Reliabilitet ved bruk av Renfrew Bus Story test

Ingvill Nordeide

Masteroppgave i spesialpedagogikk

Det utdanningsvitenskapelige fakultet

Institutt for spesialpedagogikk

UNIVERSITETET I OSLO

30.05.08.

INNHOLD

TABELLISTE

VEDLEGG

Sammendrag

I det longitudinelle forskningsprosjektet ”Child, Language & Learning: The nature and development of language and communication skills in pre-school children.” drevet ved Institutt for spesialpedagogikk (UiO), søkes økt kunnskap om norske barns språkutvikling. Foreløpig er 200 barn blitt testet med et testbatteri bestående av 18 deltester. ”The Renfrew Bus Story” er èn av deltestene som inngår i batteriet. Fokus i oppgaven er å undersøke reliabilitet ved Bus Story skårene slik testen er benyttet i forskningsprosjektet.

I og med at det kun har vært mulig med en empirisk undersøkelse av Bus Story skårenes *vurdererreliabilitet*, har oppgaven et todelt fokus. I oppgavens første del blir det drøftet hvilke feilkilder som kan tenkes å ha påvirket Bus Story skårene i forskningsprosjektet, og videre hvordan reliabiliteten kunne vært undersøkt dersom man hadde all nødvendig informasjon. Metodene som tas opp til drøfting er tradisjonelle estimeringsmetoder utviklet innen klassisk reliabilitetsteori og metoder fra Generalizability theory (G teori).

Drøftingen viser hvordan tradisjonelle estimeringsmetoder vil overestimere Bus Story skårenes reliabilitet, og det konkluderes med at estimering gjennom G teori vil være mest hensiktsmessig. G teori legger dermed grunnlaget for den empiriske estimeringen av Bus Story skårenes vurdererreliabilitet som representerer oppgavens andre del. Det blir gjort et utvalg av barn som har blitt testet og vurdert i forskningsprosjektet, og det gjøres en ny vurdering av dem ved hjelp av reskåring. Estimering av reliabilitetskoeffisienter gjennomføres og det drøftes hvordan koeffisientene kan tolkes. Det konkluderes med at Bus Story skårene har tilfredsstillende vurdererreliabilitet.

Forord

Å skrive en masteroppgave med fokus på reliabilitet har vært både spennende og krevende. Det er mange som fortjener takk.

Veileder Thor Arnfinn Kleven fortjener en stor takk. Du har bidratt med uvurderlig hjelp, støtte og inspirasjon.

Takk til forskergruppen Child, Language & Learning og deres forskningsassistenter som lot meg få tilgang til deres data.

Jeg vil også takke venn og medstudent Heidi Osa Michalsen som har vært en trofast støttespiller gjennom hele studietiden. Takk for gode diskusjoner og støtte i både medgang og motgang.

Til sist vil jeg takke min tålmodige Vegard, familie og venner for oppmuntring og gode avbrekk underveis i prosessen.

Oslo, mai 2008.

Ingvill Nordeide

1. Innledning

1.1 Bakgrunn for valg av tema

Under skriving av masteroppgaven har jeg jobbet som forskningsassistent i et forskningsprosjekt drevet ved Universitetet i Oslo, kalt ”Child Language & Learning: The nature and development of language and communication skills in pre-school children.” (heretter omtalt som språkprosjektet). Dette er en longitudinell undersøkelse som skal følge barns språkutvikling fra de er 4 til 8 år. Språkutvikling hos barn med minoritetsbakgrunn, Cochlea implantat, Down syndrom og spesifikke språkvansker skal ses opp mot en normgruppe bestående av ca 200 barn med normal språkutvikling (Child Language & Learning prosjektsøknad 2007). Som èn av ti forskningsassistenter i prosjektet har jeg testet språkferdigheter i barnegruppen med normal språkutvikling. Barna som nå er 4 år har blitt testet med et testbatteri bestående av 18 deltester. Testbatteriet er konstruert for å fange opp kompleksiteten i språkutviklingen og for å få informasjon om sentrale språklige parametere som vokabular, fonologiske ferdigheter, grammatiske og syntaktiske ferdigheter i muntlige og skriftlige modaliteter (Child Language & Learning prosjektsøknad 2007).

Arbeidet som forskningsassistent vekket interessen for reliabilitet. Opplæring i testbatteriet og gjennomføring av testene gjorde meg oppmerksom på hvor mange feilkilder som kan påvirke testresultater. Som forskningsassistent hadde jeg særlig fokus på egen rolle i prosjektet. Forstår jeg retningslinjer for administrering og skåring likt som de andre forskningsassistentene? Ville barnet fått et annet resultat dersom en av de andre forskningsassistentene skåret det? I tillegg så jeg hvordan dagsform hos barna og andre omstendigheter rundt testingen så ut til å påvirke barnas prestasjon og resultat.

Særlig èn av deltestene i testbatteriet, The Renfrew Bus Story, vekket interessen for reliabilitet. Denne testen skiller seg fra resten av testbatteriet ved at den er en narrativ test. Bus Story administreres ved at man forteller barnet en historie om en buss, og ber deretter barnet om å gjenfortelle historien. Barnets fortelling skåres i forhold til hvor mye relevant informasjon som formidles, og hvor lange setninger de produserer (Renfrew 1997).

Bruk av narrative tester både til diagnostisering og predikering av vansker er blitt anbefalt i en rekke artikler (Bishop & Edmundson 1987, Feagans & Appelbaum 1986, Paul & Smith 1993, Howlin & Kendall 1991, Botting 2002 og Pankratz m.fl. 2007). Noe av begrunnelsen for dette er at narrativ testing er preget av en åpen og naturalistisk testsituasjon som lettere fanger opp kompleksiteten i språket (Botting 2002). Samtidig hevder Brown (1996) at nettopp disse kvalitetene kan representere trusler hva angår reliabilitet. Selv når man måler relativt stabile størrelser som høyde og vekt, kan det snike seg inn målefeil, men truslene ser ut til å være enda større når man skal måle mer komplekse fenomener som for eksempel språk (Brown 1996). I følge Befring (2007) kan slike fenomener være vanskelige å måle fordi de varierer fra situasjon til situasjon og lar seg påvirke av forhold som er irrelevante i vår sammenheng. Fan og Chen (2000) understreker videre hvordan det kan være vanskelig å utforme tydelige skåringsregler av komplekse fenomener, og dermed at samme adferd resulterer i ulike skårer fordi skåringsprosessen ikke er objektiv nok. Howlin og Kendall (1991) vurderte i en av sine undersøkelser en rekke språktester, og fant at komplekse språktester kommer dårligst ut med tanke på krav til informasjon om reliabilitet. Det finnes med andre ord gode argumenter for bruk av Bus Story, men for å vite noe om skårenes pålitelighet blir det avgjørende å undersøke reliabiliteten.

Brennan (2001a) belyser fenomenet reliabilitet ved å vise til følgende eksempel; en person med èn klokke vet alltid hva klokken er, mens en person med to klokker aldri er helt sikker. Eksemplet understreker hvordan man ofte overser muligheten for at informasjonen man har, ikke er pålitelig. Overført til testing viser dette til faren ved å

ukritisk stole på testresultater uten å undersøke i hvilken grad feilkilder kan ha påvirket dem. Som Kleven (2002a) understreker, forsvinner ikke målefeilene ved å ignorere dem. Man må erkjenne muligheten for at resultatene er påvirket av feil og drøfte dem. Likevel viser flere forskningsartikler til mangelfullt fokus på reliabilitet. Forskningsfunn viser at det i mange tilfeller ikke gis noen informasjon om reliabilitet overhode (Whittington 2003, Vacha-Haase m.fl. 1999, Willson 1980). I tilfeller hvor det er oppgitt informasjon om reliabilitet, er den ofte mangelfull i den forstand at det ikke er oppgitt hvilke metoder som er brukt for å estimere reliabiliteten (Hogan m.fl. 2000) eller at det bare er henvist til reliabilitet ved tidligere bruk av samme instrument (Thompson & Snyder 1998).

Thompson foreslår at den dårlige rapporteringspraksisen kan bunne i at fenomenet reliabilitet er vanskelig tilgjengelig (Thompson 2003, Thompson & Vacha-Haase 2000). Hogan m.fl. (2000) undersøkte hvilke metoder som hyppigst ble brukt for å estimere reliabilitet, og fant en overdreven bruk av metoder fra klassisk reliabilitetsteori, også i tilfeller hvor det kunne vært mer formålstjenlig å bruke andre estimeringsmetoder. Ulike estimeringsmetoder gir gjerne ulik reliabilitet. De bygger på ulike måter å tilnærme seg reliabilitetsproblematikk, og avhenger blant annet av hvordan testresultatene skal tolkes (Brennan 2001a, Crocker & Algina 1986). Feilaktig valg av estimeringsmetoder kan dermed gi et galt bilde av reliabiliteten. Hensiktsmessig undersøkelse av reliabilitet krever dermed god kjennskap til ulike reliabilitetsteorier og estimeringsmetoder (Thompson 2003).

1.2 Problemstillinger

Problemområdene drøftet ovenfor reiser flere problemstillinger. Organiseringen av språkprosjektet og kvaliteter ved Bus Story gjør det nødvendig å undersøke reliabiliteten. For å gjøre dette på en hensiktsmessig måte, må det drøftes hvilke reliabilitetstrusler som er aktuelle. Videre vil det være nødvendig med en drøfting av ulike metoder for å estimere reliabilitet for å finne den mest passende metoden. Oppgavens første problemstilling er dermed formulert som

Hvilke reliabilitetsproblemer er aktuelle ved språkprosjektets bruk av Bus Story, og hvordan kan/bør reliabiliteten estimeres?

Ved drøfting av denne problemstillingen vil reliabilitetsteori bli sammenholdt med den praktiske gjennomføringen av språkprosjektet. Drøftingen vil dermed samtidig legge det teoretiske grunnlaget for den konkrete reliabilitetsestimeringen. For å gjennomføre estimeringen er man avhengig av empiriske data. Slik språkprosjektet av praktiske grunner er organisert har det ikke vært mulig å undersøke alle de mulige reliabilitetsproblemene empirisk. Ved første problemstilling vil drøftingen derfor måtte skje ut i fra et teoretisk perspektiv der det redegjøres for hvordan reliabilitet kunne ha vært estimert dersom de aktuelle empiriske data hadde vært tilgjengelige. Termen *estimering* er valgt for å understreke at reliabilitet strengt tatt ikke kan undersøkes men i beste fall estimeres.

De data det har vært mulig å få tilgang til innenfor prosjektet, gir bare informasjon om i hvilken grad inkonsistent vurdering har truet reliabiliteten. Dette representerer oppgavens andre problemstilling

Estimering og tolking av vurdererreliabilitet ved språkprosjektets bruk av Bus Story.

I andre problemstilling er begrepet *vurdererreliabilitet* brukt, mens det i første problemstilling kun vises til *reliabilitet*. Begrepsdifferensieringen er gjort for å tydeliggjøre at det i første del av oppgaven fokuseres på reliabilitet som helhet, mens den empiriske estimeringen bare fanger opp en del av den totale reliabilitetsproblematikken, nemlig *vurdererreliabilitet*.

I begge problemstillingene benyttes formuleringen *ved språkprosjektets bruk av Bus Story*. Formuleringen er valgt for å understreke at drøfting og estimering av reliabilitet vil være knyttet til Bus Story slik den er benyttet i språkprosjektet, og betingelsene som er lagt her. Resultatene vil avhenge av aktørene i språkprosjektet (barn og forskningsassistenter), forhold ved testen, organisering av testingen og

opplæring av assistentene. Dersom noen av disse betingelsene ble endret, ville man også kunne få andre reliabilitetskoeffisienter.

1.3 Oppbygging av oppgaven

Den videre oppgaven vil disponeres i fem følgende kapitler.

I kapittel 2 ”Språkprosjekt og Bus Story” gjøres det rede for språkprosjektet og betingelsene testingen har blitt gjort under. Det blir også gitt en innføring i Bus Story med fokus på hvordan testen administreres og skåres. Med utgangspunkt i denne informasjonen, blir det drøftet hvilke feilkilder som kan tenkes å ha påvirket Bus Story skårene.

I kapittel 3 ”Reliabilitet” gjøres det rede for tradisjonelle estimeringsmetoder og moderne metoder fra G teori. Det blir drøftet hvilke metoder som vil være mest hensiktsmessige ved estimering av Bus Story skårenes reliabilitet dersom man hadde all nødvendig informasjon. Drøftingen i kapittel 3 legger samtidig teorigrunnlaget for den empiriske estimeringen av vurdererreliabilitet i kapittel 4.

I kapittel 4 ”Estimering av vurdererreliabilitet” gjennomføres den empiriske undersøkelsen av Bus Story skårenes vurdererreliabilitet på bakgrunn av G teori. Det blir vist fremgangsmåter for estimeringen og resultatene presenteres.

I kapittel 5 ”Drøfting av resultater” drøftes det hvordan man kan tolke de estimerte reliabilitetskoeffisientene presentert i kapittel 4. Det blir drøftet i hvilken grad den estimerte vurdererreliabiliteten kan ses som tilfredsstillende og hvordan tendenser i datamaterialet kan tolkes.

I kapittel 6 ”Avslutning” avrundes oppgaven.

2. Språkprosjekt og Bus Story

Som nevnt innledningsvis vil estimering av reliabilitet være knyttet til de spesifikke betingelsene som estimeringen er gjort under. Derfor er det nødvendig å kjenne til organiseringen av prosjektet og kvaliteter ved testen. Slik informasjon vil også legge grunnlaget for å drøfte hvilke reliabilitetsproblemer/feilkilder som kan tenkes å være aktuelle.

2.1 Organisering av prosjektet

Språkprosjektet hadde oppstart høsten 2007, og i første omgang skulle det samles inn data til å danne en normgruppe. Lederne for prosjektet gjorde et utvalg av ca 200 barn som skulle inngå i denne gruppen. Barna som fikk tilbud om å delta var født innenfor perioden 01.04.03 - 01.07.04, de var altså fra 3,5 til 4,5 år. Alle barna ble hentet fra samme kommune. I og med at disse barna skulle representere normbarna, var det avgjørende å sikre at barna viste normal språkutvikling. Barn som var henvist til PPT for språkvansker eller fikk behandling av logoped/audiopedagog ble utelatt. Også barn med andre diagnoser ble utelatt dersom vanskene/funksjonshemningen hadde innvirkning på språket. Barn med klart utenlandske navn eller tospråklig bakgrunn ble heller ikke inkludert for å sikre normal språkutvikling hos denne barnegruppen.

Det ble ansatt 10 forskningsassistenter for å gjennomføre første testing av normgruppebarna. Alle forskningsassistentene var på dette tidspunktet masterstudenter i spesialpedagogikk, og hadde logopedi eller spesifikke lærevansker som fordypningsfelt. Ingen av dem hadde betydelig testerfaring fra før. Assistentene fikk opplæring i samlet gruppe i forkant av testingen. Opplæringen ble gitt av lederne for prosjektet, og det ble gitt undervisning i administrering av testbatteriet og retningslinjer for skåring. Det ble også fokusert på praktisk øvelse hvor assistentene fikk prøve materialet på hverandre. Det ble delt ut eksempler på ”ferdigskårede barn”

som kunne brukes som sammenligningsgrunnlag i egen skåring. I situasjoner hvor assistentene var usikre, kunne man henvende seg til prosjektlederne for veiledning. I perioder det ble meldt mange spørsmål, ble assistentene kalt inn til møte hvor problemstillingene ble drøftet i samlet gruppe.

Hver forskningsassistent fikk tildelt ca 20 barn de skulle teste. Tildelingen ble gjort ut i fra praktiske hensyn slik at hver assistent fikk tildelt barn som gikk i samme barnehage så langt dette var mulig å oppfylle. Alle barn som gikk i barnehage ble testet der. Barna som ikke gikk i barnehage ble testet hjemme eller på det lokale PP-kontoret. I utgangspunktet var barna alene med forskningsassistenten under testingen, men dersom det var ønskelig fikk forelder eller førskolelærer være med inn. Hvert barn ble testet over tre dager med èn økt hver dag. Deltestene var fordelt på tre testdeler som skulle administreres i stigende rekkefølge. Bus Story inngikk i den første testdelen.

2.2 Bus Story

Bus Story er en språktest for barn fra 3 til 8 år utviklet av Catherine E. Renfrew i 1969 (Renfrew 1997). Den er som nevnt en narrativ test med en standardisert gjenfortellingsoppgave. Dette har vist seg å være en god måte å undersøke barns språk på, fordi gjenfortelling er en oppgave som tapper språk på en økologisk og kompleks måte. Det kreves en rekke språklige og kognitive ferdigheter for å gjennomføre oppgaven (Paul & Smith 1993). Barnet må blant annet mestre sekvensiering av hendelser, skape og strukturere en sammenhengende tekst, overbringe ideer ved hjelp av kontekstuavhengig språk, forstå årsak-virkning forhold, ta hensyn til tilhører og hvilken bakgrunnsinformasjon denne personen har, bruke et presist vokabular og mestre morfologi og syntaks i språket (Pankratz m.fl. 2007, Paul & Smith 1993). I tillegg er oppgaven preget av å være åpen og naturalistisk ved at den bygger på narrativ aktivitet som er blant barns naturlige språkhandlinger (Botting 2002).

