Lennart Jølle

Vurderingsdialogen

En undersøkelse av tekstvurderingspraksis ved nasjonal læringsstøttende prøve i skriving

Avhandling for graden philosophiae doctor

Trondheim, juni 2015

Norges teknisk-naturvitenskapelige universitet
Det humanistiske fakultet
Institutt for språk og litteratur

NTNU
Kunnskap for en bedre verden

NTNU
Norges teknisk-naturvitenskapelige universitet

Avhandling for graden philosophiae doctor

Det humanistiske fakultet
Institutt for språk og litteratur

ISBN 978-82-326-0972-7 (trykt utg.)
ISBN 978-82-326-0973-4 (elektr. utg.)
ISSN 1503-8181

Doktoravhandlinger ved NTNU, 2015:158

Trykket av NTNU Grafisk senter

Sammendrag

I forberedelsene til det som skulle bli nasjonal utvalgsprøve i skriving, og som nå har status som nasjonal læringsstøttende prøve i skriving, ble det i 2010 etablert et semipermanent nasjonalt vurdererpanel som skal stå for vurderingene av elevtekstene. Bakgrunnen for etableringen av dette panelet var erfaringer fra 2005-forsøket på å innføre nasjonale prøver i skriving, hvor vurderingsarbeidet viste seg å være av for dårlig kvalitet (Fasting, Thygesen, Berge, Evensen & Vagle, 2009). Med dette utgangspunktet, har det vært interessant å følge lærerne i Vurdererpanelet og deres ferd fra å være tekst- og skriveinteresserte lærere, til å bli normsettende skrivevurderere på nasjonalt nivå. Det overordna målet i denne studien har således vært å undersøke hvordan uerfarne vurderere utvikler kompetanse i elevtekstvurdering.

Materialet er transkripsjoner av vurderingsdialoger samlet inn over en ettårsperiode (tre vurderingssamlinger, 2010-2011). Det er gjort opptak av de samme fem vurderere på hver samling, men siden disse har inngått i par- og gruppekonstellasjoner som stadig ble endret for å unngå utvikling av lokale vurderingskulturer, inneholder materialet dialogbidrag fra totalt 33 vurderere. 26 vurderingsdialoger ble transkribert, og hver transkripsjon inneholder 1-3 tekstvurderinger. Fokus har vært på vurdereres *vurderingspraksiser* slik de har kommet til uttrykk i par- og gruppevurderingene. Et slikt fokus på «worlding», og ikke «world» (Pollner, 1987), er metodisk forankret i etnometodologien.

Hovedspørsmålet har blitt forsøkt svart gjennom tre delstudier med tilhørende forskningsspørsmål. I første delstudie ser vi hvordan vurderernes innarbeidete vurderingspraksis forut for deltakelse i Vurdererpanelet ikke lett lar seg endre selv om vurdererne tar del i et læringsmiljø hvor endring av vurderingspraksis er en uttalt målsetting. Samtidig viser analyser av parvurderingene at når vurderingsdialogene fungerer som «utforskende samtaler» (Wegerif & Mercer, 1997), øker sannsynligheten for at vurderingene blir konsistente gjennom en form for triangulering. I andre delstudie blir det tydelig hvordan én vurderer pendler mellom intermentale (mikrososiale) og intramentale (individuelle) (Vygotsky, 1978) strategier i møte med elevtekster og ulike vurderingsressurser. Utvikling spores gjennom vurdererens internalisering av fagbegreper gjennom en målrettet remedieringspraksis (Prior & Hengst, 2010). I tredje delstudie ser vi hvordan vurdererne bruker vurderingsskjemaet til å stykke opp vurderingsarbeidet i mindre «kommunikative

prosjekt» (Linell, 2009). Strategiene for å komme til enighet om tekstkvalitet innenfor hvert prosjekt er flere og de er lite konsistente. Det er også klart at den enkelte vurderers sosiale posisjon i Vurdererpanelet i seg selv fungerer som et vektig beslutningsgrunnlag.

De ulike studiene gir anledning til å gå i dialog med annen (internasjonal) forskningslitteratur som omhandler sentralt administrerte skriveprøver. Denne litteraturen fremstiller gjerne vurdererne som trente iverksettere av en gjennomtenkt instruks, skrivevurdererne er «utførere». Det virker å være en sammenheng mellom et slikt syn på vurdererrollen og fremveksten av AES, automated essay scoring (*Assessing Writing, 18*, 2013, (temautgave)). Den foreliggende studien utfordrer en slik forståelsesramme og praksis gjennom å vise fram kompleksiteten i vurderingsarbeidet, også innenfor et praksisfellesskap der normer for ferdighetsnivå er under utvikling.

Abstract

In 2005, an attempt was made to carry out a Norwegian national writing test. The pupils' texts were assessed by local teachers. However, due to low inter-rater reliability further tests were postponed. Several explanations for the low reliability have been posited, such as the level of expertise among the raters and the lack of a shared assessment culture (Fasting, Thygesen, Berge, Evensen & Vagle, 2009). To meet these challenges, an assessment panel was established in 2010 to assess the pupils' writings for the new Learning Supportive Tests in Writing as a Basic Skill (which was intentionally to be a Norwegian Sample-based Assessment of L1 Writing as a Basic Skill). On this basis, it has been of interest to follow the assessment panel teachers in their journey from being fascinated with, and interested in, pupil writing, to becoming skilled as national standard setting raters of pupil writing. In this thesis, the aim has therefore been to investigate how novice raters develop competence within assessment of pupils' writing.

The study was conducted over three successive National Assessment Panel meetings, in June 2011, November 2011 and April 2012. Five raters where followed during this period, but since the raters continuously changed assessment partners (to stem for development of idiosyncratic assessment practices), the data consists of contributions from thirty-three raters. Twenty-six assessment dialogues were tape-recorded, each recording documents the assessment of one to three Year 8 texts. The focus has been on raters' assessment practices where such a focus on «worlding», and not «world» (Pollner, 1987), is methodologically anchored within ethnomethodology. This thesis contains three research papers, each addressing a separate research question related to the overall aim of investigating raters' writing assessment practice and it's development.

The first paper studies how the raters' incorporate new assessment procedures, made available through their participation in the National Assessment Panel, with their former assessment practice having been established through years as writing teachers. The National Assessment Panel is established to make the raters change towards shared assessment practice, but such changes are small in the overall data. However, analyses suggest that when the assessment dialogues work as "explorative talk" (Wegerif & Mercer, 1997), this is indicative of more consistent assessments due to a form of triangulation. The second paper studies how one rater changes between "intermental" and "intramental" (Vygotsky, 1978)

strategies when confronted with pupils' texts and the different assessment resources. Rater development is possible to trace through the rater's internalization of concepts through a targeted "remediation practice" (Prior & Hengst, 2010). The last paper explores how the raters use the rubric to organize the assessments within smaller "communicative projects" (Linell, 2009). The strategies used to reach agreement are several and inconsistent. Analysis also discloses that the individual rater's social position within the National Assessment Panel is important when raters are to reach agreement about text quality.

Overall, the three different studies provide an opportunity for dialogue with related international literature. This literature tends to present the raters as trained doers who are to put into effect well planned instructions. It seems to be a relation between such a view on a rater's role within writing assessment and the growth of AES, automated essay scoring (cf. *Assessing Writing, 18*, 2013). Hence, the present thesis challenges this conception and practice by offering a chink of light into the raters' complex decision-making processes.

Forord

Det er en tid for alt. Når dette avhandlingsarbeidet nå avsluttes, og jeg ser tilbake på de årene jeg har vært så heldig å få fordype meg i ett datamateriale med én bestemt problemstilling over tid, finner jeg mange å takke. Jeg vil begynne med de opplagte; de mange lærerne som til sammen utgjør det nasjonale Vurdererpanelet som står for vurderingene av læringsstøttende prøve i skriving som grunnleggende ferdighet. Da jeg møtte dem første gang, var de ikke bare vurderingsnoviser, de var også usikre på egne roller og oppdrag. Likevel gav de meg plass til å observere og gjøre datainnsamling. En slik raushet er ikke selvsagt, og jeg er de dypt takknemlig.

Videre kommer jeg ikke utenom veilederne mine, Lars Sigfred Evensen og Synnøve Matre. De har, på hver sine særegne måter, bidratt til at forskningsarbeidet har blitt drevet fremover. Lars med det kloke usagte (de som kjenner Lars skjønner hva jeg mener), og Synnøve med det kloke sagte (de som kjenner Synnøve skjønner hva jeg mener). Uten dem hadde jeg nok fremdeles stått som «Moses i ørkenen» og lurt på hvilken vei jeg skulle ta.

Studieåret 2012/2013 tilbrakte jeg ved The University of Melbourne, som «Visiting scholar». Store deler av analysearbeidet ble gjort i løpet av dette oppholdet, under veiledning av John Hattie. Jeg har sagt det tidligere, og nå slår jeg det fast; oppholdet gjorde en forskjell, også fordi jeg da ble kjent med Judy Parr som arbeider ved The University of Auckland. Hun fortjener en hjertens takk for grundige lesninger av to av mine «work-in-progress».

Jeg har også en rekke medstipendiater å takke. Først og fremst en stor takk til skrivegruppa mi; til Cherise Kristoffersen, Ingeborg Sæbøe Holten, Tove Lafton, Tone Fjogstad Langnes og Børge Skåland. Vi har lest og respondert på hverandres utallige utkast de siste årene. Til alle døgnets tider har vi stått på! Takk også til alle «kullingene» mine ved Nasjonal forskerskole for lærerutdanning, NAFOL. Jeg er overbevist om at akkurat vi, ved å dele, kritisere, utfordre og glede, skapte et unikt læringsmiljø i de fire årene vi var en del av NAFOL.

Selv om drivstoffet i denne prosessen har vært faglig engasjement og nysgjerrighet, har jeg mine verdslige behov. En stor takk rettes derfor til Høgskolen i Sør-Trøndelag, Avdeling for lærer- og tolkeutdanning for lønnsmidler og arbeidsplass. Takk også for generell raushet underveis. Takken går også til Skrivesenteret som også generøst har bidratt. På samme måte må NAFOL igjen nevnes, som ved å tilby velproporsjonerte stipendordninger, gjorde utenlandsoppholdet mulig.

Til slutt; avhandlinga har vært en del av arbeidshverdagen min lenge. To av ungene mine, Selma og Emil har blitt fire år eldre i løpet av denne perioden, og minstemann, Mikkel, har kommet til. Mye har vært sagt om det å fullføre et avhandlingsarbeid, men det skal dere vite, kjære unger, at neppe vil dere få en pappa med en så fleksibel arbeidssituasjon igjen! Det samme går for deg, Anne! Og jeg kjenner at jeg savner det allerede.

Innhold

Tabeller og figurer

Tabeller

Figurer

1 Innledning

1.1 Prolog

I 2005 arbeidet jeg som lektor i den videregående skolen i Sør-Trøndelag. Forut for innføringen av Kunnskapsløftet (www.udir.no/lareplaner/kunnskapsloftet/), ble vi lærere dette året sendt på diverse kurs for å forstå rekkeviddene av den nye læreplanen. Blant norsklærerne var det særlig ei endring som plaget, og det var det nye hovedområdet «sammensatte tekster».[1] Hva var nå det for noe? For å bøte på denne frustrasjonen, ble dette temaet for en av kursdagene. Og fra denne samlingen sitter jeg igjen med et klart minne. En kollega (ja, det er alltid en kollega!) ville diskutere vurderingsutfordringene knyttet til sammensatte tekster. I korthet gikk bekymringen hans ut på følgende: Han kunne karaktersette en tradisjonell skriftlig tekst etter én gjennomlesning. Denne evnen, som han omtalte som en ryggmargsfølelse, var totalt fraværende når det gjaldt sammensatte tekster. Hvordan skulle han da klare å karaktersette slike prestasjoner?

Opplevelsen ble en vekker for meg som norsklærer. Det min kollega brakte til torgs var sider ved vurderingsarbeidet som jeg opplevde var lite diskutert. Med begrepet «ryggmargsfølelse» kjente jeg meg igjen i de situasjoner hvor jeg satt med bunker av elevtekster som jeg relativt enkelt vurderte gjennom en holistisk vurderingspraksis. Plutselig stod det veldig tydelig for meg at vurderingskriteriene våre stort sett var *tause*. Og fordi vi manglet et språk om det kvalitative arbeidet som vurdering av tekster er, ble vurderingsarbeidet *privatisert*. Det er selvsagt flere aspekter ved dette som er svært problematisk. Jeg nøyer meg her med å peke på to. For det første fører tause og individuelle vurderingspraksiser til at det ikke etableres felles normer for hva som forventes av elevenes tekster. Det ligger med andre ord en iboende urettferdighet i en slik praksis. For det andre, og viktigere etter min oppfatning, er det at min kollegas følelse av å være i villrede ved vurdering av sammensatte tekster, er parallell til elevenes følelse av å være i villrede ved skriving av tekster. På samme måte som min tidligere kollega ikke hadde utviklet en ryggmargsfølelse for hvordan en god sammensatt tekst ser ut, har elevene ennå ikke utviklet noen ryggmargfølelse

[1] Etter hva jeg husker tok vi ikke på alvor innføringen av de grunnleggende ferdighetene. Som norsklektorer involverte vi oss ikke i diskusjoner om hvordan for eksempel skriving skulle forstås og vurderes i alle fag, men hadde nok med oss selv og det nye i læreplanen for norskfaget. Det kan også legges til at hovedområdet «sammensatte tekster» ble fjernet ved revisjonsarbeidet i 2013.

for hva som er kvalitet i tradisjonelle elevtekster. I en opplæringskontekst er det opplagt at det elevene trenger er tilbakemelding om hva som vil føre de videre som skrivere. Lærerne trenger med andre ord et vokabular for å snakke om tekster sammen med elever som virker formativt i elevenes videre arbeid. Jeg velger å begynne her fordi hendelsen inngår i min fortelling om hvordan dette avhandlingsarbeidet kom i gang. Samtidig må det understrekes at selv om min Damaskus-opplevelse var både sterk og inderlig, var den neppe i seg selv verdt et doktorgradsstipend.

Denne avhandlingen handler om vurdering av skriving. Nærmere bestemt handler den om hvordan en gruppe lærere plassert i et fellesskap vurderer elevtekster og hvordan de utvikler skrivevurderingskompetanse i dette fellesskapets etableringsfase. Fellesskapet er et nasjonalt vurdererpanel bestående av om lag 100 lærere som ble etablert i 2010 som et sentralt ledd i forberedelsene til innføringen av «læringsstøttende prøver» (udir.no/Vurdering/Laringsstottende-prover/) i skriving fra høsten 2014. Lærerne i dette panelet har siden oppstart, sammen med faggruppa[2] som har hatt ansvaret for utviklingen av prøven, arbeidet med å skape en felles forståelse av hva skriving er, hva som ligger i begrepet «skrivekompetanse», hva man kan forvente av elever på ulike trinn (etter fire og syv års opplæring), de har utviklet oppgaver og de har ikke minst vurdert elevtekster i ulike piloteringsrunder. Det overordna formålet med dette arbeidet har vært å skape et miljø med høy tekstkompetanse og felles forventningsnormer med det mål for øye å oppnå valide og reliable vurderinger av elevtekster.

Med begrepet «læringsstøttende» signaliserer Utdanningsdirektoratet at det er snakk om læringsfremmende prøver, altså prøver som skal gi lærer og den enkelte elev nyttig informasjon om veien videre i læringsprosessen. Prøvene er imidlertid også såkalte «utvalgsprøver» i det et nasjonalt representativt utvalg elever på de bestemte årstrinn gjennomfører skriveprøvene, og hvor det nasjonale Vurdererpanelet står for vurderingene av disse. Formålet er å skape og å opprettholde forventningsnormer som lokale lærerkollegier kan bruke som grunnlag for å skape tolkningsfellesskap. Vurdererpanelets posisjon som premissleverandør for skrivevurdering, og dermed også skriveopplæring, er dermed gitt (se for eksempel Hamp-Lyon (2000) og Messick (1996) om testers såkalte «washback»-effekt på

[2] Under datainnsamling til dette avhandlingsarbeidet ble faggruppa ledet av professor Lars Sigfred Evensen (fram til 2013). I tillegg bestod gruppa av professor Kjell Lars Berge, høgskolelektor Trine Gedde-Dahl, dosent Rolf Fasting og professor Ragnar Thygesen.

instruksjoner). Dette har også vært en viktig faktor for min interesse i vurdererpanelets arbeid; vurderingspraksisen som der blir etablert, er tenkt etablert i skolen som sådan. Og den må vi få økt kunnskap om.

1.2 Rasjonale

I norsk skolekontekst fikk skriving en annen rolle etter LK06. Som én av fem grunnleggende ferdigheter er ikke skriving lenger bare et norskfaglig emne, men en kompetanse som går på tvers av fag og som angår alle elever og lærere. På den måten er skriving en sentral del av literacy-tenkningen som ligger til grunn for Kunnskapsløftet; flaskehalsen for det gode liv er ikke lenger jordeiendom eller kapital, men evnen til å beherske tegnsystemer! Parallelt med oppvurderingen av skriving, vokste det fram et ønske om å skaffe til veie informasjon og kunnskap om skrivekompetansen slik den er i norsk skole, både som utgangspunkt for videre opplæring og som en måte å ansvarliggjøre skoler.[3] Det ble derfor bestemt at det skulle avholdes nasjonal prøve i skriving fra 2005. Vurderingene av elevtekstene fra prøven viste seg imidlertid å bli så upålitelige at resultatene ikke kunne formidles tilbake til elever og lærere, og videre skriveprøver ble lagt på is. Spørsmålet om reliabilitet var imidlertid dermed reist, og da det i 2009 ble bestemt at skriveprøven skulle gjeninnføres, nå som en utvalgsprøve, ble prøvens pålitelighetsproblematikk et sentralt spørsmål som det måtte finnes et adekvat svar på. Spørsmålet ble med andre ord; hvordan skaper man en valid skriveprøve der vurderere evner å foreta reliable tekstvurderinger?

Denne avhandlingen tar utgangspunkt i noen av de utfordringer skriveprøvene har møtt på i denne sammenhengen, og på den måten balanserer avhandlingen mellom to forskningstradisjoner, skriveteori på den ene siden og testteori på den andre. Skriveforskningen har etter hvert lange tradisjoner i Norge, men det samme er ikke tilfelle med skrivetestforskningen. Denne systematiske skjevheten i innsats skyldes kanskje at i Norge har det vært lærerne selv som står for vurderingene av elevenes skrivekompetanse gjennom hele den obligatoriske opplæringen, med unntak av den femtedelen av 10-årstrinnskullet som blir trukket til eksamen i norsk skriftlig hvor ekstern sensor står for vurderingene. Det som har blitt oppfatta som av interesse når det gjelder skriving skjer i undervisningsarealet, og ikke under sensurmøtet! (Men se KAL-prosjektet (Berge, Evensen, Hertzberg & Vagle, 2005)).

[3] Et ledd i dette arbeidet var opprettelsen av det nasjonale skrivesenteret i 2009 (http://skrivesenteret.no).

Arbeidet inneholder tre studier hvor forholdet mellom skrivevurdering som sosial fortolkningspraksis og skriveprøver som testkonstrukt blir diskutert og problematisert på ulike måter. Alle studiene bygger på datamateriale hvor vurderere, som alle er medlemmer av det nasjonale Vurdererpanelet, vurderer elevtekster parvis eller i små grupper. Vurdererpanelet står for vurderingene av skriveprøven, og etableringen av panelet er et svar på reliabilitetsutfordringen: Norske lærere mangler tekstkompetanse og felles vurderingskultur, og gjennom felles opplæring, trening og pilotering innenfor et praksisfelleskap er målsettingen å bøte på dette. I mitt avhandlingsarbeid har jeg vært opptatt av de parvise vurderingssamtalene vurdererne gjennomførte under et piloteringsarbeid i 2011/2012. Jeg har vært opptatt av hva vurdererne 'gjør med ord', det vil si hvordan vurdererne gjennom dialog kommer fram til beslutninger om tekstkvalitet. Siden vurdererne er plassert i en læringssituasjon, har jeg i tillegg vært opptatt av om, eventuelt hvordan, vurderingspraksisen deres endres over tid etter hvert som de blir mer kompetente.

1.3 Problemstilling og mål

Spørsmålet som har drevet avhandlingsarbeidet fremover gjennom tre delstudier, har vært:

Hvordan utvikler uerfarne vurderere kompetanse i elevtekstvurdering?

Et slikt spørsmål gjør krav om presiseringer. Hva som menes med 'å utvikle tekstvurderingskompetanse' har selvfølgelig avgjørende betydning både for hvordan man går fram og for hva man finner (av interesse). Gjennom avhandlinga er det derfor en målsetting at jeg gjennom egen posisjonering og gjennom drøftinger makter å meisle ut en bestemt forståelse av hva som kan ligge i det å utvikle tekstvurderingskompetanse som igjen kan bidra til å skape moment i videre skriveprøveutvikling. En drivkraft i arbeidet har vært å forstå mer av de avveininger som ligger til grunn for vurdereres bestemmelse av tekstkvalitet. Siden lærerne som inngår i Vurdererpanelet, og som dermed har status som nasjonale vurderere, får opplæring, har det også vært et mål å bedre forstå hvordan et slikt læringsfellesskap påvirker den faktiske vurderingspraksisen.

Delstudiene er empirisk baserte og etnometodologisk inspirerte. Det betyr at jeg har vært opptatt av hva vurdererne gjør når de vurderer elevtekster, og mindre opptatt av hva de sier at de gjør. Det betyr igjen at fokuset har vært på vurderernes beslutningsprosesser.

Samtidig har det vært et mål å forstå vurderingsarbeidet og vurdererutviklingen fra et faghistorisk og teoretisk perspektiv. Det vil bli hevdet at hvordan vi forstår vurderingskompetanse har store implikasjoner for utviklingen og gjennomføringen av vurderingsarbeid i en skriveprøvesammenheng.

1.4 Empiri og metode

Nasjonalt, og særlig internasjonalt, finnes det en rekke studier som bidrar til å *forklare* sammenhenger mellom ulike variabler som påvirker tekstvurdering. I litteraturgjennomgangen kan vi for eksempel se at ved holistiske vurderinger legger vurdererne mest vekt på tekstens innhold og organisering, mindre vekt på rettskriving, og minst vekt på språkbruk (se for eksempel Freedman, 1979a). Allerede i 1990 gjorde imidlertid Brian Huot det klart at forskningsfeltet trenger flere studier «on the raters themselves, the nature of the fluent reading process, and the process of reading according to specific guidelines, especially for the purpose of agreement» (Huot, 1990, s. 257). Vi skal senere se at Huots etterlysning av forskning på området delvis har fått gehør, og ikke minst håper jeg dette avhandlingsarbeidet bidrar til å bringe noe lys inn i det tussmørket som tekstvurderernes beslutningsprosesser må sies å være omsluttet av.

Hovedanliggende med denne avhandlinga er et ønske om å bedre *forstå* vurdererens beslutningsprosesser, ikke ut fra deres subjektive erfaringer, men ut fra deres *handlinger*. Fokus har vært på vurdereres *vurderingspraksiser* slik de har kommet til uttrykk i par- og gruppevurderingene. Dette har hatt konsekvenser for empiritilfang: Empirien har i hovedsak vært transkripsjoner av vurderingsdialoger. Et slikt fokus på «worlding», og ikke «world» (Pollner, 1987, s. 7), er metodisk forankret i etnometodologien. Delstudiene undersøker vurderernes ytringer; hvordan de (ikke) snakker om elevtekster og vurderingsressurser, hvordan de argumenterer, og hvordan de tar i bruk ulike strategier i vurderingspraksisen. Dette blir forsøkt forstått med støtte i et dialogisk syn på læring og meningsproduksjon.

1.5 Teoretisk tilnærming

Martin Buber var i første halvdel av forrige århundre sentral i utviklingen av det som fikk betegnelsen dialogfilosofi. For Buber karakteriseres ikke mennesket under substans, men *relasjon*. Det vil si at jeget er ikke-eksisterende uten et du, og således blir dialogen et *ontologisk* prosjekt (Buber, 2007). Med det markerer Buber avstand fra det individorienterte kunnskapssynet som har sine røtter hos Descartes og senere Kant, og som vi i dag blant annet

finner uttrykt i kognitiv læringsteori. Det er videre mulig å følge ei linje fra Buber til Mikhail Bakhtin, hvor sistnevntes dialogbegrep må sees på som et *epistemologisk* prosjekt. Bakhtin skriver at "[t]ruth is not to be found inside the head of an individual person, it is born between people collectively searching for thruth, in the process of their dialogical interaction" (1984, s. 110). Språkforskeren og kommunikasjonsteoretikeren Per Linell står på disses skuldre, men ønsker å gi dialogbegrepet en sterkere forankring i faktiske dialoger. For som han skriver: "Bakhtin was primarily a literary scholar (…), and was not in a position to make analyses of conversation" (Linell, 2001, s. 49).

I det foreliggende avhandlingsarbeidet støtter jeg meg på en slik empirinær dialogforståelse. Delstudiene tar på alvor vurderernes språklige ytringer i vurderingsarbeidet og holder disse for å være normerende tankeredskaper: Vurderingspraksisen er slik som den kommer til uttrykk i vurderingssamtalene.

Hvordan vurderere utvikler tekstvurderingskompetanse innenfor en testkontekst vet vi svært lite om. Ved å anvende en dialogteoretisk inngang til vurderingspraksisen er det mulig å studere denne nært i et utviklingsperspektiv: Mening skapes både i situasjoner og er samtidig situasjonsoverskridende; gjeldende praksiser peker både bakover, mot tidligere praksiser, og fremover, mot mulige nye praksiser. Ved å legge vekt på hvordan vurdererne vurderer i en opplæringsfase, er det mulig å undersøke hvilke ressurser vurdererne, med sin lærerprofesjonskompetanse, tar med seg inn til Vurdererpanelet; hvordan de argumenter og forhandler med hverandre, med doxa og med nye forventningsnormer og ny skriveforståelse; og hva de lar vinne fram som ny vurderingspraksis.

De analytiske begrepene som er sentrale i studiene, har gjort det mulig å undersøke ulike aspekter ved vurderernes arbeid: vurderernes «situerte interaksjon» (Linell, 2009, s. 98), altså parvurderingenes hva og hvordan (*referent* og *respons*); vurderernes utvikling av vurderingskompetanse (*semiotisk remediering* og *internalisering*); og vurderernes egenvaliderte vurderingstrategier (*kommunikative prosjekter* og *strategier*) (se for øvrig kapittel 5.4 for mer om analytiske (be)grep).

En dialogteoretisk tilnærming til meningsskaping er krevende fordi den søker å ivareta kompleksiteten i kunnskapsproduksjonen med en grunnforestilling om at mening initialt er et sosialt fenomen. Med forståelse av meningsskaping som noe som hele tiden skjer "imellom", vil det være avgjørende for meg å studere dette "mellomrommet". Og det er et arbeid som krever en *pragmatisk* (handlingsorientert) tilnærming.

1.6 Oversikt over delstudiene

	Artikkel 1	**Artikkel 2**	**Artikkel 3**
Tittel	Pair assessment of pupil writing: A dialogic approach to studying the development of rater competence	To become a more proficient rater: What does it take?	Rater strategies for reaching agreement on pupil text quality
Materiale	Transkripsjon av 26 par-/gruppevurderinger gjennomført i løpet av tre påfølgende Vurdererpanel-samlinger, i juni 2011, november 2011 og april 2012.	Transkripsjon av parvurderinger hvor én vurderer deltok i hver av de påfølgende Vurdererpanel-samlingene, i juni 2011, november 2011 og april 2012.	Transkripsjon av 26 par-/gruppevurderinger gjennomført i løpet av tre påfølgende Vurdererpanel-samlinger, i juni 2011, november 2011 og april 2012.
Forskningsspørsmål	Hvordan vurderer uerfarne vurderere innenfor Vurdererpanelet elevtekster, og hvordan endrer denne praksisen seg over tid mens vurdererne blir mer erfarne og kompetente? Hvilke potensielle effekter har endret vurderingspraksis på vurderingene?	Hvordan manifesterer vurderernes vurderingskompetanse seg i vurderingsdialogene over tid der vurdererne snakker seg fram til enighet om tekstkvalitet?	Hvilke kommunikative strategier er suksessfullt tatt i bruk for å nå enighet om tekstkvalitet, og hvordan skal disse forstås i vurderernes arbeid med å etablere en felles vurderingskultur?
Analytiske begreper	Referent. Responsmønster	Semiotisk remediering. Internalisering	Kommunikative prosjekt. Kommunikative strategier
Funn	Vurdererne som er medlemmer av Vurdererpanelet har med seg en vurderingspraksis som ikke lett lar seg endre selv om vurdererne tar del i et læringsmiljø hvor endring av vurderingspraksis er en uttalt målsetting. Samtidig viser analyser av parvurderingene at når de fungerer som	Vurdererne pendler mellom å ta i bruk intermentale (sosiale) og intramentale (individuelle) strategier i møte med elevtekster og ulike vurderingsressurser. Utvikling spores gjennom vurdererernes internalisering av fagbegreper gjennom en målrettet remedieringspraksis.	Vurderne bruker vurderingsskjemaet til å dele opp vurderings-arbeidet i mindre kommunikative prosjekter. Strategiene for å komme til enighet om tekstkvalitet innenfor hvert prosjekt er flere og tilfeldig anvendt. Det er også klart at den enkelte vurderers posisjon i Vurdererpanelet i seg

	«utforskende samtaler», øker sannsynligheten for at vurderingene blir konsistente gjennom en form for triangulering.	Analysene gir grunnlag for å reise spørsmålet om hvorvidt en mye brukt praksis hvor vurdererne får kort forberedelsestid, tar tilstrekkelig hensyn til skrivevurderingens kompleksitet.	selv fungerer som et vektig beslutningsgrunnlag.

1.7 Avhandlingens deler

Avhandlingen består av to deler; en kappe og tre enkeltstudier i artikkelform. I kappa presenteres kontekstuelle, teoretiske, metodiske og forskningsetiske sider ved avhandlingsarbeidet som ikke har blitt viet nok plass i enkeltstudiene. I tillegg til å vise sammenhenger mellom de ulike delstudiene, er det et mål at kappa skal vise hvordan de ulike delstudiene bidrar til å belyse hovedspørsmålet. Avhandlingsarbeidets forskningsbidrag innenfor fagfeltet blir også diskutert.

Kappa, som er avhandlingens første del, består av 6 kapitler. I innledningskapitlet (kapittel 1), som vi nærmer oss slutten av, blir scenen satt. Leseren bør nå ha en klar idé om hva som venter av leseopplevelser i det som kommer. I kapittel 2 blir sammendrag av de tre delstudiene som inngår i avhandlingen, presentert. Som en begynnende øvelse kan dermed leseren fint lese kappa uten først å lese de ulike delstudiene. Videre, i kapittel 3, blir forskningsfeltet som dette avhandlingsarbeidet ønsker å være en del av, skrevet fram. Avhandlingens relevans burde nå være tydelig. Forståelsesramme og metoderedegjørelse følger deretter i kapittel 4 og kapittel 5. Førstnevnte gir leseren innsyn i hvilken teoretisk plattform avhandlingen hviler på, noe som har vært lite kommunisert i de ulike delstudiene av sjangermessige årsaker. I metodekapitlet blir transparens et nøkkelord; transparens blir forsøkt etablert gjennom å legge fram datamaterialet og analytiske begrep og metodiske tilnærminger, samtidig som transparens blir brukt som argument for arbeidets troverdighet. I siste kapittel blir tråder nøstet opp, bidrag diskutert og visjoner avslørt.

Avhandlingens andre del består av tre artikler. De blir her presentert i den rekkefølgen de ble skrevet:

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach to studying the development of rater competence. *Assessing Writing, 20*, 37-52.

Jølle, L. (upubl.). To become a more proficient rater: What does it take?

Jølle, L. (in press). Rater strategies for reaching agreement on pupil text quality. *Assessment in Education: Principles, Policy & Practice*, DOI: 10.1080/0969594X.2015.1034087.

2 Sammendrag av artiklene

2.1 Artikkel 1

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach to studying the development of rater competence. *Assessing Writing, 20*, 37-52.

Første studie er todelt. Først presenteres en kartlegging av de ressurser vurdererne benytter seg av i vurderingsarbeidet, samt en kartlegging av en bestemt dimensjon ved interaksjonen mellom vurdererne (responsmønsteret). I denne delen inngår det også en undersøkelse av om interaksjonen og referansebruken endrer seg over i tid i et miljø hvor endring er intendert, det vil si hvor læring skal pågå. Dernest inneholder artikkelen en diskusjon om hvorvidt en slik første kartlegging kan danne utgangspunkt for å definere skrivevurderingskvalitet.

Alle lydopptak av de parvise vurderingene som er gjort i forbindelse med avhandlingsarbeidet, inngår som materiale i denne studien, det vil si opptak fra tre samlinger i perioden juni 2011 til april 2012. I materialet inngår det totalt 33[4] vurderere som er involvert i 26 ulike vurderingssamtaler hvor hver samtale tar for seg én til tre elevtekster. Tekstene som ble vurdert var skrevet av elever i starten av åttende skoletrinn og ble vurdert etter kompetansekravene etter syv års opplæring.

Forskningslitteraturen har en «slagside» når det gjelder hvordan man går fram for å undersøke tekstvurderingskompetanse; den har en tendens til å studere vurderingskvalitet ut fra psykometriske parametre, og vier i liten grad oppmerksomhet til de faktiske avveininger vurdererne foretar i vurderingsarbeidet.

Hovedmålet er å undersøke hvordan vurdererne balanserer over tid mellom egen erfaringspraksis og det nye som Vurdererpanelet tilbyr av ressurser. Studien er kvalitativ i den forstand at den undersøker sammenhengen mellom vurderernes bruk av vurderingsressurser og grad av motstand/aksept fra medvurderer. Metodisk blir dette gjort ved at samtalenes responsmønstre og vurderernes bruk av referenter (ulike vurderingsressurser) i argumentasjonen kodes og kategoriseres ved hjelp av programvaren NVivo10. Disse blir så gjenstand for komparativ analyse for å undersøke utviklingstrekk over tid.

[4] I den publiserte artikkelen har jeg feilaktig skrevet at det er 28 vurderere som inngår i det totale datamaterialet. Jeg kan ikke gjøre rede for hvordan denne feilen har oppstått.

Resultatene viser at erfarne lærere ikke har lett for å endre vurderingspraksis når de blir satt til å vurdere i en ny kontekst med et nytt skrivekontrukt og nye vurderingsressurser. De går fra å være erfarne lærere til å bli uerfarne tekstvurderere hvor mestringsstrategien i hovedsak ser ut til å være å holde seg til kjent praksis. Selv etter tre samlinger á to dager med intens opplæring og trening hadde ikke Vurdererpanelet endret vurderingspraksis i særlig grad. Dette har implikasjoner for hvordan man fremover må jobbe med konstruksjon av skriveprøver for å gjøre de valide og reliable. Studien viste imidlertid at i de situasjoner hvor vurdererne viste seg som kompetente, etterlot de seg «verbale spor» i dialogene som gjorde vurderingene etterprøvbare, og dermed mer valide.

2.2 Artikkel 2

Jølle, L. (upubl.). To become a more proficient rater: What does it take?

I denne andre studien blir én vurderers vurdererpraksis fulgt tett over de tre panelsamlingene. Siden denne vurdereren alltid vurderer sammen med ulike kollegaer, inngår det flere vurderere i materialet, men i et diakront perspektiv er det denne ene vurdereren som fremstår som studiets case.

