett «uventet» dårlig tilpasning til Guttman-strukturen, men tilpasningsanalyser gjort uten disse personene ga ingen betydningsfull forbedring for konstruktene *FNLmaj*, *FNLmin*, *INLmaj* eller *INLmin* sin overordnede tilpasning til Rasch-modellen (data ikke vist).

Målefeilen til estimatene av påstandenes affektive nivå var relativt små for alle påstandene (se tabell 8). En viktig årsak til dette er at ingen påstander hadde svært lavt eller høyt affektivt nivå i forhold til helsesøstrenes holdningsnivåer. Målefeilene til estimatene for terskelverdier lokalisert ved veldig lave eller veldig høye affektive nivåer hadde som forventet størst målefeil, fordi det er få personer med så lave og høye holdningsnivåer at informasjonen blir sparsom. Målefeilene var imidlertid relativt beskjedne og varierte mellom 0,1 og 0,3 logits.

Tabell 8: Tilpasningsstatistikk for påstandene i konstruktene *FNLmaj*, *FNLmin*, *INLmin* og *INL*. Påstandene er sortert etter affektivt nivå innenfor hvert konstrukt.

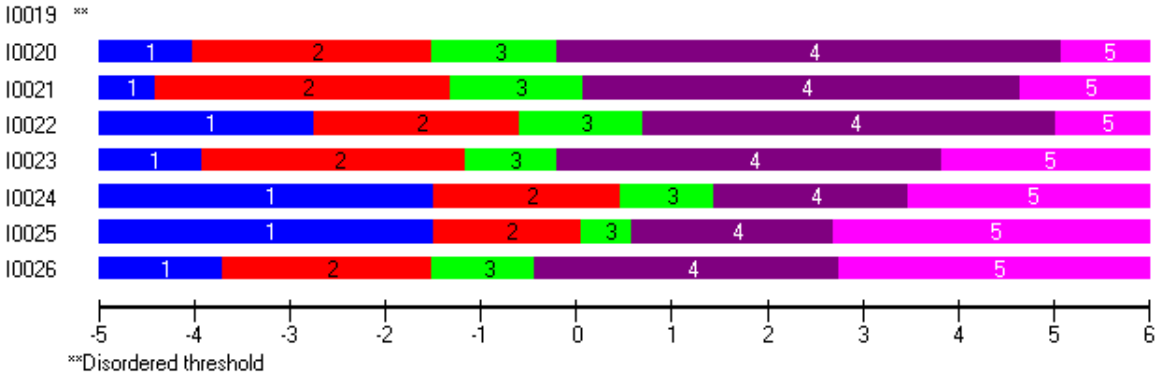
Konstrukt	Påstand	Affektivt nivå	*SE*	*z-fit*	*DF*	X^2	*DF*	*P* (X^2)	Uordnet svarkategorier	DIF
FNLmaj	1	-0,77	0,07	-2,6	525	27,5	6	**0,000**		
FNLmaj	2	-0,37	0,08	-4,0	525	22,7	6	**0,001**		
FNLmaj	5	-0,33	0,07	0,6	525	12,0	6	0,061		
FNLmaj	4	0,28	0,07	-0,2	525	18,5	6	0,005		
FNLmaj	3	1,18	0,07	2,9	525	15,9	6	0,014		
FNLmin	14 (1)	-0,64	0,06	-0,1	520	7,4	9	0,597		
FNLmin	18 (5)	-0,61	0,07	2,3	520	14,8	9	0,098		
FNLmin	15 (2)	0,01	0,07	-3,2	520	21,7	9	0,010		
FNLmin	17 (4)	0,55	0,06	-0,3	520	12,3	9	0,197		
FNLmin	16 (3)	0,69	0,06	0,0	520	21,1	9	0,012		
INLmaj	10	-1,27	0,07	-3,2	577	12,4	9	0,194		
INLmaj	8	-0,53	0,07	-3,3	577	14,9	9	0,095		
INLmaj	7	-0,46	0,08	-3,9	577	21,8	9	0,009		
INLmaj	13	-0,25	0,07	-1,5	577	8,3	9	0,507		
INLmaj	6	0,03	0,06	-0,1	577	5,5	9	0,789		
INLmaj	9	0,62	0,06	-1,0	577	10,8	9	0,293		
INLmaj	11	0,72	0,06	0,0	577	12,8	9	0,170		
INLmaj	12	1,15	0,06	4,2	577	40,4	9	**0,000***		non-unif.
INLmin	26 (13)	-0,71	0,05	1,0	578	8,0	9	0,531		
INLmin	19 (6)	-0,50	0,05	-1,7	578	14,8	9	0,096	x	
INLmin	23 (10)	-0,43	0,06	-2,4	578	32,0	9	**0,000**		
INLmin	21 (8)	-0,23	0,06	-1,7	578	27,3	9	0,001		
INLmin	20 (7)	-0,19	0,06	-0,7	578	17,7	9	0,039		
INLmin	25 (12)	0,46	0,05	7,3	578	57,6	9	**0,000***		non-unif.
INLmin	22 (9)	0,62	0,05	1,0	578	5,7	9	0,770		
INLmin	24 (11)	0,98	0,05	2,0	578	13,9	9	0,125		

Tabell 8 presenterer tilpasningsstatistikk for FNL- og INL-konstruktene avhengig av om helsesøstrene har svart på påstandene med tanke på majoritetsbrukere (*FNLmaj* og *INLmaj*) eller minoritetsbrukere (*FNLmin* og *INLmin*). Dataene er beskrevet ved påstandenes affektive nivåer, «standard målefeil/error» (*SE*), z-fit-residualer (*z-fit*), antall frihetsgrader (*DF*), kji-kvadratverdier (X^2) og kji-kvadratsannsynlighet (PX^2). Videre viser tabellen hvilke påstander som hadde uordnete svarkategorier, det vil si påstander hvor Likert-fempunktskalaen ikke fungerte tilfredsstillende. Tabellen oppgir også hvilke påstander som viste non-uniform DIF (non-unif). Påstand 12 og 25 viste begge non-uniform DIF for personfaktoren «kommunestørrelse». Underdiskriminerende påstander er merket med *, og påstandene med svakest tilpasning til Rasch-modellen er merket med

fet skrift. Påstander i parentes er de samme som i konstruktet ovenfor (påstandene 14-26 er identiske med 1-13).

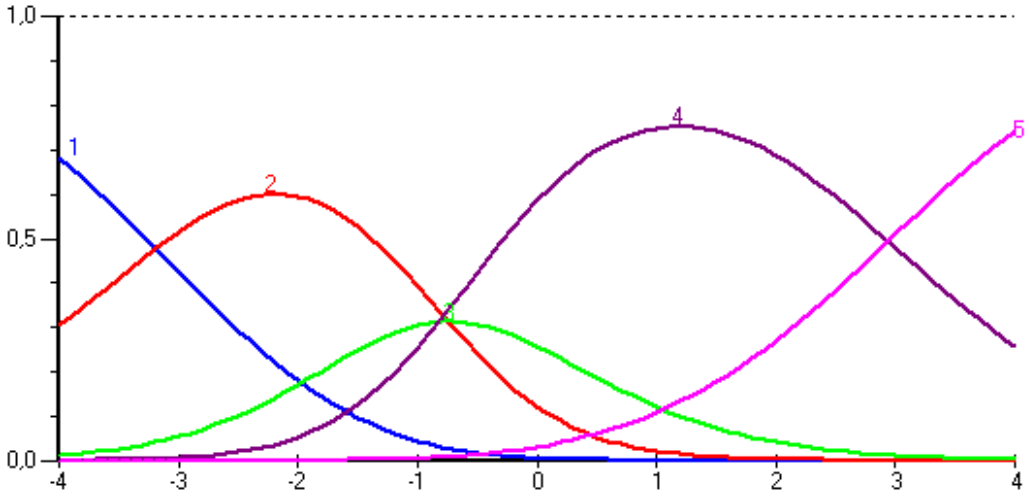
4.2.4 Terskelverdier og svarkategorier

Med unntak av påstand 19 i konstruktet *INLmin* hadde alle påstandene i de originale *FNLmin*, *FNLmaj* og *INLmaj*-konstruktene ordnete svarkategorier. Med det menes at svar i høyere svarkategorier representerer høyere holdningsnivåer og at svarkategoriene fungerer som ønsket. Figur 13 viser en oversikt over påstandenes terskelverdier i konstruktet *INLmin*, hvor ** markerer reverserte terskelverdier for påstand 19.

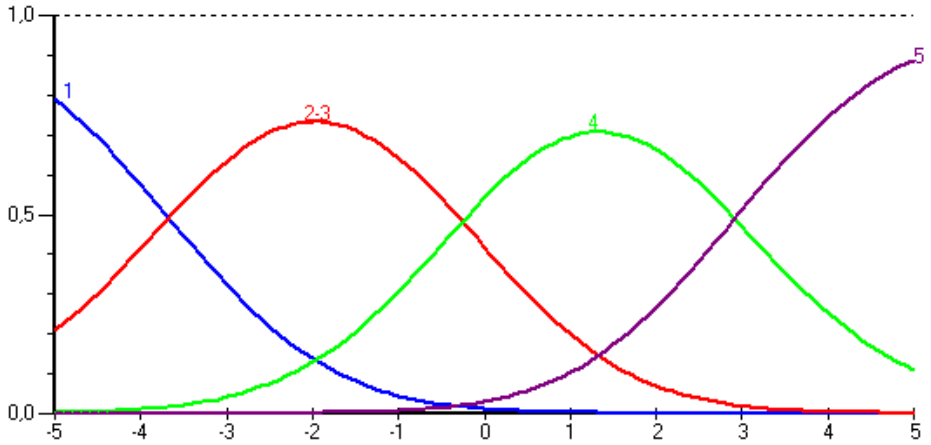


Figur 13: Oversikt over terskelverdiene til påstandene i konstruktet *INLmin*. ** symboliserer reverserte terskelverdier for påstand 19.