I flere undersøkelser har Bus Story vist seg å være en god måte å få informasjon om barns språk på. Dette gjelder både ved diagnostisering og prediksjon av fremtidige ferdigheter eller vansker hos barn. Det viser seg at barn med språkvansker har dårligere narrative ferdigheter enn barn med normal språkutvikling (Fey m.fl. 2004). Dette underbygger Paul og Smiths (1993) funn som viste at Bus Story på en god måte diskriminerer mellom barn med språkvansker og barn med normal språkutvikling. Howlin og Kendall (1991) fant også at testene som tappet språk på en kompleks måte var mer sensitive i forhold til språkvansker enn mer avgrensede språktester. Bus Story viste seg å være særlig god i denne sammenheng (Howlin & Kendall 1991).

Også den prediktive evnen til Bus Story er dokumentert gjennom flere undersøkelser. Bishop og Edmundson (1987) fant at Bus Story med høy treffsikkerhet predikerte om 4-åringer ville ha språkvansker eller ikke ett og et halvt år senere. De fant også at narrative ferdigheter var en av de ferdighetene som best predikerte skolefaglig suksess. Dette svarer til funnene fra Feagans og Appelbaums (1986) undersøkelse hvor de fant at narrative ferdigheter predikerte senere akademisk utfall som leseforståelse og matematiske ferdigheter. Pankratz m.fl. (2007) fant at Bus Story på en god måte predikerte senere språkferdigheter som lesing og skriving. Disse forskningsresultatene viser at Bus Story har god både diagnostisk og prediktiv validitet. De forklarer også hvorfor Botting (2002) anbefaler å inkludere Bus Story i testbatterier som skal teste språk.

Bus Story er ikke offisielt oversatt til norsk. Bruk av Bus Story i språkprosjektet bygget derfor på den engelske manualen utarbeidet av Renfrew i 1997. Testadministrasjonen starter ved at testleder viser barnet billedboken og de ser sammen på bildene. Når barnet har gjort seg kjent med bildene, sier testleder at hun skal fortelle barnet historien om bussen, og at etterpå skal barnet få fortelle historien til henne. Så forteller hun historien mens de sammen ser på hvert av bildene i serien. Testleder skal holde seg strengt til historien slik den er skrevet i testmanualen for at alle barn skal presenteres for den samme historien. Samtidig skal testleder tilpasse tempo og pauser i forhold til barnets alder og evne til konsentrasjon (Renfrew 1997).

Etter å ha fortalt historien bes barnet på nytt om å gjenfortelle historien, og promptes med ”det var en gang…” Dersom barnets fortelling stopper opp, kan testleder drive barnet videre ved minimal og indirekte prompting som ”og så…?” Når barnet selv mener at det er ferdig å fortelle, er testen også ferdig administrert.

Barnets gjenfortelling av historien tas opp på lydbånd slik at den kan tolkes og skåres i ettertid. Før skåring skal både barnets gjenfortelling og testleders prompting underveis transkriberes. Hvor barnets setninger starter og slutter, skal markeres i transkriberingen. Da det er stor variasjon i hvor tydelig barn markerer sine setninger, blir dette en vurderingssak for den som transkriberer. Setningsdelingen må tolkes ut fra holdepunkter som tonefall, pauser, semantikk og syntaks (Renfrew 1997).

Videre skal transkripsjonen brukes som skåringsgrunnlag. Skåringen foregår i et ferdig skåringsskjema som er en del av testmaterialet. Skåringsskjemaet ble oversatt til norsk av prosjektlederne (vedlegg 1). Kun de av barnas setninger som anses som relevante for historien føres inn i skåringsskjemaet. I skjemaet finner man en skåringsguide som viser kjerneinnholdet i historien, hvem som er aktør, og hvordan historien skal bygges opp. Første linje i skåringsguiden heter for eksempel ”buss – rampete.” Dette forteller at momentet som skal nevnes først i gjenfortellingen er at historien handler om en rampete buss. Nøkkelordet er ”rampete” og ”bussen” er aktør. Dersom barnet har startet sin gjenfortelling med ”dette er en historie om bussen som var rampete” skal denne setningen noteres på linjen bak ”buss – rampete.” Slik fylles alle relevante transkriberte setninger inn i skåringsskjemaet, og hver setning kan skåres opp i mot skåringsguiden (Renfrew 1997).

Barnets gjenfortelling leder ut til to ulike skårer; èn for informasjon og èn for setningslengde. Først skåres fortellingens informasjon. Setningene kan få fra 0 til 1 poeng eller fra 0 til 2 poeng, avhengig av hvor sentrale de anses for å være. For at en setning skal skåres til maksimal poengsum må barnets setning være lik den som står i skåringsguiden eller vurderes til å dekke det samme innholdet som setningen i skåringsguiden. For eksempel vil både ”det var en rampete buss” og ”det var en slem buss” gi full pott i forhold til ”buss – rampete.” Slik må vurdereren tolke hver av

barnets setninger og avgjøre om de dekker innholdet i skåringsguiden. Det skal også tas hensyn til barnets dialekt og uttrykk når det avgjøres om setningene er tilstrekkelig synonyme (Renfrew 1997). For setninger med maksimal skåre på 2 poeng, kan det gis 1 poeng for halv respons, dersom barnets setninger kun delvis dekker det relevante innholdet. Et eksempel på dette kan være at barnet sier ”så hoppa bussen over” mens det i skåringsguiden heter ”buss – hoppet over gjerdet.”

I tillegg til at innholdet skal være dekkende, må barnet også presentere hendelsen på riktig sted i hendelsesforløpet (Renfrew 1997). Det vil si at dersom barnet tidlig forteller at bussen faller i dammen, gis det ikke poeng for denne setningen fordi den representerer slutten på historien. Hvor avvikende plasseringen skal være for at den ikke er poenggivende, må vurderes av den som skårer testen.

Til sist må også den aktuelle aktøren presenteres for at setninger skal gis maksimal skåre (Renfrew 1997). Dersom barnet sier ”den var rampete” i stede for ”bussen var rampete” trekkes barnet 1 poeng. Det skal imidlertid ikke trekkes poeng for de påfølgende setningene hvor bussen fremdeles er aktør. Barnet får med andre ord ikke følgefeil, men blir trukket et nytt poeng dersom det ikke introduserer neste aktuelle aktør. På grunnlag av disse skåringsreglene skåres de utvalgte setningene i forhold til hver setning i skåringsguiden. Delskårene for hver linje summeres og gir en samlet råskåre for informasjon (Renfrew 1997).

Setningene som nå har fått 1 eller 2 poeng for informasjon, inngår som skåringsgrunnlag for setningslengde. Det vil si at setninger som ikke er tatt med i skåringsskjemaet eller har blitt vurdert til 0 poeng, ikke tas hensyn til når setningslengden beregnes. For å beregne denne skåren, telles ordene i de aktuelle setningene. Innledende småord som ”og, også, så, at og da” skal ikke telles med. Gjentakelser av samme ord innenfor en setning skal heller ikke telles to ganger. Det vil si at setningen ”og så kom det et tog…et tog der” regnes som en setning på 5 ord. Når ord i alle aktuelle setninger er talt opp trekkes de 5 lengste setningene ut, summeres og deles på 5 slik at skåren for setningslengde blir et mål på gjennomsnittlig setningslenge av de 5 lengste setningene (Renfrew 1997).

I følge den engelske manualen tolkes resultatene i forhold til en normering gjort på engelske barn. Om et barns resultat vurderes som tilfredsstillende eller ikke, avhenger av hvor mye barnets skåre avviker fra skåren som er normert for barnets mentale alder (Renfrew 1997). Med andre ord opereres det med en cutoff skåre. Bus Story skårene vil imidlertid bli tolket annerledes i språkprosjektet. Barnegruppen testet her skal danne et nytt normgrunnlag, som siden skal tjene som sammenligningsgrunnlag for språkutvikling hos barn med minoritetsbakgrunn, Cochlea implantat, Down syndrom og spesifikke språkvansker (Child Language & Learning prosjektsøknad 2007). Dermed vil de ikke bli tolket i forhold til engelsk normering og cutoff.

2.3 Aktuelle reliabilitetsproblemer

Til nå har det blitt fokusert på betingelsene barna er blitt testet under. Men hva ville skje med barnas Bus Story skårer dersom noen av betingelsene ble endret? Kleven (2002a) reiser tre reliabilitetsspørsmål som kan være nyttige å drøfte ved undersøkelse av feilkilder. Først stiller han spørsmålet ”i hvilken grad er resultatet avhengig av tilfeldige dag til dag svingninger i personens prestasjonsevne?” Selv om barna ble testet over tre dager, ble de bare testet èn gang med Bus Story. Bus Story skårene er altså knyttet til hva barna var i stand til å prestere akkurat denne dagen på akkurat dette tidspunktet. Så kan man spørre seg om det er sannsynlig at resultatet ville bli annerledes om barnet ble testet en annen dag? Det er flere forhold som tilsier at man må svare bekreftende på dette spørsmålet. Det var ikke fastsatt når på dagen barna skulle bli testet, og flere barn ble testet per dag. Med andre ord var det tilfeldige forhold som avgjorde når testingen startet for det enkelte barnet. Det er grunn til å tro at barna som var trette da de ble testet på slutten av dagen, presterte dårligere enn barn som ble testet mens de var opplagte. Dette kunne også være knyttet til generell dagsform og ikke bare til tid på dagen. Hva barna hadde opplevd rett før testingen, virket også til å påvirke deres prestasjoner. Noen av barna var oppskjørtet eller lei seg da de kom inn, og dette så ut til å påvirke deres evne til konsentrasjon mot oppgaven. At de ble testet på ulike dager, gjorde også at situasjonen rundt testingen varierte.

Som nevnt ble de fleste barna testet i egen barnehage. Dette var til tider et utfordrende miljø å teste i på grunn av støyfylte omgivelser og hyppige avbrytelser. Det kan tenkes at de barna som opplevde mindre forstyrrelser oppnådde høyere skårer enn barna som ble ofte ble avbrutt og forstyrret av støy. I noen tilfeller var det tydelig at barna sporet av oppgaven på grunn av hendelser utenfor testsituasjonen.

Videre reiser Kleven (2002a) spørsmålet ”i hvilken grad er resultatet avhengig av hvilke konkrete oppgaver som er gitt?” Ved testing gis det ofte en rekke deloppgaver (items) som samlet gir en testskåre. Bus Story består derimot bare av èn oppgave eller èn historie. Man bør dermed spørre seg om det er sannsynlig at barnet ville oppnå et annet resultat dersom det ble testet med en annen historie? Det er ikke urealistisk å tro at noen av barna ble mer engasjert av busshistorien enn andre barn. I og med at oppgaven er en gjenfortellingsoppgave, er det avgjørende at barna husker elementene i historien. Bedre kjennskap til historiens elementer, eller større interesse for dem, kan ha ført til at noen barn har tatt til seg og husket historien bedre enn andre barn, og at dette igjen har påvirket gjenfortellingen.

Til sist stiller Kleven (2002a) spørsmålet ”i hvilken grad er resultatet avhengig av hvem som vurderer prestasjonene?” I språkprosjektet fungerte forskningsassistentene både som testledere og som skårere/vurderere. Det vil si at spørsmålet reist her kan utvides til ”i hvilken grad er resultatet avhengig av hvem som har testet barna?” På samme måte som barna bare ble testet èn dag ved hjelp av èn historie, ble de også testet og vurdert av èn person. Det er flere forhold som taler for at barna kunne fått andre skårer dersom andre personer hadde administrert testen eller vurdert gjenfortellingen. Det kan tenkes at testledere i ulik grad lykkes i å engasjere barna ved hjelp av innlevelse, fortellerstil og tilpassing av tempo. Retningslinjene for hvor mye hjelp og prompting testleder kan gi, er av slik art at hver testleder må vurdere hva som er akseptabel grad av hjelp. En testleder som bruker god tid ved hvert bilde og som støtter og prompter barnet gjennom gjenfortellingen, vil sannsynligvis hjelpe barnet til høyere skårer enn en testleder som rakst blar gjennom historien dersom barnet ikke umiddelbart forteller. I hvilken grad testleder lykkes i å skape en trygg

testsituasjon, kan også ha påvirket barnets skåre ved at følelse av trygghet kan tenkes å påvirke fortellerlyst.

Videre er transkriberings- og skåringsreglene av slik art at vurdererne må ta i bruk egen tolkning i skåringsprosessen. De må blant annet tolke hvor barnas setninger starter og slutter, i hvilken grad de dekker målinnholdet på en tilfredsstillende måte, og hvor grensen går mellom informasjon gitt på tilfredsstillende tidspunkt og når den er gitt for tidlig eller for sent. I og med at gjenfortelling er en så åpen oppgave, har det bare vært mulig å lage noen prinsipielle skåringsregler. For å gi vurdererne noen flere holdepunkter, utarbeidet prosjektlederne i tillegg noen eksempler på ytringer og hvordan de burde skåres. Likevel er det grunn til å tro at skårene kan være påvirket av hvordan de ulike vurdererne har opplevd barna og deres fortelling. I tillegg kan unøyaktighet, summeringsfeil og vurderernes dagsform ha påvirket hvilken skåre de har gitt.

2.4 Replikasjon

I foregående kapittel ble det vist at flere typer målefeil kan tenkes å ha påvirket Bus Story skårene i språkprosjektet. For å kunne bevare i hvilken grad reliabilitetstruslene er reelle, er man avhengig av replikasjon. Brennan (2001a) beskriver replikasjon ved sitt klokkeeksempel presentert innledningsvis. Ved hjelp av èn klokke må man stole blindt på at klokken er pålitelig og viser riktig tid. Men dersom to klokker viser ulik tid, vet man at minst èn av dem er upålitelige. Overført til språkprosjektet betyr det at testing èn gang med èn historie administrert og skåret av èn person, ikke kan gi oss informasjon om Bus Story skårenes reliabilitet. Dersom dette var all tilgjengelig informasjon, måtte man blindt stole på at skårene var pålitelige og ikke påvirket av målefeil. Dette ville være en risikabel antagelse tatt i betrakting de mulige feilkildene drøftet i forrige kapittel. For å unngå en slik risikabel antagelse, vil man være avhengig av å kunne teste barna to eller flere ganger, med to eller flere historier, administrert og skåret av to eller flere personer hver gang. Men andre ord ville man være avhengig av replikasjon av testinger, historier, testledere og vurderere.

Slik språkprosjektet er organisert, var det ikke mulig å teste barna med Bus Story flere ganger. Testingen inngikk som ledd i en longitudinell undersøkelse, og retesting ville forstyrre prosjektets formål. Å teste barna med flere historier var også umulig da det bare finnes èn historie. Replikasjon av testledere var heller ikke gjennomførbart da kun èn testleder kan administrere testen hver gang. Det som imidlertid var gjennomførbart, var replikasjon av vurderere. At barna skulle bli vurdert av flere vurderere hver gang, ville ikke avhenge av verken kvaliteter ved testen eller organisering av språkprosjektet. Med andre var det kun mulig å undersøke Bus Story skårenes vurdererreliabilitet.

Replikasjon av vurderere organiseres gjerne som reskåring (Crocker & Algina 1986). Det gjøres en ny skåring av et testgrunnlag som allerede er skåret av en annen vurderer. En slik form for replikasjon kunne altså skje uten å påvirke betingelsene lagt for språkprosjektet. Likevel lå det noen begrensninger for hvordan reskåringen kunne skje. I og med at alle forskningsassistentene var masterstudenter og inne i en travel periode, var det ikke noe alternativ å be dem om reskåringsarbeid. Derfor valgte jeg å gjøre reskåringen selv. Det ble nødvendig å gjøre et utvalg av testgrunnlag som skulle reskåres, da reskåring av hele normgruppen ville bli for omfattende i forhold til masteroppgavens tidsbegrensninger. For at flest mulig av barna skulle bli vurdert av de to samme vurdererne, valgte jeg å gjøre et utvalg av forskningsassistenter, i stede for et direkte utvalg av barn. Med hver forskningsassistent fulgte alle barna assistenten hadde testet og skåret.

Det ble gjort et utvalg på 2 forskningsassistenter. Dette tilsvarte ca 40 gjenfortellinger som skulle reskåres, da hver assistent hadde testet ca 20 barn. Dette ble regent som et overkommelig omfang på datainnsamlingen. Samtidig ville et utvalg på 2 forskningsassistenter og 40 barn føre til at man oversteg 15 % av barna som var inkludert i språkprosjektet. Dette var ønskelig da flere reliabilitetsundersøkelser opererer med lignende størrelser (Fan & Chen 2000, Paul & Smith 1993, Pankratz m.fl. 2007).

Før reskåringen kunne starte måtte det avgjøres hvor i vurderingsprosessen reskåringen skulle ta utgangspunkt. Som nevnt inngår både transkribering og skåring i vurderingen av Bus Story skårene, og begge representerer mulige feilkilder. Slik kunne det være fornuftig å ta utgangspunkt i lydopptakene av barnas gjenfortelling og på grunnlag av denne gjøre en ny transkribering og skåring. En slik fremgangsmåte ville imidlertid ikke gi mulighet til å differensiere mellom målefeil grunnet transkribering og målefeil grunnet skåring. I stede ble det tatt utgangspunkt i lydopptakene ved reskåring av halvparten av barnas gjenfortellinger, mens det ved reskåring av den andre halvdelen ble tatt utgangspunkt i transkriberingen allerede gjort av den opprinnelige vurdereren. Med andre ord ble det tatt utgangspunkt i den opprinnelige transkriberingen ved reskåring av historiene som fulgte den ene assistenten, mens det ble laget en ny transkribering på grunnlag av lydopptakene ved reskåring av historiene som fulgte den andre assistenten. Hvem sine skåringer som skulle nytranskriberes og ikke, ble avgjort ved hjelp av loddtrekning. Ved reskåring kjente jeg ikke til skårene som opprinnelig var gitt barnet, for å unngå at dette skulle farge reskåringen. Etter at all reskåring var gjennomført, ble de to vurderingene av hvert barn matchet.