Heller enn å undersøke hvordan denne vurdereren bedømmer ulike tekster sammenlignet med andre vurderere, blir vurdererens arbeid og utvikling fortolket ved å dra veksler på sosial teori om læring. Analysene er et argument for å forstå utvikling av vurderingskompetanse som en kompleks sosial (intermental) og individuell (intramental) prosess. En sentral del av læringen handler om å utvikle et relevant metaspråk gjennom målrettete remedieringspraksiser.

Funnene indikerer at det eksisterer en diskrepans mellom læringsteori og skriveprøvekonstruksjon. Psykometrien har hatt sterk innflytelse på skrivevurderingen, og jakten på reliabilitet har ofte blitt skjøvet foran med en a priori definert forståelse av hva kyndighet i skrivevurdering er. Konsekvensene er blant annet en overdreven tiltro til hvor raskt det lar seg gjøre å lære å vurdere (Kondo-Brown, 2002; Lim, 2011; Swain & Le Mahieu, 2012), samt for lite diskusjon om gyldigheten av vurderingene. Funnene har sentrale teoretiske og metodologiske implikasjoner for videre forskning på skriveprøver.

2.3 Artikkel 3

Jølle, L. (in press). Rater strategies for reaching agreement on pupil text quality

Tredje studie tar i likhet med første studie for seg hele datamaterialet. Det vil si at igjen er det transkriberte lydopptak fra tre samlinger i perioden juni 2011 til april 2012, hvor totalt 33 vurderere involvert i 26 ulike vurderingssamtaler, som inngår i studien.

Vurdererne i materialet er lærere som gjennomgår opplæring og settes til å vurdere elevtekster ut fra bestemte forventningsnormer og forhåndsdefinerte kriterier. Det finnes ulike tradisjoner for hvordan man ser for seg dette skal skje. Den klassiske, tekniske, positivistiske tradisjonen understreker nødvendigheten av at vurdereren «underkaster» seg forventningsnormene og de eksplisitt gitte kriteriene for å skape pålitelige vurderinger. I denne tradisjonen ligger det dermed også en oppfatning av at dette er mulig. En annen, mer hermeneutisk orientert tradisjon vil ikke utelukke muligheten for at et sett med forhåndsdefinerte kriteria ikke på tilfredsstillende vis tar høyde for alt som inngår i vurderingen av tekstkvalitet. I denne tradisjonen vil vurderingers pålitelighet selvsagt også være viktig, men «not everything» (Sadler, 2009).

Med forankring i Linells forståelse av språkhandlingers «doble dialogisitet» (Linell, 2001), det at enhver dialog alltid er i dialog med tidligere og påfølgende dialoger, er denne siste delstudien en empirisk undersøkelse av hvilke (mer eller mindre) underliggende rekursive kommunikative strategier som ligger til grunn for å vinne fram med argumenter når beslutninger om tekstkvalitet skal tas.

I tillegg til å peke på tre hovedstrategier som synes å være tilbakevendende i arbeidet med å nå fram til enighet om tekstkvalitet, inneholder studien også en diskusjon om forholdet mellom disse. Videre blir det også påvist hvordan en vurderers posisjonering evner å trumfe strategiene i svært ulike, ofte motsvarende, sammenhenger. Skrivevurderingsforskningen er opptatt av, og oppmerksom på, at det er menneskelig fortolkning som ligger til grunn for slutningen fra elevers skriveprestasjoner til utbytte (en eller annen form for score). Denne siste studien er en understreking av at den iboende kompleksitet som ligger i en slik hermeneutisk praksis ikke blir borte selv i en testsituasjon hvor definerte kriterier og standarder er utarbeidet.

3 Bakgrunn

Litteraturen som presenteres i det følgende vil være styrt av forsøket på å sette de tre delstudiene inn i en relevant kontekst. I hvilket forskningsfelt hører studiene hjemme? Hvem er avhandlingas studier i dialog med? Og hvor er de dunkle stedene som studiene forsøker å gi lys? Disse er sentrale spørsmål som har drevet fram det foreliggende kapitlet. Og behandlet i den rekkefølgen de her er presentert, gir det ei fortelling om den store skrivevurderingsforskningsfamilien, om de heteste diskusjonene innenfor denne familien, og til slutt om et tema som her blir holdt for å være viktig, men som inntil nå har blitt stemoderlig behandlet av den samme familien.

Kapitlet er følgelig tredelt. Gjennomgangen starter med en type teori – praksis-diskusjon idet forholdet mellom teoretisk forståelse av hva skriving er (skrivekonstruktet) og praktisk gjennomføring av skriveprøver, belyses. Hvordan skriveprøver er utformet henger gjerne sammen med teoretiske forutsetninger (selv om det kanskje helst skulle vært den andre veien rundt, jf pragmatismen, eventuelt teori og praksis som forent, jf. Huot, 2002, s. 165 ff.). Hvilke syn på skriving og testing som ligger til grunn for (og former) skriveprøver, er derfor av grunnleggende interesse. Målsettingen med denne første delen er å vise hvordan aktualiteten i mitt forskningsbidrag springer ut fra en slik grunnlagsdiskusjon. Deretter vil jeg i gjennomgangens andre del presentere forskningslitteratur relatert til skriveprøvenes balansering mellom kravet om *validitet* og kravet om *reliabilitet*. Denne delen aksentuerer konfliktaksene mellom det som omtales som «two cultures» (Elliot & Perelman, 2012, s. 149 ff.) innenfor skrivevurderingsforskningen, der de to ulike kulturene, i alle fall i praksis, har vektlagt hvert sitt testkrav.[5] Og til sist, i avhandlingen har jeg vært opptatt av vurderernes arbeid innenfor en ekstern administrert og standardisert skriveprøvesammenheng. Et sentralt testkrav i en slik sammenheng er interrater reliabilitet, det vil si vurderernes evner til å utvise samsvar i vurderingsarbeidet. Eksternt organiserte skriveprøver har alltid måttet forholde seg til utfordringer knyttet til å oppnå tilstrekkelig vurderersamsvar. Siste del av litteraturgjennomgangen vier derfor plass til studier som på ulike måter har undersøkt skriveprøvens «vurderereffekt».

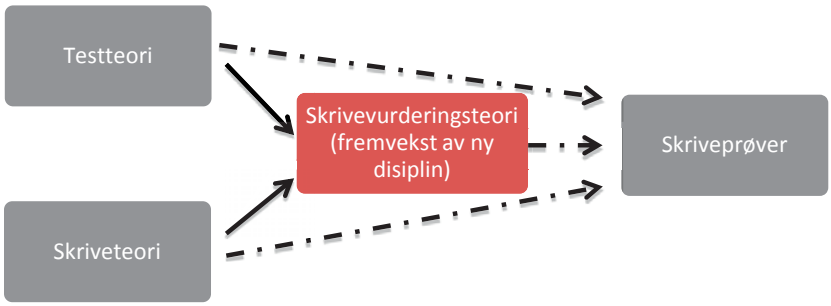
[5] Jeg velger å presentere en prinsipiell debatt i denne sammenhengen og vil i mindre grad komme inn på begrepsparets ulike dimensjoner. For en bredere og dypere forståelse av validitets- og reliabilitetsbegrepene, anbefales annen litteratur (se for eksempel Cronbach, 1988, 2004; Gipps, 1994; Kane, 2006; Messick, 1989).

I gjennomgangen vil det bli tydelig at jeg forholder meg vel så mye til engelskspråklig forskningslitteratur som til skandinavisk eller nordisk litteratur. Det er i alle fall to grunner til det. Den første årsaken er knyttet til *relevans*. Skrivevurdering som disiplin er internasjonal. Et eksempel: Selv om IEAs internasjonale undersøkelse av elevers skrivekompetanse på slutten av 80-tallet viste oss hvordan tekstkompetanse er kulturelt forankra og dermed ikke mulig å sammenligne på tvers av ulike språk og utdanningssystemer (Gorman, Purves, & Degenhart, 1988; Purves, 1992), viste den samme studien at man kan arbeide generisk med skrivevurdering som teoretisk disiplin. Den andre årsaken til avhandlingas internasjonale orientering er knyttet til relativ *nærhet*: Det meste av den engelskspråklige litteraturen som presenteres (spesielt i de to første delkapitlene) er amerikansk. Og selv om det er et hav mellom amerikanske og norske skrivere, er norsk skriveopplæring sterkt influert av amerikansk skriveteoriutvikling. Prosessorientert skrivepedagogikk som seilte fra USA inn i norske klasserom fra slutten av 80-tallet, er bare ett eksempel på denne felles historien (jf. Dysthe, 1987). I norsk sammenheng ser vi også hvordan man det siste tiåret i arbeidet med å utvikle nasjonale skriveprøver/utvalgsprøver i skriving har orientert seg internasjonalt i det teoretiske grunnlagsarbeidet. Nærliggende her er det å trekke fram valideringsfokuset som har hentet inspirasjon fra den nevnte IEA-studien.

3.1 Skriving – prøver og konstrukt

Under *1.2 Rasjonale* pekte jeg på hvordan denne avhandlingen tematisk befinner seg i et landskap som historisk sett kan sies å ha to forfedre; testteori (psykometri) og skriveteori. I en gjennomgang av henholdsvis testteoritradisjonen og skriveteoritradisjonens innflytelse på den praktiske skrivevurderingen i USA gjennom forrige århundre, viser Behizadeh og Engelhard Jr. (2011) at testteorien har hatt stor innflytelse på utvikling og gjennomføring av skriveprøver, mens ulike teorier om skriving i svært liten grad har blitt anvendt i dette arbeidet (se også Condon, 2011). Det har dermed vært en nærmest en-til-en-forhold mellom testteori og gjennomføring av skriveprøver, noe som har ført til «reliabilitetsorienterte» testformer som flervalgsprøver (indirekte målinger av skriveferdigheter) og vekt på tekstens formelementer. Dette er testformer hvor det er enklere å oppnå reliable målinger enn ved skriveprøver hvor alle dimensjoner ved skrivekompetansen vurderes. Behizadeh og Engelhard Jr. ser imidlertid en endring de senere 20 årene ved at 1980-tallets økologiske, sosiokulturelle

og prosessorienterte oppfatning av skriving har gjort krav på en mer kontekstuell forankring ved vurdering av skrivekompetanse. En ny forståelse av skriving, som ikke ser på skriving som bare et sett ferdigheter, men også som en kompetanse i å skape mening i spesifikke situasjoner, har skapt «validitetsorienterte» testformer som eksempelvis mappevurderingen. Behizadeh og Engelhard Jr. argumenterer for styrking av skrivevurdering som egen teoretisk disiplin («theories of writing assessments») hvor pålitelige målinger av skrivekompetanse går sammen med en velfundert teori om hva skriving er (jf. Huot (2002), men se særlig Fasting, Thygesen, Berge, Evensen & Vagle (2009) og Johnson & Elliot (2010) for interessante forsøk på å validere skriveprøver knyttet til så vel nasjonale prøver som ingeniørkurs ved et amerikansk universitet). Skrivevurderingdisiplinen kan således sees som et forsøk på å bygge bro mellom sine opphavsdisipliner, jf. Figur 1.



Figur 1 Strukturen som skriveprøver inngår i. Hentet fra Behizadeh & Engelhard Jr., 2011, s. 205.

Det ligger en tydelig indre spenning i å måle skriveprestasjoner. Det å oversette en prestasjon til en score på en skala forutsetter at det er målbart, altså at det er entydig, sikkert og presist. Skal noe måles kan det ikke være bra og dårlig på samme tid. Dette er kjente forestillinger innenfor den klassiske positivistiske vitenskapstradisjon etter Comte (2009). Spenningen oppstår imidlertid ved at det som skal måles i denne sammenhengen, skriving, blir helst beskrevet som komplekst, flertydig og unikt, altså på måter vi kjenner igjen innenfor den hermeneutiske vitenskapstradisjonen etter Schleiermacher (1977) og Dilthey (2010). En tekst kan innenfor et slikt paradigme nettopp være både god og mindre god på samme tid (jf. Moss, 1994).

Uavhengig av hva motivasjonen er for å teste skriving, om det er for å få et grunnlag for å ansvarliggjøre utdanningsinstitusjoner (bedrive politikk), for å få et grunnlag for å rangere og kvalifisere elever og studenter for videre utdanning, eller om det er for å få et grunnlag for den enkeltes videre opplæring, må denne indre spenningen i skriveprøven løses. Det kan virke hensiktsmessig å forske fram skrivevurdering som egen disiplin bestående av aktører som ser og har kunnskap om dette. Hva det vil si, vet vi ennå ikke. Noen vil hevde dette lar seg best gjøre innenfor testteoretiske rammer (Li, 2003; Mislevy, 2004), andre ser for seg en mellomposisjon (Condon, 2011; Evensen, 2014; Huot, 2002; Moss, 1994), mens atter andre vil mene at vi ikke evner å gjennomføre skriveprøver som baserer seg på et velfundert (det vil si valid) skrivekonstrukt før positivismens idealer er forlatt (Lynne, 2004; Wilson, 2006).

I amerikansk kontekst ser det enn så lenge ikke ut til at denne nye teoretiske disiplinen har fått praktisk gjennomslagskraft. Jefferey (2009) har vist hvordan delstatsprøver i skriving stort sett legger vekt på tekstenes formelementer, det vil si «reliabilitetsorienterte» testformer, og at prøveutviklerne nesten uten unntak unngår å gjøre rede for skriveteoretisk fundamentering. De ovafor nevnte IEA-prøvene hadde en helt annen konstruktvalidering i forkant, og det samme kan altså sies om arbeidet som er gjort i forbindelse med de norske utvalgsprøvene (jf. Berge, Evensen, Thygesen og Fasting, 2007; Fasting, Thygesen, Berge, Evensen og Vagle, 2009). IEA-prøvene viste seg imidlertid å ha liten gyldighet som internasjonal studie på grunn av skrivingens kulturelle og kontekstuelle forankring. De norske utvalgsprøvene i skriving vet vi fremdeles ikke hvordan vil utvikle seg. Men prøvedesignet har en klar kvantitativ dimensjon; kravet om reliabilitet i klassisk psykometrisk forstand ligger der. Forskningsgruppen bak utvalgsprøvene inntar altså en skrivevurderingsteoretisk mellomposisjon som beskrevet over.

3.2 Validitet og reliabilitet

Det er ulike måter å rekonstruere skrivevurderingshistorien. Over har jeg vist hvordan testteorien i stor utstrekning har prega gjennomføringa av skriveprøver, men at vi de senere tiårene har sett at skrivevurdering som egen disiplin er i ferd med å etablere seg. På den måten kan vi forstå skrivevurderingshistorien som drevet av ulike *teoretiske* disipliner. Kathleen Yancey (1999) har vist hvordan ulike linser kan brukes til å skape andre fortellinger om den samme historien. Ifølge henne kan omdreiningspunktet like gjerne være *metoden*, *skriveren*

(det vil si eleven), *forholdet mellom ekspertisen og skrivelærerkollegiet*, eller *forholdet mellom klasseromsvurdering og standardiserte vurderingssituasjoner*. Hver især inviterer disse til ulike, men sammenfallende, fortellinger om utviklingen av skrivevurdering som praksis. Huot (2002) foreslår imidlertid et annet omdreiningspunkt, nemlig *forholdet mellom validitet og reliabilitet*. Ved å la dette begrepsparet være styrende element, vil historiens dialektikk framstå slik i en amerikansk kontekst: Først dominerte reliabilitetshensyn (flervalgstester), så validitetshensyn (holistisk vurdering av essays), og nå er begge like viktige (gjerne mappevurdering). En samtids opplevelse av «å stå i syntesen» synes å være ahistorisk, og vi kjenner igjen nåtidsposisjonen også i norsk kontekst hvor balanse mellom validitet og reliabilitet altså synes å ha prioritet i utviklingen av utvalgsprøvene.

I det videre vil jeg fokusere på disse to begrepene, ikke for på nytt å dra historiske linjer, men for å aksentuere en sentral problemstilling som stadig er like aktuell. For hva vil det si å balansere validitets- og reliabilitetshensyn ved gjennomføring av skriveprøver? Jeg er redd jeg ikke kommer til å svare tilstrekkelig godt på spørsmålet, men gjør likevel følgende: Først presenterer jeg sentrale aspekt ved begrepene, deretter viser jeg fram en debatt som tydelig får fram skillelinjene mellom ulike teoretiske ståsteder som uomtvistelig har praktiske konsekvenser for skriveprøveavvikling.

I det følgende presenteres ikke en utfyllende diskusjon omkring validitetsbegrepet, men heller aspekter av det slik det er definert av Messick: ”Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (Messick, 1989, s. 13). Et sentralt poeng hos Messick er altså at validitet ikke er iboende i selve prøven, men at det er slutninger og handlinger gjort på grunnlag av testresultater som skal være gjenstand for validering. Det følger av dette at vi må ha en klar forståelse av grunnlaget for de tolkninger som blir gjort. For skriveprøver sin del; vi må en klar forestilling om hva skriving er («theoretical rationales») og hvordan de enkelte elevtekstene svarer til en slik forståelse («empirical evidence»). Et godt eksempel på at dette ikke er uproblematisk, er samspillet mellom lese- og skrivekompetanse. Dersom en skriveprøve baserer seg på forberedelser knytta til leseaktiviteter (for eksempel å lese ulike fagtekster for deretter å skrive en drøftende tekst), vil lesekompetansen nødvendigvis påvirke skriveprestasjonen. Elever med utprega lesevansker vil i slike tilfeller prestere relativt sett svakere enn elever uten lesevansker. Messick omtaler slike trusler mot validitet som

«construct-irrelevant variance» (Messick, 1989). Det er det samme Evensen poengterer når han skriver at slike eksempler «innebærer at vi må «holde tunga rett i munnen» med hensyn til begrepsforståelse» (Evensen, 2010, s. 16). I samme artikkel peker Evensen på at det finnes en annen side ved validitetsbegrepet som er knytta til breddedimensjonen. Studier har vist at elever får stor frihet til å velge skriveoppgaver de føler de behersker (Vagle & Evensen, 2005). Elevene spesialiserer seg særlig innenfor fortellende tekster. Spørsmålet blir da om man på en valid måte får vurdert elevers skrivekompetanse dersom de ikke prøves i andre skrivehandlinger? Der den førstnevnte dimensjonen handler om *begrepsvaliditet*, handler sistnevnte om *innholdsvaliditet*. Evensen peker på at det i liten grad har vært diskutert hva man mener med skrivekompetanse, og at det derfor har vært et behov for å definere et valid skrivebegrep (Evensen, 2010). I sammenheng med 2005-forsøket (nasjonale prøver i skriving), ble et slikt begrep utviklet, presentert ved en modell (Fasting m.fl., 2009; Evensen 2010). Denne modellen, eller Skrivehjulet som det blir kalt, ligger også til grunn når de læringsstøttende prøvene i skriving gjennomføres fra 2014.

Som en dimensjon av en prøves gyldighet, finner vi vurderingen av dens reliabilitet. Begrepet er tilsvarende komplekst og fanger inn flere sider ved en måling (for eksempel «holdbarhet» (Yu, 2005) og «konsistens» (Cronbach, 1951). Innenfor skriveprøver blir reliabilitetsundersøkelser foretatt for å utelukke at tilfeldigheter preger resultatene i for stor grad. Det foreliggende avhandlingsarbeidet som ser på skrivevurderingsarbeid innenfor et praksisfellesskap, har å gjøre med variasjon i vurderingene knytta til skrivevurderingens subjektive karakter. Vurderersamsvar beskriver i hvilken grad de ulike vurdererne vurderer like tekster likt (Weigle, 2002). Utvikling av et tolkningsfellesskap innenfor en skriveprøvesammenheng er nettopp uttrykk for et ønske om akseptabelt vurderersamsvar (jf Berge, 2009).

Vi skal se at mye forskning handler om skriveprøvers reliabilitetsutfordringer (jf. Berge, 2005, s. 106). Et grunnleggende problem ved store, eksternt administrerte skriveprøver er at elevteksten som skal vurderes skal «oversettes» fra noe svært komplekst og unikt til å framstå som svært enkel og sammenlignbar med andre tekster. Det er en kjensgjerning at denne oppgaven er krevende for vurdererne (jf. skriveprøvers indre spenning beskrevet over), og en rekke (eksperimentelle) studier søker å bidra til en bedre forståelse for hvorfor dette er så krevende, eventuelt også «foreskrive medisin» for å gjøre skrivevurderingene mer reliable.

En vanlig forståelse av forholdet mellom validitet og reliabilitet knyttet til forskjellige former for tester, og dermed også skriveprøver, er at det er fint mulig å tenke seg at en test er reliabel uten å være valid, mens det er umulig å se for seg en test som er valid uten å være reliabel. I førstnevnte situasjon vil testen være reliabel dersom vurderernes dommer i tilstrekkelig grad er konsistente og sammenfallende selv om de er ugyldige (in-valide), for eksempel fordi testresultatene i for stor grad inneholder «construct-irrelevant» variasjon. I den andre situasjonen vil upålitelige vurderinger simpelthen diskreditere gyldigheten av testen. Anvendt på skriveprøver: Når tilstrekkelig mange vurderere vurderer forskjellige elevtekster tilfeldig ulikt, er det vanskelig å se for seg at man er enig om hva det er som egentlig er gjenstand for vurdering. Det er denne logikken Moss utfordrer i artikkelen «Can There Be Validity Without Reliability?» (1994) hvor hun svarer bekreftende på spørsmålet som ligger i artikkelens tittel. Hun argumenterer for så vidt ikke mot at reliabilitet er sentralt, men mener at rollen pålitelighetsaspektet skal få ved en gitt prøve ikke må tas for gitt, men heller inngå i avveininga når prøven designes: «The decision about which strategy to use should depend upon the aims and consequences of the assessment in question» (s. 8). Moss lanserer så hermeneutikken som en alternativ tilnærming til vurderingsarbeidet. Hun skriver:

> A hermeneutic approach to assessment would involve holistic, integrative interpretations of collected performances that seek to understand the whole in light of its parts, that privilege readers who are most knowledgeable about the context in which the assessment occurs, and that ground those interpretations not only in the textual and contextual evidence available, but also in a rational debate among the community of interpreters. (Moss, 1994, s. 7)

I sitt tilsvar til Moss applauderer Robert Mislevy henne for å vise hvordan hermeneutikken tilbyr andre begreper og forståelsesmåter for å nå fram til gyldige dommer. Men Mislevy mener Moss opererer med et for snevert reliabilitetsbegrep (samvariasjon og interrater korrelasjon) når hun slår fast at reliabilitet ikke er en forutsetning for validitet (Mislevy, 1994). Med en bredere forståelse av hva reliabilitet er, nemlig som «credibility of evidence, *where credibility is defined as appropriate to the inference*" (Mislevy, 1994, s. 10), vil reliabilitet være en nødvendig forutsetning for en gyldig slutning. Fra en statistikers ståsted (les Mislevys) virker dette å være selvsagt. Spørsmålet da er hva uenigheten skyldes. Eller, hvordan kan Moss

få seg til å hevde at skrivevurderinger kan være gyldige uten å være reliable når psykometrien sier noe annet? Da er det interessant at Mislevy skriver at «test administrators warn test users against interpreting scores without other sources of information, but that the test users themselves are most prone to reify "traits" such as "IQ" or "writing ability" (s. 10). Her gir Mislevy uttrykk for en praksis-testteori-konflikt som er verdt å forfølge.

I 2003 går nemlig også Heng Li i rette med Moss' artikkel fra 1994. Li anklager Moss' resonnement for å inneholde alvorlige feil som må rettes opp. Han går deretter svært *testteoretisk* til verks for å vise Moss' mangler, og nettopp det at han går teoretisk til verks, er hans poeng: Li mener nemlig at det er en «tendency not to draw the distinction between reliability as an abstract concept, and reliability coefficients which operationalize and quantify this concept under specific measurement models» (Li, 2003, s. 90). Og det er denne tendensen til å benytte en operasjonell definisjon av begrepet i en teoretisk diskusjon som Li mener fører galt avsted. Mislevy (2004) støtter deretter Lis innvendinger, men igjen anerkjenner han at hermeneutiske analysetilnærminger har en utfyllende rolle sammen med psykometriske metoder ved ulike tester. Mislevy tillater seg likevel å spørre om informasjonen man får ved hermeneutiske arbeidsmetoder er verdt kostnaden (2004, s. 243).

Moss svarer deretter på kritikken framsatt av Li og Mislevy i en ny artikkel (Moss, 2004). Her presiserer hun at det nettopp var *testpraksisen* (og ikke *testteorien*) hun var opptatt av da hun skrev artikkelen i 1994. Hun går også i rette med Mislevy (og Li) og fastholder at psykometrien og hermeneutikken på sentrale områder skiller lag. For eksempel, når Mislevy påpeker at «collecting predetermined items of evidence, to be evaluated along predetermined lines, is a strategy for obtaining at relatively low cost information that previous work suggest will be useful" (Mislevy, 2004, s. 239), er dette korrekt innenfor psykometrien, men totalt uforenelig med en hermeneutisk tilnærming hvor det unike og kontekstuelle står sentralt.

Jeg finner denne teoretiske debatten som Moss innledet i 1994, og som ble vekket til live igjen 10 år senere, svært interessant. Den hører helt klart hjemme i kampen om hegemoni innenfor skrivevurderingsdisiplinen. På den ene siden er skriveforskning tradisjonelt sett opptatt av kompleksiteten ved skriving, ved det at skriving krever flere ulike kompetanser som er vanskelig å måle. På den andre siden er psykometrien heller opptatt av standardisering og generaliserbarhet, og søker å eliminere feiltolkninger. Når Moss så stiller spørsmål om en sentral maksime innenfor testteorien er gyldig(!), er det ikke annet å vente enn at det

angrepne forskningsfeltet forsvarer seg. Det interessante for dette avhandlingsarbeidet, for å komme tilbake til det, er at denne debatten viser at det ikke er selvsagt hva *balanse mellom validitet og reliabilitet* vil si i skrivevurderingssammenheng.

3.3 Vurdererens rolle ved skriveprøver

Vurdereren er skriveprøvens portvokter. Over seg har han de føringer som ligger i prøvekonstruktet; vurderingsmodell, vurderingsskjema, kriterier og vurderingsform. Under seg har han teksten; elevens ytringer formulert etter evne og motivasjon. Så er det vurdererens oppgave å plassere elevens ytringer i form av denne teksten på rett plass på en gitt skala. Lykkes flere vurderere med dette over tid, er det uttrykk for reliable vurderinger i testteoretisk forstand. Utfordringen er at på tross av omfattende opplæring ser det ut til å være svært vanskelig å eliminere en signifikant «vurdererereffekt» (Du & Wright, 1997; Engelhard, 1994; Gyagenda & Engelhard, 2009; McNamara, 1996).[6]

I litteraturgjennomgangen av vurdererens rolle i vurderingsarbeidet i delstudien «Pair assessment of pupil writing» valgte jeg å skille mellom den forskning som er «produktorientert» og den som er «prosessorientert». Den produktorienterte forskningen er reliabilitetsfokusert: Kvalitetsvurderinger er det samme som sammenfallende og konsistente vurderinger (Jonsson og Svingby, 2007). I den andre studien, «To become a more proficient rater: What does it take?», beskriver jeg hvordan arbeidet med å oppnå slike vurderinger handler om å korrigere for «vurdererereffektene» (jf. Saal, Downey og Lahey, 1980). Dette kan enten gjøres statistisk (for eksempel ved å korrigere for «hauker»/»duer»), ved utvalg (for eksempel ved å rekruttere et så homogent vurdererkorps som mulig), og/eller ved trening. En studie gjennomført av Elder, Knoch, Barkhuizen og von Randow (2005) er et eksempel på sistnevnte: En gruppe vurderere bedømmer en rekke tekster hvor vurderingene så sammenlignes med forhåndsvurderte standardtekster. Treningen består i å vurdere tekster helt til vurderingene på en reliabel måte sammenfaller med disse standardtekstene. Vurderere som vurderer tilfredsstillende på denne måten, omtales gjerne som ekspertvurderere (Lim, 2011).

[6] Innenfor et hermeneutisk paradigme vil det være vanskelig i det hele tatt å snakke om «vurdererereffekt». Innenfor et slikt paradigme vil det være en selvfølge at vurderernes tolkninger av tekstkvalitet varierer fordi selve vurderingsarbeidet er knytta til den enkelte vurdererens unike livshistorie. Å snakke om «vurdererereffekt» som man empirisk eller statistisk forsøker å eliminere, fremstår dermed som en dehumanisering av vurderingsarbeidet. Dette blir enda tydeligere når «vurdererereffekt» til tider blir erstattet med «rater error» (jf. Engelhard, 1994; Rudner, 1992).

Særlig denne avhandlingas første delstudie presenterer en rekke studier som avdekker ulike vurdererrelaterte kilder til usikre resultat, og jeg kunne her fortsatt med å presentere studier som på denne måten arbeider med å redusere «målefeilene» forbundet med at resultatene baserer seg på menneskers tolkningsarbeid. I stedet vil jeg fokusere på den forskningen som jeg omtaler som «prosessorientert». Slike studier undersøker på ulike måter selve vurderingsarbeidet, altså de meningsskapende prosessene vurdererne involverer seg i når de skal avgjøre ulike teksters kvalitet.

Et område som har interessert forskere lenge, er hvilke tekstdimensjoner vurdererne legger mest vekt på i vurderingsarbeidet. Diederich, French & Carltons studie fra 1961 har blitt stående som førende for mye skriveprøveutvikling. De fant at 53 vurderere som vurderte 300 studenttekster, gav responser som falt innenfor følgende fem dimensjoner: idé, form, smak («flavor»), rettskriving og språkbruk. I ettertid har tekstkvalitet forstått som en sammensetning av disse ulike dimensjonene stått sterkt. Et annet klassisk og eksperimentelt eksempel er Freedman (1979a, 1979b) som ved å omskrive elevtekster til å være sterke og svake i henholdsvis innhold, organisering, rettskriving og språkbruk, viser hvordan vurderere holder innhold og tekstorganisering som viktigst av de ulike tekstdimensjonene. Andre studier mer eller mindre bekrefter dette (for eksempel Breland og Jones, 1984).

Denne forskningen var viktig og har gitt oss en innsikt i vurderernes arbeid som vi i dag nærmest tar for gitt og som danner grunnlaget for hvordan man strukturerer et komplekst tekstvurderingsarbeid. Likevel gjenstår mye, og i 1990 skreiv Huot følgende: «Other than results that measure the importance of content and organisation in rater judgment of writing quality, little is known about the way raters arrive at theses decisions» (Huot, 1990, s. 258). Denne mangelen på studier om hvordan vurdererne «får jobben gjort», er senere blitt delvis gjort bot på i form av studier som søker å finne svar på hvordan vurderere, gjerne med ulik kompetanse/erfaring, kognitivt strukturerer arbeidet. Cumming (1990) fant eksempelvis i så måte at skrivevurderere tar i bruk en kombinasjon av tolkningsstrategier og vurderingsstrategier tilhørende ulike tekstdimensjoner i arbeidet, mens Cooksey, Freebody & Wyatt-Smith (2007) i sine analyser av klasseromsforankra vurdering fant strategier som i utstrakt grad er kontekstuelt forankra (både relatert til eleven, læreren og det sosiale). Det har også blitt utviklet forslag til ulike modeller. Milanovic, Saville & Shuhong (1996) er et eksempel. De mener å se at tekstvurdereren går fram på følgende måte: internalisere vurderingsskjema, scanne teksten, kjapp gjennomlesning, vurdere, modifisere og påfølgende

endelig vurdere. Crisps (2010) prosessmodell er noe mindre finmasket: (prolog – tanker om arbeidet), tolkningsfase, vurderingsfase, vurderingen, (epilog – tanker om vurderingen). Slike empirisk forankra kognitive modeller er utvilsomt viktig for å få en bedre forståelse av vurderingsarbeidet, for å bedre tilrettelegge for opplæring, og kanskje viktigst, for å bedre kunne designe en skriveprøve. Når det gjelder det sistnevnte har Barkaoui (2010) påpekt at det også her gjenstår mye arbeid. For eksempel viser han i sin studie hvordan vurderingsprosesser er relatert til valgt vurderingsform (holistisk vs analytisk): Ved bruk av analytiske vurderingsmodeller bruker vurdererne langt mer tid og kognitive ressurser på vurderingsskjema enn ved bruk av holistiske vurderingsmodeller (jf. dette avhandlingsarbeidets andre delstudie, «To become a more proficient rater: What does it take?»).

Begge disse bolkene av studier (utmeislinga av typiske tekstdimensjoner og utviklinga av kognitive vurderingsprosesser) har vært viktig i arbeidet når skrivevurderere skal forberedes til vurderingsarbeid. Opplæring, eller trening, blir nemlig holdt for å være en avgjørende komponent i enhver skriveprøveavvikling (Alderson, Clapham & Wall, 1995; Purves, 1992; Stuhlmann, Danile, Dellinger, Kenton & Powers, 1999). Med den tidligere forskningen har skriveprøveutviklere fått innsikt i hva vurdererne har og ikke har blikk for i vurderingsarbeidet. På den måten har man fått kunnskap som gir mulighet for å kompensere for kjente reliabilitetstrusler knytta til vurdereren (for eksempel hva de vektlegger ved en tekst som beskrevet over, men også deres bakgrunn, jf. Leckie & Baird, (2011), verdier, jf. Baker, (2010)) og organiseringa (for eksempel vurdereres tilbøyelighet til å sammenligne tekster etter «nærhetsprinsippet», jf. Vaughan (1991)). De undersøkelser som har blitt gjort i forhold til effekten av slik trening er imidlertid ikke lystig lesning (jf. Black, 1962; Lumley & McNamara, 1995; Weigle, 1994). McNamara slår i så måte fast at «assessment procedures which rely on single ratings by trained and qualified raters are hard to defend» (1996, s. 235).

Det synes altså å være en diskrepans mellom «sunn fornuft» og forskning på dette området. Den sunne fornuft sier at øvelse gjør mester, det vil si at vurderertrening vil gi mer valide og reliable vurderinger. Men de ulike studiene støtter ikke opp om en slik antakelse. I delstudie 2, «To become a more proficient rater: What does it take?», mer enn antyder jeg at det vil være klokt å lytte til den sunne fornuft i denne sammenhengen. Det flertydige bildet som skapes av effekten av opplæring kan nok langt på vei forklares av en svak definisjon av hva man legger i vurderertrening. Litteraturen viser eksempler hvor denne treninga har vart

fra en halv time (Barrett, 2001), til en halv dag (Kondo-Brown, 2002; Weigle, 1998), en dag (Swain & Le Mahieu, 2012; Wyatt-Smith, Klenowski & Gunn, 2010), og opp til mer omfattende seminarer som varer over flere dager (Purves, Gorman and Takala, 1988). I min delstudie foreslår jeg å se på denne forberedelsen med et læringsteoretisk perspektiv. Da er det enklere å se at det å skape endringer/forbedringer i vurderingsarbeidet er tidkrevende.