De uordnede terskelverdiene til påstand 19 var lokalisert mellom svarkategori 2 og 3 og mellom svarkategori 3 og 4. Figur 14 viser sannsynlighetskurver for svarkategoriene til påstand 19. Vi ser at svarkategori 3 «verken eller» ikke var den mest sannsynlige svarkategorien, uansett holdningsnivå. De reverserte terskelverdiene gir dermed et inntrykk av at skillet mellom svarkategori 3 og 4 (terskel 3) er lokalisert ved et lavere holdningsnivå enn skillet mellom svarkategori 2 og 3 (terskel 2). Dette kan oppfattes som at det kreves «lavere» holdningsnivå for overgangen mellom svarkategori 3 «verken eller» og 4 «i stor grad», enn for overgangen mellom svarkategori 2 «I liten grad» og 3 «verken eller». Svarkategoriene 2 og 3 ble slått sammen (inngår i analyse 10). Dette medførte at terskelverdiene ikke lenger var reverserte, og at svarkategoriene fungerte tilfredsstillende (se figur 15).



Figur 14: Sannsynlighetskurver for svarkategoriene i påstand 19. Kurvene viser at svarkategoriene ikke fungerte tilfredsstillende, siden svarkategori 3 «verken eller» ikke er det mest sannsynlige valget for noe holdningsnivå. Terskelverdiene mellom svarkategori 2 og 3 og mellom svarkategori 3 og 4 er reverserte.



Figur 15: Sannsynlighetskurver for svarkategoriene i påstand 19 etter å ha slått sammen svarkategori 2 og 3. Terskelverdiene er ikke lenger reverserte og svarkategoriene fungerer tilfredsstillende.

Påstand 6 (samme påstand som påstand 19) i konstruktet *INLmaj* hadde tilsvarende problem med midtkategorien etter at påstandene 7 og 8 var slått sammen eller splittet for å løse mulig svaravhengighet. Påstand 6 ble reskåret tilsvarende påstand 19, hvilket førte til at svarkategoriene ble ordnet på samme måte som vist i figur 15 (reskåring ingikk i analyse 6 og 8).

4.2.5 Analyse av DIF

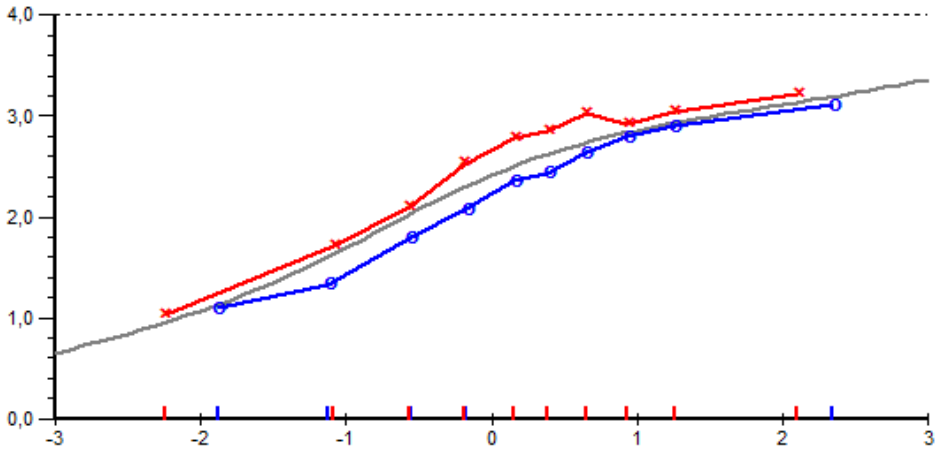
Påstander med uniform DIF ble kun observert i *INL-stacked* konstruktet. Her viste påstandene 6, 11 og 13 uniform DIF (se tabell 9).

Tabell 9: Tilpasningsstatistikk for alle påstander i konstruktene *FNL-stacked* og *INL-stacked*. Påstandene er sortert etter affektivt nivå.

Konstrukt	Påstand	Affektivt nivå	*SE*	*Z-Fit*	*X2*	*DF*	*P*(X^2)	Svar.kat.	DIF	*P*(DIF)
FNL-stacked	1	-0,88	0,04	-4,20	12	7	0,099		non-unif.	0,000
FNLstacked	2	-0,38	0,04	-7,94	30	7	0,000		non-unif.	0,000
FNLstacked	5	0,04	0,05	3,32	9	7	0,269		non-unif	0,000
FNLstacked	4	0,22	0,05	-0,97	7	7	0,384			0,709
FNLstacked	3	1,00	0,05	4,33	15	7	0,039		non-uniform	0,000
INL-stacked	8	-0,602	0,04	-5,88	18	9	0,031		non-unif	0,000
INL-stacked	7	-0,584	0,04	-5,00	15	9	0,105		non-unif	0,000
INL-stacked	13	-0,497	0,04	-0,14	6	9	0,696		uniform	0,000
INL-stacked	10	-0,416	0,04	-4,76	18	9	0,037			0,086
INL-stacked	6	-0,273	0,04	0,70	6	9	0,761		uniform	0,000
INL-stacked	9	0,452	0,04	0,48	5	9	0,868			0,106
INL-stacked	11	0,922	0,04	-0,14	5	9	0,817		uniform	0,000
INL-stacked	12	0,998	0,04	8,68	28	9	0,001		non-unif	0,000

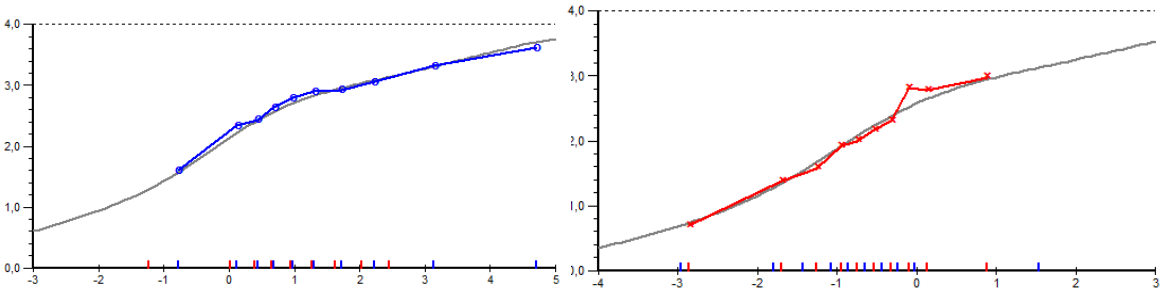
Tabell 9 beskriver tilpasningsstatistikk for påstandene i konstruktene *FNL-stacked* og *INL-stacked* ved bruk av deres affektive nivåer med standard målefeil/error (*SE*), dataenes z-fit-residualer (*z-fit*), kji-kvadratverdier (X^2), frihetsgrader (*DF*) og kji-kvadrat sannsynligheter ($P(X^2)$). Tabell 9 viser videre hvilke påstander som viste DIF, DIF sannsynlighet (*P(DIF)*), og om DIF var non-uniform (non-unif) eller uniform.

Hver av påstandene 6, 11 og 13 ble splittet i to nye «virtuelle» påstander, én for kategorien «majoritet» og en for «minoritet». Figur 16 viser ICC og helsesøstrenes observerte verdier (svar) på påstand 6. Vi kan se at påstanden diskriminerer på samme måte for de to personfaktorkategoriene langs hele «trekket» (langs førsteaksen), slik at det dannes to parallelle kurver som ikke krysser hverandre. Den røde kurven forbinder helsesøstrenes observerte verdier for minoritetsbrukere, mens den blå kurven forbinder de observerte verdiene for majoritetsbrukere. Kurvene i figur 16 viser at helsesøstrene er «mer enig» i påstanden om at minoritetsbrukere stiller flere spørsmål om helse og kosthold under konsultasjonen, enn majoritetsbrukere, uavhengig av brukernes oppfattete INL-nivå (langs førsteaksen).



Figur 16: Sammenligning av helsesøstrenes observerte verdier for majoritetsbrukere (blå kurve) og minoritetsbrukere (rød kurve) på påstand 6 «De stiller meg ofte spørsmål om helse og kosthold under konsultasjonen».

Figur 17 viser ICC og helsesøstrenes observerte verdier på de to nye «virtuelle» påstandene som ble opprettet fra påstand 6. Forskjellen i affektivt nivå mellom de to virtuelle påstandene fra påstand 6 var 0,61 logits. De affektive nivåene for de to virtuelle påstandene avhengig av om helsesøstrene hadde svart med tanke på majoritet eller minoritetsbrukere var henholdsvis 0,05 logits og -0,56 logits.