2.5 Etiske hensyn

Ved planlegging og gjennomføring av forskningsprosjekter, skal det tas en rekke etiske hensyn. I følge NESH pkt. 10 (2006) skal alle forsknings- og studentprosjekter som innebærer behandling av personopplysninger meldes. Opplysninger som navn og alder på barna og lydopptak av barnas stemmer regnes som personopplysninger, og det måtte søkes godkjenning fra Datatilsynet. Dette var allerede gjort da jeg ble ansatt som forskningsassistent i språkprosjektet. Når barn er deltakere i forskningsprosjekter skal det også tas særlige hensyn i forhold til dem. I NESH pkt. 12 (2006) er det uttrykt at barn har særlig krav på beskyttelse i tråd med alder og behov, og at deres aksept er avgjørende for gjennomføring av prosjektet. Så lenge barna er under 15 år, skal det innhentes samtykke fra foresatte. Kravet om innhenting av samtykke fra

foresatte ble ivaretatt av lederne for språkprosjektet. I tillegg pliktet hver forskningsassistent å ivareta barnet på best mulig måte i testsituasjonen. NESH stiller også i pkt. 8 og 9 (2006) krav om at samtykket skal være fritt og informert. I dette ligger det at foreldrene skal ha all nødvendig informasjon om forskningsprosjektet, og ikke kjenne seg presset til å delta. Foreldrene mottok et informasjonsskriv om prosjektet, og det ble holdt et møte mellom representanter fra språkprosjektet og de aktuelle barnehagene slik at også de fikk nødvendig informasjon.

Videre skal det også tas hensyn til mine direkte informanter; forskningsassistentene. De har i følge NESH pkt. 8 (2006) krav på informasjon om hva det innebærer å delta i prosjektet og hensikten med forskningen. Det ble sendt ut informasjon til alle forskningsassistentene om prosjektet jeg planla. Jeg ba også om deres tillatelse til å få tilgang til skårene de hadde gitt barna de testet i språkprosjektet. Også her er det krav om å innhete fritt og informert samtykke av forskningsassistentene (NESH pkt. 9 2006). I informasjonen som ble sendt ut ble det understreket at all deltakelse var frivillig, og at de som ville gi meg denne tillatelsen måtte bekrefte dette.

I NESH pkt. 14 (2006) stilles det krav til konfidensialitet. Forskningsassistentene ble informert om at de ville bli behandlet anonymt både underveis i mitt prosjekt og i fremstillingen av oppgaven. Dette kravet gjelder selvsagt også for de andre aktørene i prosjektet. Både barna og barnehagene vil bli anonymisert i oppgaven.

Forskningsassistentene som samtykket i å delta, sendte selv sine Bus Story skåringer til meg. Lydfilene ble derimot hentet hos lederne av språkprosjektet da forskningsassistentene hadde levert fra seg filene på dette tidspunktet. De kunne heller ikke overføres via internett da opptak av barnas stemmer regnes som personopplysninger og ikke skal oppbevares elektronisk (NESH pkt. 16 2006).

3. Reliabilitet

I kapittel 2 ble det drøftet hvilke feilkilder som kan tenkes å ha påvirket Bus Story skårene i språkprosjektet. Det ble også vist hvordan replikasjon er nødvendig for å kunne svare empirisk på dette spørsmålet. På grunn av kravet til replikasjon har det bare vært mulig å gjøre en empirisk undersøke Bus Story skårenes vurdererreliabilitet. Videre drøfting av problemstilling 1 må derfor skje ut ifra et teoretisk synspunkt. Det vil bli drøftet hvordan Bus Story skårenes reliabilitet *kunne vært* undersøkt dersom man hadde den nødvendige informasjonen. For å drøfte problemstillingen, blir det nødvendig å se nærmere på reliabilitetsteori.

3.1 Reliabilitetsbegrepet

Reliabilitet kan språklig sett oversettes til pålitelighet, men har en mer avgrenset betydning i forskningslitteratur (Kleven 2002a). Her forklares reliabilitet gjerne som konsistensen til en måling (Crocker & Algina 1986). Bak en slik definisjon ligger ønsket om at resultater skal være konsistente over replikasjoner. Det er for eksempel ønskelig at gjentatt testing av samme barn under samme betingelser skal gi samme resultater. Dersom en persons høyde ble målt tre dager på rad og gav tre ulike resultater, ville man tolke dette som at målingene er påvirket av feil. Konsistente resultater tolkes altså som reliable resultater. Men hva er det så som påvirker konsistensen? Det skilles grovt mellom tilfeldige og systematiske målefeil (Shavelson & Webb 1991, Crocker & Algina 1986).

Systematiske målefeil er feil som har en konsekvent effekt på barnets resultat hver gang det blir testet (Crocker & Algina 1986). Et barn som nesten ikke tør snakke fordi det kjenner seg utrygg i testsituasjoner, vil få en lav Bus Story skåre som er lite representativ for barnets egentlige språkferdighet. Neste gang barnet befinner seg i en lignende testsituasjon, vil det med stor sannsynlighet reagere på samme måte. Dette vil da representere målefeil fordi Bus Stort skårene påvirkes av forhold som testen

ikke er ment å måle. Målefeilen klassifiseres som systematisk fordi den sannsynligvis vil slå likt ut for det samme barnet hver gang. Denne typen målefeil vil ikke påvirke konsistensen til barnets skårer siden effekten er konsekvent, og vil dermed heller ikke påvirke reliabiliteten. De vil imidlertid påvirke validiteten eller gyldigheten til skårene i og med at de representerer målefeil.

De *tilfeldige* målefeilene vil derimot ha en effekt på reliabiliteten fordi denne typen feil er inkonsistente. De kan slå ut positivt eller negativt, og det er tilfeldig når og for hvem de slår ut (Crocker & Algina 1986). De fordeler seg på en tilfeldig måte, og har dermed en tendens til å jevne seg ut i det lange løp (Kleven 2002a). Rand (1971) skiller mellom fire kilder til tilfeldige målefeil; forhold ved testsituasjonen, forhold ved testen, forhold ved forsøkspersonen og forhold ved forsøksleder. Det betyr at drøftingen av Klevens (2002a) reliabilitetsspørsmål i kapittel 2.3 omhandler tilfeldige målefeil som vil påvirke Bus Story skårenes reliabilitet. Det er tilfeldig hvordan barnas dagsform, forhold ved testsituasjon, busshistorien og inkonsistent testadministrasjon og vurdering slår ut for det enkelte barnet hver gang det blir testet. Ved at det kun er de tilfeldige målefeilene som påvirker reliabilitet, kan høy reliabilitet også defineres som at skårene i liten grad er påvirket av tilfeldige målefeil (Kleven 1995). Når det videre i oppgaven vises til målefeil, menes da de tilfeldige målefeilene i og med at det kun er de som er av interesse når det fokuseres på reliabilitet.

3.2 Tradisjonelle estimeringsmetoder

Som nevnt innledningsvis kan ikke reliabilitet undersøkes eller måles direkte. Dette fremgår også av reliabilitetsspørsmålene drøftet i kapittel 2.3. Det er for eksempel umulig å få et konkret mål på hvor mye barnets dagsform har påvirket testresultatet i praksis. Reliabilitetsspørsmålene er spørsmål det i beste fall kan estimeres svar på. Derfor er det også konstruert ulike metoder for å estimere reliabilitet. De tradisjonelle estimeringsmetodene bygger på hvordan reliabilitet har blitt forstått og utviklet

innenfor klassisk reliabilitetsteori siden starten av 1900-tallet. Dette er også de mest kjente og brukte metodene (Hogan m.fl. 2000, Fan & Chen 2000, Thompson 2003).

I klassisk reliabilitetsteori, også kalt sann skåreteori, anses enhver observert skåre (testresultat) som bestående av to hypotetiske komponenter; en sann skåre og en tilfeldig feilkomponent. Dette uttrykkes gjennom ligningen ”observert skåre = sann skåre + feil” (Crocker & Algina 1986). En slik ligning og fastsetting av sann og feil skåre fungerer bare i teorien. I praksis vet man ikke hva som skyldes barnets ”sanne ferdigheter” og hva som skyldes påvirkning fra målefeil. På grunn av påvirkning fra ulike målefeil, vil barnets observerte skåre mest sannsynlig variere noe for hver gang det blir testet. For å likevel få tilgang til et barns sanne skåre, defineres den som et gjennomsnitt av alle de skårene barnet kunne fått dersom det ble testet uendelig mange ganger under uendelig mange betingelser (Crocker & Algina 1986) Fra dette ser man at sann skåre egentlig bare betyr konsistent skåre (Kleven 1995). Det er en teoretisk eller matematisk størrelse, og ikke en sannhet slik man intuitivt forstår begrepet. Hva som er ”sant” i denne sammenhengen bestemmes av den gjennomsnittlige skåren og hvordan den matematisk er regnet ut.

Tradisjonelt har man vært opptatt av reliabilitet som stabilitetsproblemer og ekvivalensproblemer (Brown 1983). Stabilitet viser til hvor konsistente skårene er over tid. Tilfeldige målefeil som for eksempel barnas dagsform, kan føre til inkonsistente skårer over tid, og dermed redusere stabilitetsreliabiliteten. Ekvivalens viser derimot til hvor konsistente skårene er over testformer eller testitems. Dersom Bus Story testen for eksempel bestod av tre historier som barna skulle gjenfortelle, og barnas skårer var konsistente fra historie til historie, ville skårenes ekvivalensreliabilitet være høy.

For å estimere skårers stabilitetsreliabilitet benyttes gjerne metoden test-retest (Brown 1983). Barnet testes med samme test på to eller flere tidspunkt, og de ulike skårene barnet oppnår korreleres med hverandre. Høye korrelasjoner gir høye reliabilitetskoeffisienter og tolkes som høy stabilitetsreliabilitet. Metoden bygger med andre ord på en antagelse om at forskjeller mellom barnets skårer ved første og andre

testing representerer målefeil (Crocker & Algina 1986, Kleven 2002a). Dette kan være en risikabel antagelse med tanke på at det også kan være andre årsaker enn målefeil som gir differanse mellom skårene. Det tas ikke hensyn til at barnas ferdighetsnivå kan ha endret seg, og endret seg ulikt, i mellomtiden. Som Kleven (2002a) påpeker, er det ikke urealistisk at slike endringer kan skje på relativt kort tid, tatt i betraktning at mye av arbeidet i vårt fagfelt fokuseres mot læringsinstitusjoner. Det bør derfor ikke gå for lang tid mellom test og retest dersom korrelasjonen skal brukes som estimat for reliabilitet. Samtidig er det forhold som taler for at tidsintervallene heller ikke bør være for korte. Dersom man tenker seg at det ble gjort en retest av normgruppebarna med et kort tidsintervall, kunne det føre til at noen av barna gjorde det bedre ved retest på grunn av at de husket historien og dermed var i stand til å få med flere av historiens elementer. Dette viser at forutsetningen om at det ikke har skjedd virkelige forandringer i løpet av tidsintervallet er meget tvilsom, og at dette må tas hensyn til ved tolkingen av reliabilitet estimert ved test-retest.

I og med at språkprosjektet er et longitudinelt prosjekt, skal barna testes med Bus Story to ganger årlig over fire år. At tidsintervallene er på ca. 6 måneder, gjør det rimelig å tolke endringer i Bus Story skårene som i hovedsak et resultat av virkelige endringer som modning og læring. Dermed kan ikke korrelasjonen mellom ulike målinger tolkes som et estimat for reliabilitet.

Liksom man bruker test-retest for å estimere stabilitetsreliabilitet, brukes gjerne parallelle former for å estimere ekvivalensreliabilitet (Crocker & Algina 1986). Barnet blir testet med to eller flere parallelle testformer, og skårene korreleres med hverandre (Brown 1983). Høy korrelasjon mellom skårene gir høye reliabilitetskoeffisienter og tolkes som høy ekvivalensreliabilitet. For å benytte seg av denne metoden er man altså avhengig av å ha tilgang til parallelle og ekvivalente testformer. Som allerede nevnt er ikke dette tilfelle for Bus Story. Det er bare utarbeidet èn historie. Dette betyr ikke at Bus Story skårer ikke kan ha redusert ekvivalensreliabilitet, men at det ikke lar seg gjøre å undersøke den med mindre det utarbeides flere parallelle historier. Som eksemplifisert av Brennan (2001a) vil det

alltid være en risiko for at den ene klokken man har viser gal tid, og man har ingen mulighet til å avkrefte dette før man har en annen klokke å sammenligne med. Aktuelle målefeil grunnet historien som brukes vil være like reelle, man har bare ikke mulighet til å studere dem så lenge det kun finnes èn historie.

Metodene test-retest og parallelle former er fokusert mot dag til dag svingninger i barnas prestasjonsevne og inkonsistens grunnet historien som er brukt. Men som drøftet i kapittel 2.3, er det grunn til å tro at Bus Story skårene også kan være påvirket av målefeil grunnet inkonsistent vurdering. I testteoribøker nevnes ofte prosentvis enighet som et mulig mål på vurdererreliabilitet (Kline 2005, Clark-Carter 2004). Her beregnes hvor stor prosentandel av skårene vurdererne er enige om. Det opereres altså kun med et skille mellom enig og uenig, og man får ikke differensiert mellom ulike grader av uenighet (Brown 1983). En mild uenighet behandles på samme måte som en sterk uenighet. Crocker og Algina (1986) hevder at prosentvis enighet kan gi verdifull informasjon, men at målet ikke kan sammenlignes med eller erstatte de tradisjonelle estimeringsmetodene. For å estimere vurdererreliabilitet benyttes da prinsippet om parallelle former beskrevet ovenfor, og ekvivalente testformer byttes ut med ekvivalente vurderere (Brown 1983). I stede for å beregne korrelasjon mellom skårer på to testformer, beregnes korrelasjon mellom skårer gitt av to ulike vurderere.

Når barn testes to ganger eller med to former/vurderere, forutsettes det i klassisk reliabilitetsteori at de to målene er parallelle. Parallellitet er definert som at barnet har samme sanne skåre på de to målene og at barnets observerte skåre har samme mean, varians og feilvarians på de to målene (Brennan 2001b). Forutsetningen om parallelle skårer som ligger til grunn for estimering av reliabilitet i klassisk reliabilitetsteori er så strenge at de synes nesten umulige å innfri i praksis (Cronbach m.fl. 1963). Bruk av tradisjonelle estimeringsmetoder vil slik bygge på antagelser man ikke har mulighet til å bekrefte i praksis.

Til nå er det blitt presentert estimeringsmetoder hvor det kun fokuseres på èn feilkilde. Når alle andre betingelser holdes konstant, vil kun èn av alle de aktuelle feilkildene for Bus Story skårene få mulighet til å redusere den estimerte

reliabilitetskoeffisienten. Dette vil etter all sannsynlighet føre til en overestimert koeffisient (Thompson 2003, Brown 1983). Man kunne selvsagt gjennomføre begge metodene. Man kan tenke seg at man fant tilfredsstillende reliabilitet i begge tilfeller. Men målefeilene fanget opp i de to reliabilitetsestimatene, kan ikke ses som to sider av samme sak. De er ulike målefeil som vil være kumulative (Thompson 2003). Med andre ord er det ikke sikkert at skårenes totale reliabilitet ville være tilfredsstillende selv om man både ved test-retest og parallelle vurderere fant tilfredsstilende reliabilitet.

Ved å kombinere metodene test-retest og parallelle former, kan man imidlertid estimere reliabilitet med fokus på flere feilkilder samtidig (Crocker & Algina 1986). På samme måte som prinsippet om parallelle former kan overføres til vurdere, kan det også overføres til testledere. I og med at testen ikke kan administreres av flere personer hver gang, ville man være avhengig av at barna testes flere ganger for å fange opp målefeil grunnet inkonsistent testadministrasjon i reliabilitetsestimatet. Ved kombinering av metodene kunne barna blitt testet to ganger, og man kunne ha byttet testleder og vurderer fra gang til gang. Skårene barna oppnådde ved testing/testleder/vurderer 1, kunne så blitt korrelert med skåren de fikk ved testing/testleder/vurderer 2. Da ville den estimerte koeffisienten påvirkes både av feil grunnet inkonsistens i prestasjonsevne, inkonsistent administrasjon og inkonsistente vurderinger. Påvirkning fra de ulike feilkildene vil imidlertid samles i èn udifferensiert feilkomponent, og man får ikke mulighet til å avgjøre hvilke målefeil som har påvirket skårene i størst grad og dermed er de mest alvorlige.