Materialet i denne avhandlinga er hovedsakelig vurderingsdialoger. Ved å studere disse dialogene har jeg kunnet trekke slutninger om hvordan vurdererne snakker om tekster i en spesifikk vurderingskontekst, hvordan de tar i bruk normerende vurderingsressurser i dette arbeidet, og også hvordan vurdererne interagerer med hverandre i denne profesjonsdialogen. Med vurderingsdialoger som datamateriale er det svært interessant at Meadows og Billington, i en omfattende litteraturgjennomgang av studier knyttet til skriveprøvers reliabilitet, således skriver:

> The process of reaching a consensus regarding the best mark for a script may serve a useful training function, improving the accuracy with which examiners apply the marking scheme. While no empirical investigation of this possibility has been uncovered in producing this review, there have been many studies of the effectiveness of other methods of examiner training. (Meadows & Billington, 2005, s. 50)

De finner altså ingen studier som undersøker hvordan vurderere sammen når fram til enighet om hva tekstkvalitet er og hva som gjelder som kjennetegn på tekstkvalitet. I tiden etter 2005 har det kommet noen studier som tar for seg modereringssesjoner hvor vurderere under ledelse av forsker/ekspert snakker seg fram til enighet om tekstkvalitet (Colombini & McBride, 2012; Wyatt-Smith, Klenowski & Gunn, 2010). Disse har gitt oss innsikt i kompleksiteten i vurderingsarbeidet og er således viktige bidrag i valideringsarbeidet som blir gjort innenfor skrivevurdering. Det samme kan sies om arbeid av Evensen (2012, 2014) og Matre og Solheim (2014). Evensen viser i sitt arbeid hvordan lærere fra ulike fagtradisjoner gjennom vurderingssamtaler kommer nærmere en felles tolkning av tekster, og han argumenterer for at det valideringsarbeidet som må gjøres i forbindelse med vurdering av skriving må komme «nedenfra», fra lærerne som har erfaring med og kunnskap om elevers skriveutvikling (jf. hva jeg i delstudie 2 omtaler som en dialogisk tilnærming til vurdereropplæring i stedet for en mer tradisjonell hierarkisk tilnærming). Matre og Solheim bruker på samme måte vurderingssamtaler i sin studie. De ser på hvordan små lærergrupper vurderer elevtekster ved

hjelp av normerende støttemateriale (jf. Matre, Berge, Evensen, Fasting, Solheim & Thygesen, 2011) og finner at lærere utfører den samme oppgaven på ulike måter. Matre og Solheim kategoriserer måten lærerne forholder seg til støttematerialet på som enten «instrumentell», «læring pågår» eller «fleksibel». De to ytterkategoriene sammenfaller med DeRemers (1998) «rubric-based evaluation» og «text-based evaluation». «Rubric-based» vurdering kjennetegnes ved at vurdereren hele veien kontrollerer egen oppfatning av teksten med formuleringer i vurderingsskjemaet. «Text-based» vurdering, derimot, kjennetegnes ved at vurdereren er familiær med vurderingsressursene før tekstlesningen tar til. Kognitiv oppmerksomhet kan da gis til teksten, men fremdeles med vurderingskriteriene som grunnlag for vurdering. Matre og Solheim finner imidlertid at det meste av vurderingsarbeidet er et sted i mellom disse to ytterkategoriene (altså under «læring pågår»).

Litteraturgjennomgangen som jeg nå har presentert fremstår som et argument for mitt avhandlingsarbeid. Skrivevurderere blir ofte mer eller mindre eksplisitt fremstilt som «nyttige idioter»[7] ved en skriveprøves pålitelighetsjag. Jeg har forsøkt å vise hvordan dette kan ha seg, men det har også vært et mål å presentere litteratur som reiser spørsmål ved en slik oppfatning. De ulike studiene i denne avhandlingen skal leses som et uttrykk for et ønske om å forstå hvordan skrivevurderere sammen når fram til beslutning om tekstkvalitet, hvordan de tar i bruk ulike hjelpemidler i dette samarbeidet, og hvordan felles opplæring virker inn på vurderingsprosessene.

[7] Huot (2002) uttrykker omtrent det samme når han er bekymret for at vurderertrening utelukkende fungerer som verktøy for å oppnå reliable vurderinger. Se også Sadler (2009).

4 Forståelsesramme

Forrige kapittel brukte jeg både til å diskutere relevant litteratur og til å posisjonere egen forskning som mer dialogorientert enn den dominerende hierarkisk orienterte forskningslitteraturen. Målsettingene med det foreliggende kapitlet er flere; først å gjøre rede for det dialogbaserte vitenskapsteoretiske grunnlaget for avhandlingsarbeidet. Deretter vil jeg drøfte sentrale begrep som er knyttet til hovedproblemstilling og datamateriale, det vil si *læringssyn*, Vurdererpanelet som *praksisfellesskap* og *vurderingsdialogene*. I siste del løfter jeg igjen blikket og diskuterer *kunnskapsbegrepet* i lys av å ha som målsetting at vurdererne skal foreta det DeRemer omtaler som «text-based evaluation» (1998). Kapitlet munner ut i presentasjon av en kunnskapsmodell som jeg argumenterer for også er gjeldende for vurderernes vurderingskompetanse. Siden dette på mange måter er gamle tanker tenkt om igjen, vil jeg i kapitlets avslutningsdel først presentere en historisk versjon av kunnskapsbegrepet i en vestlig kontekst.

4.1 En dialogbasert epistemologi

Filosofen Martin Buber var i første halvdel av forrige århundre sentral i utviklingen av det som fikk betegnelsen dialogfilosofi. For Buber karakteriseres ikke mennesket under substans, men *relasjon*. Det vil si at jeget er ikke-eksisterende uten et du, og således blir dialogen et *ontologisk* prosjekt (Buber, 2007). Andre har også på tilsvarende måte pekt på Ego-Alter-relasjonen som fundamental ikke bare for erkjennelse, men i det hele tatt for menneskelig væren (Markova, 2003, 2006; Skaftun 2002).

Med en slik ontologi følger det en tilsvarende epistemologi på kjøpet (Marková, 2006, s. 128) som er et brudd med et dominerede ”monologisk” kunnskapssyn (Linell, 2001, 2009). Dette monologiske kunnskapssynet har sine røtter hos Platon, og senere hos Descartes, og vi finner det i dag som det dominerende kunnskapssyn innenfor en rekke vitenskaper, men også, vil jeg påstå, som en gjeldende «folkelig» oppfatning av hva viten *egentlig* er. Suksessen for det kartesianske kunnskapssynet har selvsagt ligget i det faktum at det har vist seg å fungere. Modernitetens historie blir gjerne forklart som et resultat av en slik suksess (Henry, 2004; Kennington, 2004), og det vil jeg komme tilbake til mot slutten av dette kapitlet.

Med relasjonstenkningen til Buber knyttes det mye sterkere bånd mellom subjektet og ”det/den andre”, noe som fra et tradisjonelt vitenskapsteoretisk ståsted blir kritisert for å gjøre vitenskapen subjektiv og dermed usikker. Til det svarer den danske filosofen Søren

Kjørup at det på dette fundament ikke ligger noen resignasjon i forhold til begrepet ”sikker viten” som vi har lært å kjenne det de seneste hundreårene. I stedet er det snakk om et skifte av perspektiv. Han poengterer at det ligger ”i ordet ”viden” eller ordet ”erkendelse” at det er noget mennesker har eller producerer, så hvor der foreligger viden eller erkendelse, foreligger der spor af mennesker” (Kjørup, 2008, s. 182). Konsekvensene av dette er altså ikke at alt er relativt eller at sann kunnskap er en umulighet. I argumentasjonen for hva Kjørup kaller ”pragmatisk konstruktivisme”, framhever han at i stedet for å oppfatte det menneskelige perspektiv som en feilkilde, må det forstås som ”en nødvendig betingelse for at der overhovedet kan foreligge erkendelse” (2008, s. 183). Sannhet er dermed mulig å oppnå, men det kan ikke løsrives fra kultur og språk. Den samme tenkningen finner vi også hos Hans-Georg Gadamer. Om fysikkens verden spør han retorisk: ”Men er det virkelig slik at denne verdenen eksisterer i seg selv og ikke lenger er relativ til menneskets væren (…)?” (Gadamer, 2010, s. 494). Han svarer umiddelbart: ”Verken det biologiske eller det fysikalske universet kan i virkeligheten benekte at det er relativt til menneskets væren” (2010, s. 494).

Denne nedrivingen av det dikotomatiske skillet mellom subjekt og objekt viser seg i filosofien og vitenskapsteorien på mange områder. Mikhail Bakhtin understreker for eksempel hvordan humanvitenskapens studieobjekt, teksten, må sees på som en ytring (1999). Teksten forstått som ytring flytter den så å si over på samme banehalvdel som autoren og gjør den til en medspiller (forstått som bidragsyter i meningsproduksjon), ikke til noe som skal studeres. Linell har laget en kort (og ufullstendig) liste over slike kartesianske dikotomier som brytes ned innenfor en dialogisk forståelsesramme (2009, s. 391).

Som tradisjon er det nok fenomenologien, som vi kjenner den utviklet av Edmund Husserl (1970), og senere Maurice Merleau-Ponty (1994), som først, og etter hvert kanskje også grundigst, har drøftet perspektiveringens rolle ved kognisjon og persepsjon. Et sentralt poeng innenfor fenomenologien er at både hverdagslivet og vitenskapen er ufri fra fordommer og forestillinger, og at heller ikke vitenskapen dermed kan gjøre seg fri fra den samme fortolkende livsverden som vi lever våre hverdagsliv i. Det vi sanser, sanser vi fra et bestemt perspektiv. For Merleau-Ponty er all bevissthet det samme som perseptuell bevissthet (1994). Denne dreiningen fra rasjonalismens ”rene” bevissthet over til en type kroppslig erkjennelse, er en dreining fra ”å tenke” til ”å kunne”. Kunnskap er ikke bare noe man har, det er også noe man gjør.

For Merleau-Ponty blir altså kunnskapen synliggjort gjennom handlinger. Dermed ser vi også sammenhengen med en annen tradisjon som må sees som en forløper for en mer helhetlig dialogisme, nemlig pragmatismen. Utdanningsteoretikeren, den Hegel-inspirerte John Dewey, forfektet en mellomposisjon der subjektet (eleven) og objektet (kunnskapsstoffet) spiller på lag. Erkjennelse har utgangspunkt i det engasjerte subjekt som i en gitt situasjon føler på et problem som en ønsker å løse (Dewey, 1902). Dagens PBL (problembasert læring) som metode er et resultat av Deweys filosofi. Teoriers relevans gir seg ikke selv, og om virkeligheten ikke stemmer overens med forsøket på å forstå den, er det forsøket som må forkastes, ikke virkeligheten som må endres. Dette kan synes som en selvfølge, og er det vel også. Likevel; Thomas Kuhns (paradigmatiske) bok *The Structure of Scientific revolutions* (1962), hvor han lanserer begrepet *paradigmeskifte*, har gjort oss oppmerksomme på mangelfulle teoriers seiglivethet.

Det jeg til nå har tegnet opp, er et vitenskapsteoretisk bilde som forsøker å ta høyde for at det er et sansende og tenkende subjekt som skaper mening gjennom deltakelse i verden. Forkjært formulert kunne man dermed si at mennesket står i veien for «sikker» viten. Men et slikt utsagn faller på sin egen urimelighet nettopp ved at viten forutsetter mennesket (Jf. Kjørup, 2008. Se over). Det er med dette utgangspunktet, og blant annet på de ovenfor nevnte kjempers skuldre, språkviteren Linell står når han svært ambisiøst presenterer og diskuterer dialogismen som samfunnsvitenskapen og humanioras meta-vitenskapsteori (for eksempel Linell, 2001, 2009). Det er han selv som omtaler dialogismen som en meta-teori i det den favner omkring en rekke ulike dialogbaserte disipliner, eksempelvis dialogorientert filosofi, diskursorientert sosiologi og sosialt og kulturelt orientert psykologi (se for øvrig en oversikt i Linell, 2009, s. 402-403). Linell trekker frem følgende dimensjoner ved dialogismen som grunnleggende: Mennesket skaper mening gjennom å *interagere* i *kontekstuelt* forankra situasjoner ved hjelp av *semiotiske ressurser* (Linell, 2009, s. 31). Hva ligger så i det?

La oss begynne bakerst; for det første vil dialogteoretikere hevde det er et filter mellom verden (i seg selv) og det sansende og tenkende menneske. Det er gjennom dette filteret vi *konstruerer mening*, eller som dialogteoretikere vil formulere seg, *medierer mening*. Den sentrale medieringsform innenfor dialogismen er den semiotiske (Wertsch, 1991, 1998), det vil si språk og andre tegnsystemer. Vi samhandler, eller i alle fall agerer sammen, ved hjelp av *semiotiske ressurser*.

For det andre; *interaksjon er* både kommunikasjon og tenkning. Kognitiv virksomhet går ikke forut for *andre*-vendthet, den konstitueres *gjennom* en slik orientering (Potter, 1998). Denne oppfatningen, at vi så og si blir til gjennom å gå i dialog med andre mennesker, andre artefakter, tidligere opplevelser, framtidig mulige opplevelser og så videre, understreker dialogismens epistemologiske karakter. Bakhtin formulerer denne innsikten slik: ”Truth is not to be found inside the head of an individual person, it is born between people collectively searching for thruth, in the process of their dialogical interaction” (1984, s. 110).

For det tredje; det at meningsskaping er *kontekstuelt* forankra betyr at vi alltid erfarer verden i spesifikke situasjoner. Hvilke kontekstuelle faktorer som spiller inn er avhengig av hva som gjøres relevant i den enkelte situasjon (Linell, 2009, s. 17). I tillegg er det slik at når vi handler, har det kontekstuelle implikasjoner som går utover den spesifikke situasjonen. Fra et sosialkonstruksjonistisk perspektiv snakker Fairclough i så måte om hvordan disse enkeltsituasjonene på et nivå skaper og opprettholder sosiale identiteter og roller, mens de på et stadig mer abstrakt nivå også bidrar til å skape og opprettholde større kunnskapssystemer (Fairclough, 2003). Innenfor dialogismen vil man i tillegg legge vekt på at relasjonen mellom det situasjonelle og det stabiliserende er gjensidig og samtidig. Denne samtidige påvirkning har Lars Sigfred Evensen forsøkt fange inn i en diatopmodell (Evensen, 2002, s. 399). Med utgangspunkt som anvendt språkviter skriver han:

> Whereas constructionism (…) has largely treated conventional lingusitic items as a priori givens, and interactionism simply cannot account for them, dialogism will regard them as underdetermined meaning potentials (…). Linguistic items need specific contexts for their actualization and thus only get a (temporarily) fixed meaning in a specific, peopled context. The diatope is hence the symbolic locus of actual meaning, at all levels of discourse analysis”. (Evensen, 2002, s. 404)

Til nå har jeg presentert en grunnlagsforståelse for hvordan jeg posisjonerer meg epistemologisk. Likevel er det, slik Linell presiserer, at dialogisme er en meta-teori for hvordan vi forholder oss til verden, og ikke i seg selv en teori som hjelper til med å forklare mer presserende og spesifikke ”kommunikative problem”. Etter nå å ha ”ankret opp”[8] vil jeg i det

[8] En svak metafor med tanke på dialogismens vektlegging av bevegelighet. Det heraklitske ”panta rhei”, eller ordtaket ”man kan ikke to ganger gå ned i samme elv” passer vel bedre, noe som peker mot at heller enn å ankre opp har jeg kastet meg ut på dypt vann.

følgende drøfte andre begreper som har vært sentrale for dette avhandlingsarbeidet. Som en leserveiledning kan jeg først gjenta avhandlingens hovedproblemstilling: *Hvordan utvikler uerfarne skrivevurderere kompetanse i elevtekstvurdering?* Problemformuleringen gjør krav på visse presiseringer. Først og fremst ligger det ikke opp i dagen hva det vil si *å bli* kyndig i elevtekstvurdering. Hvordan lærer man seg dette? Og viktig i en forskningskontekst, hvordan undersøker man om noen har utviklet seg som vurderere? Det følger av disse spørsmålene at det er et behov for en avklaring av begrepet *læring*. I tillegg er det ikke gitt hva det vil si *å være* en kyndig tekstvurderer. Hva kan man når man er kyndig i denne sammenhengen? *Kunnskapsbegrepet* vil altså også bli diskutert. I dette arbeidet vil jeg forsøke å kontrastere mine posisjoneringer mot monologiske alternativer, og da gjerne de teorier som ligger bak den dominerende praksisen som ble presentert i litteraturgjennomgangen.

4.2 Syn på læring

I grunnbøker om læring blir gjerne tre hovedtradisjoner presentert (for eksempel Imsen, 1998; Jordan, Carlile & Stack, 2008): Behaviorisme, kognitivisme og ulike former for sosialt og kulturelt orienterte læringsteorier. De er her nevnt i tilblivelseshistorisk korrekt rekkefølge, og selv om det er en tendens til å se på rekkefølgen som representasjon for fremskritt, er det nok heller slik at de fleste praksiser hviler på en kombinasjon av ulike teorier (Baird, Hopfenbeck, Newton, Stobart & Steen-Utheim, 2014; James, 2006). Desto viktigere blir det derfor å gjøre rede for ens egen forståelse av fenomenet. Det skulle nå være klart at jeg holder interaksjon for å være en forutsetning for læring. Denne interaksjonen trenger imidlertid ikke å være en ytre interaksjon mellom to personer; å tenke kritisk innebærer for eksempel evnen til reint kognitivt å konstruere for- og motargumenter, det vil si å innta ulike posisjoner. Men denne evnen til å snakke med seg selv krever nettopp et språk. Og det språket fortsetter å bli lært gjennom sosial deltakelse.

Et slikt syn på læring bygger på Lev Vygotskys arbeid (1978, 1987). En populær framstilling av Vygotskys tenkning er at han ser på læring som en prosess som går fra det sosiale og innover mot det kognitive, i stedet fra det kognitive og utover til det sosiale. Vygotsky skriver selv: «All the higher functions [of psycological processes] originate as actual relations between human individuals» (1978, s. 57). En slik forståelse av læring som sosial gjør krav på hjelpemidler som binder den enkelte til verden rundt. Vygotsky hevder i så måte at evnen til handling på grunnlag av «høyere mentale funksjoner» er forbundet med evnen til å

beherske tekniske og psykologiske redskaper. Slike *medierende midler* (Wertsch, 1991, 1998) er menneskets forbindelsesledd med omverden og følgelig det som gjør det mulig for oss å orientere oss i verden slik vi gjør (Säljö, 2001). Vygostky holder de semiotiske ressursene, og da spesielt verbalspråket, for å være de viktigste medierende midlene vi tar i bruk i så måte. Språket er spesielt viktig på grunn av dobbeltrolla det har: Språket er både ei forutsetning for læring og et resultat av læring (Igland, 2008, s. 52).

Med inspirasjon fra Vygotsky har fokuset på «en proksimal utviklingssone» dreiet oppmerksomheten mot læringspotensialet som kan realiseres forutsatt adekvat støtte (Dysthe, 1995; Hoel, 2000) eller hjelp fra "en venn", jf Lars Sigfred Evensens tittel om Vygotskys læringsteori (Evensen, 2009). Denne sosiale og kontekstuelle forankringa av læring har gitt støtet til det vi i dag omtaler som *sosiokulturell* læringsteori. Læring skjer i et sosialt samspill mellom enkeltmennesker ved hjelp av medierende artefakter. Uten å nedvurdere viktigheten av denne siden av Vygotskys teori, unnslipper de kognitive aspektene ved Vygotskys arbeid dersom man ikke samtidig legger vekt på internaliseringsprosessene i læringsforløpet. Vygotsky mener nemlig at læring skjer i to omganger (Vygosky, 1978, s. 57). Først gjennom mikrososial interaksjon, hva han omtaler som *intermentale* prosesser. Dette skjer gjennom repetisjon og etterligning. Etter hvert vil imidlertid disse prosessene gjøre den enkelte i stand til å mestre aktiviteten på egen hånd, hva Vygotsky omtaler som *intramentale* prosesser. Det man har øvd på i sosiale settinger, gjennom støtte og egeninnsats, har man til slutt internalisert og evner å gjøre alene.

Den sosiale dimensjonen som er knyttet til læring gjør det vanskelig, for ikke å si umulig, å anse seg som utlært i noe. Læring som fenomen er altså ikke et spørsmål om å kunne eller ikke kunne. Innenfor sosiokulturelle rammer vil læring bedre forstås som å gå fra *lite* til *mer*, fra *grunt* til *dypere* (Evensen, 2013, s. 115). Evensen omtaler dette som «spirallæring» og argumenterer for at et slikt læringssyn blir «a way of life» (2013, s. 116). Nå vet ikke jeg om vurdererne i Vurdererpanelet ser på vurderingsarbeidet som et livsprosjekt, men denne måten å forstå læring på gir klangbunn i den vurderingspraksis jeg observerer de tar del i. En drøfting av dette basert på datamaterielat (dvs. vurderingsdialogene) er presentert i delstudie 2 og blir derfor ikke nærmere diskutert her.

Jeg avslutter delkapitlet med å påpeke hvordan det læringssyn som her er presentert, slett ikke er hegemonisk og selvsagt, og ikke minst vise hvordan praksis fort lar seg diktere av teori, innenfor sentralt administrerte, gjerne såkalte «high stake» skriveprøver. I en studie

presenterer Elder, Barkhuizen, Knoch & von Randow (2007) resultat fra et online opplæringsprogram hvor tekstvurderere vurderer på forhånd vurderte tekster (såkalte "benchmark texts"). Kvaliteten i vurderingsarbeidet, og dermed kvaliteten av opplæringsprogrammet, blir deretter evaluert på grunnlag av hvor nært vurdererne kommer vurderingene av de på forhånd vurderte tekstene. Treningen består i å vurdere helt til deltakernes vurderinger blir lik vurderingene av standardtekstene. Det ligger noen klare forutsetninger om hva en tekst er, hva tekstvurderernes arbeid består i, samt hvordan de går fram for å bli bedre vurderere innebygget i en slik måte å forsøke å skape bedre vurderinger på. Disse forutsetningene sammenfaller godt med følgende definisjon av monologisme: [T]he constituent theories of monologism are the information processing model of cognition, the transfer model of communication, and the code model of language" (Linell, 2009, s. 36). I Elder m.fl. sin studie er de opptatt av forbedret vurderingskvalitet, men det er helt tydelig at det er andre læringsmodeller de bruker for å designe opplæringsprogrammet. For det første er det snakk om énveis overføring av korrekt vurderingspraksis kontrollert gjennom øyeblikkelig respons på adferd i form av "korrekt" eller "feil" vurdering (behaviorisme i praksis). Og for det andre snakker forskerne om "training input" og vurderernes mulighet til å monitorere egen oppførsel (kognitivisme i praksis).

Selv om jeg her trekker fram én enkeltstudie som eksemplet for å vise hvilke forståelsesrammer denne avhandlingen ikke har, mener jeg det ikke er snakk om misbruk. Tvert imot; med støtte i litteraturgjennomgangen er det grunn til å påstå at Elder m.fl. sin studie er en prototypisk representant for en særdeles utbredt måte å behandle tekstvurderingskompetanse og utvikling av sådan på. Det dialogiske, dialektiske, det komplekse og sammensatte er ikke engang problematisert. Det samme gjelder for det kontekstuelle, som jeg nå går over til.

4.3 Vurderingsfellesskap

Det er blant andre John Dewey som har påpekt at kunnskapsutvikling skjer i situerte sosiale arenaer. Det vil si at vi lærer gjennom handling innenfor bestemte kontekster (Dewey, 2008). Kontekstens betydning, eller situasjonen som Dewey omtalte den, har senere utviklet seg til et eget forskningsområde innenfor sosiokulturell læringsteori. En sentral representant i dag er Etienne Wenger (Lave & Wenger, 1991; Wenger, 1998). Hovedinteresseområdet hans er menneskers deltakelse i praksisfellesskap.

Wenger poengterer at det for et praksisfellesskap er sentralt at den har et anliggende som skiller seg fra andre, at den etablerer et fellesskap gjennom felles læringsaktiviteter, diskusjoner og informasjonsdeling, og at gruppa etablerer felles praksis (Wenger, 1998). For Vurdererpanelet betyr det henholdsvis at de som gruppe eksklusivt vurderer elevtekster, de bygger et faglig fellesskap gjennom diskusjoner om elevteksters kvaliteter, samt gjennom kursing om skriving, tekst og vurderingsarbeid, og de etablerer felles praksis gjennom reelle og faktiske tekstvurderinger. Det knytter seg altså kontekstuelle forventninger og føringer til de kunnskapsbaserte handlingene man utfører.

Siden mitt primærmateriale er vurderingsdialogene, er jeg i delstudiene tettere på de enkelte mindre par- og gruppekonstellasjonene enn jeg er på Vurdererpanelet som helhet. I delstudie 2 drøfter jeg likevel balansen mellom deltakerautonomi og normgivende ressurser innenfor et praksisfellesskap generelt og Vurdererpanelet spesielt. Jeg skal ikke her gjenta meg, bare kommentere at denne tenkningen omkring det å være en del av noe større samtidig som man påvirker dette større, er en av dialogismens grunnpilarer slik det er beskrevet tidligere, og slik det vil bli presentert under om selve vurderingsdialogene.

Et kritisk aspekt ved vurderingsfellesskapet sett som et praksisfellesskap er knyttet til den situasjonsoverskridende normerende praksis som ligger intendert i designet. Vurdererne vurderer tekster fra et utvalg elever på femte og åttende årstrinn. Som sådan er skriveprøven ment å være normsettende og læringsstøttende (http://www.udir.no/Vurdering/Laringsstottende-prover/). Samtidig er det arbeidet som har framskaffet disse normene forankret i Vurdererpanelet; forståelsen av hva skriving er og evnen til å kjenne igjen kvalitet i tekstene er situert i et bestemt praksisfellesskap. Det går ut over rammene for denne avhandlinga, men på sikt vil det være interessant å se om tekst- og skrivesynet som er opparbeidet innenfor Vurdererpanelet når ut til elevene i de enkelte klasserom. I en praksisfellesskapsterminologi vil det være snakk om å stimulere den enkelte vurderer til å ta jobben som ”boundary spanner” (Bergenholtz, 2011) mellom ulike tekstvurderingspraksiser. (Klasserommet og Vurdererpanelet er i tilsvarende terminologi ”boundary practices” (Wenger, 1998, s. 114-115)). Skal vurderingspraksisen innenfor Vurdererpanelet lykkes i å virke normerende, er man avhengig av at det skapes mange nok forbindelser ut til skrivelærerne i skolen slik at de på tilsvarende måter kan forbedre egen tekstkompetanse. Nettopp det at skrivevurdering ikke er en instrumentell handling som bare oppskriftsmessig kan overføres fra den ene til den andre, er noe av det som går som en rød

tråd gjennom de ulike delstudiene i denne avhandlinga. Det samme vil da gjelde for lærerne ute i skolene.

4.4 Vurderingsdialogen

Selv om Vurdererpanelet markerer de sosiale rammene for denne studien, er det vurderingsdialogene vurdererne er involvert i når de vurderer den enkelte elevtekst, som er det sentrale datamaterialet. Vurdererne sitter hovedsakelig parvis, noen ganger i små grupper, og vurderer kvaliteten på de enkelte vurderingsområdene til den enkelte tekst. Grunnen til at de vurderer sammen er todelt; for det første vil hver enkelt tekst bli vurdert likt av minst to vurderere (vurderingsparet). Målet er at slik sampraksis raskere skal føre til lik(ere) vurderingspraksis med påfølgende høyere vurderingssamsvar (interrater reliability). For det andre fører slik samvurdering til kunnskapsdeling og kunnskapsutvikling. Vurdererne blir "presset" til å sette ord på hvorfor de mener en tekst er av en bestemt kvalitet. Hva som er holdbare argument blir i par- og gruppesammenheng testet mot kollegaers oppfatninger. I begge tilfeller følger det av avhandlingas dialogiske forankring at dette sosiale samspillet blir holdt for å bli best forstått som en meningskonstruerende prosess. Det vil si at tenking og kommunikasjon er samtidige prosesser i vurderingsdialogene. Folkelig heter det riktignok at man skal tenke før man snakker, men i praksis er det nok likevel slik at «snakket» fungerer som tankens redskap (jf. Vygotskys forståelse av semiotiske ressurser som forutsetning for høyere kognitive ferdigheter).

Når samtalen likevel ikke fremstår som kaotisk og usammenhengende, er det fordi det i dialogene ligger noen strukturerende prinsipper som skaper sammenheng (jf. Linell, 2001, s. 67 ff.). I vurderingsdialogen tar den enkelte vurderer ansvar gjennom deltakelse. Hvert dialogisk bidrag fungerer både som et svar på det som har vært sagt og som styrende for hva som vil bli sagt. Et bidrag peker altså både bakover og fremover og blir således meningsfullt også gjennom hvor det er plassert i en *sekvensiell kjede* av flere bidrag. Det at ei ytrings betydning slik er «videre» enn ordenes betydning (semantikken), peker mot dialogens *interaksjonelle* karakter. Det at det som blir sagt er rettet mot den andres bidrag, gjør at dialogen er en felles konstruksjon. Dette igjen, det at samtalen er en felles konstruksjon, peker mot samtalens sosialt skapte rammer, det *sjangerspesifikke* ved dialogen. Og på samme måte som det enkelte bidrag i dialogene både er en respons på noe som er sagt og et initiativ i forhold til hvordan man ønsker samtalen skal utvikle seg, er den enkelte vurderingsdialog

svarende i forhold til tidligere vurderingsdialoger og justerende i forhold til fremtidige tilsvarende dialoger.

Dialogens sekvensialitet, interaksjonalitet og sjangerspesifikke karakter skaper de rammer som gjør meningskonstruksjon mulig (jf. Markova og Linell, 1996, s. 357). I grunnleggende forstand er det nettopp disse sidene ved dialogen som har vært denne avhandlingas undersøkelsesområde. Jeg har vært interessert i hvordan vurdererne kommer til enighet (skaper mening) om tekstkvalitet i en kontekst hvor vurdererne er sentrale bidragsytere i arbeidet med å utvikle felles forståelse for hvordan nettopp kvalitet gjenkjennes i elevtekstene. Ved å undersøke vurderingsdialogen, ikke vurderingsresultatet, ligger det en anerkjennelse av at gyldigheten av vurderingsarbeidet er å spore tilbake hit, i vurderernes streben etter å gjøre et meningsfullt (holdbart) arbeid.

I delstudiene beveger jeg meg delvis på et mikronivå, delvis på et mesonivå. I delstudie 1, hvor jeg påviser sammenheng mellom triangulering av vurderingsgrunnlaget på den ene siden og et rikt og heterogent responsmøster på den andre, er jeg tett på det enkelte bidrag i dialogene. Det samme er tilfelle i delstudie 2 hvor jeg viser hvordan vurderernes remedieringspraksis avslører hvilken begivenhet (tekst) som er gjenstand for oppmerksomhet, og som dermed avdekker hvordan det å ta i bruk vurderingsressurser er en kognitivt krevende oppgave. Men fordi arbeidet vurdererne gjorde under datainnsamlinga langt på vei var et nybrottsarbeid (de var noviser i konteksten, Vurdererpanelet var nyetablert og vurderingsressursene var tentative), har det også vært interessant å studere dialogene i lys av den situasjonsoverskridende tradisjonen, altså på et mesonivå. Hvordan tar vurdererne sammen i bruk nye ko(n)-tekstuelle ressurser innenfor Vurdererpanelets rammer der hver især bringer med seg idiosynkratiske vurderingspraksiser? Hvilke diskursive strategier viser seg å fungere, det vil si lede til enighet og beslutning? Det er problemstillinger knyttet til disse spørsmålene som hovedsakelig blir diskutert i delstudie 3. Dette er interessant fordi disse strategiene vil raskt kunne etablere seg som normgivende praksis, det vil si som en egen sjanger.

4.5 Hva det vil si å kunne

I avhandlinga er jeg altså opptatt av å belyse hva vurdererne *gjør med ord* (jf. talehandlingsbegrepet (Austin, 1962; Searle, 1965); hva de holder for å være godt og dårlig i elevtekstene, hvordan de snakker seg fram til forståelse av hva ressurspersoner og

tekstressurser uttrykker, men også hvordan talehandlingene i vurderingsdialogene endrer seg med øving og erfaring. Det at vurdererne er del av et Vurdererpanel hvor vurderingsarbeid går hånd i hånd med plenumsdiskusjoner og fagrelevante forelesninger, er et uttrykk for et ønske om at vurderingsarbeidet skal ha en retning; det skal bli «bedre». Men hva er «bedre» i denne sammenhengen? Hva vil det si å ha tekstvurderingskompetanse? Jeg skal presentere et forslag til hva det kan være innenfor en dialogisk epistemologi som har vært skissert i dette kapitlet, men jeg velger å gå veien om hva det kan bety at man i det hele tatt *kan* noe.

Mennesket har vært opptatt av hva kunnskap er til alle tider, men vi har ikke alltid hatt samme svar på hva kunnskap er og hva som er verdien av kunnskap. Det epistemologiske spørsmålet «hva kan vi med sikkerhet vite?» har vært stilt sammen med mer pragmatiske og nytteorienterte spørsmål som «hva skal vi bruke kunnskapen til?». I dagens kunnskapssamfunn(!) er kunnskap gjerne knyttet til en ressursdiskurs. Om kunnskapssamfunnet kan vi lese følgende i Stortingsmelding nr. 30, Kultur for læring: «De viktigste innsatsfaktorene i arbeidslivet er ikke lenger kapital, bygninger eller utstyr, men menneskene selv. Statistisk sentralbyrå har anslått at 80 prosent av den norske nasjonalformuen består av menneskelige ressurser». Vi har altså gått fra et industrisamfunn og kapitalsamfunn til et kunnskapssamfunn (jf. Drucker, 1959; Cetina, 1999). Men hva legger vi i kunnskapsbegrepet i dag? Hva er det vi kan og hva er det som settes pris på av kunnskap? Dette er grunnleggende og sentrale spørsmål som kan virke fjernt fra det som har vært undersøkt i dette avhandlingsarbeidet, men vi skal snart se at nettopp avklaring av *kunnskapssyn* er sentralt når man søker å forstå tekstvurderingskompetanse blant utvalgte, og etter hvert skolerte, vurderere i en bestemt kontekst.

Det finnes dominerende tradisjoner vi kan følge tilbake helt til antikken. Det moderne vitenskapssyn som karakteriseres av en tiltro til sikker viten, har sine røtter i Platons formlære. Med han blir vi introdusert for den rene kunnskapen, *episteme*, som er sikker og udiskuterbar (Platon, 2001). For Platon finnes det virkelige og sanne i abstraksjonene, og ikke i den interessefylte verden slik mennesket sanser den. Episteme har en motsats i *doxa*. Doxa er uttrykk for menneskers erfaringsbaserte oppfatninger, meninger og handlinger og som dermed blir usikre, flyktige og upålitelige. Artefakter holder Platon for å være manifestasjoner av den samme doxa. Tingene rundt oss er derfor like usikre som erfaringene våre av dem. Aristoteles bygger på Platon i sin kunnskapsfilosofi og deler forgjengerens syn angående den vitenskapelig-teoretiske kunnskap, eller episteme. Uten å knytte episteme til en abstrakt

idéverden, er den også for Aristoteles en kunnskap som ikke kan være annerledes: «Viten er oppfatning om ting som er almene og som er til med nødvendighet» (Aristoteles, 1999, s. 62).[9] Følgelig holder Aristoteles episteme for å være den høyeste form for kunnskap, og eksempler vil være matematikk og geometri. Men til forskjell fra Platon holder ikke Aristoteles doxa for å være upålitelig og falsk. I stedet sier han at menneskers tilvirkninger og handlinger potensielt representerer to ulike praktiske kunnskapsformer. Den ene, *techne*, er den tilvirkningskunnskapen vi kjenner igjen fra håndverket. Vi kan kalle den for praktisk-produktiv kunnskap, og ferdigheten som er knyttet til den er *å kunne*. Den andre praktiske kunnskapen er knyttet til de *overveielser* en gjør forut for en *handling*. Den som overveier godt i enkelttilfellene besitter det Aristoteles betegner som fronesis, eller klokskap. Vi ser her at Platon og Aristoteles, på tross av forskjeller, holder én bestemt kunnskap, den vitenskapelig-teoretiske, som overlegen andre kunnskapsformer. Grunnlaget er en forestilling om et absolutt skille mellom kropp og sinn, mellom sansing og kognisjon som på bestemte områder har vist seg å være suksessfullt.