Figur 17: ICC og helsesøstrenes observerte verdier på de nye «virtuelle» påstandene basert på påstand 6. Den blå kurven viser helsesøstrenes observerte verdier når de svarte med tanke på majoritetsbrukere, mens den røde kurven viser for minoritetsbrukere.

For de virtuelle påstandene opprettet fra påstand 11 skilte det 0,21 logits i affektivt nivå mellom majoritetsbrukere (0,93 logits) og minoritetsbrukere (0,72 logits). Mens for de virtuelle påstandene fra påstand 13 skilte det 0,51 logits i affektivt nivå mellom minoritetsbrukere (-0,77 logits) og majoritetsbrukere (-0,26 logits) (data ikke vist). Når påstand 11 med størst DIF ble fjernet «forsvant» imidlertid DIF for påstand 13. Den observerte DIFen i påstand 13 kan være kunstig og komme som følge av DIF-analysene

av de andre påstandene (Andrich & Hagquist, 2012). Siden det kan være vanskelig å avgjøre hvilke påstander som viser ekte eller kunstig DIF ble det valgt å splitte alle påstandene med uniform DIF.

4.2.6 Svaravhengighet

I konstruktene *FNLmaj* og *FNLmin* hadde påstandene 1 og 2, og 14 og 15 (tilsvarende påstander) en positiv residualkorrelasjon større enn 0,2 over den gjennomsnittlige residualkorrelasjonen i konstruktene. Residualkorrelasjonene var imidlertid svært lave, kun 0,128 og 0,015 i henholdsvis *FNLmaj* og *FNLmin*. I forsøk på å løse opp den mulige svaravhengigheten ble påstandene enten slått sammen (analyse 2 og 4) eller splittet (analyse 3 og 5) (se tabell 7). Den metoden som ga dataene best tilpasning til Rasch-modellen (vurdert ut fra X^2-statistikk) ble benyttet. Begge konstruktene ga data med best overordnet tilpasning til Rasch-modellen etter at splitfunksjonen var benyttet, (p = 0,182 og 0,059) for henholdsvis *FNLmaj* og *FNLmin* konstruktene (se tabell 7).

Påstandene 7 og 8 i konstruktet *INLmaj* hadde en høy positiv residualkorrelasjon på 0,476. De «samme» påstandene, påstand 20 og 21 i konstruktet *INLmin* hadde også høy positiv residualkorrelasjon på 0,387. I forsøk på å løse opp svaravhengigheten ble påstandene enten slått sammen eller splittet (i separate analyser). Den metoden som ga dataene best tilpasning til Rasch-modellen (vurdert ut fra X^2-statistikk) ble benyttet. I konstruktet *INLmaj* ble påstandene 7 og 8 først slått sammen (analyse 6). Deretter ble påstand 12 som var underdiskriminerende slettet (analyse 7). Ingen av disse modifiseringene førte til tilstrekkelig tilpasning. Først da påstandene ble splittet (analyse 8) og påstand 12 ble slettet (analyse 9) ga *INLmaj* konstruktet data med tilstrekkelig overordnet tilpasning til Rasch-modellen (p = 0,104) (se tabell 7).

Påstandene 20 og 21 i *INLmin* konstruktet ble slått sammen (analyse 10) og de underdiskriminerende påstandene 24 og 25 ble slettet (analyse 11). Dette resulterte i at konstruktet ga data med tilstrekkelig overordnet tilpasning til Rasch-modellen (p = 0,052). Splitfunksjonen ble derfor ikke benyttet.

Det ble observert tilsvarende svaravhengighet mellom de samme påstandene i *FNL-stacked* og *INL-stacked* konstruktene. Residualkorrelasjonene var henholdsvis 0,117 (> 0,2 over gjennomsnittlig residualkorrelasjon) og 0,442 mellom påstandene 1 og 2 og 7 og 8 i konstruktene *FNL-stacked* og *INL-stacked*.

5. Diskusjon

Dette kapittelet starter med en diskusjon av utvalget/referanserammen, før enkelte resultater fra KTT blir diskutert i lys av moderne testteori. Hoveddelen av kapittelet er viet diskusjon rundt forskningsspørsmålene i oppgaven. Avslutningsvis oppsummeres hovedfunnene og en konklusjon blir gitt.

5.1 Utvalg/referanseramme

Den kvantitative delen av SOMAH-DP2 er en tverrsnittsundersøkelse som har til hensikt å beskrivelse den populasjonen som utvalget er trukket ut fra (Ringdal, 2007). Dataene gir et øyeblikksbilde av situasjonen, og disse bør ikke brukes for å trekke slutninger om «tilstanden» langt frem i tid (Ringdal, 2007).

Utvalgsmetoden i SOMAH-DP2 var en form for «sannsynlighetsutvelging med stratifiseringsstrategi» (Ringdal, 2007). Først ble helsestasjonene stratifisert etter landsdel, og deretter ble helsestasjoner tilfeldig trukket ut. En slik stratifisering sikret at alle landsdeler ble inkludert i studien.

Antallet helsestasjoner fra Sørlandet var imidlertid underrepresentert, mens Østlandet var overrepresentert. Mulige årsaker til dette kan være temamotivasjon, erfaringer eller utfordringer med helsekommunikasjon ovenfor personer med minoritetsbakgrunn. Slik at helsesøstre tilsatt ved helsestasjoner i områder med en høy andel minoritetsbrukere vil kanskje ha større «motivasjon» / se nytten av undersøkelsen enn helsesøstre tilsatt ved helsestasjoner i områder med en lav andel minoritetsbrukere. Dette kan ha betydning for funnenes «generaliserbarhet» - i hvilken grad «funnene» kan overføres fra utvalget til den populasjonen som utvalget representerer. Vi vet heller ikke hvordan helsestasjonene som ikke ønsket å delta i studien eventuelt skiller seg fra de som valgte å delta. Utvalgsmetoden kunne imidlertid vært forbedret ved at helsesøstre innad i en utvalgt helsestasjon også hadde blitt randomisert. Begrunnelsen for å gjøre dette er at helsesøstre fra samme helsestasjon *kan* ha noe mer til felles enn andre, og at antall uavhengige respondenter dermed nærmer seg eller er lik antall helsestasjoner.

Påstandene i spørreskjemaet var utviklet med bakgrunn i resultater fra kvalitative for-undersøkelser, noe som prinsipielt kan bidra til å øke spørreskjemaets innholdsrelaterte evidens (DeVellis, 2012). Bruk av kvalitative undersøkelser som intervju eller fokusgrupper i forkant av spørreskjemautvikling er anbefalt og anses å være

et viktig ledd i utviklingen av spørreskjemaer (DeVellis, 2012). Undersøkelsens innholdsrelaterte evidens er i tillegg styrket ved at FNL og INL er teoretisert og beskrevet i litteraturen (Diamond, 2007; Guttersrud et al., 2013; Silk et al., 2008). Påstandene som er brukt for å måle FNL og INL i SOMAH-studien (og dermed også i denne masteroppgaven) har derfor teoretisk forankring. Hvorvidt dataene fra påstandene faktisk fanger opp FNL og INL – «måler det de utgir seg for å måle» – er tett knyttet til dataenes konstruktrelaterte evidens. Dette er forsøkt studert empirisk gjennom Rasch-analysene i masteroppgaven.

5.1.1 Rasch-analyse

Analyse av dataenes X^2-statistikk og z-fit-residualer kan vise hvilke påstander som gir data med sterk eller svak tilpasning til Rasch-modellen. Rasch-analysene sier imidlertid ikke noe om *hvorfor* påstandene gir data med sterk eller svak tilpasning til modellen.

Sletting av påstander bør gjøres med varsomhet, siden dette kan endre eller svekke det konstruktet som undersøkelsen skal fange opp (Andrich & Marais, 2012; Lundgren-Nilsson & Tennant, 2011). I denne masterstudien ble dataene fra påstander som ga data med svak tilpasning forsøkt justert for å bedre tilpasningen. Sletting av påstander som ga data med svak tilpasning ble bare valgt som en siste løsning, og data ble slettet for å undersøke hvordan dette endret dataene og påvirket tilpasningen til Rasch-modellen.

Ved store utvalg vil selv små avvik fra Rasch-modellen kunne bli signifikante (Andrich & Marais, 2012). Mens mindre utvalg har normalt mindre «styrke» når det kommer til å «avdekke» mulige avvik fra Rasch-modellen. Og ved små utvalg vil estimatene for påstandenes affektive nivå normalt ha større målefeil (Linacre, 1994).

For å få *stabile estimater* av *påstandenes affektive* nivå, bør antall respondenter tilsvare totalt antall «terskler» (antall svarkategorier langs Likert-skalaen minus 1) i konstruktet multiplisert med en faktor av størrelsesorden 20-30[1]. For eksempel vil et konstrukt bestående av fem påstander med femdelt Likert-skala kreve mellom 5x(5-1)x20 og 5x(5-1)x30 respondenter – altså mellom 400 og 600 respondenter (Linacre, 1999, 2002; Tennant & Lundgren-Nilsson, 2013). FNL og INL konstruktene inneholder henholdsvis 5 og 8 påstander. Totalt antall terskler i FNL konstruktet vil være 5x(5-1) mens det i FNL konstruktet vil være 8x(5-1). For å oppnå tilstrekkelig stabile estimater

[1] Et eksempel for å illustrere sammenhengen mellom utvalgsstørrelse og estimering av stabilt vanskegradnivå for testoppgave: Hvis én person har svart på en flervalgsoppgave vil estimatet av p-verdi være 0 eller 1. Hvis to personer har svart kan estimatet bli 0, 0,5 eller 1.

av påstandenes affektive nivåer i *FNLmin* og *FNLmaj* bør derfor utvalgsstørrelsen være mellom 400 og 600 personer, mens utvalgsstørrelsen for *INLmin* og *INLmaj* bør være mellom 640 og 960 personer. Utvalgsstørrelsen på N = 666 i SOMAH-DP2 bør således kunne være tilstrekkelig stor til å gi stabile estimater av påstandenes affektive nivåer, særlig for *FNLmin* og *FNLmaj* konstruktene.