I og med at de tradisjonelle estimeringsmetodene bygger på korrelasjon, legger kjennetegnene ved korrelasjon som statistisk metode noen føringer for reliabilitetskoeffisientene. Korrelasjon er et mål på grad av samvariasjon mellom to variabler (Lund & Christophersen 1999). Overført til reliabilitet, viser altså korrelasjonskoeffisientene hvor sterk samvariasjon det er mellom skårer på to parallelle testinger, testledere eller vurderere. Perfekt samvariasjon trenger imidlertid ikke bety at skårene er helt like, da samvariasjon er noe annet enn likhet.

Man kan tenke seg reskårte Bus Story gjenfortellinger, der to vurderere aldri har gitt samme barn den samme skåren. Dersom de likevel er enig i rangeringen av barna, vil dette gi perfekt korrelasjon og perfekt vurdererreliabilitet. Fra dette fremgår det at korrelasjonsmål kun tar hensyn til skårenes relative plassering (Kleven 2002b). Dersom formålet med testingen var å finne de flinkeste barna, er de to vurdererne helt enige, og en høy korrelasjonskoeffisient ville ikke være misvisende. Men dersom testresultatene skal benyttes til formål hvor også absolutt plassering av skårer er av interesse, ville en perfekt korrelasjon være misvisende og tradisjonelle estimeringsmetoder lite hensiktsmessige. Bruk av to sensorkommisjoner gir et tydelig eksempel på slike tilfeller. Man kan tenke seg at den ene kommisjonen konsekvent gav èn karakter lavere enn den andre. De ville være helt enige i rangering av studentene, og tradisjonelle estimeringsmetoder ville gi perfekt korrelasjon, men studentene ville trolig være mer opptatt av absolutt plassering av karakter og ikke være likegyldig hvem av de to sensorkommisjonene som skulle vurdere deres besvarelse.

Dette understreker hvor avgjørende det er å benytte seg av estimeringsmetoder som er tilpasset formålet med skårene. Likevel er dette et problem som ofte overses og som fører til at reliabilitet overestimeres (Brown 1983). Når Bus Story skårene skal benyttes til å danne en ny normering i språkprosjektet, er det ikke nok at vurdererne har vært enig i rangering av barna. Absolutt plassering av skårer vil også ha betydning for normeringen. Dermed blir det nødvendig å ta hensyn til både de målefeil som påvirker relativ plassering og de som påvirker absolutt plassering. Følgelig vil ikke tradisjonelle estimeringsmetoder som bygger på korrelasjon fange opp alle aktuelle målefeil ved estimering av Bus Story skårenes reliabilitet.

Samlet kan man si at det er flere forhold som taler for at bruk av tradisjonelle estimeringsmetoder vil være lite hensiktsmessig ved estimering av Bus Story skårenes reliabilitet. For det første bygger metodene på noen forutsetninger som kan være problematiske. Kravene som stilles til parallellitet synes å være så strenge at de vanskelig lar seg innfri i praksis, og antagelsen om at det ikke skjer virkelige

endringer mellom test og retest er også svært risikabel. I tillegg er det sannsynlig at bruk av tradisjonelle estimeringsmetoder vil føre til overestimert Bus Story skåre reliabilitet ved at aktuelle målefeil overses. Ved bruk av test-retest eller parallelle former får bare èn blant flere aktuelle feilkilder påvirke reliabilitetskoeffisienten. Man får heller ikke fanget opp målefeil som påvirker absolutt plassering av skårer fordi metodene bygger på korrelasjon som kun tar hensyn til skårenes relative plassering. Dersom man kombinerer metoder for å fange opp flere av de aktuelle målefeilene, samles målefeilene i èn udifferensiert feilkomponent og man har ikke mulighet til å avgjøre hvilke feilkilder som representerer de mest alvorlige reliabilitetstruslene.

Det finnes imidlertid estimeringsmetoder som løser noen av problemene drøftet her. Siden G teori har en del fordeler ikke minst når det gjelder å studere reliabilitet ved vurdering, vil denne teorien bli presentert relativt grundig i neste kapittel.

3.3 G teori

G teori (generalizability theory) ble utviklet av Cronbach med medarbeidere på 1960-tallet som et alternativ til klassisk teori, og omtales gjerne som moderne reliabilitetsteori (Thompson & Vacha-Haase 2000, Crocker & Algina 1986). Klassisk reliabilitetsteori har lagt noe av grunnlaget for G teori, og slik finnes det en del likhetstrekk mellom teoriene. Samtidig tilnærmer man seg reliabilitetsproblematikk på en annen måte i G teori, og teorien løser noen av problemene hvor klassisk teori kommer tilkort. Selv om det er over 30 år siden G teori for første gang ble presentert i trykket form, er ikke teorien blitt like utbredt som metoder fra klassisk reliabilitetsteori. I grunnleggende testteoribøker vies det fremdeles plass til grundig behandling av tradisjonelle estimeringsmetoder, mens G teori ofte behandles overfladisk eller ikke nevnes i det hele tatt (Hogan m.fl. 2000). I en undersøkelse av ca. 800 artikler hvor det var estimert reliabilitet, ble det funnet omfattende bruk av tradisjonelle estimeringsmetoder, mens metoder fra G teori ikke var brukt i noen av tilfellene (Hogan m.fl. 2000).

3.3.1 Grunnprinsipper

Grunntanken i G teori bygger på generalisering. På grunn av begrensninger i tid, penger og andre ressurser, vil man ofte kun ha tilgang til testskårer fra èn testsituasjon, der èn testform, èn testleder og èn vurderer har blitt benyttet (Fan & Chen 2000, Thompson 2003). Det er imidlertid svært sjelden man tester et barn for å finne barnets skåre ved bruk av akkurat denne situasjonen, formen, testlederen eller vurdereren. Som regel ønsker man å vite noe mer generelt om barnets ferdighetsnivå. Når man tester barn med Bus Story, er man interessert i deres narrative språkferdigheter uavhengig av for eksempel barnas dagsform eller hvem som har vurdert prestasjonen. Man ønsker med andre ord å generalisere barnets skåre til å gjelde utover de konkrete betingelsene barna er blitt testet under. I G teori tas ønsket om generalisering på alvor, og man tilnærmer seg reliabilitet ved hjelp av spørsmålet ”med hvilken sikkerhet kan man generalisere fra barnets observerte skåre til barnets universskåre?” (Thompson & Vacha-Haase 2000).

G teoriens universskåre svarer til sann skåre i klassisk reliabilitetsteori, og er definert som et gjennomsnitt av de skårene som finnes i barnets univers (Shavelson & Webb 1991). At den betegnes som *universskåre,* viser til at en observert skåre ses som et tilfeldig utsnitt av barnets univers av mulige skårer. Hvilke mulige skårer som finnes i barnas univers, er avhengig av hvordan universet er definert.

Når universet defineres, definerer man samtidig hvilke feilkilder som inkluderes i reliabilitetsestimeringen. I G teori betegnes de inkluderte feilkildene som fasetter (Shavelson & Webb 1991). Ideelt sett burde det tas hensyn til alle de aktuelle feilkildene drøftet i kapittel 2.3 når universet defineres. Men i og med at det bare finnes èn historie, vil det ikke være mulig å estimere i hvilken grad inkonsistens grunnet historien har påvirket Bus Story skårene. I og med at testen kun kan administreres av èn testleder hver gang, kan heller ikke testadministrasjon defineres som en egen fasett. Like fullt vil inkonsistent testadministrasjon representere en aktuell feilkilde. Selv om samme testleder ble benyttet, har man ingen garanti for at testlederen har vært konsistent i sin administrasjon fra gang til gang. Ved å la to ulike

testledere administrere testen ved første og andre testing, kan inkonsistent testadministrasjon og inkonsistens i barnas prestasjonsevne samles i fasetten gjentatte testinger. De ulike målefeilene kan da ikke skilles fra hverandre, men alle får mulighet til å redusere den estimerte reliabilitetskoeffisienten. Til sist kan vurderere defineres som en egen fasett da to vurderere kan vurdere hvert barn hver gang ved hjelp av reskåring. I G teori har man da et to-fasett design med vurderere og gjentatte testinger som inkluderte fasetter. Universet kunne dermed vært definert til å romme alle mulige skårer barna kunne oppnådd dersom de ble testet uendelig mange ganger (med alle tenkelige testledere) og ble vurdert av alle tenkelige vurderere. Når barna i språkprosjektet kun er testet èn gang (og dermed med kun èn testleder) og er vurdert av èn vurderer, er også de oppnådde Bus Story skårene kun et utsnitt av alle skårene som finnes i deres univers.

Ved definering av fasetter, definerer man samtidig hvilke målefeil som får påvirke feilkomponenten i reliabilitetsestimeringen. Da historier ikke er blant de inkluderte fasettene, får heller ikke inkonsistens grunnet historien som brukes mulighet til å påvirke feilkomponenten, selv om dette er målefeil som i realiteten kan ha redusert Bus Story skårenes reliabilitet. Fra dette fremgår det at ”feil” slik begrepet brukes i reliabilitetsteori, ikke kan forstås som alle mulige feil, men kun som definerte feil (Brennan 2001a). I klassisk teori kan definering av feil ligge implisitt i valg av metode. Dersom test-retest ble brukt for å estimere Bus Story skårenes reliabilitet, ville feilkomponenten kun bli påvirket av målefeil grunnet dag til dag svingninger i barnas prestasjonsevne, og andre aktuelle målefeil ville bli oversett. At definering av feil må skje eksplisitt i G teori, kan medføre økt bevissthet om hvilke feilkilder som inngår i feilkomponenten, og kanskje enda viktigere, hvilke som ikke gjør det (Brennan 2001a).

3.3.2 Generalisering over fasetter

Som nevnt innledningsvis i kapitlet, er det barnas universskåre som er målet for våre generaliseringer. Man ønsker at testresultatet (eller den observerte skåren) man har

trukket ut, med rimelig høy sikkerhet skal kunne generaliseres til barnets universskåre. I og med at universskåren er definert som gjennomsnitt av skårene i universet, vil universskåren også være avhengig av hvilke fasetter som er inkludert. En generalisering fra observert skåre til universskåre, vil dermed si en generalisering over fasetter. Man ville for eksempel generalisere fra skårer gitt av utvalgte vurderere til gjennomsnittet av skårer gitt av alle vurderere i universet, og tilsvarende for testinger med ulike testledere.

Reliabilitetsestimeringen gir imidlertid ikke informasjon om hvorvidt man kan generalisere over historier. Dersom man i to-fasett designet fant høy reliabilitet, og tolket dette som at barnas skårer i liten grad var påvirket av betingelsene de ble testet under, ville man gjøre noen generaliseringer som ikke nødvendigvis ville være gyldige. Man ville da ha generalisert over historier uten å ha belegg for en slik generalisering. I følge Cronbach m.fl. (1963) har spørsmålet om gyldige generaliseringer gjerne fått ligge implisitt, og problemer med risikable generaliseringer har blitt oversett. Også i denne sammenheng trekkes eksplisitt definering av fasetter i G teori frem som en styrke som kan føre til økt bevisst rundt de generaliseringer man gjør (Cronbach m.fl. 1963).

For at generalisering over fasetter skal være mulig, stilles det noen krav til fasettene det skal generaliseres over. Utvalgte testinger og vurderere skal enten være tilfeldig trukket fra universet, eller de skal være så like de man finner i universet at man ville være villig til å bytte dem ut med andre testinger og vurderere (Shavelson & Webb 1991). Tilfeldige utvalg er kjennetegnet ved at alle i populasjonen (eller universet) har like stor sjanse til å bli trukket ut (Befring 2007, Lund 2002a, Kleven 2002c). For å gjøre et tilfeldig utvalg av vurderere og testinger, er man altså avhengig av å ha kjennskap til alle vurdererne og testingene (med alle tenkelige testledere) som befinner seg i universet slik man har definert det. Dette kan i praksis være svært problematisk. Man kan tenke seg at universet av vurderere er definert til å romme alle vurderere som er opplært i å skåre Bus Story fortellinger. Da vil universet være teoretisk definert. I praksis vil man sannsynligvis ikke ha kjennskap til alle disse

vurdererne, og dermed heller ikke ha mulighet til å gi dem lik sjanse til å bli trukket ut. Enda mer problematisk ville det sannsynligvis være å gi alle tenkelige testinger med alle tenkelige testledere lik mulighet til å bli trukket ut.

I slike tilfeller kan utvalg begrunnes ut ifra prinsippet om utbyttbarhet. Man må spørre seg om man er villig til å bytte ut utvalgte testinger/vurderere men andre testinger/vurderere fra universet. Det blir med andre ord et spørsmål om representativitet. Hvor representativt utvalget er, kan også være en vanskelig avgjørelse og må vurderes av den enkelte undersøker. Ved å definere universet til å romme utvalgte testinger/vurderere og alle lignende testinger/vurderere, vil utvalget sannsynligvis være representativt. Dette betegnes av Brennan (2000) som å definere universet som et speil av utvalget.

Samtidig understreker Shavelson og Webb (1991) at generalisering fra utvalg til univers alltid vil innebære feil. På denne måten bygger også G teori på noen antagelser som kan være problematiske. Kravene til tilfeldig utvalg eller utbyttbarhet synes imidlertid ikke like strenge og urealistiske som kravene til parallelle mål i klassisk reliabilitetsteori. Det stilles det ingen krav til skårenes spredning eller gjennomsnitt. I følge Crocker og Algina (1986) kan man ved hjelp av G teori estimere reliabilitet uten å være avhengig av å gjøre urealistiske forutsetninger som ikke kan unngås i klassisk reliabilitetsteori.

3.3.3 Skårevariasjon

I og med at det kun ville være mulig å inkludere to fasetter ved estimering av Bus Story skårenes reliabilitet, kunne som vist i kapittel 3.2 det samme være gjort ved bruk av tradisjonelle estimeringsmetoder. Ved å kombinere test-retest og parallelle former ville reliabilitetskoeffisienten ta hensyn til både gjentatte testinger, testledere og vurderere. Likevel ville det være en vesentlig forskjell mellom estimering ved to-fasett designet i G teori, og ved test-retest med parallelle former fra klassisk teori. Som nevnt ville alle målefeilene samles i èn udifferensiert feilkomponent ved kombinering av de tradisjonelle metodene. Reliabilitetsestimering i G teori bygger

imidlertid på variansanalyse i stede for korrelasjon, og gir slik mulighet til å splitte opp den samlede feilkomponenten i bidrag fra ulike feilkilder (Shavelson & Webb 1991). Der variasjon kun deles i to komponenter i klassisk teori (sann og feil skåre), deles den i så mange komponenter som mulig i G teori (universskåre og ulike kilder til målefeil).

I G teori forklares totalvariasjonen i Bus Story skårene som sammensatt av variasjon grunnet ulike målefeil og den typen variasjon man ønsker når man tester barn med Bus Story. Sistenevnte variasjonstype betegnes gjerne som ”object of measurement” (Shavelson & Webb 1991). I språkprosjektet vil barnet være målet for målingen. Det vil si at det som ønskes målt ved hjelp av Bus Story er variasjon grunnet barnet. Dersom to barn oppnår ulik skåre på Bus Story testen, er det ønskelig at denne ulikheten representerer ulik gjenfortellingskompetanse hos barna og at den ikke er grunnet målefeil.

All variasjon som ikke er grunnet ”object of measurement”, vil være der grunnet èn eller annen form for målefeil. Målefeilene kan påvirke skårene på ulike måter. Dersom vurdererne av ulike årsaker ikke har vært like strenge generelt sett (den ene vurdereren har konsekvent vært strengere enn den andre) kan dette ha ført til at de ikke har vært samstemte i sin vurdering av barna. Dermed vil barnas skåre avhenge av hvem som har vurdert deres prestasjon. I G teori betegnes denne typen målefeil som hovedeffekt fra vurderere (Shavelson & Webb 1991). I tillegg kan det hende at vurdererne ikke har vært like strenge ved vurdering av *alle* barna. En vurderer kan for eksempel ha vært strengere ved vurdering av barn 1 og 3, enn hun har vært ved vurdering av barn 2 og 4. Dette betegnes i G teori som interaksjonseffekt mellom barn og vurderer (Shavelson & Webb 1991).

Tilsvarende kan variasjon grunnet testinger deles i hoved- og interaksjonseffekter. Dersom forhold ved de to testingene (som inkonsistens i barnas prestasjonsevne eller inkonsistent testadministrasjon) har ført til at alle barna presterte litt bedre eller dårligere, vil dette være hovedeffekt fra testinger. Dersom forhold ved testingene har ført til at noen barn har gjort det litt bedre ved første testing, og at andre barn har

gjort det litt bedre ved andre testig, har forhold ved testingen påvirket barnas skårer ulikt og vil representere en interaksjonseffekt mellom barn og testinger.

Fra dette fremgår det at hovedeffektene vil påvirke absolutt plassering av skårer, men ikke rangeringen av dem. Denne typen målefeil betegnes i G teori som absolutte feil. Tilsvarende vil interaksjonseffekter mellom barn og de ulike feilkildene påvirke skårenes relative plassering, og betegnes som relative feil. (Shavelson & Webb 1991).

I tillegg til hoved- og interaksjonseffekter, opereres det i G teori med en resterende feilkomponent som består av uidentifiserte feil. I og med at det ikke er mulig å differensiere mellom interaksjonseffekt og uidentifiserte feil, ses de ofte under ett. Slik skilles det mellom tre former for variasjon; variasjon som ”object of measurement”, variasjon som hovedeffekt fra feilkilder, eller som interaksjonseffekt og uidentifiserte feil (Shavelson & Webb 1991).