Det var imidlertid langt senere, under 1600-tallets rasjonalisme, at denne forestillingen ble etablert som den absolutt dominerende epistemologi: Sikker kunnskap deduseres fra absolutt sikker viten. Med Descartes kan vi si at vi gikk fra å akkumulere kunnskap på usikkert grunnlag til å stille oss det grunnleggende spørsmålet: Hva kan vi med sikkerhet vite? Når Descartes og samtidas filosofer og vitenskapsmenn forsøkte å svare, fant de forbilder i disipliner som er løsrevet fra sanseerfaringer (matematikken og logikken). Med disse disiplinene som utgangspunkt utviklet Descartes en metode som skulle begrense muligheten for å bli lurt av sansene ved empiriske studier (som han tross alt holdt for å være de mest nyttige). Dersom ideen om objektiv og absolutt sikker viten (episteme) var det vitenskapsteoretiske utgangspunktet for den moderne vitenskap, kan vi si at med Descartes' systematiske tvil fikk den moderne vitenskap sin metode.

I Descartes arbeid ligger det en interessant "glidning" i forholdet mellom episteme, techne og fronesis som er viktig. Techne har nærmet seg episteme. Vi snakker gjerne om naturvitenskapens interesse i å *forklare* fenomener i verden og samfunnsvitenskapens interesse i å *forstå* (mennesket i) verden (jf. Dilthey, 2010). Dette skillet handler om at naturvitenskapene primært er opptatt av kausale relasjoner, mens samfunnsvitenskapene

[9] En diskusjon omkring de ulike kunnskapsformene blir mest systematisk presentert i Bok VI, side 58-71.

primært er opptatt av hermeneutiske relasjoner. Forklart på en annen måte: naturvitenskapen er objektiv på den måten at den er uavhengig av mennesket, mens samfunnsvitenskapen nettopp er opptatt av mennesket. På Descartes' tid (og frem til i dag?) var et slikt skille uproblematisk: så lenge falsifikasjon (Descartes' metodiske tvil) er idealet, er kunnskapen vi får "sann" i nærmest epistemisk forstand.[10]

En likefram forståelse av kunnskapsbegrepet er knyttet til at den vitenskapelig-teoretiske kunnskapen har hatt høy status, den praktiske og tekniske kunnskapen har dratt fordeler av å kunne løsrive seg fra det ustabile (mennesket), mens kunnskap knyttet til menneskelig adferd, forståelse og meningskonstruksjoner har hatt lav status eller har rett og slett vært oversett.[11] Konsekvensen har vært at det som har vært verdsatt i kunnskapssamfunnet (sic!) i hovedsak har vært den abstrakte og tekniske kunnskapen. Vi kjenner det igjen fra historien, gjennom opphøyelsen av disipliner som filosofien, matematikken, metafysikken, teologien og senere, naturvitenskapen. Men de fleste vil vel også kjenne det igjen fra vår egen samtid; fra tendensen til å skape "teoritunge" (les abstrakte) fagutdanninger, fra naturvitenskaplig baserte statusyrker, fra fjernsynsprogrammer som *Hjernevask* og *Folkeopplysningen*, ja kanskje endog fra den (forhåpentlige) anerkjennelse et arbeid som det du leser nå avstedkommer. Stemmer i samfunnsdebatten uttrykker det samme. I et debattinnlegg skriver Lena Lindgren at praktisk kunnskap i dag «strupes av byråkratiske styringsmodeller basert på det kvantitativt målbare. Vi ser det i utdanningssystemet i en banal dyrkelse av teoretisk kunnskap på bekostning av praktisk kunnskap» (*Forskerforum* 2014 9, s. 40-41).[12]

Bildet er selvsagt ikke så entydig, og senere har det etter hvert så tradisjonelle teoriladete kunnskapssynet blitt utfordret. På den ene siden er det blitt gjort ved at man i det hele har stilt spørsmål ved forestillingen om den vitenskapelig-teoretiske kunnskapens forrang, slik vi kanskje kan forstå Deweys forsøk på omveltning av relasjonen mellom teori og praksis i utdanningen, eller kanskje slik vi i helt nyere tid kan forstå miljøbevegelsens arbeid

[10] Det følger av utlegningen i dette kapitlet at det å anvende begrepsparet forstå-forklare til å markere skille mellom de ulike vitenskapene er problematisk så lenge mening og erkjennelse er avhengig av mennesket.
[11] Fra et annet ståsted er det den samme historien Evensen presenterer når han vurderer fagdisiplinen *anvendt språkvitenskap*s status i forhold til *teoretisk lingvistikk* (Evensen, 2013, s. 27 ff.).
[12] Innlegg holdt ved Høgskolen i Oslo og Akershus 07.10.2014 i forbindelse med lanseringen av boka *Profesjonshistorier*, redigert av Rune Slagstad og Jan Messel. Lindgren bruker her "praktisk kunnskap" til å beskrive mer håndverksmessige ferdigheter, og ikke "teknisk kunnskap" som konnoterer mer til naturvitenskapen.

med å finne politiske løsninger på klimautfordringen. På den andre siden er det blitt gjort ved at man har problematisert selve inndelingen av ulike kunnskapsformer som tilhørende ulike aktiviteter. Er det virkelig slik at kunnskapsformene tilhører ulike domener, eller er det slik at innenfor de fleste profesjoner opptrer kunnskapsformene som ulike dimensjoner ved aktiviteten? Schöns begrep "refleksjon-i-handling" (Schön 2004) kan her fungere som eksempel på et forsøk på å forstå kunnskapsdimensjonenes samtidighet. Det er i kjølvannet av disse strømningene samfunnsdebattant Lindgren mener å se konturene av en opposisjon mot episteme-hegemoniet fra profesjonsutdannede. Hun skriver: "Det er en opposisjon mot økonomisk instrumentalisme – som i narsissismens mykere vokabular gjerne kalles for visjoner, fleksibilitet, identitet, omstilling. Det er en kamp for faglighet vi ser – en kamp for fronesis" (op.cit., s.41).

4.6 Hva det vil si å være en kompetent tekstvurderer

Der Aristoteles gjerne presenterte ulike aktiviteter som kjennetegnet av ulike kunnskapsformer, ser vi altså i senere tid en tendens til å se på disse ulike kunnskapsformene som dimensjoner *innenfor* den enkelte aktivitet, eller den enkelte profesjon. Ved å vektlegge en profesjons handlinger, gis de samme handlingene moralsk karakter. Det man gjør har konsekvenser, og det er den som utfører handlingen som har ansvaret for disse konsekvensene. Spørsmålet er om det ikke er slik at det i de fleste profesjoner avkreves samtidige kunnskapsformer for å være kompetent. Hvis ikke blir det fort at man "vert fagidiot på sitt eige omkverve og vanleg idiot på alle andre" (Hellesnes 1992: 81). Jeg vil hevde at den gode vurderer slettes ikke er en idiot[13], men heller en praktiker som både er teoretisk sterk og viser evne til situasjonsvarhet.

Schöns (2004) utgangspunkt for drøftelsen av kunnskapsbegrepet er et teori-praksis-forhold i profesjonsutdanninger der gammel praksis ikke ser ut til å være vellykket. Han mener nemlig å se at profesjonsutdannerne paradoksalt nok har lite kontakt med praksis selv om profesjonsutøverne løser sine oppgaver på bakgrunn av erfaringer, intuisjon og kunnskaper som sjeldent er erverva gjennom utdanningsinstitusjonene. Schön mener dette skyldes et kunnskapssyn hos utdanningsinstitusjonene som var preget av en forståelse av profesjonell aktivitet som "en instrumentel løsning af problemer, som er regelsat gennem anvendelsen af

[13] Vi kan riktignok argumentere for at *idioten* bare er den *særegne* som de *idiografiske* vitenskaper faktisk er interessert i. Men det vil nok være forkjært forstått i denne sammenhengen.

videnskabelig teori og teknikk" (2004, s. 29). Denne tekniske rasjonaliteten mener Schön ikke kan gjøre rede for praktisk kompetanse i komplekse situasjoner.

Det er i denne sammenheng Schön presenterer sin alternative kunnskapsforståelse, uttrykt som "refleksjon-i-handling". I begrepet ligger det at det ikke er slik at man innafor en profesjonell praksis først lærer noe teoretisk for så å anvende teorien i praksis, som et sekvensielt utført handlingsmønster. I stedet er det slik at når eksempelvis en vurderer utfører en kompetent vurdering av en elevtekst, så foretar hun én handling fordi vurderingskompetansen vises ikke som kunnskap før i praktisk handling, her vurderingssituasjonen. For Schön er det slik at den kvalifiserte profesjonsutøver arbeider med utgangspunkt i et reportoar av eksempler, men samtidig som han ser det unike i lys av det kjente, lar han det ikke passe inn. Det nye vil inneholde elementer som også er forskjellig fra reportoaret, og det å se noe som unikt, vil forhåpentligvis føre til at man også handler unikt.

Det å behandle en situasjon som unik er det motsatte av å handle instrumentelt. Mønstre vil gjerne dannes på bakgrunn av enkeltepisoder og enkeltobservasjoner, men de ulike episodene og observasjonene tvinges ikke inn i noen mønstre. Det er her vi ser at en slik situasjonsvarhet krever praktisk klokskap, eller fronesis. Det er med andre ord ikke nok å vite hva som er «rett» og «sant» eller å være effektiv og funksjonell, men man må også vurdere konsekvensene av de beslutninger man gjør på grunnlag av teoretisk og praktisk kunnskap. Hva har så denne forståelsen av kunnskap å gjøre med vurderernes vurderingsarbeid? Jo, mye, vil jeg hevde. La meg illustrere med et utdrag fra datamaterialet. Her er det tre vurderere som vurderer en teksts kommunikative kvaliteter. I vurderingsskjemaet (vedlegg 2) er fem ulike støttespørsmål knyttet til denne tekstdimensjonen, og disse spørsmålene er ment å veilede vurdererne i vurderingsarbeidet. Følgende spørsmål ser imidlertid ut til å gi vurdererne problem: «*Er skriveren til stede i teksten på en relevant måte?*».

Juni 2011. Vurdererne Tom, Trine og Stein.

TOM Men jeg er litt usikker på om hun er tydelig til stede. Så må ofte tenke på det, for hun--. Hun prater på vegne av åttende trinn og prater på vegne av lærerne. Ehm:

(…)

TRINE Ja, men ville det vært positivt om hun tok og snakket på vegne av bare seg personlig?

TOM Nei, men skal vi vurdere om det er positivt eller ikke? Eller om hun--

STEIN Nei altså, er, er den som henvender seg til rektor, er hun til stede i teksten? (…) På en relevant måte, det synes jeg. Altså, tenker en ikke på at du--, at du er saklig og tydelig i det du vil formidle til rektor? Er det ikke det du tenker på da?

TOM Jeg er litt usikker hva du egentlig--. Hvordan vil hun IKKE være til stede i teksten? For eksempel.

(…)

STEIN Hm:

TOM Og hvordan ville hun da skrevet om hun skulle vært MER til stede i teksten?

TRINE Mer til stede så ville det kanskje vært litt mer muntlig språk og--. ”Når JEG ofte tar med meg--”.

TOM Ja.

STEIN Ja:

TRINE «Så synes JEG det er tungt da».

STEIN Mhm:

TRINE Ehm:

TOM ”Egen bær-”, «egen bærbar datamaskin vil hjelpe meg på mange måter».

TRINE Mhm.

TOM «Datamaskin vil hjelpe meg å tyde skriften og få med alle ordene riktig fordi jeg og mange andre har problem med slike ting».

STEIN Mhm.

TOM Da hadde hun i hvert fall vært mer til stede. Men jeg vet ikke om hun ikke er til stede allikevel.

TRINE [Mhm].

TOM [Hehe].

TRINE Ja. Om du skal--. Det er ikke ledetråden da som kanskje misviser oss litt? For vi skal jo vurdere om hun er--, om hun kommuniserer godt. Og jeg synes det er mer positivt på den måten hun har gjort det enn om hun hadde gjort det sånn som du sier.

TOM Mhm.

TRINE Og vært veldig tydelig til stede.

STEIN [Mhm].

TOM [Ja].

TRINE Eh. Så kanskje denne ledetråden ikke får fram nødvendigvis hva som er--

TOM Ja.

TRINE --det mest positive med teksten.

STEIN [Mhm].

TRINE [For] det er jo egentlig det som er interessant.

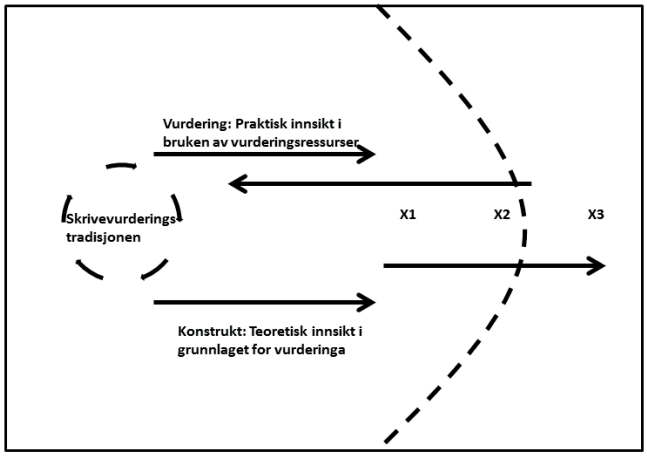
TOM Ja, må jeg velge mellom et ja eller nei, så ville jeg valgt et ja.

Tom, Stein og Trine opplever en konflikt mellom normerende vurderingsressurser, egne vurderingserfaringer og hva som vil gagne eleven. Et sentralt spørsmål vil da være: Hvordan løser en kompetent vurderer slike situasjoner? Hvilket svar man gir avhenger av hvilken rolle man anser at vurdereren har. Flere steder i delstudiene mine, og også her i litteraturgjennomgangen, har jeg problematisert det som synes å være det vanligste svaret på spørsmålet; nemlig at vurdereren skal holde seg til vurderingsressursene, skal avstå fra å problematisere den iboende forståelsen av skriving som ligger i konstruktet, og for all del skal unngå å reflektere over hva de enkelte vurderingsbeslutninger betyr for den anonyme eleven som har skrevet teksten (jf den monologisk funderte skriveprøvemodellen beskrevet av Meadows og Billington (2005, s. 48). Grunnen til at dette har vært det gjengse svaret er selvfølgelig at en slik vurdereradferd høyner muligheten for å oppnå pålitelige vurderinger. Jeg vil legge til, i alle fall på kort sikt (jf. Evensen, 2014).

I det hele tatt ser det ut til at sentralt organiserte skriveprøver ofte er administrerte og organiserte på måter som speiler godt den aristoteliske kunnskapsinndelinga. Bejar deler (2012) skrivevurderingen inn i to hovedfaser; designfasen og vurderingsfasen. Designfasen er det forskerne som har hånd om. Her valideres konstruktet, vurderingsskjema utvikles, pilotering gjennomføres og eksempeltekster velges ut. Kunnskapsformen som dyrkes synes tydelig, nemlig den teoretisk-vitenskapelige. I vurderingsfasen kommer imidlertid en annen gruppe inn, vurdererne. De er valgt ut for å vurdere tekstene i lys av det utarbeidete vurderingsskjemaet. Gjennom trening har de etablert mer eller mindre like mentale representasjoner av vurderingsskjemaet og arbeidet deres består i å jamføre disse med de mentale tekstrepresentasjonene som etableres ved tekstlesningene. Vurdererne skal *gjøre* noe, de skal omkode komplekse tekstdimensjoner til bestemte mestringsnivå. Kunnskapsformen som her beskrives kjenner vi igjen som den mer håndverksmessige. Vurdererne skal trenes opp til relevant vurderingspraksis. Det synes altså meningsfullt å forstå en slik dominerende skriveprøvepraksis i lys av en kunnskapsinndeling i Aristoteles' egentlige forstand: Bestemte kunnskapsformer følger bestemte profesjoner. Følgelig blir det tilnærmet vanntette skott mellom designerne av skriveprøvene og vurdererne. Dette forsterkes igjen ved at skriveprøvene er hierarkisk organiserte. Vurderernes praksis blir i enden godkjent,

korrigert eller underkjent av de som er lenger opp i hierarkiet (jf. Meadows og Billington, 2005).[14]

Det er mot denne forestillingen jeg foreslår at det er mer gagnlig å forstå tekstvurdere som kompetente om de blir gjort ansvarlige gjennom konstruktinnsikt, konstruktpåvirkning og profesjonsautonomi. Under (Figur 2) presenteres en kunnskapsteoretisk modell for skrivevurdererkompetanse. Modellen er en tillempet versjon av Paul Otto Brunstads generelle profesjonskompetansemodell (2007).



Figur 2: Kunnskapsteoretisk modell for skrivevurdererkompetanse (etter Brunstad, 2007).

Modellen synliggjør hvordan praktisk og teoretisk kunnskap former skrivevurderingstradisjonen og at denne kunnskapen ikke strekker til når vurderinger skal tas fordi den nye situasjonen alltid er unik (X1). For å foreta de rette valgene må en derfor være villig til å ta sjanser (X2), men heldigvis minsker risikoen ved økt evne til å forutse følgene av de valg man tar (X3). Evnene til å forutse følgene er derfor avhengig av erfaring kombinert med praktisk og teoretisk innsikt. En slik økologisk forståelse av tekstvurderingskompetanse medfører at det å forutse følgene av vurderingshandlinger (X3) like gjerne handler om å forstå følgene knyttet til vurderingskonstruktet i abstrakt forstand som til å forstå følgene av den enkelte konkrete vurderingen i en mer isolert tekstlig forstand.

Disse egenskapene, å se det unike i situasjonen, å være villig til å ta risiko og å ha evne til å forutse følger, er kjernen i å utøve klokskap. Pilen som går tilbake fra vurderingshandling

[14] Et tilleggsmoment her er at ved en slik fordeling av arbeidsoppgaver har ansvaret en tendens til å forvitre. Det moralske ved handlingene blir mindre tydelige.

til fagtradisjonen symboliserer klokskapens forankring i og påvirkning på den praktiske og teoretiske kunnskapen som danner fagtradisjonen. Det betyr at opparbeidet erfaring på grunnlag av praktisk og teoretisk kunnskap gjør at man er bedre rustet til å foreta nye vurderinger i det operative området. Modellen får fram hvordan klokskap er å handle på grunnlag av praktisk og teoretisk kunnskap uten å la denne kunnskapen hindre en i å behandle enkelttilfellet som noe unikt.

Med den triadiske kunnskapsmodellen som bilde for hva som kreves av en vurderer for å foreta holdbare vurderinger, er det klart at prosessen med å validere skriveprøver kompliseres ytterligere. Som jeg har vært inne på er det tradisjon for å begrense vurdererens rolle i dette valideringsarbeidet til det som er knyttet til reliabilitet (både indre konsistens og vurderersamsvar). Med grunnlag i det jeg har skrevet fram i dette kapitlet, vil det være riktigere å si at i kvalitetssjekken av en skriveprøve er reliabiliteten av vurderernes arbeid en nødvendig, men ikke tilstrekkelig forutsetning for validitet. Vurdererne må i tillegg ha tilstrekkelig kompetanse til å foreta selvstendige vurderinger i hver enkelt vurderingssituasjon som yter elevteksten rettferdighet. Tekstvurdering er også en moralsk praksis. For at vurdererne skal være i stand til dette, må de involveres i de prosessene som skal lede fram til kunnskapsbaserte vurderinger, det som Bejar omtaler som designfasen (Bejar, 2012). Under forberedelsesarbeidet, som gjøres i forkant av (og underveis i) vurderingsarbeidet, blir vurdererne således «subjects in learning» i stedet for «objects of teaching» (Evensen, 2013, s. 191)).[15]

[15] Det er interessant at det arbeidet som legges ned av vurdererne for å forberede seg til vurderingsarbeidet i engelsk terminologi blir beskrevet som ”rater training”. Jeg vil foreslå «rater learning» som et alternativ til «rater training». Det er forskjell på å trene for å bli god og å gjennomføre opplæring for å bli god.

5 Metode

5.1 Sammenheng mellom forståelsesramme og tilnærming

Det er ingen overdrivelse å påstå at skrivevurdering ikke er en eksakt vitenskap. Strevet etter å oppnå pålitelige vurderinger er et vitnesbyrd på det. Den enkelte vurderer finner ikke den riktige vurderingen *i* elevteksten, men gjør en vurdering ut fra kunnskap om elevtekst, normer og andre kontekstuelle forventninger. Når jeg i denne avhandlinga har vært opptatt av vurderingskompetanse, er det nettopp fordi hva vurdererne legger merke til ved elevteksten, definerer verdien av teksten. Det er gjennom hvordan vurdererne *ytrer seg* om elevteksten, teksten blir gitt mening. Det at tilgang til kunnskap og erkjennelse på en slik måte går gjennom språket, blir gjerne forklart gjennom begrepet «the linguistic turn» (jf. tittelen på 1967-antologien redigert av Richard Rorty). Det lar seg ikke her gjøre å nøste opp de idémessige forløperne til et slikt språksyn (og en slik epistemologi), og heller enn å forsøke å gjøre det, vil jeg i stedet kort påpeke hvordan en slik vending har ledet til en beslektet vending mot praksis (jf. Schatzki, Knorr Cetina & von Savigny, 2001).

Som jeg var inne på i kapittel 4 oppgir fenomenologer ideen om at vi kan ha kunnskap om verden slik den *er* og hevder at det vi kan vite noe om er verden slik den *framstår*. Det vi kan erfare, vite noe om og sanse, det erfarer, vet og sanser vi med kroppen som «filter». Theodore Schatzki omtaler de ulike teorier som deler dette synet som utgangspunkt for viten, som «praksisteorier», og han skriver om disse: «Practice theory's embrace of embodied understanding is rooted in the realization that the body is the meeting points both of mind and activity and of individual activity and social manifold» (Schatzki, 2001, s. 9). Det er praksiser, og ikke ideer om idealer og abstraksjoner som blir holdt for å være det sentrale.

I antologien *Kvalitative metoder i et interaktionistisk perspektiv* beskriver Järvinen og Mik-Meyers hvilke konsekvenser en praksisorientert forståelsesramme bør få ved observasjonsstudier (Järvinen og Mik-Meyer, 2005, s. 97 ff).[16] For det første fremhever de at man bør interessere seg for hvordan aktører gjennom samhandling konstruerer meningsfull aktivitet. Man er altså ikke ute etter hva som *egentlig* foregår, kun det som *foregår*. Det finnes en rekke studier av vurdereradferd som avslører en tiltro til at det er mulig å avsløre hva som

[16] I begrepet interaksjonisme legger Mik-Meyer og Järvinen at det er i samspillet mellom mennesker/mennesker og artefakter at handlinger blir meningsfulle (2005, s. 10). Vi kjenner også igjen interaksjon som et av de tre grunnleggende trekkene ved dialogisme som Linell fremhever, ved siden av sekvensialitet og kontekstualitet (jf. kap. 4) (Linell, 2009, s. 31).

egentlig foregår i hodene til vurdererne. Titler som «Looking behind the curtain» (Connor-Linton, 1995) og «Holistic assessment: What goes on in the rater's mind» (Vaughan, 1991) er således betegnende. I dette avhandlingsarbeidet har jeg ikke hatt ambisjoner om på denne måten å nå «backstage». Heller holder jeg det som er interessant for å være fremme på «scenen», i denne sammenhengen i vurderingsdialogene. For det andre er man med en interaksjonistisk tilnærming opptatt av forholdet mellom praksis og de institusjonelle rammene for den samme praksis. Denne "doble dialogisiteten (jf. Linell, 2001, s. 54) retter oppmerksomheten mot forholdet mellom den fortløpende meningsproduksjonen og de historiske og samtidige kontekstuelle rammene for praksisen. For det tredje er man opptatt av hvordan aktører posisjonerer seg i forhold til hverandre og i forhold til hva som er handlingas målsetting. Ulik posisjonering bringer med seg bestemte rettigheter og plikter som danner foreløpige rammer for handling. For det fjerde, og da er vi tilbake til synet på språk som jeg innledet med, ser man på ytringer som handlinger. Det å ytre seg innebærer både å si noe om noe og å gjøre noe gjennom det man sier. Språket reflekterer altså ikke en virkelighet; den konstruerer virkeligheten. Et femte, og her siste, kjennetegn ved en samhandlingsorientert metodologi er at det man søker ikke er en form for objektivitet, men heller å analysere "*den objektiverede virkelighed*" (Järvinen & Mik-Meyer, 2005, s. 105). Med det menes det å søke kunnskap om den virkelighet som mennesker sammen gjennom praksis holder for å være meningsfull (jf. "pragmatisk konstruktivisme" som beskrevet i kapittel 4 (Kjørup, 2008, s. 183 ff.)).

Järvinen og Mik-Meyers poeng er at metodevalg og databehandling må speile valgt forståelsesramme. Det er krav om sammenheng mellom teori og metode. Hva betyr det for dette avhandlingsarbeidet? Jeg har vært opptatt av *hvordan* vurdererne snakker om tekstkvalitet, det vil si *hvordan de empirisk konstruerer* bestemte tekstforståelser. Etnometodologien tilbyr et rammeverk for å gripe en slik forståelse av forholdet mellom praksis og praksisutøver i det det er en teori om *hvordan* mennesker sammen skaper og opprettholder en felles forståelse av bestemte aktiviteter. Gubrium og Holstein skriver om etnometodologien at den forener en ""phenomenological sensibility" with an abiding concern for interactional process" (1997, s. 40). Retningen er opptatt av hvordan mennesker samhandler i konstruksjonen av en meningsfull hverdag. Hvordan mennesker interagerer språklig blir da sentralt. Det er denne dimensjonen ved etnometodologien jeg henter inspirasjon fra i mitt arbeid. Også jeg er opptatt av samtalen (vurderingsdialogen), og også jeg

holder språk for å være praksis (data), og ikke kun rapport av praksis (datakilde). Til forskjell fra en «sosiologisk etnometodologi» er jeg imidlertid ikke her opptatt av folks hverdagskompetanse, men heller en spesifikk profesjonskompetanse (se Heritage og Clayman (2010) om institusjonelle samtaler).

I kjølvatnet av etnometodologien utviklet Harward Sacks, senere særlig i samarbeid med Emanuel Schegloff og Gail Jefferson, samtaleanalysen som metode (se Sacks, Schegloff & Jeffersen, 1974). Gjennom dybdeanalyser av transkriberte lydopptak av samtaler, er målet å identifisere normgivende språkhandlinger innenfor bestemte praksiser.[17] Metodisk er måten jeg har analysert datamaterialet på inspirert av samtaleanalysen. Likevel, på to ulike måter distanserer jeg meg. For det første fokuserer den klassiske samtaleanalysen utelukkende på samtale*situasjonen*. Det betyr at det er hva som blir sagt og gjort, samt kontekstuelle faktorer som trekkes inn i analysene. Samtaleanalyse holder dermed de mer omfattende sosiokulturelle aspekter ved situasjonen utenfor, og det samme gjelder foreliggende skriftlig materiell. Dersom man ser på de ulike praksisteorier som innplassert på ei kontinuumslinje mellom situert interaksjon på den ene siden og situasjonsoverskridende praksiser på den andre, vil med andre ord en klassisk samtaleanalyse tilnærmet være en rendyrka interaksjonsanalyse. Delstudiene presentert i denne avhandlinga er på tilsvarende vis ikke rene interaksjonsanalyser. Tvert imot er både de videre kulturelle referanserammene og reifikasjonene av skriftnormer sentrale i studiene. Vurderernes praksis blir drøftet både i lys av kunnskap om vurdererkultur og i lys av deres forsøk på forståelse – og anvendelse – av skriftlige vurderingsressurser.

For det andre er samtaleanalysen som metoderetning kritisert for å være monologisk i den forstand at den er for lite oppmerksom på samtalens samarbeidende dimensjoner, eksempelvis ved at en sentral analysekategori innenfor samtaleanalysen er nærhetspar («adjacency pair») bestående av et initiativ med en påfølgende respons på linje med en tradisjonell lineær kommunikasjonsmodell. Et alternativ er å se en ytring som ledd i en sekvensiell kjede som både er bakovervendt (respons på hva som har blitt sagt) og fremovervendt (initiativ i forhold til hvordan man ønsker samtalen skal utvikle seg). Der samtaleanalysen holder nærhetspar for å være tilstrekkelig for å kunne si noe om et innhold

[17] I en oversiktsartikkel om samtaleforskning deler Jan Svennevig samtaleforskningen inn i fire retninger; Conversation Analysis (etter Sacks), diskurspragmatikk, Brimingham-skolen og interaksjonell sosiolingvistikk (Svennevig, 1999). De ulike retningene representerer ulike tilnærminger til samtalen som forskningsobjekt, uten at det her er passende å presentere nærmere de ulike retningene.

er blitt oppfattet, er det innenfor en sterkere samhandlingsorientert analyse nødvendig med minimum 3 bidrag for å si noe om kommunikasjonen er vellykket (Linell, 2001, s. 159 ff.). Dette kan oppleves som en uskyldig forskjell, men ulikheten er viktig nok. 3-leddstrukturen fanger inn hvordan en meningsfull samtale er et felles prosjekt.

Begge disse presiseringene, en sterkere vektlegging av relevante situasjonsoverskridende praksiser og samtalens felles meningskonstruksjon, gjør krav på analysebegrep som er funksjonelle både for å undersøke de situerte tilfellene i seg selv (vurderingsdialogene) og i lys av den praksis vurderingsdialogene skriver seg inn i. Dette vil jeg komme tilbake til i kapittel 5.3, men nå først vil informantene, og det praksisfellesskap de er en del av, bli presentert.

5.2 Vurdererpanelet og informantene

Informantene i avhandlingsarbeidet er alle deltakere i Vurdererpanelet. Vurdererpanelet består av en gruppe lærere som står for vurderingene av læringsstøttende nasjonal utvalgsprøve i skriving som grunnleggende ferdighet. Som nevnt tidligere er motivasjonen for å etablere et panel som skal stå for all vurdering av prøven, knyttet til ønsket om at vurdererne både skal utvise større samsvar enn erfart tidligere (2005-forsøket) og at de skal utvikle tekst-/vurderingskompetanse. Rekruttering til Vurdererpanelet har hovedsakelig skjedd på to måter: Først og fremst ved å kontakte gjennom landets fylkesmenn for å nå fram til de skoler (og de lærere på disse skolene) som på ulike måter har arbeidet spesielt med skriving som grunnleggende ferdighet. På den måten har man kommet i kontakt med lærere som har vist spesiell interesse for skriving. Denne måten å rekruttere lærere til panelet på har også sørget for at panelet har representanter fra hele landet. I tillegg har lærere blitt rekruttert ved hjelp av «snowball sampling» (Biernacki & Waldorf, 1981), noe som betyr at allerede rekrutterte lærere, samt involverte forskere, har tatt i bruk sitt profesjonelle nettverk for å foreslå andre aktuelle kandidater til Vurdererpanelet.[18]

I Norge har vi ingen tradisjon for å vurdere kompetansen til lærerne. Vi vet derfor ingenting om hva den enkelte lærer kan om, i dette tilfelle, skriving og vurdering av skriving. Det har derfor vært uaktuelt å rekruttere lærere til panelet ut fra kompetansekriterier. Ved heller å rekruttere lærere på måten som er beskrevet over, har man likevel ved hjelp av

[18] Informasjon gitt av skriveprøvens daværende leder, Lars Sigfred Evensen, under en samtale i mars 2011.

indirekte mål på godt resultat, nemlig engasjement, sørget for å finne lærere som vil bidra til kvalitet og kontinuitet i Vurdererpanelet.

De to første samlingene var oppstartssamlinger; i september 2010 og mars 2011 ble to ulike grupper á om lag 40 vurderere samlet til et to-dagers oppstartsseminar: deltakerne ble presentert for oppdraget, de fikk noe «faglig påfyll» og de gjennomførte vurderingsarbeid både individuelt og i grupper som gav forskergruppa bak prøvene inngangsdata. Tredje samling, i juni 2011, var første samling hvor Vurdererpanelet som helhet var samlet, det vil si begge gruppene fra de to oppstartssamlingene. Dagen før denne siste samlinga, gjennomgikk noen nyrekrutterte samme opplegg som de andre, slik at det på denne første fellessamlinga deltok 87 vurderere fra Vurdererpanelet. Målsettingen er at panelet skal ha rundt 90-100 medlemmer til enhver tid. Det er begrenset hvor forpliktet vurdererne er til å være med i panelet, slik at Vurdererpanelet vil være et semipermanent panel som hele tiden vil være avhengig av å rekruttere nye medlemmer etter hvert som noen av ulike årsaker trekker seg fra arbeidet.

Panelet skal stå for vurderingene av både femtetrinntekster (etter fire års opplæring) og åttendetrinntekster (etter syv års opplæring). Rekrutteringen til panelet er innrettet slik at om lag halvparten av vurdererne hovedsakelig har små- og mellomtrinnerfaring, og vurderer således femtetrinntekstene, mens den andre halvparten hovedsakelig har mellom- og ungdomstrinnerfaring, og vurderer således åttendetrinntekstene. Fordi de to ulike gruppene har møtt på ulike utfordringer, har de til tider arbeidet hver for seg i panelet.[19] I de tilfeller dette skjedde valgte jeg å følge åttendetrinnvurdererne, og det er således denne gruppa som er avhandlingas informantgrunnlag.

I avhandlingsarbeidets kontekst er panelets tredje samling den første samlinga hvor data, i form av lydopptak av vurderingssamtaler, ble samlet inn. I tillegg ble det samlet inn data på de to påfølgende samlingene, i november 2011 og i juni 2012. De tre ulike delstudiene i avhandlinga tar alle utgangspunkt i dette datamaterialet. Til sammen ble det gjort 26 lydopptak av vurderingssamtaler; 3 lydopptak i juni 2011, 14 opptak i november 2011 og 9 opptak i april 2012. Hvert lydopptak inneholder vurderinger av 1-3 elevtekster.

Gruppa som vurderer åttendetrinntekster består av om lag 45 vurderere. Da jeg ba om samtykke for å gjøre lydopptak, var det imidlertid tre vurderere som ikke ønsket at det skulle

[19] Se ellers de ulike delstudiene for informasjon om hva vurdererpanelsamlingene blir brukt til.

bli gjort lydopptak. Argumentet til de som ikke ønsket at det skulle bli gjort opptak mens de vurderte, var i hovedsak at de følte seg faglig usikre. De gav uttrykk for forståelse for at de var i samme båt som de andre i panelet, men de valgte likevel å takke nei til å delta i studien. Dette ble selvsagt respektert. Disse vurderernes nei til å bli gjort lydopptak av, sammen med at vurderingskonstellasjonene kontinuerlig forandres for å unngå subkulturer innenfor Vurdererpanelet, førte til at det til sammen inngår 28 vurderere i materialet som er samlet inn.