For å få *stabile estimater* av *respondentenes holdningsnivåer* må antall terskler (avhenger av antall påstander og antall svarkategorier langs Likert-skalaen) være tilstrekkelig høyt. Størrelsen på målefeilen til estimatene av personenes holdningsnivåer avhenger også av targeting, slik det ble forklart i teorikapittelet. I følge Linacre (2002) bør antallet respondenter ved bruk av PCM tilsvare antallet svarkategorier i en påstand, multiplisert med en faktor av størrelsesorden 100 eller mer, for å gi stabile estimater av holdningsnivå. For en fempunkts Likert-skala tilsvarer dette en utvalgsstørrelse rundt N= 500.

For å ha tilstrekkelig styrke i dataene til å *separere* personer basert på holdningsnivå, bør erfaringsmessig (tommelfingerregel) antall terskler i et spørreskjemakonstrukt overstige om lag 15-20 og i en kunnskapstest 20-40, men dette avhenger i vesentlig grad av hvor snevert konstruktet er definert. Anslagene er basert på erfaringer med hvor mange påstander, spørsmål eller oppgaver som gir tilstrekkelig høy Cronbach's alpha. Høy alpha impliserer høy sann varians, som betyr god spredning i oppnådde poengsummer og dermed god evne til å separere respondentene. Lie (2010) viser også til at holdningsundersøkelser generelt kan imidlertid inneholde færre terskler innenfor et konstrukt, og likevel inneha tilstrekkelig alpha sammenlignet med kunnskapstester.

5.2 Diskusjon av resultater

I denne delen av diskusjonskapittelet diskuteres funnene fra Rasch-analysen, og eksplisitt funnene for det første og andre forskningsspørsmålet som omhandler dataenes reliabilitet og validitet. Imidlertid har jeg valgt og først, kort trekke frem enkelte resultater fra KTT, sett i lys av resultatene fra Rasch-analysene.

5.3 Resultater fra KTT sett fra et Rasch-analytisk perspektiv

En utdypende sammenligning av analyseresultater fra KTT og Rasch-analyse er ikke sentralt i denne oppgaven, men det kan være interessant å trekke frem enkelte resultater fra KTT i et Rasch-analytisk perspektiv.

Som nevnt i metodekapittelet, er faktoranalyse en korrelasjonsmodell som vurderer korrelasjonen (ladningen) mellom påstander og én eller flere komponenter/faktorer (Pallant, 2007). Påstander med høy korrelasjon på en felles komponent/faktor ofte er ansett som «gode» i faktoranalyse (Sick, 2011), men slike påstander er imidlertid vist å kunne overdiskriminere og være svaravhengige i Rasch-analyse (Sick, 2011). Dette vises også til en viss grad empirisk i dataene i denne masteroppgaven. For eksempel viste faktoranalyse at påstandene 1, 2, 7, 8, 14, 15, 20 og 21 hadde de høyeste komponentladningene i *FNLmaj, FNLmin, INLmaj* og *INLmin* konstruktene. Rasch-analyse viste at i de fleste tilfellene var påstandene overdiskriminerende, dog med unntak av påstandene 14, 20 og 21. Videre viste Rasch-analysene at påstandene 1, 2, 14, 15, 7, 8, 20 og 21 hadde mulig svaravhengighet.

Reliabilitetskoeffisientene CCA og PSI var relativt like for alle konstruktene. Reliabilitetsanalyser viste at påstandene 1, 2, 8, 10, 15, 17, 21 og 23 hadde størst innflytelse for CCA-verdien i konstruktene *FNLmaj, FNLmin, INLmaj* og *INLmin*. Med unntak av påstandene 17 og 21 overdiskriminerte alle påstandene i Rasch-analysen. Når overdiskriminerende/svaravhengige påstander ble «tatt hensyn til» gjennom sammenslåing eller splitting i Rasch-analyse, ble PSI-verdien til konstruktene redusert (se tabell 7). Dette viser hvordan overdiskriminerende/svaravhengige påstander kan øke reliabilitetskoeffisientene. I KTT er oppgaver med høy diskrimineringsevne ofte ansett som gode (Masters, 1988), og påstander som bidrar til høy CCA-verdi i konstruktet vil gjerne bli beholdt. I moderne testteori kan overdiskriminerende påstander gi data med for svak tilpasning til modellen. Få vil imidlertid anse dette som et særlig «problem».

5.4 Resultater fra Rasch-analyse

5.4.1 Konstruktenes, påstandenes og personenes tilpasning til Rasch-modellen

Ingen av de opprinnelige *FNLmaj, FNLmin, INLmaj* og *INLmin* konstruktene hadde tilstrekkelig overordnet tilpasning til Rasch-modellen. Tilpasningen ble imidlertid forbedret ved å slå sammen eller splitte påstandene med mulig svaravhengighet, splitte eller slette påstandene med DIF, og å slette underdiskriminerende påstander med svak tilpasning til Rasch-modellen.

Det var i stor grad de samme påstandene i *INLmaj* og *INLmin* konstruktene som hadde svak tilpasning. For eksempel var de likelydende påstandene 10 i konstruktet *INLmaj* og 23 i konstruktet *INLmin* overdiskriminerende, mens de likelydende

påstandene 12 i *INLmaj* og 25 i *INLmin* underdiskriminerende. Påstandene 10 (23) og 12 (25) bør bearbeides og prøves ut på nytt, for de samlet trolig inn «overflødig» informasjon eller målte for mye «annet» enn det de var tiltenkt å måle.

Slik FNL er målt i denne oppgaven, var rangeringen av påstandene etter affektivt nivå i *FNLmaj* og *FNLmin* konstruktene relativt lik. For eksempel, hadde påstand 1(14) «*De leser god*t» lavest affektivt nivå, og påstand 3(16) «*De gir inntrykk av å ha lest helse- og kostholdsbrosjyrene som jeg gir dem*» høyest affektivt nivå i begge konstruktene (dvs. i datasett 1 og datasett 2). En mulig tolkning er at helsesøstrene mener at begge brukergruppene har grunnleggende leseferdigheter, men at det å lese «vanskeligere» tekster som for eksempel kostholdsbrosjyrer, krever et «høyere FNL-nivå». I litteraturen er det vist at kun en liten andel brukere av ulike helsestasjoner leser utlevert materiell som de får på helsekonsultasjoner (Szwajcer, Hiddink, Koelen & van Woerkum, 2009). Dette representerer en annen mulig forklaring på det høye affektive nivået til påstand 3(16).

5.4.2 Svarkategorier og terskelnivå

Den «nøytrale» midtkategorien «verken eller» forårsaket uordnete svarkategorier for påstand 19 i konstruktet *INLmin*. Påstand 6 i konstruktet *INLmaj*, som er tilsvarende påstand 19 i *INLmin*, indikerte samme problem med nevnte midtkategori. Reskåring til en firedelt Likert-skala ga ordnete terskelverdier.

Helsesøstrenes svarmønstre viste at den midtre svarkategorien hadde relativt få observerte verdier i forhold til tilstøtende svarkategorier. Årsaken til de reverserte terskelnivåene kan derfor skyldes at helsesøstrene ikke diskriminerte sterkt nok mellom ordlyden i de tilstøtende svarkategoriene «i liten grad», «verken eller», eller «i stor grad» og «verken eller» for akkurat denne påstanden. Påstanden bør følgelig omformuleres og testes ut på nytt.

Andre studier har også rapportert problemer med kategoriske måleskalaer som har en «nøytral» midtkategori (Comins et al., 2007; Guttersrud et al., 2013). For å unngå problemer med «nøytrale» midtkategorier i Likert-skalaer, kan det heller benyttes en *partallskala* med for eksempel 4 eller 6 svarkategorier. Måleskalaer med mellom 7 og 9 svarkategorier har høyere reliabilitet enn skalaer med lavere antall svarkategorier (Preston & Colman, 2000). Dette skyldes at flere svarkategorier øker variansen til fordelingen av personestimatene og dermed evnen til å separere personer. En utfordring med et høyt antall svarkategorier i Rasch-analyse er imidlertid at det kan øke sannsynligheten for

reverserte terskelnivåer, fordi respondentene har problemer med å diskriminere mellom kategoriene når antall kategorier blir for høy (Tennant & Lundgren-Nilsson, 2013).

Svarkategorienes tilhørende ordlyd i et spørreskjema kan forårsake problemer fordi de kan oppfattes og tolkes forskjellig fra person til person. For eksempel betyr ikke nødvendigvis «sjelden», «ofte», «uenig» eller «enig» det samme for alle respondenter. Alternativt kan bare skalaens ekstreme kategorier beskrives med ord. Da vil respondentene i større grad føle at det dreier seg om en «gradert» holdning mellom ytterpunktene (Preston & Colman, 2000).