3.3.4 Krysset og nestet design

Det har blitt konstatert at Bus Story skårenes feilvarians kan bestå av både hoved- og interaksjonseffekter. I hvilken grad man har mulighet til å skille ut hva som er hva, avhenger imidlertid av hvordan replikasjonen har foregått. Dersom alle barna ble testet to ganger og vurdert av to vurderere hver gang, ville man ha det som i G teori betegnes som et krysset design (Shavelson & Webb 1991). Testinger ville være krysset med vurderere. Når alle barn er vurdert av de to samme vurdererne hver gang, kan man skille ut variasjon grunnet ulik strenghet mellom vurdererne (hovedeffekter). Tilsvarende når alle barna er testet to ganger, kan man skille ut hovedeffekter fra testinger (Shavelson & Webb 1991).

I noen tilfeller vil det ikke være mulig å få til kryssede design. Dersom man tenker tilbake til replikasjon av vurderere i språkprosjektet (drøftet i kapittel 2.4), hadde hver assistent vurdert ca. 20 barn hver. For å få til reskåring av mer enn 20 gjenfortellinger, var det ikke gjennomførbart at alle barna skulle vurderes av de

samme to vurdererne. To assistenter hadde allerede skåret hver sin halvdel av den utvalgte barnegruppen, og reskåringen ble gjort av èn assistent. Dette tilsvarer et nestet design i G teori (Shavelson & Webb 1991). Alle barna er vurdert av to vurderere, men ikke de samme to vurdererne. I slike tilfeller har man ikke lenger mulighet til å skille mellom hovedeffekt og interaksjonseffekt. Det lar seg ikke gjøre å identifisere om den ene vurdereren konsekvent har vært strengere enn den andre når ikke alle barna har blitt vurdert av samme vurderere. På grunn av denne begrensningen anbefales bruk av kryssede design i G teori (Shavelson & Webb 1991). Når det likevel åpnes for bruk av et nestet design i G teori, betyr ikke dette at reliabilitetskoeffisientene ikke påvirkes av hovedeffekter (som i klassisk teori), men at man ikke har mulighet til å identifisere hva som er hovedeffekter og hva som er interaksjonseffekter.

3.3.5 G studie og D studie

I G teori skjer den konkrete reliabilitetsestimeringen gjennom en G studie (generalizability study) og en D studie (decision study). I det følgende vil det bli vist hvordan Bus Story skårenes reliabilitet kunne blitt estimert i to-fasett designet gjennom disse to studiene. For å lette fremstillingen, vil det bli konstruert et eksempel med tenkte tallstørrelser.

I G studien studeres variasjonen eller variansen i skårene. Ved hjelp av variansanalyse deles totalvarians opp i varianskomponenter (Shavelson & Webb 1991). Med andre ord deles Bus Story skårenes variasjon opp i hovedeffekter for barn, testinger og vurderere, og interaksjonseffekter dem imellom. Slik blir resultatet av en G studie varianskomponenter som viser hvor stor andel eller prosent av den totale variansen som må tilskrives de ulike kildene til variasjon (Brennan 2000). I tabell 1 er de aktuelle variasjonskildene presentert med tenkte størrelser for varianskomponentene.

Tabell 1. Tenkt G studie av Bus Story skårer

Variasjonskilde	σ^2	EVK	%
Barn (b)	σ_b^2	0,550	55
Testinger (t)	σ_t^2	0,020	2
Vurderere (v)	σ_v^2	0,040	4
Barn x Testinger (bt)	σ_{bt}^2	0,050	5
Barn x Vurderere (bv)	σ_{bv}^2	0,100	10
Testinger x Vurderere (tv)	σ_{tv}^2	0,040	4
Barn x Testinger x Vurderere (btv,e)	$\sigma_{btv,e}^2$	0,200	20

Fra tabellen fremgår det at Bus Story skårenes varians kan deles i syv varianskomponenter når de to fasettene testinger og vurderere fokuseres. Det er altså forutsatt et krysset design i den tenkte estimeringen, da det som nevnt ikke ville være mulig å skille hovedeffekter fra interaksjonseffekter i et nestet design. De tre øverste radene viser til hovedeffekter grunnet barn, testinger og vurderere. De tre neste radene viser interaksjonseffekter mellom hver av hovedeffektene. I tabellenes siste rad vises interaksjonseffekt mellom alle de tre variasjonskildene samlet med uidentifiserte feil (betegnet som e for error). Som vist betegnes varianskomponentene med tegnet σ^2. Videre er det oppgitt tenkte størrelser for hver av de estimerte varianskomponentene (EVK). Formler og fremgangsmåter for å estimere varianskomponentene, vil bli vist i kapittel 4 hvor man har empiriske tall å ta utgangspunkt i.

Når totalvariansen er delt opp i varianskomponenter, kan man vurderere hvilke feilkilder som er mest omfattende og som dermed har påvirket Bus Story skårene i størst grad. Dersom tallene fra tabell 1 var reelle, kunne man tolket dem som at kun halvparten av Bus Story skårenes variasjon (55 %) er grunnet variasjon i barnas

gjenfortellingskompetanse. Man kunne også lest fra tabellen at inkonsistent vurdering har vært en større reliabilitetstrussel enn inkonsistens grunnet testinger. Skårene er med andre ord mer avhengige av hvem som har vurdert prestasjonene enn de er avhengige av når barna ble testet og hvem som testet dem.

Når varianskomponentene fra G studien er estimert, tjener disse som grunnlag for estimering av reliabilitetskoeffisienter i D studien (Shavelson & Webb 1991). I G teori finnes det to ulike reliabilitetskoeffisienter, èn for relative avgjørelser og èn absolutte avgjørelser. Ved estimering med fokus på relative feil alene, betegnes reliabilitetskoeffisienten som ”generalizability coefficient” (notert som $E\rho^2$). Når det tas hensyn til både absolutte og relative feil, kalles koeffisienten ”index of dependability” (notert som ϕ) (Shavelson & Webb 1991). Undersøkeren må med andre ord avgjøre hvilke feil eller effekter som skal inkluderes i reliabilitetsestimeringen.

Dersom formålet med Bus Story testingen var å finne de barna som skåret best eller dårligst, ville en reliabilitetsestimering som kun inkluderte relative feiltyper være mest hensiktsmessig. Om vurdererne ikke hadde vært like strenge generelt sett, eller om forhold ved testingene førte til at alle barna skåret litt bedre eller dårligere, ville ikke spille noen rolle så lenge målefeilene ikke påvirket rangeringen av barnas skårer.

Dette er imidlertid ikke tilfelle slik Bus Story skårene er tenkt brukt i språkprosjektet. I og med at skårene skal brukes til å lage en normering, vil både relative og absolutte feil påvirke skårenes reliabilitet. Dersom vurdererne ikke har vært like strenge generelt sett, og forhold ved testingene har ført til at alle barna skåret litt bedre eller dårligere, vil dette påvirke barnas skårer som normeringsgrunnlag. Dersom vurdererne i tillegg har vært inkonsistente i sin vurdering fra barn til barn, og forhold ved testingene har påvirket barnas skårer ulikt fra barn til barn, vil også dette være målefeil som reduserer reliabiliteten til skårene som normeringsgrunnlag.

Fra dette fremgår det at en estimering som lar både de relative og absolutte målefeilene påvirke reliabilitetskoeffisienten, vil gi et mer korrekt bilde av Bus Story

skårenes reliabilitet enn man ville få ved kun å inkludere relative feil. Der tradisjonelle estimeringsmetoder ville oversetimere Bus Story skårenes reliabilitet ved kun å ta hensyn til relative feil, kan man i G teori velge hvilke feiltyper som skal inngå. Feilvariansen estimeres ved hjelp av følgende formler.

$$\sigma^2_{Rel} = \frac{\sigma^2_{bt}}{n_t} + \frac{\sigma^2_{bv}}{n_v} + \frac{\sigma^2_{btv,e}}{n_t n_v}$$

$$\sigma^2_{Abs} = \frac{\sigma^2_{t}}{n_t} + \frac{\sigma^2_{v}}{n_v} + \frac{\sigma^2_{bt}}{n_t} + \frac{\sigma^2_{bv}}{n_v} + \frac{\sigma^2_{tv}}{n_t n_v} + \frac{\sigma^2_{btv,e}}{n_t n_v}$$

Formlene viser tydelig hvordan kun interaksjonseffektene mellom barn og feilkilder inngår i estimering av relativ feilvarians. I absolutt feilvarians (som inkluderer både relative og absolutte feil) inngår derimot seks av de syv varianskomponentene fra tabell 1. Den eneste varianskomponenten som ikke inngår, er hovedeffekt fra barn som ikke er en feilkilde men ”object of measurement”.

Formlene viser også at hver varianskomponent skal divideres på ønsket antall testinger og vurderere. For å estimere reliabilitet for bruk av flere testinger og vurderere, benyttes varianskomponentene fra G studien som parametre og man deler på ønsket antall testinger/vurderere. Dette betyr at jo flere testinger/vurderere man deler på, jo mindre blir feilvarianskomponentene som inngår i reliabilitetsestimeringen, og størrelsen på reliabilitetskoeffisienten øker. Som man intuitivt ville anta, vil altså generaliseringer fra gjennomsnitt av flere testinger/vurderere være sikrere enn generalisering fra èn testing og èn vurderer.

I språkprosjektet ble barna kun testet èn gang og vurdert av èn vurderer. En reliabilitetsestimering av Bus Story skårene i språkprosjektet bør derfor også estimere reliabilitet for èn testing og èn vurderer. Dersom man ikke finner tilfredsstillende reliabilitet for bruk av èn testing og èn vurderer, kan det imidlertid også være interessant å estimere reliabilitet for bruk av flere testinger og vurderere. Slik kan man få praktisk informasjon om hvor mange testinger eller vurderere som må benyttes for å oppnå tilfredsstillende reliabilitet. Det bør nevnes at man også kan få

tilsvarende informasjon ved bruk av Spearman-Brown formelen fra klassisk teori (Fan & Chen 2000).

Når feilvariansen så er estimert, kan reliabilitetskoeffisientene beregnes ut fra følgende formler

$$E\rho^2 = \frac{\sigma_b^2}{\left(\sigma_b^2 + \sigma_{Rel}^2\right)}$$

$$\phi = \frac{\sigma_b^2}{\left(\sigma_b^2 + \sigma_{Abs}^2\right)}$$

De to koeffisientene beregnes på samme måte; universvarians delt på observert varians (som består å univers- og feilvarians). Det eneste som skiller dem fra hverandre er altså hvilken feilvarians som inkluderes. Resultatene av reliabilitetsestimeringen på grunnlag av de tenkte størrelsene er oppgitt i tabell 2. En mer grundig fremstilling av koeffisientestimeringen vil bli gitt i kapittel 4 hvor man har tilgang til empiriske tall.

Tabellen er vist i sin helhet på neste side.

Tabell 2. Tenkt G studie og D studie av Bus Story skårer

Kilde	G studie	Alternative D studier				
n_t =	1	1	2	2	1	2
n_v =	1	2	1	2	3	3
Barn (b)	0,550	0,550	0,550	0,550	0,550	0,550
Testinger (t)	0,020	0,020	0,010	0,010	0,020	0,010
Vurderere (v)	0,040	0,020	0,040	0,020	0,013	0,013
Bt	0,050	0,050	0,025	0,025	0,050	0,025
Bv	0,100	0,050	0,100	0,050	0,033	0,033
Tv	0,040	0,020	0,020	0,010	0,013	0,007
Btv,e	0,200	0,100	0,100	0,050	0,067	0,033
σ^2_{Rel}	0,350	0,200	0,225	0,125	0,150	0,092
σ^2_{Abs}	0,450	0,260	0,295	0,165	0,197	0,122
$E\rho^2$	.61	.73	.71	.81	.79	.86
ϕ	.55	.68	.65	.77	.74	.82

I tabell 2 vises de samme varianskomponentene som i tabell 1 som tjener som parametre for den videre estimeringen. Varianskomponentene er delt på antall testinger og vurderere i de ulike D studiene. For hvert av D studiene er det estimert absolutt og relativ feilvarians, samt koeffisienter for absolutte og relative avgjørelser.

Dersom tallene fra G studien var reelle, kunne resultatene først og fremst vært tolket som at bruk av èn testing og èn vurderer gir en reliabilitetskoeffisient (ϕ) på .55. Kun 55 % av variansen i Bus Store skårene ville vært knyttet til ulik fortellerkompetanse hos barna. Det vil si at de oppnådde Bus Story skårene bare delvis kunne generaliseres til barnas universskårer, fordi de i betydelig grad ser ut til å være påvirket av målefeil grunnet dag til dag svingninger i barnas prestasjonsevne,

inkonsistent testadministrasjon og inkonsistent vurdering. Samtidig kan det være vanskelig å avgjøre hvor høye koeffisientene bør være før de anses som tilfredsstillende. I følge Brown (1983) finnes det ingen fasitsvar på dette spørsmålet. Ideelt sett skulle 100 % av totalvarians vært assosiert med universvarians, men dette vil svært sjelden være tilfelle i praksis. Dermed må hver undersøker avgjøre hvor mye feilvarians man vil være villig til å akseptere.

D studiene i tabell 2 viser hvor mange ganger barna måtte bli testet og hvor mange vurderere det måtte benyttes for å oppnå ulike reliabilitetskoeffisienter. Dersom man tenker seg at kritisk grense for reliabilitet var satt til .80, måtte barna blitt testet to ganger og vurdert av tre vurderere for å oppnå tilfredsstillende reliabilitet ($\phi < .80$) for Bus Story skårene. Slik kan D studier gi praktisk informasjon om hvilke betingelser som må legges for testingen for å oppnå reliable resultater.

Men selv høye reliabilitetskoeffisienter måtte her bli tolket med forbehold. I og med at inkonsistens grunnet historien ikke er fanget opp, kan det tenkes at skårenes reliabilitet ville bli lavere dersom også disse målefeilene hadde fått mulighet til å redusere reliabilitetskoeffisientene.

3.3.6 Oppsummering

Det har blitt drøftet hvilke reliabilitetsteorier som vil være mest hensiktsmessige for å forstå Bus Story skårenes reliabilitet, og hvilke metoder det vil være mest hensiktsmessig å bruke ved estimering av reliabiliteten. Gjennom drøftingen har det blitt pekt på en rekke fordeler ved bruk av G teori. For det første kan reliabilitetsestimering i G teori skje uten at man må bygge den på urealistiske forutsetninger om parallellitet og fravær av virkelige endringer mellom test og retest som legges i klassisk reliabilitetsteori.

I tillegg åpner G teoriens bruk av variansanalyse for en mer raffinert reliabilitetsestimering enn den man finner i klassisk teori. Ved å ta hensyn til både testinger og vurderere i samme reliabilitetsestimering, får man fanget opp både

hovedeffekter fra feilkildene, interaksjonseffekter mellom barn og feilkildene, og interaksjonseffekt mellom de to feilkildene. I og med at Bus Story skårene skal danne grunnlag for normering og vil påvirkes av både relative og absolutte feiltyper, betyr dette at man ved bruk av G teori kan fange opp flere av målefeilene som i realiteten kan ha redusert Bus Story skårenes reliabilitet, enn man får fanget opp ved bruk av tradisjonelle estimeringsmetoder.

Bruk av G teori gir ikke bare mulighet til å fange opp en større del av aktuell feilvarians i estimeringen, men gir også en differensiert feilkomponent som viser omfang av de ulike feilkildene. Der tradisjonelle metoder kun estimerer størrelsen på èn udifferensiert feilkomponent, kan G teori gi verdifull informasjon om hvilke feilkilder som representerer de mest alvorlige truslene. Det skal riktignok nevnes at man ikke får mulighet til å skille inkonsistens i barnas prestasjonsevne fra inkonsistent testadministrasjon, som begge inngår i fasetten gjentatte testinger slik designet er definert her. Likevel vil det være fordelaktig at G teori differensierer feilkomponenten i så mange kilder som mulig.

Til sist ser G teori ut til å gi et godt begrepsverk for å forstå reliabilitet. Som drøftet kan reliabilitet være et fenomen det er vanskelig å begripe, og usikkerhet og misforståelser kan føre til lite hensiktsmessig estimering og tolkning av reliabilitet. Ved bruk av G teori tvinges man til å eksplisitt definere hvilke fasetter som skal inkluderes. Dette kan øke undersøkers bevissthet om hvilke målefeil som får påvirke reliabilitetskoeffisientene, og kanskje enda viktigere hvilke målefeil som *ikke* gis anledning til å redusere reliabilitetskoeffisientene. I G teori kommer det også tydelig frem at man kun kan generalisere over fasetter som har vært inkludert i estimeringen, og som kan hevdes å være random ut fra karvene som stilles i G teori. Slik kan eksplisitt definering av fasetter kan også tenkes å øke undersøkers bevissthet rundt de generaliseringer man ønsker å gjøre fra barnas testresultater til universskårer.

Tatt i betraktning fordelene som finnes ved bruk av G teori, kan det synes overraksende at de tradisjonelle estimeringsmetodene fremdeles er de mest anvendte metodene (Fan & Chen 2000, Hogan m.fl. 2000). Som understreket av Jaeger (1991)

fører bruk av G teori til at testbrukere ikke lenger må basere sine reliabilitetsestimeringer på foreldede metoder som man finner i klassisk reliabilitetsteori.