Selv om alle tre delstudier bruker det samme datamaterialet, blir dette gjort noe ulikt i de enkelte studiene. Av de 28 vurdererne er det fem som det er gjort lydopptak av fra alle tre samlingene. I delstudie 1, hvor utviklingsdimensjonen ved aspekter ved vurderingsdialogenens innhold og form blir undersøkt og drøftet, blir disse fem vurdererne derfor presentert som hovedinformanter (Jølle, 2014). I delstudie 2 følger jeg derimot én vurderers vurderingspraksis gjennom de tre samlingene tettere i en typisk casestudie (Jølle, upubl. a). Fordi jeg også her ønsket et utviklingsperspektiv, stod valget mellom en av de fem som ble dokumentert gjennom hele datainnsamlingsperioden. Da valget falt på «Trine» har det sammenheng med hennes aktive rolle i enhver parsammensetning. Siden denne studien primært er opptatt av remedieringspraksiser, er det viktig at vurdereren som følges tar i bruk vurderingsmateriell på en aktiv måte. Som sådan ble «Trine» valgt både av bekvemmelighetshensyn og av teoretiske årsaker (jf. «theoretically sampling», Mason, 1994). I tredje delstudie er hele grunnmaterialet en del av studien uavhengig av tidsdimensjonen (Jølle, upubl. b). Heller enn at noens bidrag er tillagt større vekt i analysene og den påfølgende diskusjonen, har det bildet som tegner seg i lys av forskningsspørsmålet på grunnlag av alles bidrag her blitt vektlagt.

5.3 Datamaterialet

5.3.1 Datainnsamling

Alt materiale er samlet inn i løpet av 3 Vurdererpanelsamlinger 2011/2012. Jeg fikk tilgang til panelet gjennom skriveprøveprosjektets daværende leder, Lars Sigfred Evensen. Jeg bestemte meg for å delta på ei samling før oppstart av datainnsamling (mars 2011), både for å gjøre meg kjent med hva som foregikk og for at vurdererne i panelet kunne få et første møte med meg uten at jeg trengte å virke invaderende. Denne første observasjonsrunden var også viktig for å planlegge hvordan jeg burde gå fram for å få gode data. Mot slutten av denne samlinga fikk

jeg presentert meg og prosjektet mitt for vurdererne, og jeg delte også ut et samtykkeskjema (se vedlegg 1) hvor jeg ba dem om å ta stilling til om de ville tillate at jeg gjorde lydopptak av vurderingssituasjonene (i tråd med NESH sine krav om informert samtykke, https://www.etikkom.no/forskningsetiske-retningslinjer/Samfunnsvitenskap-jus-og-humaniora/). I tillegg spurte jeg også om tillatelse til eventuelt å gjøre intervju senere i prosessen, men dette ble aldri aktuelt.

Feltloggen ble imidlertid viktig underveis i datainnsamlingsperioden og i etterkant i analysene og drøftingen av observasjonsmaterialet. Jeg brukte feltloggen på hovedsakelig to måter. For det første hadde jeg den alltid fremme der jeg satt bakerst i salen under panelets felles diskusjonsøkter. Det å gjøre «naive beskrivelser» (Tjora, 2010, s. 57) under disse øktene viste seg å bli viktig for å forstå utfordringene til den enkelte vurderer og til Vurdererpanelet som praksisfellesskap. I begynnelsen var jeg mest opptatt av hva de snakket om. Hva var det de sleit med? Hva var det de ikke forstod eller ikke var enige med forskergruppa i? Det var tydelig at det var «de erfarne lærerne» som tok ordet i en, for dem, ny kontekst hvor de var hensatt til å være «novisevurderere». Det å ta fram feltloggen med blikk for disse spørsmålene hjalp meg i det senere arbeidet med å forstå de transkriberte vurderingsdialogene som «kamper» mellom gammel og ny vurderingspraksis. Etter hvert ble jeg også mer oppmerksom på hvem som tok ordet under plenumsøktene. Det som framkom her av eksempler på diskursiv makt og arbeidsfordeling, skjerpet mitt fokus på samme fenomener i vurderingsdialogene. For det andre brukte jeg feltloggen til å gjøre underveisvurderinger av det som kom fram i ulike sammenhenger, samt å sette fram hypoteser som skulle styre mitt blikk når jeg senere skulle undersøke vurderingsdialogene. Reint praktisk gjorde jeg disse notatene i situasjoner hvor vurdererne var opptatte med vurderingsarbeid og jeg kunne bruke tid til å fordøye inntrykk fra disse fellesøktene. Med hjelp av feltloggen var analysearbeidet altså godt i gang mens datainnsamlinga fremdeles pågikk.

Det er likevel vurderingsdialogene som er det sentrale datamaterialet i avhandlinga, og det var hvordan jeg skulle gå fram for å få gjort tilstrekkelig lydopptak av vurderingssituasjonene som opptok meg mest i forkant av Vurdererpanelsamlingene. Her møtte jeg på en utfordring knytta til timing: Da jeg presenterte meg for panelet første gang, var alle på vei hjem. Jeg ba derfor vurdererne om å sende meg deres svar på samtykkeskjemaet per epost. På første datainnsamlingssamling, juni 2011, var det imidlertid relativt mange av vurdererne som ikke hadde tatt stilling til samtykkespørsmålet. For ikke å ta

for mye oppmerksomhet bort fra årsaken til at de var samlet (det vil si ta del i skriveprøveprosjektet), konsentrerte jeg meg om å gjøre opptak av de vurdererne som hadde gitt tillatelse. Dette førte imidlertid til at det ble gjort forholdsmessig få opptak under denne samlinga sammenlignet med de påfølgende (3, mot henholdsvis 14 og 9). Mellom første og andre datainnsamling sørget jeg for at samtlige vurderere svarte på samtykkeskjemaet, og det var altså 3 vurderere som ikke ønsket at det ble gjort lydopptak av de vurderingsdialogene de deltok i. Med godt samarbeid med forskergruppa som organiserte samlingene, og som også på forhånd fordelte vurdererne i ulike par-/gruppekonstellasjoner, bød ikke dette på problemer. Det at jeg på forhånd visste par-/gruppekonstellasjonene, gjorde at jeg kunne sørge for at det ikke dukket opp lydopptaker der de som ikke ønsket opptak var. Mitt inntrykk er at på denne måten var det ingen av de andre vurdererne som i det hele tatt la merke til at noen av kollegaene ikke ønsket å bli gjort opptak av.

Forut for andre vurdererpanelsamling, bestemte jeg meg også for å gjøre et grep for å minske «støyen» ved min tilstedeværelse. I stedet for at jeg skulle løpe rundt og starte og stoppe de ulike lydopptakerne, instruerte jeg vurdererne i bruk av opptakerne slik at de selv kunne styre disse. Fordelene var flere. Jeg slapp å være til stede flere steder samtidig (de ulike parene/gruppene startet opp vurderingssesjonene samtidig). Det ville vært vanskelig å få dokumentert hele vurderingsforløpet for alle parene/gruppene i så tilfelle. Viktigere var det kanskje likevel at ved å la vurdererne selv styre lydopptakeren ble min tilstedeværelse mindre synlig. Jeg skriver «mindre synlig» for som Robert Aunger skriver, er det slik at selv «noninvasive data collection techniques, which do not involve the presence of an observer (e.g. through the use of cameras or tape recorders), are likely to have some effect on what subjects say and do» (Aunger, 2004, s. 35).

En hjelp for min del var at Vurdererpanelet var en fersk institusjon hvor roller stadig var til forhandling. Vurdererne vurderte elevtekstene, men de ble hele tida oppfordret til å delta i utviklinga av skriveprøvene sammen med gruppa bak prøvene. Faggruppa på sin side, forsøkte å avveie innspill fra vurdererne opp mot de de føringer som allerede lå i prøvekonstruktet. Samtidig stod de for all organisering og bidro med «faglig påfyll» sammen med ulike gjesteforelesere. Selv observerte jeg altså under plenumsmøter og sørget for at de rette vurdererne fikk full-ladete lydopptakere på rett tidspunkt. Fellesnevneren var at vi alle var genuint interesserte i tekstvurdering generelt, og skriveprøveprosjektet spesielt. Samtidig var jeg ikke en del av deres snevrere fellesskap; det var bare de som vurderte tekstene, det

var de som gav meg data. Jeg var forskeren som fort kunne bli oppfattet som «Vurdereren av vurderne». Mitt inntrykk er likevel at jeg ble akseptert som en som hørte hjemme i panelet. Vi hadde ulike roller, ja, men ellers, i småpausene, under måltidene og på kveldstid pratet vi om faglige og ikke-faglige emner uten tanke på de ulike rollene vi ellers hadde under samlingene.

5.3.2 Oversikt over datamaterialet

I tabellen under (Tabell 1) presenteres en skjematisk oversikt over de transkriberte lydopptakene av vurderingsdialogene som inngår i studien. Her framkommer det at majoriteten av konstellasjonene er parvise (15), en stor minoritet består av sammensetninger med tre vurderere (8), mens noen er større grupper med fire eller fem vurderere (3). Det er i utgangspunktet ingen forskjell mellom de parvise konstellasjonene og treer-gruppene. Disse består av to eller tre vurderere ene og alene av praktiske hensyn. Da forskergruppa satt sammen konstellasjonene var det ikke alltid at tallene gikk opp i par samtidig som det alltid oppstod endringer på grunn av sykdom og lignende. Det gav en del treer-grupper. Når det gjelder firer- og femmer-gruppene er disse spor av forskergruppas ønsker om noen ganger å slå sammen vurderergrupper til hjelp i normeringsarbeidet. Det er altså snakk om en injeksjon av «felles vurderingspraksis».

Selv om alle de transkriberte lydopptakene er del av avhandlinga, har noe av materialet framstått som mer sentralt enn annet. Med fokus på utvikling, særlig i delstudie 1 og 2, har de vurderingsdialogene hvor jeg har opptak av de samme vurderne fra alle datainnsamlingstidspunktene kommet i relieff. I tabellen er disse merket gult.

Tabell 1 Skjematisk oversikt over datamaterialet

	Vurdererkonstellasjoner – antall vurderte elevtekster – lengde på lydopptak
Juni 2011	V3[20], V14 og V25 – 1 tekst – 49:34 V7 og V15 – 1 tekst – 36:55 V8, V21, V23 og V24 – 1 tekst – 30:05

[20] I artiklene, som er engelskspråklige, er bokstavforkortelsen V (for «vurderer») byttet ut med R (for «rater») for alle informanter. Det enkelte tall, som er koden til vurdereridentitet, er selvsagt beholdt.

November 2011	V1 og V2 – 1 tekst – 11:10 V1 og V28 – 1 tekst – 28:41 V1, V23, V26, V27 og V28 – 1 tekst – 51:15 V3 og V4 (x2) – 2 tekster – 44:50 V3 og V19 – 1 tekst – 25:30 V6, V7 og V29 – 1 tekst – 24:49 V7 og V10 – 1 tekst – 27:55 V7 og V11 – 1 tekst – 26:05 V6 og V8 – 3 tekster – 36:15 V10 og V11 – 1 tekst – 36:07 V12 og V13 – 1 tekst – 12:21 V12, V13 og V14 – 1 tekst – 28:46 V15 og V16 – 1 tekst – 11:11
April 2012	V3, V35 og V36 – 1 tekst – 25:00 V7 og V31 – 2 tekster – 20:10 V7, V18, V37 og V38 – 1 tekst – 22:26 V14, V21 og V34 – 1 tekst – 20:51 V14, V32 og V33 – 1 tekst – 13:32 V6 og V15 – 1 tekst – 11:29 V15, V23 og V24 (x2) – 2 tekster – 45:29 V19 og V30 – 1 tekst – 13:01
Totalt	33 vurderere – 26 vurderingsdialoger

5.3.3 Behandling av datamaterialet – praktisk og etisk

Etter at opptakene var gjort satt jeg igjen med et stort materiale som måtte transkriberes. Dette arbeidet ble satt bort til en profesjonell transkriptør. Jeg valgte likevel å samarbeide

med transkriptøren under transkripsjonene av de to første lydopptakene for på den måten sette en standard for den øvrige transkripsjonen. Jeg valgte å gjengi dialogene med en ortografisk skrivemåte uten blikk for alle detaljer i det språklige uttrykket. Transkripsjonskodene som er presentert under, viser hva som ble markert på hvilken måte. Transkriptøren sendte meg transkripsjoner i bolker, noe som gjorde at jeg fikk god anledning til å lese gjennom transkripsjonene med tilhørende lydspor på øret. Noen ganger hendte det at min kjennskap til de kontekstuelle forholdene gjorde at jeg hørte «bedre» hva som ble sagt under vurderingsdialogene. Noen passasjer som på den måten var markert som «ikke gjenkjennbar prat», kunne jeg høre innholdet av og dermed korrigere for i transkripsjonen.

Transkripsjonskoder:

Hun sa [det er sant	Overlapp starter ved [
(xxx)	Ikke gjenkjennbar prat
(…)	Pause i mer enn 3 sekunder
Ja:	Forlenging av en lyd
Vurde-	Avbrutt ord
I tilfelle--	Avbrutt setning
JA	Taler med emfatisk trykk
˚nå˚	Tale med lav stemme
?	Stigende intonasjonskontur (spørrende)
((telefonen ringer))	Kontekstuell informasjon

Når det gjelder de forskningsetiske sidene ved behandling av datamaterialet, har det selvsagt vært svært viktig å anonymisere informantene. Det er ingenting av det jeg skriver her som den enkelte vurderer skal måtte stå til ansvar for. Derfor utarbeidet jeg en kodenøkkel hvor hver vurderer ble tildelt et tall som den skjematiske oversikten over datamaterialet illustrerer (se

5.3.2 Oversikt over datamaterialet).[21] Kodenøkkel og det øvrige materialet har videre blitt behandlet i henhold til de krav Norsk samfunnsvitenskapelig datatjeneste (NSD) stiller til informasjonssikkerhet. Prosjektet har også søkt og fått godkjenning av NSD for gjennomføring av datainnsamling på dette grunnlaget.

5.4 Analytiske grep

Hovedmålet for avhandlingsarbeidet er å studere tekstvurderere i arbeid over tid for å få bedre innsikt i hva tekstvurdererkompetanse består i og hvordan vurdererne utvikler slik kompetanse i et gitt praksisfellesskap. Tilnærmingen for å undersøke dette har vært dialogisk fundert; sam-handling har vært et stikkord. I de ulike analysene i de ulike delstudiene anvender jeg tilnærmingsmåter og analytiske begrep som på ulike måter reflekterer en slik forståelsesramme. I dette delkapitlet vil jeg kort rekapitulere ved å presentere de analytiske grep som blir gjort i de ulike delstudiene for a) å vise hvilken felles teoretisk «himmel» de tilhører og b) å vise hvordan de likevel er egnet til å undersøke ulike aspekter knyttet til avhandlingens målsetting og c) å vise hvilke mulighetsbetingelser de tilbyr som skiller seg fra mye av det om er blitt gjort tidligere på feltet (jf. kapittel 3 Bakgrunn).

5.4.1 Delstudie 1: Analytiske grep og mulighetsbetingelser

Vurderingsdialogene kan beskrives som en variant av hva Ragnar Rommetveit omtaler som "temporarily shared social reality" (TSSR) (Rommetveit, 1974, s. 29ff). Det vil si at vurdererne, under vurderingsdialogen, etablerer og holder oppe en felles forståelse av den kommunikative situasjonen. Ved en ideell vurderingsdialog vil vurdererne ha en lik forståelse av hvordan de skal nå fram til enighet om tekstkvalitet; de vil oppfatte situasjonen likt. Men slik er det selvsagt ikke. Vurderernes ulike forutsetninger og målsetninger gjør at oppfattelsen av vurderingssituasjonen alltid bare er delvis delt (jf. Linell, 2009, s. 221ff).

Med dette som bakteppe er målsettingen i den første studien å få bedre innsikt i – og oversikt over – det som foregår i de ulike vurderingsdialogene. Med utgangspunkt i forståelsen av vurderingsdialogene som varianter av TSSR, blir derfor aspekter ved den situerte interaksjonens 'hva' og 'hvordan' studert; det vil si *responsmønster* (aspekt av dialogens 'hvordan') og *referenter* (aspekt av dialogens 'hva'). Ved å fokusere på hvordan vurdererne

[21] Jeg kunne valgt heller å gi den enkelte informant et mer personlig pseudonym, men antallet informanter tilsa at tall var mer hensiktsmessig. I delstudie 2, hvor jeg følger en vurderer tettere over tid, valgte jeg imidlertid å gi vurdererne pseudonymer. Pseudonymet Trine, som går igjen gjennom alle utdragene i studien, er vurderer V3. Videre er Tom V14, Stein er V25, Ulla er V4, Marit er V35 og Gro er V36, jf. tabellen over.

responderer på kollegaers initiativer (og bruk av vurderingsressurser), blir det mulig å analysere både hva vurdererne holder som valide argumenter i vurderingsarbeidet og betydningen av *dialogen* i dette arbeidet. Overordnet er det all grunn til å anta at vurderingsdialogene bidrar positivt mot en felles vurderingskultur. I den enkelte vurderingsdialog finner en slik antakelse styrke fordi studien viser hvordan dialogen utløser både *støtte* i form av korte bifall, *styrke* i form av referanser til andre ressurser (triangulering) og *motstand* (avvisning). Et heterogent responsmønster avdekker vurderingsdialoger hvor årsakene til en gitt vurdering kommer til overflaten; den blir synlig som argumentasjon i vurderingsdialogen. Samtidig tilslører den påviste dominerende akseptkultur blant vurdererne bildet en god del (jf funnene i studien). Tidligere studier mer enn antyder at norske skrivevurderere ikke tar del i en felles vurderingskultur (Fasting, Thygesen, Berge, Evensen & Vagle, 2009). Dette hjelper oss til å forstå akseptkulturen som et fenomen knyttet til selve interaksjonen (jf 'hvordan' over), og ikke som et uttrykk for reell enighet (det vil si vurderersamsvar) (jf 'hva' over). Ved å studere vurderingsdialogene på denne måten, gis man mulighet til å drøfte både dialogenes betydning for beslutninger om tekstkvalitet og dens relevans i lys av et mål om sterkere vurdererfellesskap med større vurderingssamsvar.

5.4.2 Delstudie 2: Analytiske grep og mulighetsbetingelser

Når jeg i andre delstudie går tettere på for å undersøke én vurderers kommunikative bidrag i vurderingsdialogene over tid, blir dette gjort ut fra en forståelse av at den enkelte vurderer er i dialog med kollegaer så vel som (normerende) vurderingsressurser. Tuftet på en sosial forståelse av læring undersøkes denne ene vurdererens *remedieringspraksis* i lys av hva som fremstår som dialogens «begivenhet» (jf begrepet «focal event», Goodwin & Duranti, 1992). Næranalyser viser hvordan vurdererne bruker mye tid på å gjøre seg familiær med – det vil si *internaliserer* – de normerende vurderingsressursene (vurderingsskjemaet med støttespørsmål (vedlegg 2) og vurderingsveiledninga (vedlegg 3)), og hvordan de da skyver oppmerksomhet bort fra elevteksten og over til vurderingsressursene. De spør seg ikke hva elevteksten kommuniserer, men heller hva vurderingsressursene kommuniserer. Derimot, når ressursene fremstår som familiære, retter vurdererne igjen oppmerksomheten mot elevteksten. Ved å se på vurderernes anvendelse av vurderingsressurser i et remedieringsperspektiv, blir reifikasjoner av normer underlagt kritisk granskning.

Implikasjoner er fremhevet i delstudien; vurderingsressurser kan ikke bare gjøres tilgjengelig og forklares, de må også bli forstått.[22]

5.4.3 Delstudie 3: Analytiske grep og mulighetsbetingelser

I tredje studie er begrepsparet "kommunikative prosjekt" og «kommunikative strategier» sentrale. Førstnevnte refererer til meningsfulle og –skapende handlinger som to eller flere samhandler om gjennom dialogen. Formålet for *prosjektet* er å løse et "problem" (Linell, 2001, s. 218). «Problem" må her ikke forstås dit hen at alle kommunikative prosjekt er problemfylte, men heller på den måten at alle prosjekt kommer i gang fordi noen ønsker å si noe til noen andre for å løse store og små oppgaver. Det kan være å overbringe informasjon, få ut frustrasjon, underholde og mye mer. For vurderne i Vurdererpanelet vil oppgaven være å komme til enighet om hva som gjelder som kvalitet ved elevtekstene, kjenne igjen disse kriteriene i tekstene, samt avgjøre hvilket nivå den enkelte elevtekst er på innenfor ulike vurderingsområder. Når det gjelder størrelsene på prosjektene, er det riktignok slik at de kan være både svært små og svært omfattende, men som regel, og i alle fall i denne studien, er de noe midt imellom. Linell skriver om kommunikative prosjekt at de "may be seen as bridging the gap between elementary contributions and local sequences, on the one hand, and the global, and more abstract notions of activity types and communicative genres" (Linell, 2001, s. 233).

Det er flere sider ved begrepet som gjør at det klinger med innenfor en dialogisk forståelsesramme. For det første peker siste del av begrepet, «prosjekt», mot at dialogen er tentativ og prosessuell. Kommunikative prosjekter planlegges, utspiller seg og fullføres i praksis blant involverte deltakere med alt det innebærer av usikkerhet, foreløpige ideer, misforståelser, motstand, ulike mål og ulik deltakelse (for å nevne noe). For det andre understreker første del av begrepet, «kommunikativ», at handlingen er sosial, den er «other-oriented» (Linell, 2001, s.219).

Der dette første begrepet, «kommunikative prosjekt», bidrar med rammer til hvordan vurderingsdialogene blir forstått, åpner det andre begrepet, «kommunikative strategier», for å undersøke hvordan vurderingsdialogen etablerer seg som en institusjonell samtale. Med andre ord, hvordan vurdererne rekursivt går fram for å løse det «kommunikative problemet».

[22] Dette er en argumentasjon som har mange likhetstrekk med kritikken av anvendelse av rubrikker i vurderingsarbeid i klasserommet (se for eksempel Wilson, 2006).

Dette er spesielt interessant i lys av Vurdererpanelets intenderte normdannende rolle for vurderingspraksis. Gjennom utvalg vurderer panelet elevtekster som går inn i en ressursbank tilgjengelig for lærere som ønsker å gjennomføre en læringsstøttende skriveprøve. Slik blir (i beste fall?) Vurdererpanelets vurderingspraksis norsk skriveopplærings peilestav.

5.4.4 Bruk av programvare

I analysene av datamaterialet har jeg benyttet meg av såkalt CAQDAS – computer-assisted analysis of qualitative data. Programvaren jeg har benyttet er NVivo 10 (http://qsrinternational.com), og den var et sentralt hjelpemiddel under både delstudie 1 og 2. I den første studien arbeidet jeg med koder og kategorier i NVivo fra begynnelsen av. Det betyr at jeg analyserte materialet fortløpende ved å undersøke ulike sammenhenger mellom ulike koder og mulige kategorier (jf. Saldaña, 2013). Fordelen med CAQDAS er at man har mulighet til å stille mange spørsmål til kodet materiale og få umiddelbare svar. Ved å få kjappe tilbakemeldinger fra programvaren kan man justere kurs; se andre forbindelseslinjer mellom ulike koder og se ulike koder som tilhørende andre kategorier. I en første, utprøvende fase var dette viktig. Materialet presentert i tabellene i delstudie 1, er resultatet av dette arbeidet. Programvaren blir altså brukt både som fortløpende analyseverktøy og som grunnlag for drøfting av materialet. Brukt på denne måten kan man også hevde at CAQDAS bygger bro mellom kvalitative og kvantitative metoder (jf. Wegerif & Mercer, 1997).

I den samme studien viser jeg også et skjermbilde av et utdrag fra en vurderingsdialog slik det framstår i NVivo 10 (Jølle, 2014, s. 48). Skjermbildet synliggjør hvordan programvaren visuelt får fram sammenhengen mellom vurderingsdialogens responsmønster og ulik bruk av argumenter (referenter). Det er for eksempel på dette grunnlaget jeg argumenterer for en sammenheng mellom et heterogent responsmønster og triangulering av vurderingsressurser. Eksemplet er også illustrerende for hvordan jeg tok i bruk NVivo 10 i den påfølgende delstudien. Trine, som blir fulgt i delstudie 2, er valgt på grunnlag av at hun fremstår som en «rik» informant; i de dialogene hun deltar i, avslører programvaren at kodet materiale finnes i stort monn. Det er derfor selvsagt ikke snakk om representativitet når Trine er valgt som informant. I stedet argumenterer jeg i artikkelen for at det er et valg gjort på grunnlag av analytisk generaliserbarhet (Yin, 2009. Se også Silverman, 2005, s. 202 ff. om teoriutvikling med hjelp av CAQDAS).

5.5 Studiens gyldighet og pålitelighet

Et sentralt tema i dette avhandlingsarbeidet har vært hvordan gyldighet og pålitelighet skal forstås i en skrivevurderingssammenheng (se særlig kapittel 3 Bakgrunn). Samtidig er det klart at jeg har måttet forholde meg til det samme begrepsparet i mitt eget avhandlingsarbeid. Gyldighetskrav og pålitelighetskrav stilles til ethvert forskningsarbeid. Transparens er her et stikkord. I det følgende vil jeg derfor løfte fram aspekter ved henholdsvis datamaterialets kvalitet, tolkningskvalitet og min egen forskerrolle for nettopp å synliggjøre forutsetninger, valg og prosesser knyttet til arbeidet.

5.5.1 Økologisk validitet

Ved studier av skrivevurderingsprosesser er det normale å samle inn data ved å intervjue vurderere, ved å be vurdererne skrive protokoll over hva de gjør mens de vurderer, og ved å be dem tenke høyt mens de vurderer (Baume, Yorke, & Coffey, 2004). Alle disse prosedyrene har imidlertid sine begrensninger i det at de ikke får fram vurderernes beslutningsprosesser per se, men heller ulike mer eller mindre rasjonaliserte versjoner av prosessene. Nettopp av den grunn blir nok høyttenkningsprosedyren («think aloud protocols», TAP) ofte foretrukket siden man her tilsynelatende er nærmest den faktiske vurderingsprosessen. Metoden blir med andre ord holdt for å ha solid «økologisk validitet» (jf. Cicourel, 1997, 2007).

I en empirisk undersøkelse av bruken av TAP som datainnsamlingsmetode ved 16 ulike skrivevurderingsstudier, problematiserer Khaled Barkoui denne antakelsen (Barkaoui, 2011). Han finner en tendens til at forskerne anerkjenner svakhetene ved metoden samtidig som de behandler svakhetene som så ubetydelige at de ikke truer datamaterialets validitet. Barkoui skriver:

> These studies tend to acknowledge the incompleteness of TAPs, while insisting on their validity. Claims for the validity of TAPs in investigating essay rating processes, as Lumley (2005) noted, are often based on (a) the literature on their advantages (e.g., Cohen 1998), (b) the precedent of other studies in using this technique to investigate essay rating or other problem-solving behaviors (e.g., writing), and/or (c) the work of Ericsson and Simon (e.g., 1987, 1984/93) as a justification for using this technique (see, for example, Wolfe et al., 1998, p. 470). (Khaled Barkoui, 2011, s. 57)

Den sterke tendensen til å ubetydeliggjøre svakhetene ved TAP som metode, holder Barkoui for å være tvilsom så lenge vi mangler kunnskap om i hvilken grad metoden evner å få fram alle sider ved vurderingspraksisen (*veridicality*) og om i hvilken grad metoden endrer selve vurderingspraksisen (*reactivity*) (Barkaoui, 2011, s. 58). Hans egen studie indikerer at det å be vurdererne tenke høyt når de vurderer nettopp ikke gir et helhetlig og sannferdig bilde av vurderingspraksisen. Barkoui finner også at det ikke bare er kognitiv overbelastning hos vurdererne (jf. Ericsson & Simon, 1993) som eventuelt fører til et forkjært datamateriale, men at eksempelvis også graden av erfaring (novise vs erfaren) og vurderingsmåte (analytisk vs holistisk) er variabler som påvirker forholdet mellom TAP og faktisk vurderingspraksis. Jeg skylder å legge til at Barkoui likevel ikke benytter funnene til å advare mot bruk av TAP, til det er fordelene for mange. I stedet oppfordrer han til mer forskning på området og til større bevissthet om potensielle skjevheter i datamaterialet som følge av TAP.

Jeg nevner dette fordi datamaterialet i denne avhandlinga er samlet inn på en annen måte: lydopptak av faktiske (ikke simulerte) vurderingsdialoger gir tilgang til vurderernes beslutningsprosesser uten at vurdererne selv har muligheter til å «redigere» egen vurderingspraksis. Forhandlingene mellom vurdererne i dialogene er de faktiske vurderingene. Som forsker har jeg ikke manipulert en aktivitet ved å be vurdererne vurdere «som om» situasjonen var helt normal. Siden vurdererne i min studie allerede skulle arbeide sammen, vurdere i par/små grupper, har jeg altså ikke vært nødt til å ta hensyn til TAP som en potensiell feilkilde.

En mulig feilkilde i materialet kan derimot være knytta til vurderernes oppmerksomhet mot lydopptakeren. Det å vite at det man sier blir tatt vare på, kan prege de fleste. Som forsker var jeg klar over denne problemstillingen, og jeg gjorde to tiltak for å minimere denne feilkilden. Det viktigste var å være tilstede i Vurdererpanelet så mye som mulig slik at de ble kjent med meg og ble vant til min tilstedeværelse. Når jeg presenterte meg for panelet under en fellessesjon, la jeg vekt på at jeg ikke var interessert i den enkeltes vurderinger, men at jeg var opptatt av panelets arbeid som helhet. Vurdererne var selv svært klar over at de stod i en situasjon hvor «flyet ble bygd mens de fløy», og jeg understreket at det var nettopp denne situasjonen som gjorde det interessant å undersøke hvordan de som vurderere løste oppgavene. Det ligger i denne «insider-strategien» at jeg deltok sammen med vurdererne på sosiale aktiviteter under samlingene. Et mer konkret tiltak i forhold til opptakssituasjonene var at jeg ikke selv stod for opptakene. I stedet viste jeg vurdererne hvordan de kunne starte

og stoppe lydopptakerne slik at de kunne administrere dette på egen hånd. På den måten unngikk jeg at jeg måtte være tilstede for å sette i gang de enkelte vurderingsdialogene. Ved å unngå situasjoner hvor jeg som forsker ved tilstedeværelse tydelig markerer at det foregår datainnsamling, var målsettingen å minimere potensiell «reactivity» (Barkaoui, 2011, s. 58). Uformelle samtaler med vurdererne tyder på at dette var et vellykket grep. For dem ble det å ta fram lydopptakeren og sette på opptak en del av forberedelsesfasen på linje med å ta fram blyant og papir.

Kombinasjonen av dette, det vil si materialet som er vurderingsdialoger slik de foregår i sitt naturlige miljø og de forhåndsregler som er tatt i forbindelse med lydopptak, gir grunnlag for å hevde at studiens økologiske validitet er de facto sterk.

5.5.2 Empirisk forankring

Kvaliteten på datamaterialet har gitt gode forutsetninger for avhandlingsarbeidet. *Forskningskvaliteten* er imidlertid på avgjørende vis forbundet med hvordan analyser, tolkninger og slutninger er forankret i dette datamaterialet. Sosiologen Clive Seale skriver:

> Methodological awareness involves a commitment to showing as much as possible to the audience of research studies (…) the procedures and evidence that have led to particular conclusions, always open to the possibility that conclusions may need to be revised in the light of new evidence. (Seale, 1999, s. x)

Det handler om å gjøre forskningen transparent slik at leseren selv kan vurdere kvaliteten på arbeidet som er blitt gjort. Dette er selvsagt langt mer krevende enn det høres ut; på et nivå handler forskning om å fortette, kategorisere og systematisere. Det vil si at man med nødvendighet etterlater noe av/store deler av materialet «bak» kategoriene og systemet. En kjent fare ved kvalitativ forskning er at man således kan bli anekdotisk (jf. Bryman, 1988, s. 77. Gjengitt etter Silverman, 2005, s. 211); som forsker trekker man frem det som skaper «den gode forskningshistorien», og ikke nødvendigvis det som materialet som helhet forteller.

Det finnes imidlertid måter å arbeide med datamaterialet på som bidrar til å øke studiens troverdighet. I dette avhandlingsarbeidet har, for det første, datamaterialet som helhet blitt presentert gjennom tabeller som tallfester tilfeller innenfor ulike kategorier (se særlig Jølle, 2014). Silverman holder denne måten å framstille materialet på i kvalitative studier som en av få muligheter for å styrke arbeidets troverdighet. Han skriver at ”[i]nstead

of taking the researcher's word for it, the reader has a chance to gain a sense of the flavour of the data as a whole" (Silverman, 2005, s. 220).

For det andre er materialet kodet innenfor kategorier som i liten grad krever fortolkning. I studiene har jeg således argumentert for at det har vært snakk om å kode «manifest content», og ikke «projective content» (Potter & Levine-Donnerstein, 1999). Seale framsnakker i tilsvarende sammenheng «low-inference descriptors» (Seale, 1999), og med det mener han nettopp beskrivelser som i stor grad «ligger opp i dagen», og som man dermed lettere kan enes om. Eksempler på dette i avhandlingen er på den ene siden de fire ulike responsalternativene (aksept, justering, oppfølgingsspørsmål og avvisning), og på den andre siden referenter som eksempelvis 'sammenligning med andre tekster' og 'vurderingsveiledningsreferanse'. Jeg vil hevde at dette er eksempler på «low-inference indicators» som datamaterialet relativt enkelt lar seg kode innenfor.

For det tredje er leseren tilbudt mange utdrag fra vurderingsdialogene i de ulike delstudiene. Analyse- og tolkningsarbeidet har vært nært knyttet opp til disse. Siden selve dialogsamspillet har stått sentralt i analysene, har dette vært viktig. Akkumulert, ved å vise fram mange utdrag, har det i tillegg vært mulig å vise trekk og tendenser mellom de ulike utdragene. Og videre, ved å sette alle disse utdragene i sammenheng med framstillingen av hele materialet, er det mulig å argumentere mot at det har foregått en anekdotisk (mis)bruk av datamaterialet.

5.5.3 Forskerrollen

Innenfor humaniora og samfunnsvitenskapene har det allerede lenge vært vanlig å anse meningsskaping som en sosial konstruksjon (se for eksempel Lock & Strongs innføringsverk *Sosial konstruksjonisme – teorier og tradisjoner* (2014)). Innenfor dette vitenskaps*paradigmet* (NB!, nok en konstruksjon) blir erkjennelse knyttet til mennesket. Erkjennelsen(!) av at forskeren ikke lenger har en selvsagt tilgang til en "the view from nowhere" (Nagel, 1989) har fått enkelte til å hengi seg til relativismen eller skeptisismen (se for eksempel Burr, 2003). Som blant andre Kjørup er inne på, synes det imidlertid som mer fornuftig å se på mennesket som en forutsetning – og ikke en feilkilde – for i det hele tatt å ha kunnskap om noe som helst:

> [Det ligger] i ordet «viden» eller «erkendelse» at det er noget mennesker har eller producerer, så hvor der foreligger viden eller erkendelse, foreligger der spor af mennesker. Den der efterlyser viden uden spor

af mennesker, efterlyser altså noget det er lige så umuligt at finde som gifte ungkarle». (Kjørup, 2008, s. 182)

Kjørup argumenterer for det han omtaler som pragmatisk konstruktivisme: Ja, erkjennelse er sosialt konstruert, men det er tilfelle uten at vi mister evnen til å skille noe riktig fra noe uriktig, til å skille falskt fra sant. Sittende med hver vår kaffekopp i hånda ser vi koppene fra hver vår side uten at det får oss til å betvile at det er kaffekopper vi ser. Ulike perspektiver og posisjoneringer viser oss med andre ord ikke at alt er relativt, men heller at ting er relasjonelle (Kjørup, 2008, s. 175).

Jeg deler denne oppfatningen av at det aldri verdinøytrale mennesket står som et filter mellom "virkeligheten" og vår erkjennelse av den, samtidig som at det ikke trenger å føre oss inn i verdiløsheten. Likevel er det slik at tapet av den fordelaktige gudsposisjon fører til at jeg som forsker så å si må trå ned blant likemenn hvor den ene måten å oppfatte erfaringer på og å nå inn til ny erkjennelse på er potensielt like kraftfull som den andre. Desto viktigere blir det da å klargjøre hvilken "rustning" jeg har ikledd meg på denne argumentasjonens arena.