5.4.3 Reliabilitets estimater

FNLmaj, FNLmin, INLmaj og *INLmin* konstruktene hadde tilsynelatende tilstrekkelig relabilitet. Reliabilitetskoeffisienten Cronbach's alpha i konstruktene kan imidlertid fremstå som noe forhøyet på grunn av de observerte bruddene på lokal uavhengighet gjennom svaravhengighet. De nye estimatene i de «modifiserte» konstruktene fra analysene 2-11 i tabell 7 er sannsynligvis et riktigere mål på *indre konsistens reliabilitet*, ettersom det er tatt hensyn til avhengigheten i dataene.

5.4.4 Targeting

Konstruktet *INLmin* hadde god «targeting», men for konstruktet *FNLmin* var helsesøstrenes gjennomsnittlige holdningsnivå lavere enn påstandenes affektive nivå. I konstruktene *FNLmaj* og *INLmaj* var helsesøstrenes gjennomsnittlige holdningsnivå høyere enn påstandens gjennomsnittlige affektive nivå.

Spredningen av terskelnivåene var større enn spredningen av påstandene og den høyeste og laveste terskelverdien for hver påstand hadde i de fleste tilfeller størst målefeil. Dette skyldes at de var lokalisert i områder langs det latente trekket hvor det var få respondenter.

For å sammenligne helsesøstrenes holdningsnivåer brukes estimatene fra konstruktene *FNL-stacked* og *INL-stacked* fremfor holdningsestimatene fra *FNLmaj, FNLmin, INLmaj* og *INLmin.* Sammenligninger av helsesøstrenes holdningsnivå tyder på at helsesøstrene har en tydelig oppfatning om at spedbarnsmødre og gravide tilhørende majoritetsbefolkningen i gjennomsnitt har høyere kompetanse innen *INL* enn spedbarnsmødre og gravide tilhørende minoritetsbefolkningen. Tilsvarende

sammenligning av helsesøstrenes holdningsnivå kunne ikke gjøres for *FNL-stacked* konstruktet siden 4 av 5 påstander viste non-uniform DIF.

På den annen side kan helsesøstrenes ulike oppfatning av majoritets- og minoritetsbrukere sine INL-nivåer skyldes selve studiedesignet. Helsesøstrene ble bedt om å svare på to likelydende påstandssett, hvor det første omhandler majoritetsbrukere og det andre omhandlet minoritetsbrukere. Det kan således tenkes at, når helsesøstrene ble bedt om å «tenke på» minoritetsbrukere når de skulle svare på de samme påstandene én gang til, så skapte det en slags forventning om at minoritetsbrukernes FNL- og INL-nivå måtte fremstilles lavere enn hos majoritetsbrukerne.

5.4.5 DIF-analyser

I følge Bjorner and Pejtersen (2010) kan DIF svekke spørreskjemaets konstruktrelaterte evidens, siden en persons svar er påvirket av andre faktorer enn personens holdningsnivå og dermed vil ikke påstandene «fungere likt» for alle personene (Kreiner, 1999).

I DIF-analysene av *FNLmaj, FNLmin, INLmaj* og *INLmin* konstruktene ga ikke helsesøstrenes alder utslag. Det var heller ingen påstander som viste DIF for om helsesøstrene arbeidet i en by eller ikke. Derimot hadde størrelsen på arbeidskommunen noe å si for helsesøstrenes svar på påstandene 12 og 25 i *INLmaj* og *INLmin* konstruktene. Men DIF var non-uniform og vi kan derfor ikke si noe om at enkelte kommunestørrelser ble «favorisert» fremfor andre, bare at påstandene ikke funger «likt» mellom kommunestørrelsene.

DIF-analysene av påstand 6 i konstruktet *INL-stacked* viste at helsesøstrene var mer «enige» i at minoritetsbrukerne i større grad enn majoritetsbrukerne stilte spørsmål under konsultasjonene. En mulig årsak til dette kan være at mange minoritetsbrukere har mangelfulle norskspråklige ferdigheter, slik at de må stille flere spørsmål for å kunne forstå hva helsesøstrene faktisk snakker om. En annen mulig forklaring kan være at de norske brukerne som besøker helsestasjonen føler at de har «gode nok» kunnskaper om ernæring, og følgelig spør sjelden.

Kvalitative data fra andre delprosjekter i SOMAH viste at flere innvandrerkvinner opplevede at kostrådene de mottok var forvirrende, fordi de ikke samsvarte med kostråd fra deres hjemland eller var i tråd med deres matkultur (Garnweidner-Holme, 2013). Tilsvarende er også sett i en norsk studie av pakistanske innvandrere med diabetes type 2 som hadde vanskeligheter med å følge kostråd fordi de ikke var i overenstemmelse med

kostråd og matkulturen fra hjemlandet (Fagerli, Lien & Wandel, 2005). Samtidig viser kvalitative fokusgruppeintervjudata fra SOMAH-prosjektet at helsesøstrene opplever forskjeller mellom majoritet- og minoritetsbrukerne sin språkforståelse, og at dette skaper helsekommunikasjonsutfordringer for mange helsesøstre (Helsesøster og PhD-student Bettina Fagerlund, HiOA, muntlig kommunikasjon, 17. oktober 2013).

DIF-analysen av påstand 11 viste at helsesøstrene opplevde at majoritetsbrukere, i større grad enn minoritetsbrukere var opptatt av å diskutere informasjon de hadde fått gjennom media om hva som er sunn mat for barn. Samtidig viste DIF-analysen av påstand 13 at helsesøstrene oppfattet at minoritetsbrukerne, i større grad enn majoritetsbrukerne, var interessert i å få vite hva som var regnet for å være et sunt kosthold for deres barn. En mulig forklaring kan være at de norske brukerne av helsestasjonen generelt sett er mer opplyst om mat og kostholdsinformasjon, for eksempel via media.

Forskjeller i majoritet og minoritetsbrukere sitt syn på helsevesenet og pasientroller kan eventuelt også være av betydning. Majoritetsbrukere er vant til et helsevesen hvor pasientrettigheter og autonomi er viktige prinsipper (Heløe, 2012). Et utjevnet autoritetsforhold mellom helsepersonell og pasient kan tenkes å gi større rom for dialog i konsultasjonene. Minoritetsbrukere kan imidlertid være vant til mer «paternalistiske» helsesystemer med lite autonomi og pasientrettigheter, samt stort autoritetsskille mellom helsepersonell og pasient. Det kan tenkes at helsesøstre som møter minoritetskvinner med slike forventninger til helsevesenet, kan oppleve at brukerne gjerne vil stille spørsmål og få «ekspert» -råd, men ikke nødvendigvis gå i aktiv dialog med helsepersonell.

Å utføre DIF-analyser kan føre til at enkelte påstander viser kunstig DIF (Andrich & Hagquist, 2012). Påstand 13 i *INL-stacked* i konstruktet mistenkes for å kunne ha kunstig DIF som følge av DIF i påstandene 11 og eller 6 i *INL-stacked* konstruktet. Begrunnelsen for dette er at DIF i påstand 13 «forsvinner» når påstand 11 eller 6 fjernes, men ikke motsatt. Det er imidlertid komplisert å avgjøre hvilke påstander som har ekte og kunstig DIF og dette ligger utenfor omfanget av denne oppgaven.

Påstandene med uniform DIF (6, 11 og 13 i *INL-stacked*) kunne beholdes ved å splitte de mellom kategoriene «majoritet» og «minoritet». Flere påstander viste imidlertid non-uniform DIF (1, 2, 3, 7, 8 og 12 i *FNL-stacked og INL-stacked*) og måtte forkastes. Siden relativt mange påstander viste DIF mellom majoritet- og minoritetsbrukere, tyder det på at spørreskjemaet kan være lite robust når det gjelder å måle invariant på tvers av kulturell bakgrunn.

5.4.6 Svaravhengighet

Etter å ha slått sammen eller splittet påstander med mulig svaravhengighet avtok PSI-verdiene for konstruktene *FNLmaj, FNLmin, INLmaj* og *INLmin*, og dette styrker mistanken om svaravhengighet (Marais & Andrich, 2008a, 2008b). Den totale X^2-verdien til konstruktene sank og X^2-sannsynligheten økte, noe som betyr at dataenes overordnete tilpasning til Rasch-modellen økte når avhengigheten var tatt hensyn til. Videre styrkes mistanken om svaravhengighet ved at det ble observert mulig svaravhengighet mellom de *samme* påstandene i *FNLmaj* og *FNLmin* og *INLmaj* og *INLmin* konstruktene. Påstandene 7 (20): «*De klarer å fortelle meg om sin eventuelle helsebekymring på en klar og forståelig måte*» og 8 (21) «*De klarer å fortelle meg om sine eventuelle bekymringer om matvaner på en klar og forståelig måte*» måler for mye av det samme og samler dermed inn «overflødig» informasjon. Mest sannsynlig opplever helsesøstrene at brukere som klarer å fortelle om sine helsebekymringer nødvendigvis også klarer å fortelle om sine kostholdsbekymringer. For påstandene 1 (14) «*De leser godt*» og 2 (15) «*De forstår innholdet i helse- og kostholdsbrosjyrene som jeg gir dem*» kan det være slik at helsesøstrene oppfatter at personer med god leseferdighet i større grad forstår innholdet i kostholdsbrosjyrene enn personer med begrenset leseferdighet.