4. Estimering av vurdererreliabilitet

I oppgavens første problemstilling ble det reist spørsmål om hvordan Bus Story skårenes reliabilitet burde estimeres, tatt i betraktning hvilke feilkilder som kan tenkes å ha påvirket skårene. Etter drøfting av ulike metoder, kan det konkluderes med at det vil være fordelaktig å estimere Bus Story skårenes reliabilitet gjennom G teori. Slik vil også G teori ligge til grunn ved besvarelse av oppgavens andre problemstilling; estimering og tolking av Bus Story skårenes vurdererreliabilitet.

Drøfting av oppgavens første problemstilling har vist at inkonsistent vurdering kun er en liten del av alle målefeil som kan tenkes å ha påvirket Bus Story skårenes reliabilitet. Målefeil grunnet inkonsistens i barnas prestasjonsevne, inkonsistens grunnet historien og inkonsistent testadministrasjon vil ikke få påvirke reliabilitetskoeffisienten ved estimering av vurdererreliabilitet. Disse målefeilene holdes konstante ved at de to vurderingene skjer på grunnlag av det samme vurderingsgrunnlaget (reskåring), og differanse i skårene gitt av de to vurdererne kan kun tolkes som inkonsistens grunnet vurderere. Inkonsistensen kan videre tolkes som ulik strenghet (relativ eller absolutt), at vurdererne har forstått skåringsreglene ulikt, et resultat av unøyaktighet/summeringsfeil/slurv, eller at vurderernes dagsform har fått påvirke hvilken skåre de har gitt. Ved estimering av vurdererreliabilitet kan man dermed studere i hvilken grad to vurderere har vært konsistente i sin vurdering, men kan ikke uttale seg om skårenes totale reliabilitet på grunnlag av estimeringen da det finnes en rekke aktuelle feilkilder det ikke tas hensyn til.

4.1 Design

I og med at det kun var mulig med replikasjon av vurderere, er dette også den eneste feilkilden som kan inkluderes i reliabilitetsestimeringen. I G teoriens termer har man da et en-fasett design (Shavelson & Webb 1991). Som drøftet tidligere vil det også være nødvendig å inkludere både relative og absolutte feil i estimeringen, i og med at

begge disse feiltypene vil redusere skårenes reliabilitet når de skal inngå som normeringsgrunnlag. I G teori betegnes dette som design for absolutte avgjørelser, fordi man i praksis velger om de absolutte feilene skal inkluderes eller ikke (Shavelson & Webb 1991).

I en-fasett designet vil universet kun bestå av vurderere, og det må defineres hvilke vurderere man vil være villig til å generalisere over. I første omgang er det ønskelig å generalisere over de ti forskningsassistentene som har vurdert barnas gjenfortellinger i språkprosjektet. Som tidligere nevnt ser assistentene ut til å være relativt like med hensyn til utdanning, tidligere testerfaring og opplæring. I en slik homogen gruppe kan èn assistent ses som like representativ for gruppa som en annen, og de to utvalgte assistentene kunne vært byttet ut med de andre forskningsassistentene. Samtidig kan universet defineres videre ved å definere det som et speil av de utvalgte assistentene (Brennan 2000). Da vil universet romme forskningsassistentene og alle andre vurderere med tilsvarende egenskaper. Slik ivaretas G teoriens krav om utbyttbarhet, og vurderere kan ses som en tilfeldig (random) fasett selv om de ved en slik definering ikke lenger vil være tilfeldig trukket.

Som drøftet i kapittel 3.3.4, anbefales bruk av kryssede design i G teori i og med at dette er en forutsetning for å kunne skille hovedeffekter fra interaksjonseffekter. Men på grunn av praktiske hensyn var det ikke mulig å oppnå at alle barna ble vurdert av de samme to vurdererne. I stede måtte det benyttes et nestet design, der de to utvalgte forskningsassistentene allerede hadde vurdert hver sin halvdel av barnegruppen, og undersøker gjennomførte reskåring. Dermed er alle barna vurdert av to vurderere, men ikke de samme to vurdererne. Når hele datamaterialet analyseres samlet, bygger altså estimeringen på et nestet design med de begrensninger det innebærer. For å likevel kunne skille mellom hoved- og interaksjonseffekter, kan imidlertid datamaterialet ses som to kryssede design. Dersom man studerer hver halvdel av barnegruppen for seg, har alle barna i hver halvdel blitt vurdert av de to samme vurdererne. Slik kan man slå fast at reliabilitetsestimeringen bygger på et random en-fasett og krysset/nestet design for absolutte avgjørelser.

4.2 Koeffisienter

Selv om det er fastslått at reliabilitetsestimering bør ta hensyn til både relative og absolutte feil (ϕ), vil det i tillegg bli presentert koeffisienter som kun tar hensyn til relative feil. I alle en-fasett design vil den relative koeffisienten fra G teori ($E\rho^2$) være analog til koeffisienten man finner ved hjelp av tradisjonelle estimeringsmetoder i klassisk reliabilitetsteori (α). Dette er gjort for å tydeliggjøre hvordan lite hensiktsmessig valg av metode også kan føre til overestimert reliabilitet.

I tillegg må det avgjøres hvor mange vurderere reliabiliteten skal estimeres for. I følge Fan og Chen (2000) overestimeres ofte vurdererreliabilitet i undersøkelser som denne. I og med at det brukes to vurderere for å realisere replikasjon, estimeres også vurdererreliabiliteten for bruk av to vurderere. Reliabiliteten funnet ved bruk av gjennomsnitt fra to vurderere blir så feilaktig generalisert til resten av skårene som kun er gitt av èn vurderer, og resultatet blir en overestimert reliabilitetskoeffisient (Fan & Chen 2000). I og med at barna i språkprosjektet opprinnelig ble vurdert av èn vurderer, og at skåren denne ene vurdereren har gitt vil gjelde uavhengig av reskåring, bør det estimeres reliabilitet for bruk av èn vurderer. Også her vil det presenteres koeffisienter for bruk av flere vurderere for å vise hvordan koeffisientene blir ulike.

Bus Story skårene kan analyseres på ulike måter i og med at det for informasjonsskårene og setningslengdeskårene beregnes både delskårer og sumskårer. Delskårene for hver linje i skåringsarket summeres og gir en sumskåre. I reliabilitetsestimeringen vil det imidlertid kun tas utgangspunkt i sumskårene for informasjons- og setningslengdeskårene. Dette er hovedsakelig gjort fordi tolkningen av Bus Story skårene gjøres ut fra sumskårene og ikke delskårene. I tillegg kan ikke delskårene ses som uavhengige av hverandre. Slik skåringsreglene er definert, kan de ha en smittende effekt på hverandre. All estimering av reliabilitet vil dermed skje på grunnlag av sumskårene. Det vil imidlertid alltid bli skilt mellom informasjonsskåre og setningslengdeskåre da det ikke gir mening å slå disse skåretypene sammen.

Til sist skilles det mellom opprinnelig og ny transkribering i estimeringen. Skillet er gjort for å undersøke i hvilken grad det kommer ytterligere feilvarians inn ved transkribering. Skillet mellom opprinnelig og ny transkribering representerer også et annet skille som bedre kan betegnes som et skille mellom vurdererpar. Ved å betegne forskningsassistentene som gjorde den opprinnelige skåringen som vurderer A og B, og assistenten som gjorde reskåringen som vurderer C, kan man skille mellom vurdererparene AC og BC. Ved å studere vurdererparene hver for seg, har man som nevnt to kryssede design i stede for ett nestet design. Følgelig kan det estimeres koeffisienter for både absolutte og relative feil så lenge gruppene analyseres hver for seg. Når hele materialet ses samlet, kan det kun estimeres koeffisienter for absolutte feil i G teori. Det vil imidlertid også bli presentert reliabilitetskoeffisienter fra klassisk reliabilitetsteori for å vise hvordan lite hensiktsmessig valg av metode kan føre til overestimert reliabilitet i et nestet design.

4.3 Informasjonsskårer

For estimering av reliabilitet i G teori, er det utarbeidet et spesialprogram kalt GENOVA som utfører de nødvendige beregningene (University of Iowa 2008). Dette kan være svært nyttig når man arbeider med store datamaterialer. Ved bearbeiding av eget datamateriale har jeg imidlertid valgt å bruke SPSS som utgangspunkt og gjøre de enkle beregningene manuelt. Dette er gjort med hensyn til at datamaterialet er lite, og for å øke egen forståelse av de prosesser som inngår i reliabilitetsestimeringen. Estimeringsprosessen vil bli nøye presentert ved de første koeffisientene, men sløyfes ved estimering av de resterende koeffisientene da den vil skje ut i fra de samme prinsippene hver gang.

4.3.1 Vurdererpar AC - Opprinnelig transkribering

Ved estimering av reliabilitet for informasjonsskårene med opprinnelig transkribering har alle barna blitt vurdert av de samme to vurdererne (vurdererpar AC), og designet

er krysset. Først estimeres varianskomponentene i G studien. Ved å gjøre en variansanalyse i SPSS får man følgende tabell.

Tabell 3. Variansanalyse av informasjonsskårer med opprinnelig transkribering (AC)

Kilde	SS	Df	MS
Barn (b)	780,059	16	48,754
Vurderere (v)	18,382	1	18,382
Bv,e	11,118	16	0,695

Fra tabell 3 ser man at det finnes tre kilder til variasjon i en-fasett designet; barn (object of measurement), hovedeffekt fra vurderere og interaksjonseffekt mellom barn og vurderere med uidentifiserte feil. Sum of Squares verdiene (SS) viser til omfanget av varians for hver av kildene. SS er delt på frihetsgrader (degrees of freedom som er lik n - 1) for å få Mean Square verdier (MS). I gruppen informasjonsskårer med opprinnelig transkribering er det 17 barn som har blitt reskåret, og frihetsgradene for personer blir 16. Frihetsgradene for vurderere vil være 1 i og med at det er 2 vurderere, og de to nevnte frihetsgradene multipliseres for å beregne frihetsgrader for barn x vurderere. MS verdiene fungerer som parametre når varianskomponentene skal estimeres. Varianskomponenten for $\sigma^2_{bv,e}$ vil tilsvare MS verdien for $\sigma^2_{bv,e}$. Når denne størrelsen er kjent, kan ligningen for vurderere og barn løses.

$$\sigma^2_b = \sigma^2_{bv,e} + n_v\sigma^2_b$$

$$\sigma^2_v = \sigma^2_{bv,e} + n_b\sigma^2_v$$

$$\sigma^2_{bv,e} = 0{,}695$$

Estimert varianskomponent for vurderere blir dermed

$\sigma_v^2 = \sigma_{bv,e}^2 + n_b\sigma_v^2 \rightarrow 18{,}382 = 0{,}695 + 17\,\sigma_v^2 \rightarrow 17\,\sigma_v^2 = 17{,}687 \rightarrow \sigma_v^2 = 1{,}040$

Tilsvarende blir estimert varianskomponent for barn

$\sigma_b^2 = \sigma_{bv,e}^2 + n_v\sigma_b^2 \rightarrow 48{,}754 = 0{,}695 + 2\,\sigma_b^2 \rightarrow 2\,\sigma_b^2 = 48{,}059 \rightarrow \sigma_b^2 = 24{,}030$

De estimerte varianskomponentene kan nå sette inn i en ny tabell hvor all informasjonen fra G studiet er presentert.

Tabell 4. G studie av informasjonsskårer med opprinnelig transkribering (AC)

Kilde	SS	df	MS	EVK	%
Barn (b)	780,059	16	48,754	24,030	93,3
Vurderere (v)	18,382	1	18,382	1,040	4,0
Bv	11,118	16	0,695	0,695	2,7

Tabellen er nå utvidet til å inneholde de estimerte varianskomponentene (EVK) og hvor mange prosent de utgjør av den totale variansen. Ved omregning til prosent kan de lettere tolkes i forhold til hverandre. Fra tabell 4 ser man at en betydelig del av variansen (93,3 %) er assosiert til barna og kan tolkes som ulik gjenfortellingskompetanse. Komponentene for feilvarians viser at 4,0 % av totalvariansen er assosiert med feil enten på grunn av at vurdererne ikke har vært like strenge eller ikke har brukt skalaen likt (hovedeffekt) og videre at 2,7 % av totalvariansen skyldes at vurdererne ikke har vært like strenge mot alle barn (interaksjonseffekt) eller på grunn av uidentifiserte feil.

På grunnlag av informasjonen som er fremkommet i G studien, kan D studien gjennomføres. Ved estimering av feilvarians med fokus på både relative og absolutte feil, benyttes formelen for σ_{Abs}^2 vist i kapittel 3.3.5. I en-fasett design vil det kun være

to varianskomponenter som inngår, èn for hovedeffekt (σ_v^2) og èn for interaksjonseffekt ($\sigma_{bv,e}^2$).

$$\sigma_{Abs}^2 = \frac{\sigma_v^2}{n_v} + \frac{\sigma_{bv,e}^2}{n_v} = \frac{1{,}040}{1} + \frac{0{,}695}{1} = 1{,}735$$

Her er varianskomponentene delt på èn, slik at reliabiliteten også estimeres for bruk av èn vurderer. Når størrelsen på estimert feilvarians så er kjent, kan den plasseres i formelen for index of dependability (ϕ) som er G teoriens reliabilitetskoeffisient for absolutte avgjørelser.

$$\phi = \frac{\sigma_b^2}{\left(\sigma_b^2 + \sigma_{Abs}^2\right)} = \frac{24{,}030}{24{,}030 + 1{,}735} = .933$$

Vurdererreliabiliteten for Bus Story informasjonsskårer med opprinnelig transkribering vurdert av vurdererpar AC, er altså estimert til .933. Reliabilitetskoeffisienten gjelder bruk av èn vurderer og absolutte avgjørelser. I tabell 4 presenteres også koeffisienter beregnet for to vurderere, relative koeffisienter beregnet i G teori ($E\rho^2$) og koeffisienter estimert ved hjelp av klassisk reliabilitetsteori og korrelasjon (α).

Tabell 5. D studie av informasjonsskårer med opprinnelig transkribering (AC)

Antall vurderere	ϕ	$E\rho^2$	α
1	.933	.972	.972
2	.965	.986	.986

Fra tabellen ser man at reliabilitetskoeffisientene for bruk av to vurderere som ventet er høyere enn for bruk av èn vurderer. Man ser også at $E\rho^2$ på .972 er større enn ϕ på .933. Hovedeffekten fra vurderere er variansen som skiller dem fra hverandre. Jo større forskjell mellom ϕ og $E\rho^2$, jo mer av feilvariansen er knyttet til at vurdererne generelt sett ikke har vært like strenge. Som fremkommet i G studien, kan

inkonsistens mellom vurderer A og C i større grad forklares som at de generelt sett ikke har vært like strenge (4 %), enn det kan forklares som at hver vurderer har vært inkonsistent over barn (2,7 %). α tilsvarer $E\rho^2$ så lenge det kun tas hensyn til èn feilkilde og designet er krysset. I dette tilfellet ville altså bruk av både $E\rho^2$ og α føre til overestimert reliabilitet for Bus Story skårene.

4.3.2 Vurdererpar BC - Ny transkribering

Analysen av nytranskriberte informasjonsskårer vil skje etter samme prinsipper og formler som i kapittel 4.3.1, i og med at det fremdeles kun er halve materialet som studeres og designet er krysset. Det vil derfor ikke vises til fremgangsmåter og utregninger. Analysen gav følgende varianskomponenter og reliabilitetskoeffisienter

Tabell 6. G studie og D studie av informasjonsskårer med nytranskribering (BC)

Kilde	SS	Df	MS	EVK	%
Barn (b)	1988,139	17	116,949	58,020	98,5
Vurderere (v)	0,028	1	0,028	-0,049	0
Bv,e	15,472	17	0,910	0,910	1,5

Antall vurderere	ϕ	$E\rho^2$	α
1	.985	.985	.985
2	.992	.992	.992

Vurdererreliabilitet (ϕ) for nytranskriberte informasjonsskårer er estimert til .985, altså høyere vurdererreliabilitet enn for vurdererpar AC. Fra tabell 6 ser man at ϕ sammenfaller med $E\rho^2$ og α. Forklaringen på dette finner man i at varianskomponenten for hovedeffekt fra vurderere praktisk talt er null (se tabell 6). Varianskomponenten for vurderere er blitt estimert til -0,049. Shavelson og Webb (1991) tolker negative komponenter som samplingsfeil på grunn av små utvalg. Siden

negative komponenter ikke kan brukes i G teori, er en mulig løsning å sette dem til 0 slik det er blitt gjort her. Shavelson og Webb (1991) mener dette ikke er en tilfredsstilende løsning, men at den likevel benyttes i mangel på andre løsninger. I tabell 6 er det negative estimatet ført opp, men satt til 0 i kolonnen over prosent for å vise hva som er blitt gjort. I og med at hovedeffekt er det eneste som skiller ϕ fra $E\rho^2$, sammenfaller de når hovedeffekten er 0.

Når hovedeffekten er satt til 0, tolkes det som at vurderer B og C har vært nøyaktig like strenge generelt sett. Den eneste inkonsistensen dem imellom er grunnet at vurdererne ikke har vært like strenge mot alle barna, men også denne feilvariansen er liten (1,5 %). Derfor har også reliabilitetskoeffisienten blitt så høy som .985.