Det første jeg vil understreke er nettopp dette at jeg holder avhandlingsarbeidet for å være et særegent argument. Som sådan plederer jeg ingen Sannhet, og aksepterer at holdbarheten til dette argumentet vil bli testet gjennom forskningsfellesskapets validitetskontroll, den kritiske dialogen. Litteraturgjennomgangen i kapittel 2 er eksempelvis ingen "state of the field"-gjennomgang. I stedet er denne delen strukturert slik at den viser vei fram til et mørklagt rom med behov for lyssetting. Litteraturgjennomgangen er altså i så måte et argument som taler for at det har vært fornuftig av meg å beskjeftige meg med avhandlingas problemstilling og de ulike forskningsspørsmål. Men der det fremsettes et argument, oppstår det gjerne et motargument. Noen motargumenter er med i avhandlinga, kanskje særlig i form av forbehold i de ulike delstudiene, men oftest er det nok slik at de største innvendingene kommer i dialogene med forskerkollegiet. Det er i denne tradisjonen av "the better argument" (Habermas, 1989) og «diskurskriteriet (Larsson, 2005) jeg forstår gehalten i mitt eget arbeid, og som jeg dermed inviterer til diskusjon i lys av (jf. Rienecker og Jørgensen, 2013).

Det andre som bør på bordet er at jeg er meg(!). Med det mener jeg at jeg har hatt mine grunner for å gå inn i dette arbeidet; jeg bringer med meg en forforståelse. Det er med denne forforståelsen jeg har tatt fatt på og balet med denne avhandlinga. Gadamer skriver:

> Det utleggende ordet er den utleggendes ord, og ikke den utlagte tekstens språk og ordforråd. Tilegnelsen er dermed ingen ren reproduksjon, og slett ingen ren videreformidling av den overleverte teksten, men forståelsens nye frembringelse. Når man med rette hevder at all mening er relatert til jeget, så innebærer dette for det hermeneutiske fenomenet at overleveringens mening blir konkretisert i relasjon til det forstående jeget som forstår meningen, og ikke i en rekonstruksjon av det jeget som har den opprinnelige meningsintensjonen. (Gadamer, 2010, s. 516)

Det vil derfor være interessant for leseren å vite noe om hvem forfatteren er. Derfor har jeg i innledninga i kappa forsøkt å gjøre rede for noe av min biografi som kan ha betydning for hvordan jeg har forstått det materialet jeg har arbeidet med. Sentralt er det selvsagt at jeg har ei tid bak meg som norsklærer. De erfaringene jeg da gjorde meg som tekstvurderer har sine tydelige avtrykk i egen posisjonering.

Et tredje perspektiv som er relatert til min egen forskerrolle, epistemologisk grunnsyn og generell akademisk akribi, er selvsagt hvorvidt arbeidet oppfattes som troverdig. Ambisjonen er at dette metodekapitlet, hvor jeg har forsøkt å «vrenge ut» arbeidsprosessen (som gjerne blir tildekket og sminket i resultat- og drøftingsdelen), bidrar til tilstrekkelig transparens (jf. *Standards for Reporting on Empirical Social Science Research in AERA Publications*, 2006). Og selv om datakvalitet og tolkningskvalitet er disktuert over, er det ikke slik at dette i seg selv er nok til å overbevise den kritiske leseren. Jeg håper derfor jeg har lykkes med å vise styrker og svakheter i mitt arbeid, både gjennom drøftinger på grunnlag av eget datamateriale og andres forskning og gjennom posisjoneringsmarkører og ulike modalitetsformer som uttrykk for grader av sikkerhet (og tvil).

6 Konklusjon

Innledningsvis i denne kappeteksten ble avhandlingas målsetting presentert: Jeg har ønsket å bidra til økt innsikt i hvordan en bestemt gruppe uerfarne tekstvurderere vurderer elevtekster og hvordan de går fram over tid for å bli mer kompetente. Målsettingen har dels vært knyttet til at vurdererne er erfarne lærere som ved å bli konfrontert med en bestemt tekstforståelse og et bestemt vurderingskonstrukt, blir gjort til vurderingsnoviser i en bestemt kontekst. Og dels har målsettingen vært knyttet til at det praksisfellesskapet disse vurdererne inngår i, Vurdererpanelet, blir gjort til normleverandør av tekstforståelse og vurderingspraksis gjennom å være vurdererne av de nasjonale læringsstøttende prøver i skriving.

Den overordna problemstillinga har jeg søkt svar på gjennom tre delstudier med sine tilhørende forskningsspørsmål. Det er ikke her sted for å repetere de svar som de ulike analysene gav. I stedet vil jeg rette oppmerksomheten mot de mer overordna og generelle bidrag av empirisk og teoretisk art som arbeidet som helhet har kommet med. Deretter vil jeg peke på begrensninger knyttet til avhandlingsarbeidet, før jeg helt avslutningsvis foreslår noen veier å følge for videre undersøkelser.

6.1 Studiens bidrag

Ved flere anledninger har jeg trukket fram hvordan skrivevurdererforskning deler seg i to hovedgreiner: den forskning som er resultatorientert og den forskning som er prosessorientert. Den resultatorienterte forskningen har jeg i noen grad gått i dialog med og diskutert relevansen til (se kapittel 3 Bakgrunn), men avhandlinga er selv en del av den prosessorienterte skrivevurdererforskningen. Denne prosessorienterte skrivevurdererforskningen tar som regel i bruk høyttenkningsmetoder for å samle inn empirisk materiale; den enkelte vurderer blir bedt om å tenke høyt mens han/hun vurderer. Til forskjell fra dette, baserer denne avhandlinga seg på et datamateriale samlet inn over tid hvor vurderere gjennom diskusjon kommer fram til enighet om tekstkvalitet innenfor et normstyrt vurderingsarbeid. Det å arbeide med et slikt materiale har noen fordeler: For det første er det *vurderingsprosessene per se* som undersøkes. For det andre kan man undersøke hva *vurdererfellesskapet* holder for å være *holdbare vurderingskriterier*. Og for det tredje kan man undersøke utvikling av *vurdererkompetanse* og trekke noen slutninger om *vurdereropplæring*. En gjennomgang av forskningslitteraturen viser at denne måten å undersøke skrivevurdereres vurderingspraksis på ikke er gjennomført tidligere (jf. Meadows & Billington, 2005), og

avhandlinga må således sees på som å komme med et viktig empirisk bidrag til forskningsfeltet.

I avhandlinga har jeg operert med et skille mellom dialogisk og monologisk epistemologi (Linell, 2001, 2009). Videre har jeg plassert avhandlingsarbeidet innenfor en dialogisk forståelsesramme, og kontrastert det med det jeg oppfatter som en dominerende monologisk forståelsesramme innenfor det internasjonale skrivevurderingsfeltet. Jeg ser på den monologiske skrivevurderingsforskningen som et paradigme med etablerte tatt-for-gittheter som har vidtrekkende konsekvenser. Jeg skal her peke på én: Tidligere har jeg vist til Bejars (2012) presentasjon av skrivevurderingens to faser, designfasen og vurderingsfasen. Jeg opplever modellen som nyttig, men som begrensende fordi det i praksis er stor fare for at det etableres én-veis kommandolinjer mellom fasene hvor designerne, som er høyest i hierarkiet, bokstavelig talt setter standarden, mens vurdererne, som er nederst i hierarkiet, utfører et stykke arbeid som evalueres i forhold til den etablerte standarden (jf. Meadows & Billington, 2005). Gjennom en monologisk optikk, blir vurdererne gjort til "utførere", til trente iverksettere av en gjennomtenkt instruks. Det er etter min oppfatning en sammenheng mellom et slikt syn på skrivevurdererrollen og fremveksten av bruken av såkalt AES, automated essay scoring. Dersom skriveprøvedesignere kan forhåndsdefinere alle relevante kriterier og kjennetegn på ulik tekstkvalitet, er det ikke en fremmed tanke at maskiner kan erstatte mennesker. Når det anerkjente fagtidsskriftet *Assessing Writing* i 2013 dedikerer et volum spesifikt til AES-forskning, er det et signal om at computervurdering er blitt «stuereint» (*Assessing Writing*, 18, 2013. Se også Shermis & Hamner, 2013).

Heller enn å forstå Bejars to skrivevurderingsfaser som monologiske, vil jeg foreslå å se på disse to skrivevurderingsfasene som dialogpartnere; designfasen virker inn på vurderingsfasen, mens erfaring fra vurderingsfasen virker inn på designet. Vurdererne er her ikke maskiner eller «utførere», men profesjonsutøvere (som beskrevet i kapittel 4.6). Datamaterialet har gjentatte ganger vist hvordan vurdererne ikke blindt følger skriveprøvedesignet, men at de tar utgangspunkt i at skriving og tekstskaping er (forsøk på) meningsfull dialog. En negativ måte å forstå dette på er at i utviklinga av skriveprøvene «ofres» reliabilitet til fordel for hva vurdererne betrakter som *meningsfullt* i den enkelte vurderingssituasjon. Et slikt perspektiv er imidlertid tufta på at det er reliabilitet som er en skriveprøves rettesnor. Denne avhandlinga er å forstå som en kritikk av et slikt perspektiv. En positiv måte å forstå vurderernes kritiske tilnærming til vurderingsoppgaven er dermed at de

gjennom en slik praksis forsøker å speile et valid skrivekonstrukt i vurderingsarbeidet. Ved at vurdererne på denne måten tar elevtekstene på alvor, blir det også tydelig hvorfor computerbasert vurdering er problematisk. Maskiner kan programmeres til å imitere atferd, men de kan ikke det som er aller viktigst, nemlig lese *mening* ut av en tekst (Perelman, 2012, 2014).

Dialogfokuset i avhandlinga har bidratt til innsikt på to nivåer; kunnskap om hvordan vurdererne samarbeider om å komme til enighet om tekstkvalitet, og kunnskap om vurderernes rolle i en skriveprøvesammenheng. Dette doble perspektivet har bidratt til å aktualisere forholdet mellom gyldighet og pålitelighet, mellom validitet og reliabilitet. Ulike validitetskrav (jf. Kane, 2006) er det primære ved en prøve, og det er ved å ta hensyn til disse kravene at prosedyrer etableres for å skape reliable prøver. Når vi i praksis ser at dette ofte er motsatt, at reliabilitet er det første som undersøkes for å teste en prøves gyldighet, gir det et forkjært utgangspunkt. Avhandlingsarbeidet kan således også sees på som en stemme for å aktualisere en diskusjon om så vel innholdet i et skriveprøverelatert reliabilitetsbegrep og reliabilitetskrav knyttet til skriveprøver som bedre stemmer med skrivevurdererens fortolkningspraksis.

6.2 Begrensninger og forslag til videre undersøkelser

Underveis i denne teksten har jeg forsøkt å være åpen på at studien som helhet er å betrakte som et argument i den akademiske debatten som omhandler skriveprøver og skrivevurdering. Dette innebærer ikke bare at jeg skriver fram en bestemt forståelse av hva skrivevurdering er, men også at studiens design reflekterer et slikt argument. Når jeg således har arbeidet med vurderingsdialoger som empirisk materiale, har det i seg selv bidratt med «vann på mølla» slik at vurderingsdialogene faktisk framstår som viktige. Kritiske røster vil eksempelvis kunne hevde at denne studien ikke dokumenterer om vurderingene blir «bedre» – i betydningen av større grad av vurderersamsvar – siden den ikke diskuterer vurderingsdialogene i lys av vurderingsresultatene. Kritikken er relevant, men kan imøtegås. Denne studien har ikke hatt som målsetting å undersøke vurderingskvalitet i form av score-studier, men heller i form av innholdet i vurderingsdialogene betraktet som beslutningsprosesser.

Andre begrensninger knyttet til at datamaterialet er hentet fra et bestemt historisk tidrom, er åpenbare. Datamaterialets representativitet kan aldri etterprøves; Vurdererpanelets første fase kan aldri gjenskapes. Som forsker fikk jeg tilgang til et

praksisfellesskap som fungerte på en bestemt måte betinget av faktorer som i dag er annerledes.[23] Jeg vil hevde at all forskning rammes av å være situert på en slik måte, og det eneste botemiddelet er åpenhet om a) hva man har undersøkt (og hva man ikke har undersøkt), b) hvilken måte man har gått fram på, og c) sammenhengen mellom data, analyser og slutninger. Leseren må vurdere om jeg her har lykkes.

På tross av disse begrensningene, vil jeg hevde at studien etablerer et grunnlag for videre undersøkelser som kan hjelpe oss med det større prosjektet denne avhandlingen skriver seg inn i, nemlig etableringen av skrivevurderingsforskning som en egen disiplin med egne krav til og forståelser av hva som er kjennetegn på kvalitet i skriveprøver (jf. Behizadeh og Engelhard jr, (2011)). En løs ende å ta tak i er således nettopp manglende undersøkelse av vurderersamsvar; det vil helt klart være svært interessant å undersøke sammenhengen mellom vurderingsdialogens hva og hvordan og påliteligheten til resultatene. Dersom heterogent responsmønster og bruk av referenter i vurderingsdialogene positivt korrelerer med vurderersamsvar, vil vi ha kommet langt i retning av å ha funnet gyldige indikatorer på vurderingskvalitet som hensynstar både kompleksiteten i skriving og behovet for symbolsk verdsetting av teksten i form av en score. Det er et scenario verdt nærmere undersøkelser.

Dersom man går nærmere inn på hva som har blitt undersøkt i de ulike delstudiene, vil vi se at sentrale spørsmål som har blitt forsøkt besvart er; hvordan vurdererne samarbeider for å nå fram til enighet om tekstkvalitet, hva de holder som relevante argument for å nå fram til enighet, hvordan de går fram for å lære seg å vurdere elevtekster med hjelp av bestemte normerende vurderingsressurser, hvilke vurderingsstrategier som viser seg å føre fram til beslutning om tekstkvalitet, og hvordan dialogisk asymmetri virker inn på beslutningsprosessene. De fleste av disse spørsmålene er knyttet til studiens eksplorerende forankring i *dialogen*. Det å studere vurderingskvalitet med dialogen som utgangspunkt, har åpnet for teoretiske og metodologiske perspektiver som på langt nær tidligere er forfulgt tilfredsstillende. I særlig grad er vurderingsdialogenes diskursive fenomener underundersøkt.

I denne kappa startet jeg med å fortelle om *tausheten* som innhyller skrivevurderingspraksisen eksemplifisert gjennom min tidligere kollega som brukte ryggmargsfølelsen som peilestav i vurderingsarbeidet. Det er mye bra å si om ryggmargsfølelsen, men velegnet til kunnskapsdeling og læring er den ikke. Som en motsats

[23] Men datamaterialet i seg selv kan selvsagt underlegges kontrollstudier, og konklusjonene i dette avslutningskapitlet kan på et slikt grunnlag alltid utfordres.

har jeg derfor oppholdt meg ved dialogen, som, hvis vellykket, nettopp kjennetegnes av samhandling, kunnskapsdeling og læring. Og akkurat her finner vi nok det ved avhandlinga som har størst verdi: Ved å fokusere på det som ligger forut for bestemmelsen om å plassere en gitt tekstdimensjon på et bestemt sted på en bestemt skala, har vi fått innsikt i hva vurdererne ser og ikke ser når de vurderer elevenes tekster.

Vedlegg 1 Samtykkeerklæring

Lennart Jølle
Høgskolen i Sør Trøndelag,
Avdeling for lærer- og tolkeutdanning
7004 Trondheim
e-post: lennart.jolle@hist.no
Tlf 73 55 90 77 / 95 94 56 25

Trondheim 20.04.12

Til medlemmer av vurderingspanelet

Forskningsprosjekt om vurdering av skriving innenfor Vurderingspanelet

Jeg ønsker å undersøke hvordan lærere i denne sammenheng (dvs. Vurderingspanelet) (videre-)utvikler kompetanse i vurdering av elevtekster og hva denne kompetansen består i. Arbeidet skal munne ut i en doktorgradsavhandling om emnet vurdering av skriving.

For å få undersøkt dette vil jeg være til stede under Vurderingspanelets samling 23.-25. april 2012. (Jeg har allerede vært til stede på alle samlingene til Vurderingspanlet fra høsten 2010 og fram til i dag). I den anledning vil jeg be om tillatelse til å gjøre lydopptak fra flere grupper/par under vurderingsarbeidet.

Materialet vil bli behandlet på en slik måte at det ikke vil være mulig for noen å kjenne igjen enkeltmedlemmer av Vurderingspanelet. Informantene vil bli anonymisert og får nye navn i materialet.

Hvis noen vil vite mer om hva jeg vil, eller skal bruke det innsamlede materialet til, er det bare å ta kontakt med meg på telefon 95 94 56 25 / 735 59 077 eller epost: lennart.jolle@hist.no.

Du gir meg tillatelse til å gjøre lydopptak av vurderingsarbeidet ved å fylle ut vedlagte svarslipp. Det er selvsagt helt frivillig å delta, og man kan senere på et hvilket som helst tidspunkt trekke seg fra å delta i prosjektet uten å måtte oppgi noen begrunnelse.

Vennlig hilsen

Lennart Jølle

Doktorgradsstipendiat

Alder	> 30 år:	30-49 år:	≤ 50 år:
Målform (privat)	Nynorsk:	Bokmål:	
Målform (arbeid)	Nynorsk:	Bokmål:	
Undervisningserfaring	> 5 år:	5-9 år:	≤ 10 år:
Antall år på ulike trinn	1. – 4. årstrinn:	5. – 7. årstrinn:	8. – 10. årstrinn:

Erklæring om samtykke i forbindelse med Lennart Jølles prosjekt *UTVIKLING AV VURDERINGSKOMPETANSE I ET PRAKSISFELLESSKAP*

Jeg GIR / GIR IKKE (stryk det som IKKE passer) tillatelse til at Lennart Jølle gjennomfører lydopptak av situasjoner der undertegnede i grupper/par foretar vurderinger av elevtekster.

Forutsetningen for tillatelsen er at lydopptak og annet innsamlet materiale blir behandlet med respekt og blir anonymisert. Bruken av materialet skal ikke på noen måter føre til ulemper for undertegnede. Prosjektet vil ellers følge gjeldende retningslinjer for personvern.

…….. (for- og etternavn i blokkbokstaver)

Dato og sted:

Underskrift:

Utvalgsprøver i skriving: Vurderingsskjema for elever ved inngangen til 8. årstrinn (mars. 2012).

VurdererID: ☐☐☐☐ VurdererID: ☐☐☐☐ ElevID.: ☐☐☐☐☐ Kjønn: ☐ Jente ☐ Gutt Første språk: ☐ Norsk ☐ Annet enn norsk

Vurderingens primærtrekk:

Å forestille seg *å ha vært med på en ekspedisjon til et spennende sted med en spesiell utfordring underveis til klassen.*

Vurdering og nivåbeskrivelser: De fleste elevtekster vil kunne plasseres på nivået *«Som det kan forventes ved starten av årstrinnet»*. Dette nivået er gitt betegnelsen *mestringsnivå 3* (M.3), og det er plassert på det midterste trinnet i skalaen. Ytterpunktene, *mestringsnivå 1 og 5*, er betegnet som *«Meget* lav grad av mestring…» og *«Meget* høy grad av mestring…» vurdert i forhold til det *«som kan forventes ved starten av årstrinnet»*. Vurderingsveiledningen og normene for femte årstrinn danner grunnlaget ved vurderingen.

Vurderingsområder:	Nivåbeskrivelser:	Mestringsnivå: (sett kryss):		Støttespørsmål ved vurderingen:	Vurdert som: - / 0 / +
1) Kommunikasjon Hvor godt *kommuniserer* elevens *forestilling* med en leser i klassen?	*Meget* lav grad av mestring innen vurderingsområdet	M. 1:		**Henvender forestillingen seg til en leser i klassen på en hensiktsmessig måte?**	
	Lav grad av mestring…	M. 2:		Formidles det en klar hensikt med teksten?	
	Som det kan forventes ved starten av årstrinnet	M. 3:		Benyttes kommunikasjonsfremmende spørsmål eller vendinger?	
	Høy grad av mestring…	M. 4:		Har forestillingen en tydelig forfatter?	
	Meget høy grad av mestring innen vurderingsområdet	M. 5:		-	
2) Innhold Har elevens *forestilling* et dekkende *innhold*?	*Meget* lav grad av mestring innen vurderingsområdet	M. 1:		**Presenteres innholdet *bredt* gjennom f.eks. flere momenter?**	
	Lav grad av mestring…	M. 2:		Er innholdet relevant med tanke på tenkt ekspedisjon?	
	Som det kan forventes ved starten av årstrinnet	M. 3:		Er innholdet tilstrekkelig utdypet?	
	Høy grad av mestring…	M. 4:		-	
	Meget høy grad av mestring innen vurderingsområdet	M. 5:		-	
3) Tekstoppbygning Hvor hensiktsmessig er *komposisjonen* i *forestillingen*?	*Meget* lav grad av mestring innen vurderingsområdet	M. 1:		**Er teksten som helhet bundet sammen på en god måte?**	
	Lav grad av mestring…	M. 2:		Er teksten hensiktsmessig strukturert?	
	Som det kan forventes ved starten av årstrinnet	M. 3:		Er setningene bundet sammen på en god måte (forbindere)?	
	Høy grad av mestring…	M. 4:		-	
	Meget høy grad av mestring innen vurderingsområdet	M. 5:		-	
4) Språkbruk Hvor god er *setningsstrukturen og ordvalget* i *forestillingen*?	*Meget* lav grad av mestring innen vurderingsområdet	M. 1:		Har teksten velformulerte setninger?	
	Lav grad av mestring…	M. 2:		**Er setningsoppbygningen variert på en hensiktsmessig måte?**	
	Som det kan forventes ved starten av årstrinnet	M. 3:		Viser teksten riktig bruk av grammatiske kontraster som og - å, da - når?	
	Høy grad av mestring…	M. 4:		**Er ord og uttrykk variert på en hensiktsmessig måte?**	
	Meget høy grad av mestring innen vurderingsområdet	M. 5:		-	

Utvalgsprøver i skriving: Vurderingsskjema for elever ved inngangen til 8. årstrinn (mars. 2012).

Vurderingsområder:	Nivåbeskrivelser:	Vurdert nivå: (sett kryss):		Støttespørsmål ved vurderingen:	Vurdert som: - / 0 / +
5) Rettskriving Hvor riktig er *rettskrivingen*?	*Meget* lav grad av mestring innen vurderingsområdet	M. 1:		**Skrives lydrette ord, inkl. ord med dobbeltkonsonant, riktig?**	
	Lav grad av mestring...	M. 2:		**Skrives *høyfrekvente*, ikke-lydrette ord riktig** (begynnende ortografisk skriving)?	
	Som det kan forventes ved starten av årstrinnet	M. 3:		Skrives *de fleste* ikke-lydrette ord riktig (ortografisk skriving)?	
	Høy grad av mestring...	M. 4:		-	
	Meget høy grad av mestring innen vurderingsområdet	M. 5:		Antall feil: *Feilfritt; „Få skrivefeil per side; „> 8 skrivefeil per side:*	

	Side nr:	Antall feil:
Antall feil på besvarelsens *første side* som er skrevet helt ut. Om ingen av besvarelsens sider er fylt helt ut, markeres dette med en strek i begge noteringsfeltene		

6) Tegnsetting: Hvor riktig er *tegnsettingen*?	*Meget* lav grad av mestring innen vurderingsområdet	M. 1:		**Er det riktig bruk av punktum, utropstegn og spørsmålstegn?**	
	Lav grad av mestring...	M. 2:		**Er det riktig bruk av punktum komma ved oppramsing.**	
	Som det kan forventes ved starten av årstrinnet	M. 3:		Er det riktig bruk av komma mellom helsetninger?	
	Høy grad av mestring...	M. 4:		Brukes komma etter foranstilt leddsetning?	
	Meget høy grad av mestring innen vurderingsområdet	M. 5:		**Markeres direkte tale på en entydig måte?**	

Kommentarer:

Vedlegg 3 Vurderingsveiledning april 2012

Vurderingsrettleiing for lærarar
Utvalsprøver i skriving for 8. trinn

Denne vurderingsrettleiinga for Utvalsprøvene i skriving for 8. trinn
har to delar:

Del I inneheld generell informasjon om utvalsprøvene i skriving. I denne delen blir det gitt informasjon om utforminga av skriveprøvene som er den same frå år til år. Teksten i Del I er lik i vurderingsrettleiingane for skriveprøvene på 5. og 8. trinn.

Del II er ei vurderingsrettleiing som presenterer den aktuelle skriveprøva på 5. eller 8. trinn, og gir retningslinjer for vurderinga av prøva. Prøva skal vurderast innanfor seks forskjellige "vurderingsområde". Dei gir samla eit inntrykk av kvaliteten på elevteksten.
Norma for vurdering av dei enkelte områda blir illustrert med ulike "referansetekstar". Dei gir autentiske teksteksempel på dei ulike kvalitetsnivåa.

I Generelle opplysningar om utvalsprøvene i skriving

Kva er grunnleggjande ferdigheiter i skriving?

Utvalsprøvene i skriving er prøver av skriving som grunnleggjande ferdigheit på tvers av fag. Å kunne ytre seg forståeleg og på ein høveleg måte i og gjennom skriftlege tekstar er ei grunnleggjande ferdigheit som alle elevar har behov for både i skulen og i livet utanfor og etter skulen.

Det er fire overordna føremål med skriving som grunnleggjande ferdigheit i skulen:

- Grunnleggjande ferdigheit i skriving skal gi støtte og hjelp til fremming av læring i faga.
- Grunnleggjande ferdigheit i skriving skal gi tilgang til ressursar for demokratisk deltaking.
- Grunnleggjande ferdigheit i skriving skal gi tilgang til ressursar for deltaking i yrkeslivet.
- Grunnleggjande ferdigheit i skriving skal vere ein reiskap for utviklinga av eigen identitet.

Ved hjelp av og gjennom skriftlege tekstar kan vi kommunisere med folk som ikkje er til stade samtidig med oss. Når vi skriv, kan vi ta vare på tankane våre og sortere dei. Vi kan bli betre kjent med oss sjølve når vi skriv. Dessutan kan vi studere våre eigne idear og idear frå andre kritisk og utvikle ny kunnskap. Vi kan sjå framfor oss det som enno ikkje har vore eller nokon gong vil bli. Vi kan uttrykke meiningane våre og argumentere for dei, og slik delta i demokratiske samfunnsaktivitetar. Følgeleg er skriveferdigheita samansett. Skriveprøvene tar utgangspunkt i at det å skrive er ei handling som skal oppfylle eitt eller fleire føremål.

Skriftspråket er ein "semiotisk" ressurs på linje med det munnlege språket. Den kompetente skrivaren må ha utvikla visse motoriske ferdigheiter i å bruke blyant, penn, tastatur og digitale skriveprogram. Ein kompetent skrivar må også beherske reglane for teiknsetting og rettskriving.

Korleis skal skriveprøva vurderast?

Om primærtrekkvurdering

Utvalsprøvene i skriving legg til grunn at ei overordna skrivehandling skal vurderast på kvar prøve. Dette blir kalla *primærtrekkvurdering*. Oppgåva er altså formulert slik at skrivehandlinga som eleven skal utføre, er framheva spesielt. For eksempel kan skrivehandlinga vere "å beskrive" fotosyntesen der føremålet er å systematisere kunnskapen om korleis dette krinslaupet i naturen fungerer. Eller så kan føremålet med den beskrivande handlinga vere å påverke lesaren til å vere meir miljøbevisst. Eit anna eksempel kan vere skrivehandlinga "å reflektere". Eitt føremål med reflekterande skriving kan vere å utvikle innsiktsfull forståing av seg sjølv, for eksempel i ei dagbok. Men eit anna føremål kan vere å invitere andre til ei tilsvarande forståing, for eksempel i eit brev eller i eit lesarinnlegg i ein ungdomsspalte. I så fall er føremålet samhandling.

Kva skal vurderast i skriveprøva?

Om skrivehandlinga

Dei fem skrivehandlingane som kan prøvast i dei utvalsprøvene, er:

- Å reflektere over eit eller anna
- Å beskrive eit eller anna
- Å utforske eit eller anna
- Å sjå for seg eit eller anna
- Å overtyde andre medmenneske

Om vurderingsområda

Elevtekstane - svara dei gir på skriveoppgåva - skal vurderast innanfor seks vurderingsområde. Kvart vurderingsområde bidrar til å gi gyldig informasjon om og innsikt i kvaliteten på skriveferdigheita til eleven. Dei seks vurderingsområda er:

- Kommunikasjon
- Innhald
- Tekstoppbygging
- Språkbruk
- Rettskriving
- Teiknsetting

Vurderingsskjemaet til bruk ved vurderinga av elevtekstane er fordelt på dei seks vurderingsområda. På vurderingsskjemaet blir fokuset i kvart vurderingsområde konkretisert for den aktuelle skrivehandlinga. Det blir gjort ved hjelp av støttespørsmål. Føremålet med støttespørsmåla er å rette merksemda til vurderarane mot dei kvalitetane i elevtekstane som er relevante for den aktuelle skrivehandlinga og det enkelte vurderingsområdet.

Referansetekstar som reiskap i vurderinga av skriveprøvene

For å konkretisere bruken av vurderingsskjemaet på dei ulike vurderingsområda bruker vi *referansetekstar* i Del II i rettleiinga. Referansetekstane gir konkrete eksempel på kvalitetsnivå innanfor kvart vurderingsområde.

Referansetekstane er oppgåvesvar som er samla inn gjennom forundersøkingar. Fleire uavhengige vurderarar er einige om vurderingane av referansetekstane. Dei eksemplifiserer kvalitetsnivåa i vurderingsskalaen: "Som ein kan forvente ved starten av årstrinnet", "Mykje høg eller høg grad av meistring på vurderingsområdet" og "Mykje låg eller låg grad av meistring på vurderingsområdet".

Kva for vurderingsskala skal brukast i vurderinga av skriveprøvene?

På kvart av vurderingsområda skil vi mellom fem kvalitetsnivå. Dei fem kvalitetsnivåa er:

- Mykje låg meistring på vurderingsområdet
- Låg meistring på vurderingsområdet
- På eit meistringsnivå som ein kan forvente ved starten av årstrinnet
- Høg meistring på vurderingsområdet
- Mykje høg meistring på vurderingsområdet

II *Informasjon om utvalsprøva i skriving for 2012*

Utvalsprøva, pilotprøve frå 2012 for 8. trinn

Oppgåva:

Menneska har gjennom alle tider hatt lyst til å utforske verda gjennom utfordrande og farefulle ekspedisjonar. Nokon har reist på slike ekspedisjonar og møtt eksotiske og framande kulturar eller farlege utfordringar. Nokon har kome heim att med ny kunnskap og innsikt i kva verda er. (Denne delen blir lesen opp av læraren.)

Oppgåveformuleringa på arket til eleven:
Sjå for deg at du har vore med på ein ekspedisjon til ein spennande plass. Skriv ein tekst til klassen om ei spesiell utfordring de møtte undervegs.

Skrivehandlinga som blir prøvd: ***Å sjå for seg***

Oppgåva er forankra i desse formåla og kompetansemåla frå læreplanane i samfunnsfag og naturfag:

Formål:

- Naturvitenskapen har vokst fram som en følge av menneskers nysgjerrighet og behov for å finne svar på spørsmål om sin egen eksistens, liv og livsformer og vår plass i naturen og i universet og er på den måten en del av vår kultur. (...) Naturfag skal bidra til at barn og unge utvikler kunnskaper og holdninger som gir dem et gjennomtenkt syn på samspillet mellom natur, individ, teknologi, samfunn og forskning. (Naturfagplanen)

- Faget skal stimulere til utvikling av kunnskap om det kulturelle mangfaldet i verda i fortida og samtida, og til forståing av forholdet mellom naturen og dei menneskeskapte omgjevnadene (Samfunnsfagplanen)

Mål for opplæringen er at elevene skal kunne:

- bruke historiske kart og framstille oppdagingsreiser som europearar gjorde, skildre kulturmøte og samtale om korleis dei ulike kulturane opplevde møtet (Samfunnfagsplanen)
- skape forteljingar om menneske i fortida og bruke dei til å vise korleis menneske tenkjer og handlar ut frå samfunnet dei lever i (Samfunnsfagplanen)

Dette skal vurderast innanfor dei ulike vurderingsområda i skriveprøva 2012

Referansetekstar:

- tekstar som viser mykje høg eller høg meistring på vurderingsområdet
- tekstar som viser meistring på eit nivå som ein kan forvente ved starten av årstrinnet
- tekstar som viser mykje låg eller låg meistring på vurderingsområdet.

Vurderingsområde 1: Kommunikasjon

Under *Kommunikasjon* blir det vurdert om teksten formidlar innhaldet på ein relevant måte i forhold til lesaren.

Skrivaren skal:

- vende seg til delvis kjende, heilt ukjende eller tenkte lesarar
- tilpasse informasjonsmengda til lesarane
- gi teksten ei relevant overskrift
- vere tydeleg til stades i teksten.

Eksempel på meistringsnivåa mykje høgt eller høgt på vurderingsområdet

Referansetekst 2: Teksten har ei tydeleg forteljarstemme. Overskrifta er villeiande. Skrivaren tilpassar informasjonsmengda suksessivt slik at lesarforventningane blir bygde opp. Teksten har eit godt utvikla forteljar-eg.

Eksempel på det meistringsnivået som ein kan forvente

Referansetekst 1: Teksten kommuniserer godt og føremålet med teksten kjem klart fram gjennom overskrifta og i forteljinga. Men informasjonsmengda er ikkje så godt tilpassa lesaren, det er mykje informasjon om utsjånaden til deltakarane som ikkje er relevant til dømes.

Eksempel på meistringsnivåa mykje lågt eller lågt på vurderingsområdet

Vurderingsområde 2: Innhald

Under *Innhald* vurderer ein kor relevant og godt utdjupa innhaldet er.

Skrivaren skal:

- velje eit innhald som er relevant
- tilpasse mengda av innhald til føremålet
- undersøkje og uttrykkje kunnskapar om noko anna enn det privat erfarte
- undersøkje og uttrykkje eigne erfaringar, tankar og kjensler for å prøve ut eigen identitet, roller, meiningar og posisjonar
- trekke også andres erfaring, synspunkt, kjensler og råd inn i utvikling og formidling av eigne kunnskapar, tankar, synspunkt og kjensler
- fantasere i tekst.

Eksempel på meistringsnivåa mykje høgt eller høgt på vurderingsområdet

Eksempel på det meistringsnivået som ein kan forvente

Referansetekst 2: Teksten er sprikande når det gjeld innhaldsdimensjonen. Innhaldet er relevant i forhold til oppgåveformuleringa og i forhold til å utrykkje eigne og andres erfaringar. Men teksten utrykkjer i liten grad innsikt i ekspedisjonar. Mengda av innhald er godt tilpassa føremålet.

Eksempel på meistringsnivåa mykje lågt eller lågt på vurderingsområdet

Referansetekst 1: Innhaldet er relevant, men mengda av dei ulike innhaldsmomenta er ikkje godt tilpassa føremålet med teksten. Det er i liten grad utdjupa kva ekspedisjonen skulle gå ut på. Teksten inneheld mykje om sjølve reisa men bruker i liten grad kunnskap, og prøver i liten grad ut andre roller enn den erfarte.

Vurderingsområde 3: Tekstoppbygging

Under *Tekstoppbygging* vurderer ein tekststrukturen eller komposisjonen, samanhengen mellom dei enkelte delane av teksten og i dei enkelte delane av teksten.