5.5 Begrensninger

Det er viktig å ha i mente at dataene som foreligger i masterstudien er *indirekte* sammenlignende målinger av FNL- og INL-nivå hos majoritet og minoritetsbrukere av helsestasjoner. Målingene baserer seg på helsesøstrenes *oppfatning* av brukernes FNL- og INL-nivå. Helsesøstrenes refleksivitet, kulturforståelse og kommunikasjonsevne kan en betydning for denne oppfattelsen. Som diskutert tidligere *kan* selve spørreskjema-designet som er utviklet av de ansvarlige for SOMAH-studien ha bidratt til uheldig «underbevisst degradering» av innvandrerkvinners FNL- og INL-nivå.

Det antas at utvalgsstørrelsen på N=666 er tilstrekkelig for kunne å detektere DIF på 0,50 logits med to personfaktorkategorier av 300 personer (Linacre, 2013). Utvalgsstørrelsen kan imidlertid være for liten til å detektere DIF for personfaktorkategorien «kommunestørrelse» siden denne inneholder fire kategorier.

Den mulige svaravhengigheten mellom påstandene 1 (14) og 2 (15) i *FNLmaj* og *FNLmin* konstruktene var basert på relativt lave residualkorrelasjoner på om lag 0,2 over

gjennomsnittlig residualkorrelasjon. En høyere positiv residualkorrelasjon ville følgelig indikert en sterkere svaravhengighet mellom påstandene. På den annen side gir data med relativt god tilpasning til Rasch-modellen ofte lave (nær 0) eller negative residualkorrelasjoner (Lundgren-Nilsson et al., 2013). Dermed kan selv små residualkorrelasjoner indikere svaravhengighet (Tennant & Lundgren-Nilsson, 2013). Det er mulig å estimere styrken på svaravhengigheten også mellom polytome påstander (Andrich, Humpry, et al., 2012), men dette er arbeidskrevende analyser og ikke så sentralt i denne sammenheng hvor et instrument blir validert.

6. Oppsummering og konklusjon

Dataene fra konstruktene *FNLmaj, FNLmin, INLmaj og INLmin* har tilsynelatende tilstrekkelig reliabilitet. Men reliabilitetskoeffisientene til konstruktene kan fremstå som noe høye grunnet mulig svaravhengighet. Helsesøstrenes svar på enkelte påstander ser ut til å kunne være påvirket av brukernes kulturelle bakgrunn. Spørreskjemaet ser derfor ut til å være lite robust for å måle invariant mellom ulike brukergruppers FNL- og INL-nivå.

Enkelte påstander i spørreundersøkelsen er «underdiskriminerende», hvilket gir data med svak tilpasning til Rasch-modellen. Disse påstandene måler tilsynelatende «for mye av noe annet enn det de utgir seg for å måle». Påstandene bør følgelig videreutvikles språklig eller erstattes med andre påstander. Det samme gjelder avhengige påstander som samler inn «overflødig» informasjon. Eventuelle nye påstander bør bidra til å justere konstruktenes affektive nivå, slik at påstandenes gjennomsnittlige affektive nivå og er bedre tilpasset helsesøstrenes faktiske holdningsnivå. Arbeidet med å videreutvikle spørreskjemaet lå imidlertid utenfor arbeidsrammen til denne masteroppgaven.

Videre antas det at bruk av fire- eller seks-delte Likert-skalaer kan løse utfordringer knyttet til uordnete svarkategorier. Seksdelte skalaer vil mest sannsynlig øke reliabiliteten til målingene gjennom å øke variansen til fordelingen av personenes holdningsestimater.

Det er *ikke* «psykometrisk evidens» for at konstruktene *FNLmaj* og *FNLmin* på en reliabel og valid måte kan beskrive forskjeller mellom ernæringsfremmende allmenndannelse hos spedbarnsmødre og gravide i majoritet- og minoritetsbefolkningen. Konstruktene *INLmaj* og *INLmin* ser derimot ut til, i modifisert form, å gi data som kan brukes til å beskrive slike forskjeller.

Denne studien viser at Rasch-analyser av data fra utprøvinger av måleinstrumenter har stort potensiale hva angår bearbeiding og videreutvikling av instrumentene *før* hovedinnsamling av data skjer. Dette gjelder både eksisterende og nyutviklete konstruktbaserte spørreskjemaer som har til hensikt å måle personers holdninger innen forebyggende folkehelsearbeid generelt og ernærings- og helsefremmende allmenndannelse spesielt.

7. Litteraturliste

Aarnes, S. B. (2009). *Utvikling og utprøving av et spørreskjema for å kartlegge nutrition literacy: assosiasjon til kjønn, utdannelse og fysisk aktivitetsnivå.* Masteroppgave, Høgskolen i Akershus). Lillestrøm.

Abebe, S. (2010). Public health challenges of immigrants in Norway: A research review. NAKMI report 2:2010. Hentet 03.11.13 http://www.migrasjonsforskning.no/site-no/04-Publikasjoner/pdf/Public%20Health%20Challenges.pdf

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-574.

Andrich, D. (1988). *Rasch models for measurement*: Beverly Hills: Sage Publications.

Andrich, D. (2005). The Rasch model explained. I S. Alagumalai, D. D. Curtis, & N. Hungi (Red.), *Applied Rasch Measurement: A book of Exemplars* (s. 308-328). Hungi: Springer.

Andrich, D. & Hagquist, C. (2004). *Detection of Differential Item Functioning Using Analysis of Variance*. Paper presentert på: Second International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch Models. Murdoch University, Perth, Australia.

Andrich, D. & Hagquist, C. (2012). Real and Artificial Differential Item Functioning. *Journal of Educational and Behavioral Statistics, 37*(3), 387-416.

Andrich, D., Humpry, S. M. & Marais, I. (2012). Quantifying Local, Response Dependence Between Two Polytomous Items Using the Rasch Model. *Applied Psychological Measurement,*(36), 309-324.

Andrich, D. & Kreiner, S. (2010). Quantifying Response Dependece Between Two Dichotomous Items Using the Rasch Model. *Applied Psychological Measurement,*(34), 181-191.

Andrich, D. & Marais, I. (2012). *Introduction to Rasch measurement of modern test theory – course notes*, Perth: University of Western Australia.).

Rumm 2030: Rasch Unidimensional Measurement Models (software). RUMM Laboratory Perth, Western Australia.

Araï, D. (2010). Moderne testteori: Rasch-modellen og utvidelse av modellen. I M. Martinussen (Red.), *Kvantitativ forskningsmetodologi i samfunns- og helsefag*. Bergen: Fagbokforlaget.

Azizi, M., Aghaee, N., Ebrahimi, M. & Ranjbar, K. (2011). Nutrition knowledge, the attitude and practices of college students *Physical Education and Sport, 9*(3), 349-357.

Baker, F. (2001). *The basics of Item Response Theory* (2. utg.).

Belvedere, S. L. & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of clinical epidemiology*, *63*(12), 1287-1297. Hentet 18.11.13 fra http://www.ncbi.nlm.nih.gov/pubmed/20971422

Bjorner, J. B. & Pejtersen, J. H. (2010). Evaluating construct validity of the second version of the Copenhagen Psychosocial Questionnaire through analysis of differential item functioning and differential item effect. *Scandinavian Journal of Public Health*, *38*(90), 90-105.

Bland, J. M. & Altmann, D. G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal*, *310*(170).

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in Human Science* (2. utg.). Mahwah, New Jersey 07430: Lawrence Erlbaum Associates.

Brodersen, J., Meads, D. & Kreiner, S. (2007). Methodological aspects of differential item functioning in the Rasch model. *Journal of Medical Economics*,(10), 309-324.

Brown, J. D. (2000). What is construct validity. *JALT Testing & Evaluation*, *4*(2), 8-12.

Comins, J., Brodersen, J., Krogsgaard, M. & Beyer, N. (2007). Rasch analysis of the Knee injury and Osteoarthritis Outcome Score (KOOS): a statistical re-evaluation. *Scandinavian Journal of Medicine & Science in Sports*, *18*(3), 336-345.

Creative Research System. (2012). *Sample size calculator*. Hentet 03.11.2013 fra http://www.surveysystem.com/sscalc.htm

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Dalane, J. Ø. (2011). *Nutrition literacy hos sykerpleierstudenter*, Masteroppgave, Høgskolen i Oslo og Akershus). Lillestrøm.

Davis, T. C., Crouch, M. A., Long, S. W., Jackson, R. H., Bates, P. & George, R. B. e. a. (1991). Rapid assessment of literacy levels of adult primary care patients. *Family Medicine*, *23*(6), 433-435.

DeAyala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Demars, C. (2010). *Item response theory*. New York: Oxford University Press.

DeVellis, R. F. (2012). *Scale Development Theory and Applications* (3. utg.). Thousand Oaks, California: SAGE Publications Ltd.

Diamond, J. J. (2007). Development of a reliable and construct valid measure of nutritional literacy in adults. *Nutrition Journal, 6*(5).

Doyle, G., Cafferkey, K. & Fullam, J. (2012). *The european health literacy survey: results from Ireland*.