4.3.3 Samlede skårer

Til nå har det blitt estimert vurdererreliabilitet innen hver av halvdelene i datamaterialet. Når det skal estimeres reliabilitet for datamaterialet samlet, har man som drøftet et nestet design i stede for to kryssede. Dette gir en noe annerledes fremgangsmåte enn den som er blitt vist i de to foregående delkapitlene. I G studien ble det funnet følgende verdier.

Tabell 7. G studie av samlede informasjonsskårer

Kilde	SS	Df	MS	EVK	%
Barn (b)	4482,086	34	131,826	65,270	98,1
V,bv,e	45,000	35	1,286	1,286	1,9

Som det fremgår av tabellen kan det kun skilles mellom to variasjonskilder i nestet en-fasett design; object of measurement og feil. Man må forholde seg til en udifferensiert feilkomponent i og med at det ikke kan skilles mellom hoved- og interaksjonseffekter. Feilkomponenten inneholder altså hovedeffekt fra vurderere, interaksjonseffekt mellom barn og vurderere samt uidentifiserte feil (v,bv,e). Som en

følge av dette er også formlene for estimering av varianskomponenter noe annerledes enn i de forrige kapitlene. Når man kjenner størrelsen for $\sigma^2_{v,bv,e}$ på 1,286, var det kun ligningen for barn (σ^2_b) som måtte løses.

$\sigma^2_b = \sigma^2_{v,bv,e} + n_v\sigma^2_b \rightarrow \sigma^2_b = \sigma^2_{v,bv,e} + 2\sigma^2_b = 1{,}286 + 2\sigma^2_b \rightarrow$ 131,826 = 1,286 + 2 $\sigma^2_b \rightarrow$

$2\sigma^2_b$ = 130,540 $\rightarrow$ σ^2_b = 65,270

Tabell 7 viser at nesten all variansen (98,1 %) er assosiert til barna. Den eneste informasjonen man får om feilvarians, er at 1,9 % av totalvariansen er assosiert med feil enten på grunn av at vurdererne ikke har vært like strenge generelt sett (hovedeffekt), at vurdererne ikke har vært like strenge mot alle barn (interaksjonseffekt) eller uidentifiserte feil. Formelen for estimering av feilvarians blir da

$$\sigma^2_{Abs} = \frac{\sigma^2_{v,bv,e}}{n_v} = \frac{1{,}286}{1} = 1{,}286$$

Estimert feilvarians satt inn i formelen for index of dependability (ϕ) gir følgende resultat

$$\phi = \frac{\sigma^2_b}{\left(\sigma^2_b + \sigma^2_{Abs}\right)} = \frac{65{,}270}{65{,}270 + 1{,}286} = .980$$

Vurdererreliabilitet for samlede informasjonsskårer er altså estimert til .980. I tabell 8 er også koeffisienter beregnet for 2 vurderere, relative koeffisienter beregnet i G teori ($E\rho^2$) og koeffisienter beregnet i klassisk teori (α) presentert.

Tabell 8. D studie av samlede informasjonsskårer

Antall vurderere	ϕ	$E\rho^2$	α
1	.980	.980	.984
2	.990	.990	.992

Tabell 8 viser hvordan α blir høyere enn ϕ. Den bygger på klassisk teori som forutsetter et krysset design og overestimerer dermed reliabiliteten dersom den benyttes i et nestet design. I dette tilfellet er ikke overestimeringen dramatisk fordi koeffisientene i utgangspunktet er så høye. De relative koeffisientene estimert i G teori ($E\rho^2$) tilsvarer ϕ, fordi den udifferensierte feilkomponenten også har inngått i estimering av $E\rho^2$.

4.4 Setningslengdeskårer

Ved estimering av reliabilitet for setningslengdeskårene, vil det ikke bli vist utregninger som i kapittel 4.3, da prinsippene vil være de samme også her. I det følgende presenteres det derfor tre tabeller, èn for vurdererpar AC med opprinnelig transkribering, èn for vurdererpar BC med ny transkribering og til sist en tabell for samlede setningslengdeskårer. Tabellene vil bli kommentert samlet til slutt.

Tabell 9. G studie og D studie for setningslengdeskårer med opprinnelig transkribering (AC)

Kilde	SS	Df	MS	EVK	%
Barn (b)	198,882	16	12,430	5,669	79,4
Vurderere (v)	7,529	1	7,529	0,379	5,3
Bv,e	17,471	16	1,092	1,092	15,3

Antall vurderere	ϕ	$E\rho^2$	α
1	.793	.838	.838
2	.885	.912	.912

Tabell 10. G studie og D studie av setningslengdeskårer med nytranskribering (BC)

Kilde	SS	Df	MS	EVK	%
Barn (b)	140,139	17	8,243	3,967	92,8
Vurderere (v)	0,250	1	0,250	-0,003	0
Bv,e	5,250	17	0,309	0,309	7,2

Antall vurderere	ϕ	$E\rho^2$	α
1	.928	.928	.928
2	.962	.962	.962

Tabell 11. G studie og D studie av samlede setningslengdeskårer

Kilde	SS	Df	MS	EVK	%
Barn (b)	482,486	34	14,191	6,66	88,4
Bv,e	30,500	35	0,871	0,871	11,6

Antall vurderere	ϕ	$E\rho^2$	α
1	.885	.885	.900
2	.939	.939	.948

De tre tabellene presentert her, viser resultater fra G studier og D studier for setningslengdeskårene. På samme måte som i forrige kapittel, er det først estimert reliabilitet for halvdelene hver for seg, og deretter for hele materialet samlet.

Sammenlignet med informasjonsskårene, er det funnet lavere koeffisienter for setningslengdeskårene enn det ble funnet for informasjonsskårene. Ved skårer med opprinnelig transkribering vurdert av vurdererpar AC, ble det funnet en ϕ på .933 for informasjonsskårene og .793 for setningslengdeskårene. Ved skårer med ny transkribering vurdert av vurdererpar BC, ble det funnet en ϕ på .985 for informasjonsskårene og .928 for setningslengdeskårene. Samlet ble reliabiliteten estimert til .980 for informasjonsskårene og .885 for setningslengdeskårene. Dette betyr at vurderne har vært mindre samstemte i skåring av setningslengde enn de har vært ved skåring av informasjon.

Reliabilitetskoeffisienten for vurdererpar BC (.928) er høyere enn den er for vurdererpar AC (.793) på setningslengdeskårene. Man ser altså den samme tendensen til at vurdererpar BC har vært mer samstemte enn vurdererpar AC som også ble sett ved informasjonsskårene. Som vist i tabell 10, ble hovedeffekt for vurdererpar BC satt til 0 da den ble estimert til -0,003. Også ved informasjonsskåring ble hovedeffekt fra vurdererpar BC satt til null. Det vil si at vurderer B og C vært like strenge generelt sett både ved informasjons- og setningslengdeskåring. De har imidlertid vært mer

inkonsistente fra barn til barn (interaksjonseffekt) ved skåring av setningslengde (7,2 %) enn de var ved informasjonsskåring (1,5 %).

For vurdererpar AC må inkonsistensen forklares annerledes. Ved skåring av informasjon var den største delen feilvarians assosiert med ulik strenghet generelt sett (4 %), og en mindre del av den (2,7 %) var assosiert med at de ikke var like strenge i vurdering av alle barna. Ved skåring av setningslengde, var derimot denne interaksjonseffekten assosiert med en betydelig større del av feilvariansen (15,3 %), mens 5,3 % var knyttet til ulik strenghet generelt sett.

5. Drøfting av resultater

I kapittel 4 er det blitt estimert en rekke reliabilitetskoeffisienter. Det er skilt mellom informasjonsskårer og setningslengdeskårer, mellom nytranskribering og opprinnelig transkribering og mellom vurdererpar. Datamaterialet er også analysert samlet. Det er skilt mellom koeffisienter for relative og absolutte avgjørelser, og mellom bruk av èn og to vurderere. Etter en slik grundig analyse av materialet, kunne det identifiseres noen tendenser som ble presentert avslutningsvis i kapittel 4. I det følgende kapitlet vil det drøftes hvilken informasjon de ulike koeffisientene kan gi.

5.1 Hovedfunn

Reliabilitetsestimering på grunnlag av samlede informasjonsskårer og setningslengdeskårer gir koeffisienter på henholdsvis .980 og .885. Som drøftet kan det være vanskelig å fastsette en grense for når reliabiliteten kan anses som tilfredsstillende, og en slik grense vil ikke bli satt ved drøfting av disse resultatene. Vurdererreliabilitet på .980 for informasjonsskårene viser at kun 2 % av variansen i bus Story skårene er assosiert med feil grunnet inkonsistent vurdering. Som Brown (1983) hevder vil man aldri finne perfekt reliabilitet, men i dette tilfellet ligger ikke koeffisienten lagt unna, og må med sikkerhet kunne tolkes som høy vurdererreliabilitet. Reliabilitetskoeffisienten for setningslengdeskårene er imidlertid noe lavere (ϕ = .885) og viser en feilvarians på 11,5 %. Vurdererne har altså vært mindre samstemte ved skåring av setningslengde enn de var ved skåring av informasjon. Setningslengdeskårene er ikke like reliable som informasjonsskårene, men vurderes likevel til å ha rimelig høy vurdererreliabilitet. De høye koeffisientene for vurdererreliabilitet kan i G teoriens termer tolkes som at barnas observerte skårer med rimelig sikkerhet kan generaliseres til deres universskårer.

Dette gjør det interessant å spørre seg hva som kan ha ført til så høye koeffisienter. Det hevdes at man gjerne får høy vurdererreliabilitet dersom vurdererne har fått

intensiv opplæring og gjennomgått trening i skåringsarbeidet (Brennan 2001b, Brennan 2000, Dunbar m.fl. 1991). Resultatene kan dermed tolkes som at forskningsassistentene i språkprosjektet har mottatt god opplæring. Som tidligere nevnt ble all opplæring gitt i samlet gruppe. Alle assistentene skåret gjenfortellinger hver for seg for så å drøfte usikkerheter med de andre assistentene og lederne i språkprosjektet. Lederne for prosjektet har også vært tilgjengelige underveis i skåringsprosessen for å svare på spørsmål som dukket opp. Dette ser ut til å ha ført til bred enighet om hvordan gjenfortellingene skal skåres.

Videre hevdes det at tydelig skåringsmanual er et kritisk punkt for vurdererreliabilitet. Bruk av tester som tydelig definerer skåringsregler gir vanligvis høye reliabilitetskoeffisienter (Brennan 2000). Dette gjelder i særlig grad der det er trukket frem eksempler på respons og vist konkret hvordan responsene skal skåres (Dunbar m.fl. 1991). Innledningsvis ble mistanken om at det vanskelig lot seg utforme tydelige skåringsregler for Bus Story drøftet. Dersom denne mistanken stemmer, må de høye reliabilitetskoeffisientene tolkes som at lederne for språkprosjektet tross i utfordringene har utformet tydelige skåringsregler for sine assistenter. Prosjektlederne la mye arbeid i tydeliggjøring av skåringsmanualen. Assistentene fikk også tilgang til protokoller hvor ulike gjenfortellinger var ferdig skåret og som kunne tjene som maler. I tillegg til god opplæring, kan dette være blant forklaringene på den høye vurdererreliabiliteten.

5.2 Transkribering som feilkilde

Når halve materialet ble nytranskribert før skåring, var dette gjort for å undersøke hvor i vurderingsprosessen de mest alvorlige feilkildene kom inn; ved transkribering eller skåring. For informasjonsskårene ble det funnet en reliabilitetskoeffisient på .933 ved opprinnelig transkribering, og .985 ved ny transkribering. Tilsvarende ble det for setningslengdeskårene funnet en reliabilitetskoeffisient på .793 ved opprinnelig transkribering og .928 ved ny transkribering. Koeffisientene er altså høyere der det ble gjort ny transkribering enn der det ble skåret på grunnlag av

opprinnelig transkribering. Man ville ventet et motsatt resultat med tanke på at ny transkribering representerer nye mulige uenigheter mellom vurderere. Særlig med tanke på at vurdererne her må avgjøre hvor barnas setninger starter og slutter, ville det ikke vært overraskende om det kom ekstra feilvarians inn ved transkribering. Reliabilitetskoeffisientene funnet her, må imidlertid tolkes som at transkribering ikke ser ut til å medføre ekstra feilvarians. Samtidig må denne konklusjonen tas med forbehold. Det kan tenkes at koeffisientene ville vært enda høyere enn .985 og .928 dersom det ikke kom inn noe feilvarians ved transkribering. Det kan i alle fall fastslås at den eventuelle feilvariansen grunnet transkribering er liten.

Reskåring av fortellingene som opprinnelig ble skåret av vurderer A, ble gjort med utgangspunkt i den opprinnelige transkriberingen. Reskåring av fortellingene som opprinnelig var skåret av vurderer B, ble gjort med utgangspunkt i lydopptakene som dannet grunnlag for en ny transkribering og skåring. At det ble tatt utgangspunkt i vurderer A sin transkribering og i vurderer B sine lydopptak, var tilfeldig. Det kunne med andre ord like gjerne vært motsatt. Da er det fristende å gjøre et tankeeksperiment angående hvordan koeffisientene kunne blitt tolket dersom det heller ble tatt utgangspunkt i vurderer B sin transkribering og vurderer A sine lydopptak. Med lavere koeffisienter ved nytranskribering enn opprinnelig transkribering, ville det være nærliggende å tolke dette som at transkriberingen i betydelig grad medførte ekstra feilvarians. Dette understreker hvor lett reliabilitetskoeffisienter kan feiltolkes.

5.3 Vurdererpar

Tankeeksperimentet ovenfor viser at koeffisientene kan tolkes på ulike måter. Man bør spørre seg hvilke andre forklaringsmåter som kan være aktuelle. Hvordan kan forskjellen mellom koeffisienten for opprinnelig- og nytranskribering forklares dersom man ser bort ifra transkribering som feilkilde? For å drøfte dette kan det være hensiktsmessig å legge bort termene opprinnelig og ny transkribering, og heller fokusere på vurdererpar. Ved hjelp av denne redefineringen kan man si at vurdererpar

BC er mer enig enn vurdererpar AC både ved skåring av informasjon og setningslengde. Om forskjellen i enighet mellom de to vurdererparene er signifikant, ligger utenfor oppgavens problemstilling å undersøke. Det vil imidlertid være interessant å undersøke hva forskjellen i enighet kan bestå i.

Først og fremst kan forskjellen i enighet tolkes som ulikhet hos vurdererne. De høye koeffisientene funnet ved vurdererpar BC viser at de to vurdererne i stor grad har vært samstemte. En hovedeffekt på tilnærmet null ved skåring av både informasjon og setningslengde kan tolkes som at de to vurdererne har forstått skåringsreglene relativt likt, at de har brukt skalaen likt og generelt sett vært like strenge. Vurdererpar AC har imidlertid ikke vært like samstemte. Særlig ved skåring av setningslengde har de vært inkonsistente med en feilvarians på 20,6 %. Dette kan bety at de har forstått reglene for skåring av setningslengde ulikt, eller at de har vært inkonsistente i bruk av dem.

I tillegg til forhold ved vurdererne, vil det også være aktuelt å ta hensyn til barna vurdert av de to vurdererparene når forskjellen mellom dem skal tolkes. Som vist i kapittel 4, er reliabilitetskoeffisientene estimert ved å dividere sann varians (eller universvarians) på observert varians (som er summen av sann- og feilvarians). Det vil si at koeffisientene vil være farget av både sann varians og feilvarians. Koeffisientene vil altså bli påvirket av den konkrete gruppen som er studert (Crocker & Algina 1986). For å undersøke hvordan de to gruppene vurdert av vurdererpar AC og BC kan ha påvirket utfallet av reliabilitetskoeffisientene, kan man se nærmere på Bus Story skårene i de to gruppene.

Tabell 12. Distribusjon av informasjonsskårer og setningslengdeskårer

	Informasjonsskårer		Setningslengdeskårer	
Vurdererpar	AC	BC	AC	BC
Snitt av mean	7,8	17,7	2,9	5,8
Snitt av SD	4,9	7,7	2,6	2,1

I tabell 12 får man informasjon om middelverdier og spredning av skårene, både for hver av vurdererne og som snitt for hvert vurdererpar. Ut fra denne informasjonen kan man drøfte hvordan distribusjon av skårer i de to gruppene kan ha påvirket reliabilitetskoeffisientene.

For informasjonsskårene ser man at standardavviket er høyere i gruppen vurdert av BC enn i gruppen vurdert av AC. Standardavviket viser hvor vidt skårene er spredt omkring mean (Brown 1983). Man finner altså at informasjonsskårene gitt av vurdererpar BC, er mer spredt enn de gitt av vurdererpar AC. I følge Crocker og Algina (1986) vil gruppens homogenitet påvirke reliabilitetskoeffisienter. En gruppe med liten spredning av skårer regnes for mer homogen enn en gruppe med mer større skårespredning. Hvordan dette påvirker reliabilitetskoeffisientene kan forklares med utgangspunkt i formelen som benyttes for å estimere reliabilitet. Som tidligere nevnt er det kun de tilfeldige feilene det tas hensyn til når man estimerer reliabilitet. En følge av at feilene oppfører seg tilfeldig, er at de er uavhengig av de sanne skårene og dermed vil feilvariansen være av noenlunde samme størrelsesorden uansett om gruppen har stor eller liten spredning av sanne skårer. Når man finner ulik spredning i to grupper, vil det derfor være de sanne skårene som er ansvarlig for ulikheten i total spredning. Følgelig vil en gruppe med stor spredning, bety stor spredning av sanne skårer. Når de sanne skårene så inngår i formelen for estimering av reliabilitet, vil spredning av sanne skårer utgjøre en relativt sett større del av de observerte skårene i gruppen med stor spredning. Derfor vil man som regel finne høyere koeffisienter i en heterogen gruppe enn man vil finne i en homogen gruppe (Crocker & Algina 1986). Dette betyr at større spredning av informasjonsskårer i gruppen vurdert av

vurdererpar BC, kan være blant forklaringen på at det er høyere reliabilitetskoeffisienter for BC enn det er funnet for AC.