Skrivaren skal:

- setje saman teksten med innleiing, hovuddel og avslutning
- strukturere teksten tematisk
- bruke relevante struktureringsmåtar, til dømes argumenterande, forklarande, forteljande og beskrivande
- bruke variert kopling mellom setningar.

Eksempel på meistringsnivåa mykje høgt eller høgt på vurderingsområdet

Referansetekst 2: Forteljingsstrukturen er fleksibel og variert og byggjer opp mot eit høgdepunkt som kjem heilt til slutt. Den korte innleiande situasjonsskildringa og kvar etterfølgjande del peikar framover mot høgdepunktet. Den forteljande strukturen i teksten er kopla på ein måte som er i samsvar med ein velbygd forteljing. Teksten er karakterisert ved

korte setningar i historisk presens. Sidan det er ein hendingsrekkefølgje, er eksplisitte koplingar ikkje nødvendige.

Eksempel på det meistringsnivået som ein kan forvente

Eksempel på meistringsnivåa mykje lågt eller lågt på vurderingsområdet
Referansetekst 1: Teksten har ein kronologisk forteljande struktur, men byggjer ikkje opp spenning mot eit høgdepunkt. Innleiinga tek svært mykje plass og det gjer at forteljinga ikkje kjem i gang. Hendingane i forteljinga er vilkårlege og har lite utbyggingspotensial fordi dei er lite relevante for det eigentlege poenget i forteljinga: ekspedisjonen til Afrika.

Vurderingsområde 4: Språkbruk

Under *Språkbruk* vurderer ein ordval og setningsbygning.

Skrivaren skal:
- bruke eit relevant ordtilfang og relevante språklege uttrykksmåtar
- variere språkbruken ved bruk av ulike setningstypar, ved hjelp av ordklassar som adjektiv og adverb
- bruke relevante stilistiske uttrykksformer, for eksempel skildring, retoriske spørsmål i ein overbevisande tekst, eller karakteriserande dialog i en forteljande tekst.

Eksempel på meistringsnivåa mykje høgt eller høgt på vurderingsområdet
Referansetekst 2: Variert ordtilfang tilpassa forteljinga. Nokså konvensjonelle skildringar som "kikket ut", stiv av skrekk" osv. Bruk av passande, stemningsskapande ordval som "børsa" og "gjenkjenne lyden" og "avfyre skudd", men også kvardagsleg ordtilfang som "Hun begynte å stresse." Svært god bruk av dramatiserande dialog for å understreke handlingsaspektet og oppbygginga mot høgdepunktet.

Eksempel på det meistringsnivået som ein kan forvente

Eksempel på meistringsnivåa mykje lågt eller lågt på vurderingsområdet
Språkbruken er enkel, med mange heilsetningar. Det er ein del bruk av skildrande adjektiv, men ikkje meir komplekse nominalfrasar. Det vekslar mellom bruk av subjekt og tidsadverbial i framfeltet. Det er ikkje brukt karakteriserande dialog, og sparsamt med skildringar.

Vurderingsområde 5: Rettskriving

Skrivaren skal:
- meistre korrekt ortografi i høgfrekvente ord som ikkje er lydrette, til dømes ord med dobbel konsonant og stumme konsonantar

- meistre orddanning på bakgrunn av morfologisk informasjon (bøyingar og avleiingar) og vise innsikt i ordklassar

Eksempel på meistringsnivåa mykje høgt eller høgt på vurderingsområdet
For å bli plassert på meistringsnivået *mykje høgt* (nivå 5), skal skrivaren skrive tilnærma alle orda rett, men nokre "slengarfeil" og talemålsrelaterte feil kan førekome.

Skrivaren meistrar ortografisk skriving, men har ein del slengarfeil.

Eksempel på det meistringsnivået som ein kan forvente
For å bli plassert på dette meistringsnivået (nivå 3) skal skrivaren meistre korrekt rettskriving og ha få skrivefeil per side.

Eksempel på meistringsnivåa mykje lågt eller lågt på vurderingsområdet
Når skrivaren har meir enn 8 feilskrivne ord per side, blir skrivaren plassert på meistringsnivå *mykje lågt* (nivå 1).

Referansetekst 1: Det er ein god del ortografiske feil i teksten når det gjeld bruk av dobbel konsonant, o-å, stor – liten bokstav og orddeling.

Vurderingsområde 6: Teiknsetting

Skrivaren skal:
- bruke komma mellom heilsetningar og etter framforstilt leddsetning
- markere direkte tale med replikkstrek eller kolon og hermeteikn.

Eksempel på meistringsnivåa mykje høgt eller høgt på vurderingsområdet
Referansetekst: Det er stort sett rett bruk av store og små skiljeteikn. Dessutan svært avansert bruk av utropsteikn, spørsmålsteikn og gjennomgåande korrekt bruk av sitatteikn.

Eksempel på det meistringsnivået som ein kan forvente

Eksempel på meistringsnivåa mykje lågt eller lågt på vurderingsområdet
Referansetekst 1: Det er gjennomgåande rett bruk av punktum i teksten. Men det er sjeldan brukt komma mellom heilsetningar og etter framforstilt leddsetning. Og det er ikkje komma når ledd blir gjentatt: "Marie hun …"

Her er to elevtekster fjernet fra
dokumentet da det ikke er gitt samtykke
til at de kan brukes i forskningsøyemed.

Litteratur

Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Aristoteles (1999). *Den nikomakiske etikk*. Oslo: Bokklubben dagens bøker.

Aunger, R. (2004). *Reflexive Ethnographic Science*. Walnut Creek: AltaMIra Press.

Austin, J. L. (1962) *How to do Things with Words: The William James Lectures delivered at Harvard University in 1955*. Oxford: Clarendon Press.

Baird, J.-A., Hopfenbeck, T. N., Newton, P., Stobart, G. & Steen-Utheim, A.T. (2014). *Assessment and Learning: State of the Filed Review*. Oslo: Knowledge Center for Education.

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*, 133-153.

Bakhtin, M. M. (1984). *Problems of Dostoevsky's Poetics*. Minneapolis: University of Michigan Press.

Bakhtin, M. M. (1999). Towards a methodology for the human sciences. I C. Emerson & M. Holquist (red.), *Speech Genres and Other Late Essays* (159-172). Austin: University of Texas Press.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54-74.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*, 51-75.

Barrett, S. (2001). The impact of training in rater variability. *International Education Journal*, *2*, 49-58.

Baume, D., Yorke, M. & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29*, 451-477.

Behizadeh, N. & Engelhard jr., G. (2011). Historical view on the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*, 189-211.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*, 2-9.

Berge, K. L. (2005). Studie 3: Skriveprøvens pålitelighet. I K. L. Berge, L. S. Evensen, F. Hertzberg & W. Vagle (red.), *Ungdommers skrivekompetanse. Bind 1* (s. 101-113). Oslo: Universitetsforlaget.

Berge, K. L. (2009). Er tolkningsfellesskap mulig å oppnå i skriveprøver? I O. K. Haugaløkken, L. S. Evensen, F. Hertzberg & H. Otnes (red.), *Tekstvurdering som didaktisk utfordring* (s. 44-54). Oslo: Universitetsforlaget.

Berge, K. L., Evensen, L. S., Hertzberg, F. & Vagle, W. (red.) (2005). *Ungdommers skrivekompetanse. Bind 1 og 2*. Oslo: Universitetsforlaget.

Berge, K. L., Evensen, L. S., Thygesen, R. & Fasting, R. B. (2007). *Sluttrapport: nasjonale prøver i skriving som grunnleggende ferdighet*. Stavanger: Universitetet i Stavanger, Lesesenteret ved det humanistiske fakultet.

Bergenholtz, C. (2011). Knowledge brokering: spanning technological and network boundaries. *European Journal of Innovation Management, 14*, 74-92.

Biernacki, P. & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research, 10*, 141-163.

Black, E. L. (1962). The marking of G.C.E. scripts. *British Journal of Educational Studies, 11*, 61-71.

Breland, H. M. & Jones, R. J. (1984). Perception of writing skills. *Written Communication, 1*, 101-119.

Brunstad, P. O. (2007). Faglig klokskap – mer enn kunnskap og ferdigheter. *Pacem 10*, 59-70. Lastet ned 1. november 2014, fra http://www.pacem.no/2007/2/2klokskap/5brunstad/

Buber, M. (2007). *Jeg og Du*. Oslo: Cappelen.

Burr, V. (2003). *Social Constructionism*. New York: Routledge.

Cetina, K. K. (1999). *Epistemic Cultures*. Cambridge, MA: Harvard University Press.

Colombini, C. B. & McBride, M. (2012). "Storming and norming": Exploring the value of group development models in addressing conflict in communal writing assessment. *Assessing Writing, 17*, 191-207.

Comte, A. (2009). *A General view of positivism.* Cambridge: Cambridge University Press.

Condon, W. (2011). Reinventing writing assessment: How the conversation is shifting. *Journal of the Council of Writing Program Administrator, 34*, 162-182.

Connor-Linton, J. (1995). Looking behind the curtain: What Do L2 Composition Ratings Really Mean? *TESOL Quarterly, 29*, 762-765.

Cooksey, R. W., Freebody, P. & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation: An International Journal on Theory and Practice, 13*, 401-434.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*, 1-21.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. I H. Wainer & H. I. Braun (red.), *Test Validity* (s. 3-17), Hillsdale: L. Erlbaum Associates.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*, 391-418.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5*, 7-29.

Dewey, J. (1902). *The Child and the Curriculum*. Chicago: The University of Chicago Press.

Dewey, J. (2008). *Democracy and Education. An Introduction to the Philosophy of Education*. Lastet ned 10. november 2014, fra http://www.gutenberg.org/files/852/852-h/852-h.htm

Diederich, P. B., French, J. W. & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service. (ERIC Document Reproduction No. ED 002 172).

Dilthey, W. (2010). Ideas for a descriptive and analytic psychology. I R. A. Makkreel & F. Rodi (red.), *Wilhelm Dilthey: Selected Works, Volume 2: Understanding the Human World* (s. 115-210). Princeton, NJ: Princeton University Press.

Drucker, P. F. (1959). *Landmarks of Tomorrow*. New York: Harper & Brothers. Lastet ned 15. november 2014, fra http://documents.irevues.inist.fr/bitstream/handle/2042/30294/XX_CNE-LIPSOR_1197.pdf.txt?sequence=3

Du, Y. & Wright, B. D. (1997). Effects of student characteristics in a large-scale direct writing assessment. I M. Wilson, G. Engelhard Jr. & K. Draney (red.), *Objective measurement: Theory into practice* (Vol. 4, 1-24). Stamford, CT: Ablex.

Dysthe, O. (1987). *Ord på nye spor. Innføring in prosessorientert skrivepedagogikk*. Oslo: Det Norske Samlaget.

Dysthe, O. (1995). *Det flerstemmige klasserommet*. Oslo: Ad Notam Gyldendal.

Elder, C., Barkhuizen, G., Knoch, U. & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*, 37-64.

Elder, C., Knoch, U., Barkhuizen, G. & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*, 175-196.

Elliot, N. & Perelman, L. (2012). Strategies in contemporary writing assessment. Bridging the two cultures. I N. Elliot & L. Perelman (red.), *Writing Assessment in the 21st Century* (s. 149-156). New York: Hampton Press, Inc.

Engelhard jr., G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge: MIT Press.

Evensen, L. S. (2002). Convention from below: Negotiating interaction and culture in argumentative writing. *Written Communication, 19*, 382-413.

Evensen, L. S. (2009). 'With a little help from my friends'? Theory of learning in applied linguistics and SLA. *Journal of Applied Linguistics, 3*, 333-353.

Evensen, L. S. (2010). En gyldig vurdering av elevers skrivekompetanse? I J. Smidt, I. Folkvord & A. J. Aasen (red.), *Rammer for skriving* (s. 13-32). Trondheim: Tapir forlag.

Evensen, L. S. (2012). Underveis mot et tolkningsfellesskap: lærerstemmer om elevtekster. I S. Matre, D. K. Sjøhelle & R. Solheim (red.), *Teorier om tekst i møte med skolens lese- og skrivepraksiser* (s. 151-160). Oslo: Universitetsforlaget.

Evensen, L. S. (2013). *Applied Linguistics. Towards a New Integration?* London: Equinox.

Evensen, L. S. (2014). Men kan det komme noe godt fra lærerværelset? Om validitet i læreres vurdering. I R. Hvistendahl & A. Roe (red.), *Alle tiders norskdidaktiker* (s. 245-257). Oslo: Novus Forlag.

Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research*. London: Routledge.

Fasting, R., Thygesen, R., Berge, K. L., Evensen, L. S. & Vagle, W. (2009). National assessment of writing proficiency among Norwegian pupils in compulsory schools. *Scandinavian Journal of Educational Research, 53*, 617-637.

Freedman, S. W. (1979a). How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology, 71*, 328-338.

Freedman, S. W. (1979b). Why do teachers give the grades they do? *College Composition and Communication, 30*, 161-164.

Gadamer, H.-G. (2010). *Sannhet og metode. Grunntrekk i en filosofisk hermenetutikk*. Oslo: Pax Forlag.

Gipps, C. V. (1994). *Beyond testing: towards a theory of educational assessment*. London: The Falmer Press.

Goodwin, C. & Duranti, A. (1992). Rethinking context: an introduction. I Goodwin, C & A. Duranti (red.), *Rethinking context: Language as an interactive phenomenon* (s. 1-42). Cambridge: Cambridge University Press.

Gorman, T. P., Purves, A. C. & Degenhart, R. E. (red.) (1988). *The IEA study for written composition 1: The international writing tasks and scoring scales*. Oxford: Pergamon Press.

Gubrium, J. F. & Holstein, J. A. (1997). *The New Language of Qualitative Method*. Oxford: Oxford University Press.

Gyagenda, I. S. & Engelhard jr., G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences. *Journal of Applied Measurement, 10*, 225-246.

Habermas, J. (1989). *The structural transformation of the public sphere: an inquiry into a category of bourgeois society*. Cambridge Mass.: MIT Press.

Hamp-Lyon, L. (2000). Social, professional and individual responsibility in language testing. *System, 28*, 579-591.

Hellesnes, J. (1992). Ein utdana mann og eit dana menneske: framlegg til eit utvida daningsomgrep. I E. L. Dale (red.), *Pedagogisk filosofi* (s. 79-103). Oslo: AdNotam Gyldendal.

Henry, J. (2004). Science and the coming of Enlightenment. I M. Fitzpatrick, P. Jones, C. Knellwolf & I. McCalman (red.), *The Enlightenment World* (s. 10-26). Abingdon: Routledge.

Heritage, J. & Clayman, S. (2010). *Talk in action. Interactions, identities, and institutions*. Hoboken, NJ: Wiley-Blackwell.

Hoel, T. L. (2000). *Skrive og samtale. Responsgrupper som læringsfellesskap*. Oslo: Gyldendal Akademisk.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

Huot, B. (2002). "(Re)articulating writing assessment for teaching and learning". *All USU Press Publications*. Book 137. Lastet ned 20. september 2013, fra http://digitalcommons.usu.edu/usupress_pubs/137

Husserl, E. (1970). *The Crisis of European Sciences and Transcendental Philosophy*. Evanston: Northwestern University Press.

Igland, M.-A. (2008). *Mens teksten blir til. Ein kasusstudie av lærarkommentarar til utkast*. Det utdanningsvitenskapelige fakultet, Universitetet i Oslo.

Imsen, G. (1998). *Elevens verden: Innføring i pedagogisk psykologi*. Oslo: Tano Aschehoug.

James, M. (2006). Assessment and learning. I S. Swaffield (red.), *Unlocking Assessment. Understanding for reflection and application* (s. 20-35). Abingdon, UK: Routledge.

Järvinen, M. & Mik-Meyer, N. (red.) (2005). *Kvalitative metoder i et interaktionistisk perspektiv: interview, observationer og dokumenter*. København: Hans Reitzels Forlag.

Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing, 14*, 3-24.

Johnson, C. & Elliot, N. (2010). Undergraduate technical writing assessment. *Programmatic Perspectives, 2*, 110-151. Lastet ned 11. desember 2014, fra http://cptsc.org/pp/vol2-2/johnson_elliot2-2.pdf

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing, 20*, 37-52.

Jølle, L. (upubl.). To become a more proficient rater: What does it take?

Jølle, L. (in press). Rater strategies for reaching agreement on pupil text quality. *Assessment in Education: Principles, Policy & Practice*, DOI: 10.1080/0969594X.2015.1034087

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review, 2*, 130-144.

Jordan, A., Carlile, O. & Stack, A. (2008). *Approaches to Learning: A Guide for Teachers*. Columbus: McGraw-Hill Education.

Kane, M. T. (2006). Validation. I R. L. Brennan (red.), *Educational measurement* (s. 17-64). Wetpoint, CT: Praeger Publisher.

Kennington, R. (2004). *On Modern Origins: Essays in Early Modern Philosophy*. Lanham, Maryland: Lexington Books.

Kjørup, S. (2008). *Menneskevidenskaberne. Humanistiske forskningstraditioner 2*. Roskilde: Roskilde universitetsforlag.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*, 3-31.

Kuhn, T. (1962). *The Structure of Scientific revolutions*. Chicago: The University of Chicago Press.

Larsson, S. (2005). Om kvalitet i kvalitativa studier. *Nordisk Pedagogik, 25*, 16-35.

Lave, J. & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Leckie, G. & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater Experience. *Journal of Educational Measurement, 48*, 399-418.

Li, H. (2003). The resolution of some paradoxes related to reliability and validity. *Journal of Educational and Behavioral Statistics, 28*, 89-95.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*, 543-560.

Lindgren, L. (2014). Praktisk kunnskap under press. *Forskerforum, 9*, s. 40-41.

Linell, P. (2001). *Approaching dialogue: Talk, interaction and contexts in dialogical perspectives*. Amsterdam: John Benjamins Publishing Company.

Linell, P. (2009). *Rethinking language, mind, and world dialogically: Interactional and contextual theories of human sense-making*. Charlotte, NC: Information Age Publishing.

Lock, A. & Strong, T. (2014). *Sosial konstruksjonisme – teorier og tradisjoner*. Bergen: Fagbokforlaget.

Lumley T. & McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 12*, 54-71.

Lynne, P. (2004). *Coming to Terms: A Theory of Writing Assessment*. All USU Press Publications. Book 149. Lastet ned 21. august 2014, fra http://digitalcommons.usu.edu/usupress_pubs/149

Markova, I. (2003). Constitution of the self: Intersubjectivity and dialogicality. *Culture & Psycology, 9*, 249-259.

Markova, I. (2006). On 'the inner alter' in dialogue. *International Journal for Dialogical Science, 1*, 125-147.

Markova, I. & Linell, P. (1996). Coding elementary contributions to dialogue: Individual acts versus dialogical interactions. *Journal for the Theory of Social Behaviour, 26*, 353-373.

Mason, J. (1994). Linking qualitative and quantitative data analysis. I A. Bryman & R. G. Burgess (red.), *Analyzing Qualitative Data* (s. 89-110). New York: Routledge.

Matre, S., Berge, K. L., Evensen, L. S., Fasting, R. B., Solheim, R. & Thygesen, R. (2011). *Developing national standards for the teaching and assessment of writing. Rapport frå forprosjekt Utdanning 2020*. Trondheim: HiST, Skrivesenteret.

Matre, S. & Solheim, R. (2014). Lærarsamtalar om elevtekstar – mot eit felles fagspråk om skriving og vurdering. I R. Hvistendal og A. Roe (red.), *Alle tiders norskdidaktiker* (s. 219-244). Oslo: Novus Forlag.

McNamara, T .F. (1996). *Measuring second language performance*. London: Longman.

Meadows, M. & Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Lastet ned 17. september 2014, fra https://cerp.aqa.org.uk/research-library/review-literature-marking-reliability

Merleau-Ponty, M. (1994). *Kroppens fenomenologi*. Oslo: Pax Forlag.

Messick, S. (1989). Validity. I R. L. Linn (red.), *Educational measurement* (s. 13-103). New York: American Council on Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256.

Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of the decision-making behavior of compositon markers. I M. Milanovic & N. Saville (red.*), Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Colloquium* (s. 92-114). Cambridge: Cambridge University Press.

Mislevy, R. J. (1994). Can there be reliability without "reliability?" *Educational Testing Service*. Research Memorandum RM-94-18-ONR. Lastet ned 10. oktober 2014, fra http://www.google.no/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CB8QFjAA&url=http%3A%2F%2Fwww.education.umd.edu%2FEDMS%2Fmislevy%2Fpapers%2FReliabilityWithoutReliability.doc&ei=l1SlVJv2H4SuygPw9YDYDw&usg=AFQjCNFoEAZMI2GQ6jpQyj_UnexKcgsT3Q&bvm=bv.82001339,d.bGQ&cad=rja

Mislevy, R. J. (2004). Can there be reliability without "reliability?". *Journal of Educational and Behavioral Statistics, 29*, 241-244.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5-12.

Moss, P. A. (2004). The meaning and consequences of "reliability". *Journal of Educational and Behavioral Statistics, 29*, 245-249.

Nagel, T. (1989). *The View From Nowhere*. Oxford: Oxford University Press.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). I C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers & A. Stansell (red.), *International Advances in Writing Research: Cultures, Places, Measures* (s. 121-150). Fort Collins, CO: WAC Clearinghouse & Parlor Press.

Perelman, L. (2014). When "state of the art" is counting words. *Assessing Writing, 21*, 104-111.

Platon (2001). *Platon: Samlede verker bind V*. Oslo: Vidarforlaget

Pollner, M. (1987). *Mundane reason: reality in everyday life and sociological discourse*. Cambridge: Cambridge University Press.

Potter, J. (1998). Cognition as context (whose cognition?). *Research on Language and Social Interaction, 31*, 29-44.

Potter, W. J. & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research, 27*, 258-284.

Purves, A. C. (1992). A comparative perspective on the performance of students in written composition. I A. C. Purves (red.), *The IEA study for written composition 2* (s. 129-152). Oxford: Pergamon.

Purves, A. C., Gorman, T. P. & Takala, S. (1988). The development of the scoring scheme and scales. I T.P. Gorman, A. C. Purves & R. E. Degenhart (red.), *The IEA study for written composition 1* (s. 41-58). Oxford: Pergamon.

Rienecker, L. & Jørgensen, P. S. (2013). *Den gode oppgaven. Håndbok i oppgaveskriving på universitet og høyskole*. Bergen: Fagbokforlaget.

Rommetveit, R. (1974). *On Message Structure. A Framework for the Study of Language and Communication*. Hoboken, NJ: John Wiley & Sons.

Rorty, R. (red.) (1967). *The Linguistic Turn. Essays in Philosophical Method*. Chicago: University of Chicago Press.

Rudner, L. M. (1992). Reducing errors due to the use of judges. *ERIC/TM Digest*. Lastet ned 10. desember 2014, fra http://files.eric.ed.gov/fulltext/ED355254.pdf

Saal, F. E., Downey, R. G. & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*, 413-428.

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*, 159-179

Saldaña, J. (2013). An introduction to codes and coding. I J. Saldaña (red.), *The Coding Manual for Qualitative Researchers* (s. 1-41). Los Angeles, Ca: Sage.

Säljö, R. (2001). *Læring i praksis: Et sosiokulturelt perspektiv*. Oslo: Cappelen Akademisk.

Schatzki, T. (2001). Introduction: practice theory. I T. R. Schatzki, K. Knorr Cetina & E. von Savigny (red.), *The Practice Turn in Contemporary Theory* (s. 1-14). New York: Routledge.

Schatzki, T. R., Knorr Cetina, K. & von Savigny, E. (red.) (2001). *The Practice Turn in Contemporary Theory*. New York: Routledge.

Schleiermacher, F. (1977). *Hermeneutics: The Handwritten Manuscripts*. Atlanta: Scholars Press.

Schön, D. (2004). *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books.

Seale, C. (1999). *The Quality of Qualitative Research*. London: Sage.

Searle, J. R. (1965). What is a speech act? I M. Black (red.), *Philosophy in America* (s. 221-239). London: Allen and Unwin.

Shermis, M. D. & Hamner, B. (2013). *Contrasting State-of-the-Art Automated Scoring of Essays: Analysis*. Lastet ned 18. desember 2014, fra: http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Silverman, D. (2005). *Doing Qualitative Research*. London: Sage.

Skaftun, A. (2002). Dialogen som paradigme. *Nordlit, 12*, 137-152.

Standards for Reporting on Empirical Social Science Research in AERA Publications (2006). *Educational Researcher, 35*, s. 33-40.

Stuhlmann, J., Danile, C., Dellinger, A., Kenton, R. & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology, 2*, 107-127.

Svennevig, J. (1999). Innledning: Samtaleforskning og språkvitenskap. *Norsk Lingvistisk Tidsskrift, 17*, 3-13.

Swain, S. S. & Le Mahieu, P. (2012). Assessment in a culture of inquiry: The story of the national writing project's analytic writing continuum. I N. Elliot & L. Perelman (red.), *Writing Assessment in the 21st Century* (s. 45-67). New York: Hampton Press, Inc.

Tjora, A. (2010). *Kvalitative forskningsmetoder i praksis*. Oslo: Gyldendal Akademisk.

Vagle, W. & Evensen, L. S. (2005). Studie 5: Oppgavesettene og elevenes oppgavevalg i KAL-årene. I K. L. Berge, L. S. Evensen, F. Hertzberg og W. Vagle (red.), *Ungdommers skrivekompetanse. Bind 1* (s. 161-204). Oslo: Universitetsforlaget.

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? I L. Hamp-Lyons (red.) *Assessing second language writing in academic contexts* (s. 111-125). Norwood, NJ: Ablex.

Vygotsky, L. (1978). *Mind in Society. The Development of Higher Psycological Processes*. Cambridge, MA: Harvard University Press

Vygotsky, L. (1987). Thinking and speech. I R. W. Rieber & A. S. Carton (red.), *The collected works of L. S. Vygotsky: Volume 1: Problems of general psychology* (s. 39-288). New York: Plenum.

Wegerif, R. & Mercer, N. (1997). Using computer-based text analysis to integrate qualitative and quantitative methods in research on collaborative learning. *Language and Education, 11*, 271-286.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

Wertsch, J. V. (1991). *Voices of mind: A sociocultural approach to mediated action*. Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.

Wilson, M. (2006). *Rethinking Rubrics in Writing Assessment*. Portsmouth, NH: Heinemann.

Wyatt-Smith, C., Klenowski, V. & Gunn, S. (2010). The centrality of teachers´ judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*, 59-75.

Yancey, K. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Ccommunication, 50*, 483-503.

Yin, R. K. (2009). *Case Study Research. Design and Methods*. Los Angeles, Ca: Sage.

Yu, C. H. (2005). Test-retest reliability. I K. Kempf-Leonard (red.), *Encyclopedia of Social Measurement,* vol. 3 (s. 777-784). Lastet ned 12. august 2014, fra http://arsmath.org/msl/Library/statistics/common/Encyclopedia-of-Socia-%20Measurement-Vol-3-%28P-Z%29.pdf

Artikkel 1

Pair assessment of pupil writing: A dialogic approach to studying the development of rater competence

Abstract

This paper reports on rating during the development of a Norwegian sample-based national assessment of L1 writing as a key competency. This assessment is to be officially introduced in August 2014. Novice members of a national rater panel to assess Year 8 pupils' texts were studied during three of their successive training sessions; in June 2011, November 2011 and April 2012. My purpose was to conduct an exploratory investigation into how the rating practice of novice raters might develop during such a preparatory stage. The raters in this study mainly assessed in pairs, and data sources were assessment dialogues. The analysis of transcripts showed that rater behaviour changed only to a minor extent towards an increased use of shared assessment resources. The quality of the assessment dialogues did not change much either, leaving the impression that raters often reached consensus without much discussion. Since reliable scoring is a collective task, however, it is argued that a more balanced use of resources, both those attained from teacher practice and those attained from being a member of a national panel, may together with more exploration-oriented dialogue be necessary to achieve sufficient reliability.

1. **Introduction**

1.1 **Background**

In Norway, writing development and expectations of writing proficiency for compulsory school pupils are of current interest among policymakers, researchers and teachers. In 2006, Norway introduced a national curriculum where different key competencies were integrated in and adapted to each subject, and *writing* is one of these competencies (Knowledge promotion, 2007). Policymakers also wanted information about the pupils' writing performance, and in 2005 a first attempt to test pupils' proficiency in Norwegian (L1) writing as a key competency was made through a national writing test (Thygesen, Berge, Evensen & Fasting, 2007). However, due to low interrater reliability the results were of such dubious value that they could hardly be reported back to the pupils or teachers, and further tests were postponed.

Several explanations for the low reliability have been posited, such as the *level of expertise* among the raters and the *lack of a shared assessment culture* (Fasting, Thygesen, Berge, Evensen & Vagle, 2009). The raters, who were all experienced teachers, had to deal with an unfamiliar theoretical construct for writing[24] when making their assessments, and this made the task doubly difficult for them. Even the most experienced teachers were turned into novice raters.[25] The lack of a common assessment culture might also be due to vague formulations in the various subject curricula as to what to expect from pupils at different Year levels. Furthermore, there has been little research on the assessment of writing in a Norwegian context (Evensen, 2009).

Therefore, when it was decided to reintroduce a writing test from 2014, a number of critical factors needed to be addressed. By designing the test as a sample-based national assessment it became possible, both financially and practically, to establish a national assessment panel (The Norwegian Assessment Panel, NAP) to carry out the assessment. During a two-year period before the re-launching of the test approximately 90 teachers were recruited to the panel.[26] NAP is now a semi-permanent panel that meets for three days on two occasions each year. The theoretical writing construct are presented to the raters on the panel, they participate in plenary and group discussions both on criteria for assessment and what to expect from pupils texts at different year levels, and they assess pupils' writings in pairs or groups according to a primary trait model. Using a five-level scale[27], the following six text domains are assessed: Communication, Content, Composition, Use of language, Spelling and Punctuation. Bearing the lack of shared standards in mind, the goal is not only to encourage the raters to use standards in common ways when assessing, but also to continue refining and developing these standards. A panel model, as used in this context, is expected

[24] The Norwegian curriculum from 2006, where writing is introduced as a key competency, has encouraged Norwegian writing researchers to develop a theoretical writing construct that forms the basis of how to understand writing. A model, called Wheel of writing, displays the basic concepts in this construct, where acts and purposes of writing are central (Fasting, Thygesen, Berge, Evensen & Vagle, 2009; Wheel of writing, 2012).

[25] It has been important to move the conceptual understanding of writing from exclusively belonging to Norwegian language instruction (L1) to an understanding of writing as a key competency that makes it relevant in every subject for a number of purposes (DeSeCo., 2005). This turn has been difficult for teachers and raters to adapt to. What is meant by proficient writing in science has for instance not been a frequent topic of discussion among assessors of writing.

[26] Since the NAP was in an early stage when the data material was compiled, the total number of raters increased from one session to the next as more raters were recruited. In June 2012, when the last tape recordings were made, the panel consisted of 90 raters.

[27] The scale ranges from M1 (much lower than expected from most students at the actual level), through M3 (as expected), up to M5 (much better than expected from most students at the actual level).

to improve the reliability of assessments, that is, changes should take place (AERA, APA, & NCME, 1999**;** Breland, Bridgeman & Fowles, 1999).

The present study should be understood with this background in mind. The purpose is to conduct a qualitative investigation into how the rating practice of novice raters within the panel develops over time.

1.2 Review of literature

For over a century researchers have been concerned about the trustworthiness of rater scoring (see Starch & Elliot, 1912). A major breakthrough came in the 1960s when two scoring models, i.e. analytic (Diederich, French, and Carlton, 1961) and holistic scoring (Godshalk, Swineford, and Coffman, 1966), were developed to improve the quality of the assessments. But even though the models still dominate the various writing test designs, the problems attached to a mark's trustworthiness have not disappeared. On the contrary, it seems that (inter)rater reliability issues are widely and vigorously discussed both among and between scholars and lay people in contemporary debates (Yancey, 1999). Bearing this in mind, it would not be an exaggeration to claim that writing assessment has turned out to be a difficult endeavour (Berge, 2005; Fasting et al., 2009; McNamara, 1996). A very recent illustration of this is found in Sweden where low interrater reliability in national writing assessments has led the government to evaluate the pros and cons in allowing performance writing assignments to be a part of the national testing system (Skolinspektionen, 2013).

When it comes to assessment, there are two main methods for studying reliability: 'product-oriented' and 'process-oriented'. The 'product-oriented' method evaluates the quality of the scoring according to its outcomes. That is, accurate and consistent scoring equals quality scoring (Jonsson & Svingby, 2007). On the other hand, the 'process-oriented' method is more concerned with the decision-making processes and scoring strategies used when assessing writing (Crisp, 2010; Suto & Greatorex, 2008). It might be fair to say that while the 'product-oriented' method has its roots in the field of educational measurement, the ancestors of the 'process-oriented' method are likely to be found in the field of writing research.[28] This study is situated within the latter tradition.

When examining the rating process two main variables can be manipulated: scoring construct and rater quality (Suto, 2012). When it comes to the construct variable, research

[28] See Huot (2002) for a presentation of the development of writing assessment.

literature shows, for instance, how *marking methods* (i.e. analytic versus holistic methods) affect the scoring (Barkaoui, 2011a; Knoch, 2007), and also how different *prompts* may lead to rater related variance in scoring (Weigle, 1999). Likewise, when examining rater quality, it is pointed out that *rater background* (Cumming, 1990; Leckie & Baird, 2011; Lim, 2011; Wiseman, 2012), *rater 'style'* (Eckes, 2008; Vaughan, 1991), *rater training* (Brown, Glasswill & Harland, 2004; Purves, 1992; Stuhlmann, Daniel, Dellinger, Kenton & Powers, 1999; Weigle, 1994, 1998) and also *rater values* and *expectations* (Baker, 2010) are variables that may influence scoring. One branch of the research related to rater quality comprises the efforts to develop a clearer model of a rater's cognitive processes when scoring (Crisp, 2010; Cumming, Kantor, & Powers, 2002; Lumley, 2002; Sakyi, 2000).

Despite this comprehensive research, there are still areas where further exploration is needed. Little is known, for example, about how raters relate their rating practice to standards. Sadler's seminal work on this topic points to the importance of exemplars and verbal descriptors as methods for overcoming the generally accepted way of assessing, where the rater´s tacit and individually held standard is likely to lead to arbitrary scoring (Sadler, 1987, 2011). Empirical studies indicate that raters pay too little attention to common frameworks and that, rather, they tend to rely on their tacit knowledge when scoring (Barrit, Stock & Clark, 1986; Wyatt-Smith, Klenowski & Gunn, 2010). These findings lead to the following question: Within an assessment panel, how do novices´ rating practices relate to standards over time as they gradually become experienced and proficient raters? Working with this research question will help us to gain a better understanding of how a joint assessment culture can potentially arise (Evensen, 2012; Lim, 2011).

1.3 Research aims

There is a general interest in gaining knowledge of the development from being a novice to becoming a proficient rater. Studies have shown how novice raters can learn to rate appropriately relatively quickly (Lim, 2011; Weigle, 1998), but while these studies investigate the quality of the scoring according to the outcomes ('product-oriented'), they do not look into the nature of the assessment practice itself ('process-oriented'). There is a need to investigate how raters make decisions when assessing and how their rating behaviour changes as they gain more experience. At stake is also the transparency of assessments. It is important to gain knowledge about raters' assessment practices with high-stake and/or national tests since their understanding and use of standards influences both the teaching of writing and

teacher writing assessment throughout school classrooms.[29] The present study aims to address the following research questions:

(1) How do novice raters within an assessment panel assess pupils' writing, and how does this praxis change over time as the raters gain experience and proficiency?
(2) What effects do changes in the assessment praxis potentially have on the assessments?