Easton, P., Entwistle, V. A. & Williams, B. (2013). How the stigma of low literacy can impair patient-professional spoken interactions and affect health: insights from a qualitative investigation. *BMC Health Serv Res, 13*, 319. doi:10.1186/1472-6963-13-319

Engelhard, G. (2013). *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences*. New York: Routledge.

Engelhard Jr, G. (2008). Historical Perspectives on Invariant Measurement: Guttman, Rasch and Mokken. *Measurement, 6*, 155-189.

Fagerli, R. A., Lien, M. E. & Wandel, M. (2005). Experience of dietary advice among Pakistani-born persons with type 2 diabetes in Oslo. *Appetite, 45*(3), 295-304. doi:10.1016/j.appet.2005.07.003

Finbråten, H. S. & Pettersen, S. (2009a). Diabetessykepleiere i Norge sine oppfatninger av pasienters health literacy. *Vård i Norden, 32*(105), 47-52.

Finbråten, H. S. & Pettersen, S. (2009b). Kunnskap er egenmakt. *Sykepleien, 97*(5), 60-63.

Finbråten, H. S. & Pettersen, S. (2012). En norsk pilotstudie av helsesøstres oppfatninger av pasienters health literacy: helsefremmende allmenndannelse. *Nordisk Tidsskrift for Helseforskning, 1*(8).

Forskning.no. (2011). *PISA-testen favoriserer noen land*. Hentet 18.11.13 fra http://www.forskning.no/artikler/2011/mai/287473

Frisbie, D. A. (1988). Reliability of scores from teacher-made test. *Educational Measurement: Issues and Practice, National Council on Measurement in Education., 7*(1), 25-35.

Garnweidner-Holme, L. M. (2013). Mulige utfordringer ved å formidle kostråd til en flerkulturell befolkning. *Norsk Tidsskrift for ernæring*,(3), 32-36.

Garnweidner, L. M., Pettersen, S. K. & Mosdol, A. (2013). Experiences with nutrition-related information during antenatal care of pregnant women of different ethnic backgrounds residing in the area of Oslo, Norway. *Midwifery*. doi:10.1016/j.midw.2012.12.006

Goldstein, H. (1970). Consequences of Using the Rasch Model for Educational Measurement. *British Educational Research Journal, 5*(2), 211-220.

Gosse, M. (2012). *Rasch analysis of food choice survey data (abstract)*. Paper presentert på: Australian Consortium for Social and Political Research Incorporated (ACRISPI) RC33 Eighth International Conference on Social Science Methodology. Sydney, Australia.

Guttersrud, Ø., Dalane, J. Ø. & Pettersen, S. (2013). Improving measurement in nutrition literacy research using Rasch modelling: examining construct validity of stage-spesific' critical nutrition literacy' scales. *Public Health Nutrition*,(11), 1-7.

Guttman, L. A. (1950). The basis for scalogram analysis. I S. A. Stouffer, L. A. Guttman, E. A. Schuman, S. Lazarfelds, A. Star, & J. A. Clausen (Red.), *Measurement and prediction* (Bind 4, s. 60-90). Princeton NJ: Princeton University Press.

Hagquist, C. & Andrich, D. (2004). Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch modeling. *Personality and Individual Differences, 36*, 955-968.

Hanson-Divers, E. C. (1997). Developing a medical achievement reading test to evaluate patient literacy skills: A preliminary study. *Journal of Health Care for the Poor and Underserved*, *8*(1), 56-59.

Helsedirektoratet. (2010). *Utviklingsstrategi for helsestasjons- og skolehelsetjenesten (IS-1798).* Oslo: Helsedirektoratet.

Helsedirektoratet. (2011). *Forebygging, utredning og behandling av overvekt og fedme hos voksne: Nasjonale retningslinjer for primærhelsetjenesten.* Oslo: Helsedirektoratet.

Heløe, L. A. (2012). Fra paternalisme til pasientrettigheter. *Tidskrift for Den norske legeforening, 4*(136), 434-436.

Hendriks, J., Fyfe, S., Styles, I., Skinner, S. R. & Merriman, G. (2012). Scale construction utilising the Rasch unidimensional measurement model: A measurement of adolescent attitudes towards abortion. *Australas Med J, 5*(5), 251-261. doi:10.4066/AMJ.2012.952

Hjellset, V. (2010). *A culturally adapted lifestyle intervention with main focus on blood glucose regulation improved the rsik profile for type 2 diabetes in Pakistani immigrant women. They are not aliens.*, Det Medisinske fakultet, Universitetet i Oslo). Oslo.

Hjörleifsdóttir, K. S. (2013). *Diabetes among Turkish immigrants in Sweden. A study of prevalence and risk factors.*, Karolinska Institutet). Sverige. Hentet 18.11.13 fra http://publications.ki.se/xmlui/bitstream/handle/10616/41592/Spikblad.pdf?sequence=2

Holbert, R. L. & LaMarre, H. L. (2011). Expanding the Use of Structural Equation Modeling (SEM) in Political Communication. I E. P. Bucy, & L. Holbert (Red.), *The Surcebook for Political Communication Research Methods, Measures, and Analytical Techniques*. New York: Routledge.

IBM SPSS Statistics for Windows. Armonk, New Tork: IBM Corp.

Ishikawa, H., Nomura, K., Sato, M. & Yano, E. (2008). Developing a measure of communicative and critical health literacy: a pilot study of Japanese office workers. *Health Promotion International, 23*(3), 269-274.

Ishikawa, H., Takeuchi, T. & Yano, E. (2008). Measuring Functional, Communicative, and Critical Health Literacy among Diabetic Patients. *Diabetes Care, 31*(5), 874-879.

Jenum A, K., Holme, I., Graff-Iversen, S. & Birkeland, K. (2005). Ethnicity and sex are strong determinants of diabetes in an urban Western society: implications for prevention. *Diabetologia*,(48), 435-439.

Jenum, A., K., Mørkrid, K., Sletner, L., Vange, S., Torper, J., L., Nakstad, B., . . . Birkeland, K., I. (2012). Impact of ethnicity on gestational diabetes identified with the WHO and the modified International Association of Diabetes and Pregnancy Study Groups criteria: a population-based cohort study. *European Journal of Endocrinology*,(166), 317-324.

Johannessen, A. (2009). *Introduksjon til SPSS* (4. utg.). Oslo: Abstrakt forlag.

Kadam, P. & Bhalerao, S. (2010). *Sample size calculation*. Hentet 03.11.13 fra http://www.ncbi.nlm.nih.gov/pubmed/20532100

Kjærnsli, M., Lie, S., Olsen, R. V. & Roe, A. (2007). *TID FOR TUNGE LØFT. Norske elevers kompetanse i naturfag, lesing og matematikk i PISA 2006*. Oslo: Universitetsforlaget AS.

Kjærnsli, M. & Roe, A. (2009). *On the right track Norwegian Students' proficiency in Reading, Mathematics and Science Literacy in the PISA Assessment 2009*. www.pisa.no.

Kjøllesdal, J. G. (2009). *Nutrition literacy: utvikling og utprøving av et spørreskjema som måler grader av nutrition literacy*, Masteroppgave, Høgskolen i Akershus). Lillestrøm.

Kommuneforlaget. (2010). Hentet 03.11.13 fra http://www.kommuneforlaget.no/kf/hjem/

Kreiner, S. (1999). Validation of index scales for analysis of survey data - the Symptom Index. I K. Dean (Red.), *Population Health research - linking theory and methods* (s. 116-144). Beverly Hills: Sage.

Kreiner, S. (2011). *Is the foundation under PISA solid? A critical look at the scaling model underlying international comparisons of student attainment.* Hentet 18.11.13 fra http://ncm3.ncm.chalmers.se/media/ncm/dokument/pisa_kreiner_.pdf

Kreps, G. L. & Sparks, L. (2008). Meeting the health literacy needs of immigrant populations. *Patient Education and Counseling, 71*(3), 328-332.

Larsen, I. F. (2000). Diabetes hos ikke-vestlige innvandrere. *Tidskrift for Den norske legeforening,*(120), 2804-2806.

Lie, S. (2010). Måling av kunnskap og holdninger i et krysskulturelt perspektiv. I M. Martinussen (Red.), *Kvantitativ forskningsmetodologi i samfunns- og helsefag*. Oslo: Fagbokforlaget.

Linacre, J. M. (1994). Sample Size and Item Calibration (or Pearson Measure) Stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (1997). KR-20 / Cronbach Alpha or Rasch Person Reliability: Which Tells the "Truth"? *Rasch Measurement Transactions, 11*(3), 580-581.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement, 3*(1), 85-106.

Linacre, J. M. (2013). Differential Item Functioning DIF Sample Size Nomogram. *Rasch Measurement Transactions, 26:4 p. 1391*. Hentet 18.11.13 fra http://www.rasch.org/rmt/rmt264a.htm

Looveer, J. & Mulligan, J. (2009). The efficacy of link items in the construction of a numeracy achievement scale from Kindergarten to Year 6. *Journal of applied measurement, 34*, 181-192.

Lundgren-Nilsson, A., Jonsdottir, I. H., Ahlborg, G., Jr. & Tennant, A. (2013). Construct validity of the Psychological General Well Being Index (PGWBI) in a sample of patients undergoing treatment for stress-related exhaustion: a Rasch analysis. *Health Qual Life Outcomes, 11*, 2. doi:10.1186/1477-7525-11-2

Lundgren-Nilsson, A. & Tennant, A. (2011). Past and present issues in Rasch measurement analysis: The functional independence measure (FIM) revisited. *Journal of Rehabilitation Medicine, 43*, 884-891.