Dersom man ser på setningslengdeskårerne, ser man imidlertid ikke den samme forskjellen i standardavvik som på informasjonsskårene. Standardavvikene er rimelig like, men litt høyere for AC enn BC. At standardavvikene for setningslengdeskårene er så like, kan synes uventet i og med at det er på setningslengdeskårene man finner den største forskjellen mellom reliabilitetskoeffisienter for de to vurdererparene. Det kan tolkes som at det må være andre årsaker i tillegg til spredning som forklarer reliabilitetsforskjellen mellom vurdererparene.

Dersom man ser på mean i tabell 12, ser man en tydelig forskjell i mean mellom skårene gitt av vurdererpar AC og BC. For både informasjonsskårene og setningslengdeskårene er mean for BC omtrent dobbelt så høy som for AC. Dette viser at førstnevnte barnegruppe har oppnådd betydelig høyere skårer enn den andre barnegruppen. Dette kan forklares som at barna vurdert av BC simpelthen var flinkere enn barna vurdert av AC. Samtidig synes det lite trolig at så store forskjeller kan forklares bare ut fra ulikt ferdighetsnivå hos barna. En annen faktor som kan ha påvirket mean i de to gruppene, er testadministrasjon. Dersom vurderer/testleder A og vurderer/testleder B har administrert testen på ulik måte, kan dette ha ført til at barna testet av den ene assistenten har fått høyere skårer enn barna testet av den andre assistenten. Det er grunn til å tro at Bus Story skårene er sensible i forhold til testadministrasjon i og med at testleder har en så sentral rolle i administreringen. Testledere som lykkes i å få barna til å kjenne seg trygge, og forteller historien på en tilpasset og engasjerende måte, vil sannsynligvis få barna til å fortelle mer enn en testleder som ikke lykkes på disse punktene. Videre kan testlederne i ulik grad prompte barnets gjenfortelling, og en testleder som prompter mye og bruker god tid ved hvert bilde vil kanskje få barnet til å fortelle mer enn en testleder som rask blar til neste side dersom barnet ikke sier noe umiddelbart.

Eksemplene her skal ikke forstås som kritikk av forskningsassistentene. Det kan være like galt å gi for mye hjelp som for lite hjelp. Poenget er at reglene for testsituasjon

muligens ikke er tydelige nok, og at inkonsistent testadministrasjon kan se ut til å ha påvirket mean i de to barnegruppene. Dersom dette er tilfelle, kan det tenkes at det ikke nødvendigvis trengs tydeligere retningslinjer for vurdering i og med at vurdererreliabiliteten er rimelig høy, men derimot at utbedrede retningslinjer for testadministrasjon kunne økt skårenes totale reliabilitet.

5.4 Validitet

Reliabilitetsundersøkelsen har vist at Bus Story skårene i språkprosjektet har høy vurdererreliabilitet. Det har blitt drøftet hva som kan forklare at koeffisientene er blitt så høye, og hvilke forhold som kan ha påvirket reliabilitetskoeffisientene. For å drøfte hvilken betydning høy vurdererreliabilitet har, kan det være nyttig å sette reliabilitet i en større kontekst.

Gjennom drøfting av G teori har ønsket om å generalisere fra testresultat til barnets universskåre (resultat uten påvirkning fra tilfeldige feil) blitt fokusert. Dersom man fører ønsket om generalisering ett steg videre, kan man si at man ønsker å generalisere fra testresultat til barnets narrative språkkompetanse. Det er jo nettopp ønsket om slike generaliseringer som ligger bak bruk av narrative tester som Bus Story. Dersom man kunne gjøre slike generaliseringer med høy grad av sikkerhet, ville det ikke bare bety at testresultatet var reliabelt, men også valid. Validitetsspørsmål betegnes gjerne som et spørsmål om resultatenes gyldighet. Mer presist forklarer Lund (2002b) validitet som gyldigheten til de *slutninger* vi trekker på grunnlag av resultatene. Hvor gyldige slutninger vi kan trekke, påvirkes av både tilfeldige og systematiske målefeil. Dette viser Kleven (1995) ved hjelp av følgende ligning: observert skåre = valid skåre + systematiske feil + tilfeldige feil. Slik tydeliggjør han at både systematiske og tilfeldige målefeil har anledning til å redusere validiteten.

Å undersøke Bus Story skårenes validitet, har ligget utenfor oppgavens fokus. Spørsmålet om validitet bringes likevel opp for å understreke at reliabilitet kun er en

del av det store bildet. Dersom man ikke har lykkes i å måle det man ønsker å måle eller det man tror man har målt, hjelper det ikke om resultatene i liten grad er påvirket av tilfeldige målefeil. Fra Klevens (1995) ligning fremgår det at god reliabilitet ikke nødvendigvis betyr god validitet. Lik vurdering er med andre ord ingen garanti for god eller ”riktig” vurdering.

Som poengtert gjennom oppgaven, viser også vurdererreliabilitet kun en del av det totale reliabilitetsbildet. Brennan (2001b) understreker dette ved å si at vurdererreliabilitet er et godt mål på i hvilken grad man har lykkes i å utvikle gode vurderere, men at dette også er den eneste informasjonen man får fra koeffisienten. Vurdererreliabilitet er altså kun en del av reliabilitet, og reliabilitet er igjen kun en del av validitet. Slik er det en rekke både tilfeldige og systematiske målefeil som kan ha redusert skårenes reliabilitet og validitet selv om vurdererreliabiliteten er høy.

Den høye vurdererreliabiliteten funnet i språkprosjektet, er altså ikke noen garanti for god validitet. Samtidig kommer det frem av Klevens (1995) ligning at lav reliabilitet vil begrense validitet. Når tidligere undersøkelser har funnet at Bus Story har god prediktiv validitet (Pankratz m.fl. 2007, Bishop & Edmundson 1987, Feagans & Appelbaum 1986), sier dette samtidig noe om reliabiliteten. Undersøkelser har vist at Bus Story-skårer oppnådd på et tidlig tidspunkt korrelerer høyt med skårer fra tester av språkferdigheter og andre skolefaglige ferdigheter på et senere tidspunkt. For at denne korrelasjonen skal være høy, er man avhengig av at skårene har hatt rimelig høy reliabilitet. Det vil si at Bus Story skårenes gode prediktive validitet, kan ses som et tegn på at Bus Story skårenes reliabilitet totalt sett må være rimelig god. Det er altså liten grunn til å tro at feilkildene som ikke er undersøkt her, er særlig alvorlige. Samtidig må det understrekes at det alltid er noe usikkerhet knyttet til slike overføringer av resultater.

6. Avslutning

Gjennom oppgaven har det blitt drøftet ulike målefeil som kan tenkes å påvirke testresultater. Målefeilene drøftet her, vil ikke bare være aktuelle i språkprosjektet, men også i en rekke tilfeller hvor Bus Story og andre tester brukes. I spesialpedagogers arbeid kan testing være en stor del av hverdagen. Tester brukes både i sakkyndighetsarbeid, i utredning og diagnostisering, og i arbeid med barn og unge. Det tas med andre ord viktige avgjørelser på grunnlag av testresultater. Dersom testresultatene er påvirket av målefeil, vil både reliabiliteten og validiteten være redusert, og informasjonen de gir oss kan være lite pålitelig og i verste fall misvisende. Dermed bør resultatenes reliabilitet studeres ved all testing for å undersøke om de gir pålitelig informasjon.

Det finnes testbrukere som tolker testresultater som en sannhet (Brown 1983). Tatt i betraktning aktuelle målefeil, vil dette være svært risikable slutninger som kan være mer til skade enn til hjelp. På den andre siden finnes det personer som hevder at man ikke bør bruke tester fordi man ikke kan stole på informasjonen de gir (Brown 1983). Ved å studere reliabilitet, kan man imidlertid få informasjon om resultatene er pålitelige eller ikke. Da kan man benytte seg av informasjon fremkommet ved testing der resultatene kan anses som pålitelige, og man kan unngå risikable slutninger på grunnlag av resultater som ikke er pålitelige. I tillegg kan man få informasjon om hvilke betingelser som bør legges for testingen slik at man i fremtiden kan få mest mulig pålitelige testresultater. Jeg håper at oppgaven kan fungere som en påminner til lesere og testbrukere om at man ikke kan ta for gitt at testresultater kun er et resultat av de forhold man ønsker eller tror man har fanget opp ved testing. Dette fikk jeg selv erfare da jeg viste en rekke bilder til en liten jente som hun skulle benevne. Som logoped fulgte jeg godt med på alle språklyder for å identifisere hvilke lyder som var på plass. Jeg hørte raskt at r-lyden ikke var ferdig utviklet. Etter å ha benevnt alle bildene reiste jenta seg, og på vei ut sa hun med flotte r-lyder ”hørte du at jeg lekte at jeg var fra Amerika da jeg lagde sånne rare r-er?”

Kildeliste

Befring, E 2007, *Forskningsmetode med etikk og statistikk*, Det norske samlaget, Oslo.

Bishop, D V M & Edmundson, A 1987, 'Language-impaired 4-year-olds: distinguishing transient from persistent impairment', *Journal of speech and hearing disorders*, vol. 52, p. 156-173.

Botting, N 2002, 'Narrative as a tool for the assessment of linguistic and pragmatic impairments', *Child language teaching and therapy*, vol. 18, no. 1, p. 1-21.

Brennan, R L 2000, 'Performance assessments from the perspective of Generalizability theory', *Applied psychological measurement*, vol. 24, no. 4, p. 339-353.

Brennan, R L 2001a, 'Some problems, pitfalls and paradoxes in educational measurement', *Educational measurement*, vol. 20, p. 6-18.

Brennan, R L 2001b, 'An essay on the history and future of reliability from the perspective of replications', *Educational measurement*, vol. 38, no. 4, p. 295-318.

Brown, J D 1996, *Testing in language programs*, Prentice Hall Regents, Upper Saddle River, New Jersey.

Brown, F G 1983, *Principles of educational and psychological testing*, 3rd ed, Holt, Rinehart & Winston, New York.

Child Language & Learning prosjektsøknad 2007, *The nature and development of language and communication skills in pre-school children*. Søknad om midler til frittstående prosjekter – samfunnsvitenskap (FRISAM), Norges forskningsråd.

Clark-Carter, D 2004, *Quantitative psychological research: A student`s handbook*, Psychology Press, Hove.

Crocker, L & Algina, J 1986, *Introduction to classical & modern test theory*, Holt, Rinehart & Winston, Fort Worth.

Cronbach, L J, Rajaratnam, N & Gleser, G C 1963, 'Theory of generalizability: A liberalization of reliability theory', *The british journal of statistical psychology*, Vol. XVI, part 2, p. 137-162.

Dunbar, S B, Koretz, D M & Hoover, H D 1991, 'Quality control in the development and use of performance assessments', *Applied measurement in education*, vol. 4, p. 289-303.

Fan, X & Chen, M 2000, 'Published studies of interrater reliability often overestimate reliability: computing the correct coefficient', *Educational and psychological measurement*, vol. 60, no. 4, p. 532-542.

Feagans, L & Appelbaum, M I 1986, 'Validation of language subtypes in learning-disabled children', Journal of educational psychology, vol. 78, no. 5, p. 358-364.

Fey, ME, Catts, HW, Proctor-Williams, K, Tomblin, JB & Zhang, X 2004, 'Oral and written story composition skills of children with language impairment', *Journal of speech, language and hearing research*, vol. 47, no. 6, p. 1301-1318.

Hogan, T, Benjamin, A & Brezinski, K L 2000, 'Reliability methods: A note on the frequency of use of various types', *Educational and psychological measurement*, vol. 60, p. 523-531.

Howlin, P & Kendall, L 1991, 'Assessing children with language tests – which tests to use?' *British journal of disorders of communication*, vol. 26, p. 355-367.

Jaeger R M 1991, 'Foreword', i Shavelson R J & Webb N M, *Generalizability theory: A primer*, Sage Publications, Newbury Park.

Kleven, T A 1995, 'Reliabilitet som pedagogisk problem', *Rapport ved Universitetet i Oslo, Pedagogisk forskningsinstitutt*, no. 9, Universitetet i Oslo.

Kleven, T A 2002a, ’Hvordan er begrepene operasjonalisert? Spørsmålet om begrepsvaliditet’, i *Innføring i pedagogisk forskningsmetode. En hjelp til kritisk tolkning og vudering*, Unipub forlag, Oslo.

Kleven, T A 2002b, ’Statistikk’, i *Innføring i pedagogisk forskningsmetode. En hjelp til kritisk tolkning og vudering*, Unipub forlag, Oslo.

Kleven, T A 2002c, ’Hvilken kontekst er resultatene gyldige i? Spørsmålet om ytre validitet’, i *Innføring i pedagogisk forskningsmetode. En hjelp til kritisk tolkning og vudering*, Unipub forlag, Oslo.

Kline, T J B 2005, *Psychological testing: A practical approach to design and evaluation*, Sage Publications, Thousand Oaks.

Lund, T 2002a, ’Generaliseringsproblematikk’, i *Innføring i forskningsmetodologi*, Unipub forlag, Oslo.

Lund, T 2002b, ’Innledning’, i *Innføring i forskningsmetodologi*, Unipub forlag, Oslo.

Lund, T & Christophersen, K-A 1999, *Innføring i statistikk*, Universitetsforlaget, Oslo.

NESH: Forskningsetiske retningslinjer for samfunnsvitenskap, humanoria, juss og teologi 2006, *De nasjonale forskningsetiske komiteer*, lesedato 10. april 2008, http://www.etikkom.no/retningslinjer/NESHretningslinjer/06.

Pankratz, ME, Plante, E, Vance, R & Insalaco, DM 2007, ‘The diagnostic and predictive validity of the Renfrew Bus Story’, *Language, speech and hearing services in schools*, vol. 38, p. 390-399.

Paul, R & Smith, RL 1993, ‘Narrative skills in 4-year-oldswith normal, impaired and late-developing language’, *Journal of speech and hearing research*, vol. 36, no. 3, p. 572-598.

Rand, G 1971, *Elementær testteori*, Universitetsforlaget, Oslo.

Renfrew, C 1997, *Bus Story Test*, 4th ed, Winslow Press Limited, Bicester.

Shavelson, R J & Webb, N M 1991, *Generalizability theory: A primer*, Sage Publications, Newbury Park.

Thompson, B 2003, 'A brief introduction to generalizability theory', i *Score reliability: Contemporary thinking on reliability issues*, Sage Publications, Thousand Oaks.

Thompson, B & Snyder, P A 1998, 'Statistical significance and reliability analyses in recent journal of counselling & development research articles', *Journal of counselling & development*, Vol. 76, no. 4, p. 436-441.

Thompson, B & Vacha-Haase, T 2000, 'Psychomertics is datametrics: The test is not reliable', *Educational and psychological measurement*, vol. 60, p. 174-195.

University of Iowa 2008, *Center for advanced studies in meausrement and assessment*, lesedato 11. mars 2008, http://www.education.uiowa.edu/casma/GenovaPrograms.htm.

Vacha-Haase, T, Ness, C, Nilsson, J & Reetz D 1999, 'Practices regarding reporting of reliability coefficients: a review of three journals', *The journal of experimental education*, vol. 67, no. 4, p. 335-341.

Whittington, D 2003, 'How well do researchers report their measures?' i Thompson, B (red.) *Score reliability: Contemporary thinking on reliability issues*, Sage Publications, Thousand Oaks.

Willson, V L 1980, 'Researcher techniques in AERA articles: 1969 to 1978', *Educational researcher*, vol. 9, p. 5-10.

Informasjon; SKÅRINGSGUIDE	**Info SKÅRE**	**TRANSKRIPSJON; BUSSFORTELLINGEN**	**SETN.-LENGDE**
Uthevet tekst = 2 p (1 poeng om halv respons) Ikke uthevet tekst = 1 p			
Buss rampete Sjåfør reparere/fikse Buss **kjørte/dro sin vei/rømme** **møtte/kjørte smn med tog** laget grimaser kjørte om kapp Tog **i tunnel** Buss **alene** **inn til byen/gata** **møtte/så politimann** Politimann blåste fløyte **Sa stopp** Buss lot som ikke hørte **kjørte videre/stoppet ikke** **ut på landet** **trøtt/lei av veien** **hoppet over gjerde/port** **møtte/så ku** Ku mø trodde ikke hva den så			
Buss **kjørte nedover** **så vann** **forsøkte/prøvde å stoppe** **visste ikke hvordan bremse/stoppe** **falt i vannet** Plask satt fast i søla Sjåfør **fant bussen/den** **ringte etter** **heisekran/kranbil** Heisekran **løftet ut/dro opp** Buss tilbake på veien			
	Info SKÅRE		**TOTAL**
			A5LS

KOMMENTARER;
Testsituasjon;

Artikulasjon;

Syntaks;