Marais, I. & Andrich, D. (2008a). Effects of Varying Magnitude and Patterns of Response Dependence in the Unidimensional Rasch Model. *Journal of applied measurement, 9*(2), 105-124.

Marais, I. & Andrich, D. (2008b). Formalizing Dimension and Response Violations of Local Independence in the Unidemensional Rasch Model. *Journal of applied measurement, 9*(3), 200-215.

Masse, L. C., Wilson, M., Baranowski, T. & Nebeling, L. (2006). Improving psychometric methods in health education and health behavior research. *Health*

education research, 21, 1-3. Hentet 18.11.13 fra http://www.ncbi.nlm.nih.gov/pubmed/17122186

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement, 25*, 15-29.

Masters, G. N. (1993). Undesirable item discrimination. *Rasch Measurement Transactions, 7*(2), 289.

Masters, G. N. (2005). Objective Measurement. I S. Alagumalai, D. Curtis, D.,, & N. Hungi (Red.), *The Rasch model explained. Applied Rasch Measurement: A book of Exemplars* (s. 14-25): Springer.

Messick, S. (1989a). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1989b). Validity. I R. L. Linn (Red.), *Educational Measurement* (s. 13-103). New York: Macmillian Publishing Company.

Naqv, A. R. (2012). *Nå slår Diabetesforbundet alarm*. Hentet 02.04.13 fra http://www.utrop.no/Nyheter/Innenriks/22929

NOU 2007: 12. (2007). *Offentlig innkreving*. Oslo: Fornyings-, administrasjons- og kirkedepartementet. Hentet 18.11.13 fra http://www.regjeringen.no/nb/dep/fad/dok/nouer/2007/nou-2007-12/12.html?id=487754

Nutbeam, D. (2000). Health literacy as a public health goal: a challenge for contemporary health education and communication strategiesinto the 21st century. *Health Promotion International, 15*, 259-267.

Paasche-Orlow, M. K., Parker, R. M., Gazmararian, J. A., Nielsen-Bohlman, L. T. & Rudd, R. (2005). Reviews. The Prevalence of Limited Health Literacy. *Journal of General Internal Medicine, 20*, 175-184.

Pallant, J. (2007). *SPSS survival manual. A step by step guide to data analysis using SPSS.* Maidenhead: Open University Press.

Parker, R. M., Baker, D. W., Williams, M. V. & Nurss, J. R. (1995). The test of functional health literacy in adults: a new instrument for measuring patients' literacy skills. *Journal of General Internal Medicine, 10*(10), 537-541.

Parmenter, K. & Wardle, J. (1999). Development of a general nutrition knowledge questionnaire for adults. *European Journal of Clinical Nutrition, 53*, 298-308.

Pettersen, S. (2003). Er også naturfagdidaktikk godt for helsen? I:. I B. B: Bungum, & D. Jorde (Red.), *Naturfagdidaktikk. Perspektiver – Forskning – Utvikling* (s. 273-288). Oslo: Gyldendal Akademisk.

Pettersen, S. (2007). *Health Claims and Scientific Knowledge. A study of How Students of Health and Sciences, their Teachers, and Newspaper Journalists Relate to Health Claims in Society.* , University of Oslo). Oslo.

Pettersen, S. (2009). Kostholdsinformasjon og annen helseinformasjon. I A. Holthe, & B. U. Wilhelmsen (Red.), *Mat og helse i skolen. En fagdidaktisk innføring* (s. 87-100). Bergen: Fagbokforlaget.

Preston, C. C. & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1-15.

Quest back. (2013). Hentet 03.11.13 fra http://www.questback.no/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test.* Copenhagen: Danish Institute for Educational Research.

Resnicow, K., Baranowski, T., Ahluwalia, J. S. & Braithwaite, R. L. (1999). Cultural sensitivity in public health: defined and demystified. *Ethn Dis, 9*(1), 10-21. Hentet 18.11.13 fra http://www.ncbi.nlm.nih.gov/pubmed/10355471

Ringdal, K. (2007). *Enhet og mangfold. Samfunnsvitenskapelig forskning og kvantitativ metode* (2. utg.). Bergen: Fagbokforlaget Vigmostad & Bjørke AS.

RUMM. (2012). Extending the RUMM2030 Analysis (8 utg.). Duncraig, Australia: RUMM Laboratory Pty Ltd.

Ryan, J. P. (1983). Introduction to latent trait analysis and item response theory. I E. D. Hataway (Red.), *Testing in the schools. New Directions for Testing and Measurement, 19*. San Francisco: Jossey-Bass.

Scott, I. & Mazhindu, D. (2009). *Statistics for Health Care Professionals*. London: SAGE Publications.

Sick, J. (2011). Rasch Measurement and Factor Analysis. *JALT Testing & Evaluation, 15*(1), 15-17.

Silk, K. J., Sherry, J., Winn, B., Keesecker, N., Horodynski, M. A. & Sayir, A. (2008). Increasing nutrition literacy: testing the effectiveness of print, web site, and game modalities. *Journal of Nutrition Education and Behavior, 40*(1), 3-10.

Sjøberg. (2009). *Norsk skole lar seg styre for mye av Pisa-undersøkelsen*. Hentet 18.11.13 fra http://www.tu.no/jobb/2012/11/12/-norsk-skole-lar-seg-styre-for-mye-av-pisa-undersokelsen

Smith, R. M. (2000). Fit Analysis in Latent Trait Measurement Models. *Journal of applied measurement, 1*(2), 199-218.

SOMAH. (2013). *Velkommen til SOMAH-prosjektet*. Hentet 18.11.13 fra http://akkordio.net/somah/

Sosial- og helsedirektoratet. (2003). *Kommunenes helsefremmende og forebyggende arbeid i helsestasjons- og skolehelsetjenesten (Veileder til forskrift av 3 april 2003 nr 450).* Oslo: Direktoratet.

Statistisk sentralbyrå. (2013). *Folkemengd og kvartalsvise befolkningsendringar, 2. kvartal 2013*. Hentet 18.11.13 fra http://www.ssb.no/befolkning/statistikker/folkendrkv/kvartal/2013-08-16#content

Steckelberg, A., Hülfenhaus, C., Kasper, J., Rost, J. & Mühlhauser, I. (2009). How to measure critical health competences: development and validation of the Critical Health Competence Test (CHC Test). *Advances in Health Science Education, 14*(1), 11-22.

Szwajcer, E. M., Hiddink, G. J., Koelen, M. A. & van Woerkum, C. M. (2009). Written nutrition communication in midwifery practice: what purpose does it serve? *Midwifery, 25*(5), 509-517. doi:10.1016/j.midw.2007.10.005

Sørensen, K. (2012). Health literacy and publich health: A systematic review and integration of definitions and models. *BMC Public Health, 12*(80).

Sørensen, K., Van den Broucke, S., Fullman, J., Doyle, G., Pelikan, J., Slonska, Z. & Brand, H. (2012). Health literacy and publich health: A systematic review and integration of definitions and models. *BMC Public Health, 12*(80).

Tennant, A. & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and rheumatism, 57*(8), 1358-1362. Hentet fra http://www.ncbi.nlm.nih.gov/pubmed/18050173

Tennant, A. & Conaghan, P. G. (2007). The Rasch Measurement Model in Rheumatology: What Is It and Why Use It? When Should It Be Applied, and What Should One Look for in a Rasch Paper? *Arthritis & Rheumatism (Arthritis Care & Research), 57*(8), 1358-1362.

Tennant, A. & Lundgren-Nilsson, A. (2013). *Introductory Rasch Analysis - course notes*, University of Leeds).

Tesio, L., Simone, A. & Bernardinello, M. (2007). Rehabilitation and outcome measurement: where is Rasch analysis-going? *Europa medicophysica, 43*(3), 417-426. Hentet 18.11.13 fra http://www.ncbi.nlm.nih.gov/pubmed/17921966

Traub, R. E. & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, National Council on Measurement in Education., 10*(1), 37-45.

Van Wyke, J. & Andrich, D. (2006). *A typology of polytomously scored mathematics items disclosed by the Rasch model: implications for constructing a continuum of achievement* (Report No. 2 ARC Linkage Grant LP0454080: Maintaining Invariant Scales in State, National and International Level Assessments). Murdoch University: Western Australia.

Weiss, B. D., Mays, M. Z., Martz, W., Castro, K. M., DeWalt, D. A., Pignone, M. P. & al., e. (2005). Quick assessment of literacy in primary care: The Newest Vital Sign. *Annals of Family Medicine, 3*(6), 514-522.

Wilson, M., Allen, D. D. & Li, J. C. (2006). Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. *Health education research*, *21*, 19-32. Hentet 18.11.13 fra http://www.ncbi.nlm.nih.gov/pubmed/16880221

Wright, B. D. (1996). Time 1 to Time 2 (Pre-test to Post-test) comparison: Racking and Stacking. *Rasch Measurement Transactions, 10*(1), 478.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, Winter,*(52), 33-45.

Wright, B. D. & Mok, M. (2000). Rasch Models Overview. *Journal of applied measurement, 1*(1), 83-106.

Wright, B. D. & Stone, M. H. (1979). *The measurement model. Best Test Design. Rasch Measurement*. Chigaco: Mesa Press.