2. **Methodology**

2.1 The dialogic aspect of joint assessment

The pair or group assessments add an exploratory dimension to this study, as most research to date focuses on rater(s) rating individually (Lumley, 2005; Penny, Johnson & Gordon, 2000). In this study, however, the focus is on the nature of discussions when raters are working together. The current context presupposes that raters engage in genuine dialogue, and do not make merely coincidental monological contributions on the topic (Linell, 2009). Wegerif and Mercer distinguish between disputional, cumulative, and exploratory talk (Wegerif & Mercer, 1997). Disputional talk is confrontational and monological in the sense of not taking the other raters' contributions into consideration when making judgments; cumulative talk builds consensus in the sense of acknowledging others' utterances through accumulation; and exploratory talk is complex in the sense that the raters engage constructively and critically in the talk, willing to "attune to the 'attunement' of the other" (Rommetveit, 1992, p. 20). Even if all three types of talk can contribute to effective communication, the third form is qualitatively different because knowledge is not only shared through exploratory talk, but also critically assessed. Here "knowledge is made more publicly accountable and reasoning is more visible in the talk" (Wegerif & Mercer, 1997, p. 53), which are preconditions for developing a shared understanding of the hallmarks of quality performance in writing within a community of practice, and hence maximise reliable assessments. This transparency makes it likely that judgments are repeated by raters later on when facing similar assessment situations.

In this study, the quality of the assessment dialogue is operationalized in terms of the raters' response patterns, building on Grauman's (1990) findings of possible response options

[29] The Norwegian writing test has a formative dimension, i.e. providing the pupils with information about their proficiency in different text domains in relation to known standards. The classroom teachers then need to know the writing construct and the standards of expectations so they can instruct the pupils accordingly.

with respect to an initiative, that is *reject*, *yes-but*, *follow-up question* and *accept*. The dominance of acceptance responses (cumulative talk) is to be expected (Hogarth, 1987), but such dominance raises questions about the rationale behind an assessment dialogue, since it would seem unnecessary to have a dialogue if the only thing that raters do is to confirm each other. The quality of judgments may also be at stake in such a situation. Through exploratory talk, on the other hand, when raters are asked to justify their judgments, it is more likely that rater reliability will increase (Baurne, Yorke & Coffey, 2004). Often it is the presence of resistance in raters' responses that then provokes justification, and Evensen (2012) showed how disagreement between raters in a dialogue may be a starting point for reaching shared understanding granted mutual openness and respect.

2.2 Context, participants and approach

The study was conducted over three successive National Assessment Panel meetings; June 2011, November 2011 and April 2012. During this period five raters were tape-recorded when assessing 'Year 8 texts', and they are this study's main informants. But to discourage pairs from developing their own assessment practices within the panel, the pair composition continuously changed, giving overall data contributions from 28 raters. The five main informants are presented as R3, R7, R14, R15 and R23 in the excerpts below.

The data collection commenced at such an early stage of the panel's existence (established 2010) that it is fair to say that all the raters in the data-collection period progressed from being novice raters to being more experienced raters. All in all, 26 assessment dialogues were recorded where each recording relates to the assessment of one to three pupil texts. The main data sources include transcripts of the tape-recordings of the assessment dialogues, documents used by the panel and the researcher´s logs. All transcripts were coded using NVivo 10 (http://www.qsrinternational.com).

Even though the raters had read the texts before working in pairs, the pair assessments were not a comparison of already marked papers. In fact, the raters were explicitly told by the test designers to discuss and negotiate text quality and not only where to place the text on a given scale. It is therefore important to point out the social design of the writing assessments; the pair/group work is not primarily aimed at enabling a more reliable individual rating practice later on, rather the assessment *per se* is in focus.

Table 1. Referents used by the raters when assessing writing within the National Assessment Panel:

	Categories	Definitions	Examples from data
Familiar assessment practice prior to NAP membership (Known referents prior to joining NAP)	A. Reference to text	Referring to text in order to make a judgment	"She doesn't mention herself, she just talks, she talks on behalf…"
	B. Citation of text	Citing text in order to make a judgment	"I don't think it's appropriate to *"Dear Tor"*, right."
	C. First scoring	Referring to a score made after a first reading before the pair/group assessment	"Yes, I have marked it "as expected" [M3], but that's fine really [to give M4]."
	D. Comparing	Comparing texts within the same corpus	"You're not supposed to compare, but if you look at the other one here…"
	E. Text knowledge and classroom practice	Referring to text knowledge and own classroom practice	"But a typical introduction when writing an argumentative text is to make a claim."
New NAP-related assessment practice (NAP-related referents)	F. Meta-discussion	Discussing different aspects of the scoring task, e.g. the quality of the assessment tools	"Isn't it the support question [in the scoring rubric] that misleads us?"
	G. Plenary discussion	Referring to statements from plenary discussions	"But we talked about that a little bit yesterday, and that, eh, if you don't have any mistakes in punctuation…"
	H. Guidelines	Referring to the normative document "Guidelines"	"Look, to get an M5 the writer has to get close to writing all the words correctly, but some random errors may occur."
	I. Expert	Referring to one of the designers of the writing test or guest lecturer	"And I think *they* mentioned that as well; when it comes to spelling it's not that easy to use."

2.3 Coding scheme

Since the raters assessed in pairs or groups and, at the same time, were part of a "community of practice" (Wenger, 1998), it was important to study the raters' assessment talk itself together with their use of referents when making judgments about pupils' text quality. To

answer the research questions, the transcripts were coded according to two main categories: *Referents* and *Responses*. *Referents* made it possible to take a closer look at specific instances within the 'discourse universe' and *Responses* created the possibility to study a dimension of the 'interactive situation' (Linell, 2009, p. 98). The following is a short presentation of the two categories together with the rationale for using them.

The first main category includes the referents that the raters used and depended on when making judgments about pupils' text quality (Table 1). The coding scheme was partially derived from related research (Cooksey, Freebody, & Wyatt-Smith, 2007; Wyatt-Smith, Klenowski, & Gunn, 2010) and was partially grounded in the data (Glaser & Strauss, 1967; Hammersley & Atkinson, 1995). The referents in use fell into two main groups. One comprises referents held to be known by the raters prior to their sitting on the panel. The raters, who all are teachers, clearly relate the use of these referents to their classroom assessment practices. Five referents where found to belong to this main group (referents A-E). Examples would be *referring to* and *citing* the writing that is being scored. The other group is new to the raters since the referents were developed/created within the NAP context. Here four referents where found in the data (referents F-I). An example here would be reference to *plenary discussions* held within the NAP.

The nature of the referents is very different. For instance, the 'Guidelines' (referent H) comprise a detailed document often discussed among the raters, while other referents were never found to be problematic, for instance the practice of citing pupils' texts (referent B). By coding the assessment dialogues according to referents used by the raters, it was possible to evaluate the grounds on which judgments of text quality were made and, furthermore, it was possible to study changes in this practice over time. This made it possible to assess whether raters increased their use of NAP-related referents through experience and practice. Such an increase was held to be a reasonable assumption when efforts were taken to improve interrater reliability by establishing a community of practice involving pair assessments, plenary discussions and lectures.

The other main category used in this study is the raters' responses to colleagues' initiatives, building on Graumann´s empirically based categorising of possible responses (Graumann, 1990) (cf. 1.2 The dialogic aspect of the assessment). In the current study, the way the responses distribute between rejections, yes-buts, follow-up questions and acceptance is seen as an indicator of the quality of the assessment dialogue. It is expected

there would be a more heterogeneous response pattern over time (indicating an increase in the use of exploratory talk) and this pattern would be less dominated by responses indicating acceptance due to increased assessment competence and a more familiar social environment.

The material that was coded in this study comprises *instances* (when it comes to referents) and turns (when it comes to responses) that are relevant to the coding scheme. It follows from this that some talk between the raters is not coded. Examples would be talk that is irrelevant to the scoring ('small talk') and utterances that do not function as responses to what has just been said, so-called 'free initiatives' (Linell, 2001, p. 175). Such an utterance occurs, for example, when a rater suddenly decides to start a new topic.

When one rater uses a specific referent in her/his argumentation, it is coded as one incident even when (s)he is 'disrupted' by a consenting utterance from the rating partner. An illustration of this is when rater R7, with reference to a pupil text, assesses spelling. Here R7's two utterances are coded only once:[30]

April 2012. Rater R7 and R37:

1. R7 Yes. Everything is basically-
2. R37 Yes, it-
3. R7 - correct, most of it.

On the other hand, an example is given below where R38 cites a text several times and where each citation is coded separately, in other words, each citation is regarded as a separate argument in the justification of a given mark:

April 2012. Rater R38, R7 and R18

1. R38 "It was as if the head just short-circuited," full stop.
2. R7 Yes, there it might suggest that…
3. R38 "And mom and dad panicked when they saw me."
4. R18 Yes.
5. R7 Yes, it's…

[30] The dialogues and the pupils' writing were all originally in Norwegian. For the purpose of this study relevant data has been translated. Every effort has been made to keep the translation as close to the original as possible.

6. R38 “They had not managed to…,” “therefore they had to carry me down.” So,
7. this is a very good paragraph, starting with that creepy feeling, over to the
8. explanations.

Peer coding is preferable as a means of establishing reliable identification of emergent themes (Neuendorf, 2002), especially in those cases where what is coded is ‘projective content’ in the text corpus (Potter & Levine-Donnerstein, 1999). In this study, however, what is coded is considered ‘manifest content’, visible on the surface of the text. The interpretation is limited to linking specific relevant utterances to their respective categories as well as defining what should be considered a single referent and what belongs to a chain of separate referents, as illustrated above. Peer coding has therefore been considered to be unnecessary. Instead, to make the single coder coding transparent, several steps have been taken: The various categories used to code the data are provided along with definitions and examples (Table 1). Elaborated examples of coded data are also given through a generous amount of excerpts, including a screen shot from coded data in NVivo 10 (Excerpt 6). To further increase the transparency between data and analyses, the analyses are closely linked to the excerpts through line references.

As is the case with almost all qualitative research, this study might also be criticised for the criteria used to select the samples to illustrate the larger material. By categorising and coding the data in NVivo 10, an attempt has been made to address this anticipated criticism: Characteristics from the coded data are used as the basis for highlighting specific instances of interest, as will be evident in the next section. In that way, it is argued, the excerpts become stories inherent in the material, and not something that by chance creates a story.

3. Results

3.1 Overall findings

The first research question relates to how novice raters on the National Assessment Panel assess pupils’ writing, and how this praxis changes over time as the raters gain experience and competence. In order to discuss this question main findings from each of the three NAP meetings are presented, as well as an analysis of the development over the months that the data material was collected.

In Table 2, assessment dialogues are shown where the five key raters have participated. The table displays the extent to which different assessment referents are used. Symbols A-E refer to referents known to the raters before becoming a member of the panel, while F-I are referents that are new to them (cf. Table 1).

Table 2: Raters use of different referents in their pair/group assessments at different times with the number of instances in parenthesis

	Referents known by the raters prior to NAP					**New referents (NAP)**			
	A	B	C	D	E	F	G	H	I
Jun-11	**66% (469)**					**34% (246)**			
	48% (341)	7% (51)	5% (37)	4% (27)	2% (13)	24% 172	1% (8)	2% (15)	7% (51)
Nov-11	**73% (327)**					**27% (122)**			
	56% (252)	9% (39)	0% (1)	3% (13)	5% (22)	21% (94)	0% (1)	6% (27)	0% (0)
April -12	**61% (218)**					**39% (142)**			
	41% (146)	9% (34)	8% (27)	2% (7)	1% (4)	25% (92)	2% (6)	11% (39)	1% (5)

Total coded use of referents during the three NAP meetings: 715 (June 2011), 449 (November 2011) and 360 (April 2012). Assessment referent codes: A: Reference to text, B: Citation of text, C: First scoring, D: Comparing, E: Text knowledge and classroom practice, F: Meta-discussion, G: Plenary discussion, H: Guidelines, I: Expert.

The table highlights several features about the raters' behaviour over the ten-month period. First, it is both interesting and surprising to register how relatively small the changes in use of different assessment referents were over the three sessions, despite the raters' participation in a learning environment over a relatively extensive period of time. Second, the majority of utterances where assessment referents were used are in categories A and F, 'Reference to pupil text' and 'Meta-discussion within the pair or group'. Whereas raters' use of 'Meta-discussions' (F) indicates an awareness of new dimensions of the assessment task (cf. 3.2.1

Meta-discussions), traditional ways of finding arguments for making judgments of the quality of pupil writings were dominant and remained dominant throughout the period (i.e. mainly category A but also categories B-E). Third, it is noticeable that despite small changes in rater behaviour, the use of the 'Guidelines' is the only category of referents that shows considerable change over time (2%-6%-11%, at the three points in time, respectively).

Table 3 provides an overview of response patterns of the five key informants during the three NAP meetings. The responses indicate that acceptance was the most frequent form of response (79%-80%-76%, at each of the three points in time, respectively). The distribution of different response alternatives was stable over the three meetings with a slight increase in use of the two categories 'rejection' and 'yes, but' (8%-10.5%-12.5%, when the two categories are merged at each of the three points in time, respectively).

Table 3: Raters' response patterns in their pair/group assessments at different times with the number of turns in parenthesis

	Rejection	**Yes, but-**	**Question**	**Acceptance**
Jun-11	2.5% (8)	5.5% (16)	13.5% (39)	78.5% (229)
Nov -11	2% (5)	8.5% (23)	9.5% (26)	80% (215)
April -12	4% (10)	8.5% (21)	11.5% (27)	76% (182)

The overview provides information about the raters' development as a group that can be summarized as follow:

- Most notable is the raters' relatively stable use of referents and type of responses, which is indicative of little change in rater behaviour over the three NAP meetings.
- The raters make their judgments on text quality predominantly by using traditional referents.
- The raters have extensive meta-discussions about their task (24%-21%-25% of the coded instances at the three points in time, respectively).
- The raters' use of the normative document 'Guidelines' increases for each NAP meeting.

The two first bullet points are hard to explain without pointing to the raters' double task; they are both assessing according to given standards and working critically with these standards in

order to improve them. Even though the raters have had a reasonable amount of time to work together, the lack of an authoritative 'narrative' seems to slow down the development of a shared understanding and praxis of assessing writing. When entering a new role as a rater in this situation, it is likely that the teachers find it safer to relate their judgments to former praxis (cf. Timperley & Robinson, 2001, about changing teachers' schema). This is congruent with the notion that the raters are recruited into the panel as proficient teachers and, as such, they are likely to have strong opinions about what counts as good writing. They bring their teacher praxis with them into their new roles as raters. However, in order to succeed, the teachers have to become raters, and to do so they need to put aside irrelevant teacher praxis, such as basing judgments on individually held criteria, and build rater competence, such as gaining knowledge about and skills in using shared standards. The two last bullet points, raters' meta-discussions and their increase in the use of the 'Guidelines', are positive signs in this regard.

3.1.1 An example of praxis: Assessment dialogue as consensus orientated and based on traditional resources

The overall findings reveal that the raters' assessment praxis only to a minor extent changed towards an increase in the use of shared assessment resources and that they seldom questioned each other's utterances. The following excerpt (Excerpt 1), where the raters R15 and R7 are undertaking one of their first assessments as members of the panel and are both to be considered novices within the NAP context, work as an example of how this may look. In the excerpt they are about to assess a text's 'Use of language':

Excerpt 1. June 2011, pair assessment. Raters R7 and R15:

1. R7 Then there's the sentence structure.
2. R15 Not much wrong there.
3. R7 No, it has the-. The text is coherent. "*It is one important effort that might*
4. *strengthen-,"* "*In addition to that*-." Yes, it is actually-, That part refers to-,
5. and that part refers to this.
6. R15 Mhm. So this is rather straightforward.
7. R7 So he actually has [cohesion. He has better cohesion than I first saw.

8. R15 [So it's on M3 then, perhaps?

9. R7 Yes, because he has both linking sentences and linking paragraphs

10. [actually.

11. R15 [Mhm. But just lacks content.

12. R7 Yeah, not many have cohesion from one paragraph to the next.

Even though there are traces of new ways of assessing pupil writing in the excerpt – they assess text domains without making an explicit holistic assessment first, and they use the new assessment scale provided – there are several other traces indicating that the raters are novices in this particular community of practice. We can see how the interactive situation is totally dominated by acceptance of one another´s utterances. Every initiative is met with agreement from the rating partner. When it comes to explicit use of assessment resources, they refer to (lines 2-5, 9, 11) and cite (lines 3-4, only rater R7) the pupil text they score, and R7 also refers to other pupils´ achievements (line 12). New NAP referents are not used at all, and their judgments about the quality of the pupil text are made entirely on the basis of the raters' 'doxa', which is the practical knowledge that constitutes the "universe of the undiscussed/undisputed" (Bourdieu, 1977, p. 164).

In the excerpt, novice raters are making the assessment. However, the overall findings (cf. Table 2) indicate that the way raters assess the texts after two/three sessions does not differ much from what we see here. Due to this, it is hard to find a logical connection between rater judgment of text features and the way they mark each of the text domains (for an example, see lines 7-8, Excerpt 1). Excerpt 1 illustrates the raters' tendency to give unreserved consent to colleagues' initiatives (cf. Tables 2 & 3).

3.2 Traces of change in rater behaviour

The second research question is whether changes in rater behaviour have consequence for rating quality. This is an important question even though the overall findings show that changes so far are minor. A closer look at changes in rater behaviour might be interesting in the future strive towards a shared understanding, and a common application of given standards. The research question will be answered by presenting examples from the data where raters' assessment dialogues show that they rely on more than their traditional

assessment resources when making judgment about text quality. By putting into action available shared resources, such as 'the Guidelines', changes in rating quality are likely to happen.

3.2.1 Meta-discussion

As pointed out in the introduction, one of the main reasons for establishing an assessment panel was a documented lack of shared standards among Norwegian raters assessing writing. One of the ways of ensuring that more reliable and valid judgments would be made by the panel, it was contended, was to have the pairs or groups discuss not only the texts at hand but also the construct the assessment should follow. Meta-discussions are therefore both wanted and required. The transcripts show several ways in which this is being done. The assignment is discussed, as are the different text domains (cf. Excerpt 2, below), the scoring rubric's support questions, the scale, the 'Guidelines' and the assessment strategies. The broader themes, such as the ideology underpinning the test and norms of expectations (standards), are also discussed. The raters use these discussions to reach understanding, to criticise, to point out things that are missing and also to find a hierarchy among the various resources.

Excerpt 2 is an example of how the raters discuss what belongs to certain text domains according to the primary trait model in use:

Excerpt 2. June 2011, group assessment. Raters R8, R21, R23 and R24:

1.	R23	But I have one objection when it comes to the content in the text.
2.	R21	Okay.
3.	R23	And that is that I find it-, that she repeats [some points, so she
4.		could-, the content could be-
5.	R24	[Yes, she does.
6.	R21	But-, but isn't that about 'composition'?
7.	R23	The que-, that is the question, if it should [be assessed another
8.		place. I believe she could have economised on her use of words.
9.	R21	[That it is a bit
10.		unorganised
11.	R24	Mhm, but that should be assessed somewhere else.

12. R23 That should perhaps be assessed somewhere else, yes. Because here
13. she repeats an argument, so that is perhaps more about [structure.
14. R24 [No,
15. because 'communication' and [2'content'2 are almost the same.
16. R21 [2I think-2
17. R24 Mhm, because if you communicate well then it would be an
18. appropriate content as well, right?

The dialogue shows that the work the raters have been instructed to do is not straightforward. Here, the different text domains seem to confuse rater R23, seeing a 'composition' issue as relevant for the text's 'content'. However, because rater R21 is confident enough to explore R23's critique (line 6), they reach a shared understanding of where this specific text feature should be assessed (lines 11-13). As such, this is an example of 'exploratory talk' where an 'attunement' takes place. An illustration of the challenges the raters are facing is shown here with an additional issue to be addressed when the first problem is solved. The excerpt ends with R24 expressing uncertainty about the difference between the two domains 'Communication' and 'Content', indicating that it is difficult to distinguish between these domains (lines 15, 17-18). The excerpt shows the importance of dialogue in an effort not only to understand important normative documents in shared ways (here the 'Guidelines') but also to have "similar recognition of performances that demonstrate those standards" (Maxwell, 2001, p. 6). However, the same excerpt even discloses how fragile the assessments are when it comes to reliability. It is likely that a feature in one specific text is being judged within different text domains by different pairs/groups of raters when these raters clearly find it challenging to separate the text domains from one another (both 'content' and 'composition', and 'content' and 'communication').

3.2.2 The 'Guidelines'

While meta-discussions were comprehensive and stable throughout the three NAP meetings, the use of the 'Guidelines' was the only referent which increased considerably during the year

the data material was collected (cf., Table 2, resource H). This increase was both wanted and expected as the 'Guidelines' document was presented to the raters by the designers of the writing test as the most important normative document the raters would have at hand. This document includes general information and a section devoted to the specific writing test. The general information presents the writing test within the framework of writing as a key competency and gives a short presentation of the primary trait model. The latter section presents the assignment, the 'act of writing' that is to be tested, the parts of the Norwegian curriculum to be scored against, a specification of given standards within the text domains and benchmark texts. Whereas the general section seemed to function as background for the assessments, presented by the designers of the test in plenum without being discussed among the raters while they were assessing texts, the section devoted to the specific assignment was mainly used in two ways: First, but not the most common occurrence, the raters used the 'Guidelines' as an entrance gateway to their work, a strategy where they read through the document to both interpret it and to have it present in their work. The following excerpt is an example of this:

Excerpt 3. April 2012, group assessment. Raters R3, R35 and R36:

1.	R3	Next is 'Content.'
2.	R35	Then there was the content yes. (xxx).
3.	R3	Eh, "Select content that is relevant, adjust the amount of content for the
4.		purpose. Examine and express knowledge about other things than the
5.		privately experienced." They all have to do that.
6.	R36	Mhm.
7.	R3	"Examine and express own experiences, thoughts, feelings in order to
8.		challenge their own identity, roles, opinions and positions."
9.	R35	Yes.
10.	R3	But not everyone writes about their thoughts.
11.	R36	No, that´s true.
13.	R3	Some just write about experiences.

14 R35 Yes, and then the text becomes more like a summary.

15. R36 Mhm.

16. R3 So there-. This is perhaps where we differ.

In this excerpt the raters collaboratively interpret how to understand the 'Guidelines'. R3 starts by citing this document (lines 3-5) whereby all three raters contribute to the discussion on how to understand it. R3's concluding remark (line 16) should be understood as showing that they have found a specific text quality that distinguishes between good and average expected performances. Having reached this conclusion, they are ready to assess the content in the next pupil text in accordance with this newly obtained understanding of this specific part of the 'Guidelines'.

Second, and the most common occurrence, was the raters' use of the document as a direct support for validating their judgments, and again there were two ways in which this was done. Most often they referred to the specifications of given standards within different text domains. Excerpt 4 is an example of this usage:

Excerpt 4. November 2011, group assessment. Raters R12, R13 and R14:

1. R14 Yes, but it is written here[31]: "Shall master orthography in

2. common words that are not orthophonetic."

3. R12 Yes.

4. R14 They should be able to do that.

5. R12 Yes. 'In-.'

6. R13 'I-.'

7. R14 That is the norm.

8. R12 She writes 'intresange.'[32]

9. R14 Yes.

10. R12 She doesn't master it.

[31] "Here" refers to the 'Guidelines'.
[32] Example of a highly frequent Norwegian word that is not orthophonetic and that is spelled incorrectly by the pupil.

11. R14 No.

The raters also used the document's benchmark texts, as shown in Excerpt 5. By comparing the text at hand with a benchmark text the raters find support for a given mark.

Excerpt 5. November 2011, group assessment. Raters R12, R13 and R14:

1. R13 I would have marked that M3.
2. R14 Let me see, for low mastery there is no introduction. The text includes a
3. reader here in a way: "To Veronica. To develop a nice class environment or
4. to have a nice school year you all have to try to not be a bully."[33]
5. R12 That is low.
6. R14 That is low.
7. R12 That is low. It just goes straight to the main part. She has one sentence: "It is
8. nice of you to ask," to create-. It is a sort [of-.[34]
9. R14 [The benchmark text has no
10. conclusion either.
11. R12 I find it hard to give her M4; that this is a high degree of mastery. Then there
12. is supposed to be an introduction, a main part and a conclusion. In this text
13. the introduction is only this one sentence.
14. R13 Mhm.
15. R14 I-, yes. I agree.

In all three excerpts presented in this section (Excerpts 3-5) the raters' judgments are likely to be reproduced by other raters given a constant use of the 'Guidelines'. The random connection between text features and the provided mark shown in Excerpt 1 (cf. 3.2.1 above) is here substituted by traceable judgments based on a normative document, a prerequisite for reliable assessments.

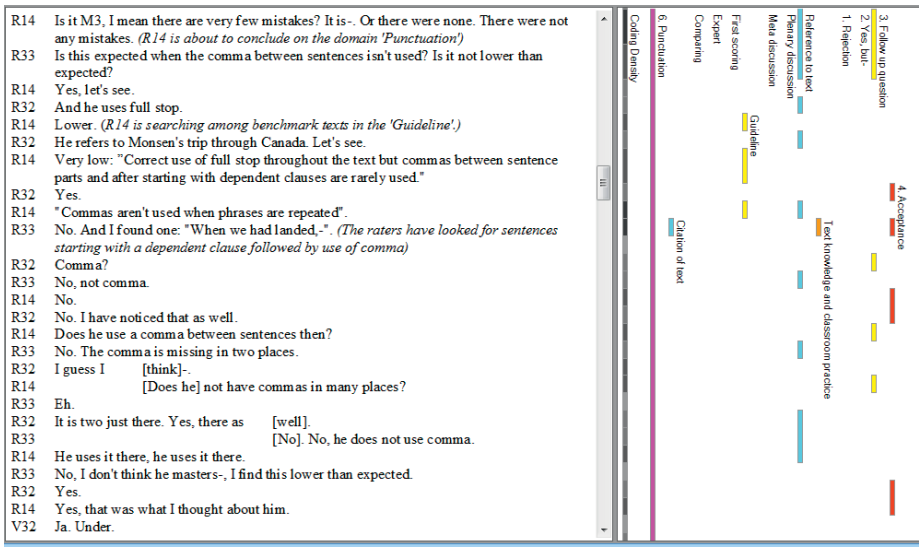
[33] R14 refers to and cites a benchmark text.
[34] R15 is now referring to and citing the text at hand that they are assessing.

3.2.3 The power of dialogue between competent raters

What is a qualitative assessment dialogue? This is not an easy question to answer due to the varying assessment situations. Sometimes the outcome is so obvious that a nod from a colleague to an utterance is enough. However, this would appear to be the exception. More often, exploratory talk is needed where raters are critically involved and both challenge each other's judgments and willingly surrender to 'the better argument' (Habermas, 1992). In this way the raters leave traces after their work, traces that are then possible to follow so similar judgments in similar situations can be replicated.

The following excerpt is an example of how this could be done and, in that sense, is an example of the power of dialogue between competent raters (Excerpt 6; the excerpt is a screen shot from NVivo 10). Rater R14 opens and closes this short excerpt, opening by suggesting a mark for the domain ('punctuation') as M3 but ending up with the conclusion that M2 is a better mark. This change of opinion is traceable to a series of follow-up questions that trigger the search for a better answer. In this search the raters turn to the 'Guidelines', to the text at hand which they refer to and cite and also to a verbalisation of their professional competencies through the use of professional terms. We see a *triangulation* between the raters when they are making their judgments that is likely to lead to more consistent assessment. This triangulation is both dependent on resistance in the dialogue and on competence in using a range of assessment referents, both new and traditional ones.

Excerpt 6. April 2012, group assessment. Raters R14, R32 and R33:



Using NVivo 10. The screen is divided into two main parts. On the left we find the transcription of the assessment dialogue. On the right we find the coding of the dialogue shown as coloured stripes. The position and length of the stripes makes a horizontal match to the dialogue. The excerpt illustrates how the two main coding categories, response pattern and assessment referents, are intertwined. At the far right an exchange between 'follow up questions' (yellow stripes) and 'acceptances' (red stripes) leads the raters to different assessment resources (i.e. the 'Guidelines' (yellow stripes) and 'Reference to text' (light blue stripes)). Coding the data in this way in NVivo 10 reveals in a visual way the close relationship between the categories: A colourful response pattern gives a colourful referent pattern.

4. Discussion

In the present study, the aim has been to explore how novice raters arrived at their judgment of text quality within a specific community of practice and also to investigate if their practice changed over time as they gained experience (ref. research question 1). The stability in *what the raters talked about* (cf. Table 2) and *how they talked* (cf. Table 3) in order to reach decisions when rating could be assessed positively in terms of being interpreted as consistent rater behaviour. However, given the raters' lack of both rating expertise and a shared understanding of what to expect from the pupils' writing (cf. 1. Introduction), such a conclusion is unlikely. On the contrary, as members of the National Assessment Panel, the raters were expected to gain proficiency over time, that is, the intention was that changes should take place. The findings indicate a need for the raters to work more in alliance with determined and authoritative shared standards in their endeavours to assess well together. After the data material was collected for this study, the team working with the writing test

gave the 'Guidelines' such a clearer authoritative status. Preliminary findings from assessments conducted after this fixation of shared standards indicate positive effects on interrater reliability.[35]

The success of the Norwegian sample-based national assessment of writing as a key competency depends on the trustworthiness of the work carried out by the members of the panel. Wyatt-Smith and colleagues' (2010) research on teachers' use of shared standards suggests that this is not straightforward. Their study shows that teachers' use of shared standards is limited, and that tacit knowledge and values interfere in the judgment process to a great extent when assessing writing. The researchers state that teachers, when acting as raters, need "practical, unambiguous advice (…) about desired judgement practice" and "that the challenge for transparency [in writing assessment] is to understand the relationship between these [explicit and tacit knowledge]" (Wyatt-Smith, Klenowski & Gunn, 2010, 72). Both statements are manifested in different ways in the last excerpts in this paper (Excerpts 4-6 in particular). They indicate that when the assessment dialogues are both exploratory and anchored to shared assessment referents, the quality of the assessment improves in terms of being more visible and verifiable, and therefore, potentially more consistent (ref. research question 2).

5. Conclusion

The pair assessments were introduced to add an exploratory dimension to the study and set the dialogue in relief (cf. 2.1 The dialogic aspect of joint assessment). The analyses suggest this is a relevant and valid category when investigating rater proficiency. A cautious inference due to methodological concerns is therefore to regard this as a contribution in addition to the more established approaches. To delve into the 'inner workings' of the rating process, studies often ask raters to assess 'as if' it were a normal situation while they verbalise their thinking (i.e. think-aloud protocols, cf. Ericsson & Simon, 1993). Even if this method is well recognised, it has some well-known shortcomings due to its artificial dimension, as described by several researchers (e.g. Barkaoui, 2011b; Long & Bourg, 1996). Hence, when the raters assess the Norwegian sample-based writings in pairs or groups, it is possible to use the dialogue as the focal point when investigating rater practice as it actually turns out to be exactly that. This

[35] These findings are for the time being not publicly available (personal communication with the project leader, Prof. Lars S. Evensen, October 14, 2013).

theoretical and methodical aspect is something that should be pursued to gain more knowledge about the relationship between rater discourse and their judgments when assessing pupil writing.

It is recognised that, as an exploratory study, the present investigation has been limited in several respects. The raters have been given the opportunity to develop shared standards together with the researchers. The idea has been that the raters have had, if only individually, first-hand knowledge about norms of expectations through their everyday teaching, but also that this kind of involvement will give the raters a feeling of ownership over the standards that in the long term will be beneficial both for the test's validity and its reliability. However, the provisional status of the standards is likely to have affected the raters' use of the same standards, leading to relatively little use of shared resources when assessing the texts (in particular referents G-I, Table 1). It is also clear that the way the raters value the different assessment resources changes over time. For instance, when raters compare texts (referent D, Table 1) to decide about text quality during the first data collection session (June 2011), the comparison is often accompanied by an additional comment indicating the referent's low value, typically "we are not supposed to compare, *but*…"). Such comments are not present during the last data collection session (April 2012), a change that suggests a development towards a validation of this referent as a resource to define norms of expectations.[36]

Further; the exclusively qualitative investigations that were undertaken in this study help to clarify the grounds on which raters arrive at decisions about text quality in the pupils' writings, and also how they develop their judgment practice over time, but it is not possible from this to determine what effects changes in practice have on the tests' reliability. A follow-up study is called upon to investigate whether, and potentially how, the raters' assessment dialogues change together with more trustworthy assessments. Such a follow-up study, where the quality of the assessment dialogues are studied together with quantitative evidence of rater scoring, would also make claims stronger.

One important aspect of the current study is related to rater proficiency and its potential development. The research literature indicates the need for a better understanding

[36] The 2006 Norwegian curriculum underscored that pupils should be assessed relative to given criteria (criterion-referenced), and not relative to their co-pupils (norm-referenced). The raters' devaluation of "comparing" as an assessment resource in the beginning has to be understood in this context. The example is also useful for clarifying that the categorization of referents might conceal different understandings of the various resources. Such differences are not investigated in any greater detail in this study.

of what it means to be(come) an expert rater. Sometimes such a rater is presented as an experienced rater (Cumming, 1990), a comparison that is questionable as experience is no guarantee of proficiency. Other times expertise is defined as the rater's ability to assess 'well' in terms of severity and consistency (Lim, 2011), which is also a problematic way to define expertise because it uses result and not performance as the only indicator. At other times, the term is recognised as so inaccurate that it has to be placed within quotation marks (Hamp-Lyons, 1994). Research on raters' decision-making processes, of which this study is an example, makes it possible to assess the degree of rater expertise based upon raters' actions rather than the results of their actions. It is argued that to define expertise on the basis of interrater reliability alone is to underestimate the importance of the interdependency between a test's validity and its reliability (Moss, 1994; Williamson, 1994).

References

AERA, APA, & NCME (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: AERA.

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*, 133-153.

Barkaoui, K. (2011a). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*, 51-75.

Barkaoui, K. (2011b). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*, 279-293.

Barritt, L., Stock, P. L., & Clark, F. (1986). Researching Practice: Evaluating Assessment Essays. *College Composition and Communication, 37,* 315-327.

Baurne, D., Yorke, M., & Coffey, M. (2004). What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29,* 451-477.

Berge, K. L. (2005). Studie 3: Skriveprøvens pålitelighet [Study 3: A writing test's reliability]. In K. L. Berge, L. S. Evensen, F. Hertzberg, & W. Vagle (Eds.), *Norsksensuren som kvalitetsvurdering* [*The Norwegian (L1) examination result as a quality assessment*] (101-113). Oslo: Universitetsforlaget.

Bourdieu, P. (1977). *Outline of a Theory of Practice*. Cambridge: Cambridge University Press

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing Assessment in Admission to Higher Education: Review and Framework*. College Entrance Examination Board, New York. Retrieved April 10, 2013, from https://www.ets.org/Media/Research/pdf/RR-99-03-Breland.pdf.

Brown, G., Glasswill, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9,* 105-121.

Cooksey, R. W., Freebody, P, & Wyatt-Smith, C (2007). Assessment as Judgment-in-Context: Analysing how teachers evaluate students' writing. *Educational Research & Evaluation, 13*, 401-434.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*, 1-21.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision Making while Rating ESL/EFL Writing Tasks: A Descriptive Framework. *The Modern Language Journal, 86*, 67-96.

DeSeCo. (2005). *The Definition and Selection of Key Compentencies*. Retrieved April 18, 2013, from http://www.oecd.org/pisa/35070367.pdf.

Diederich, P. B., French, J. W., and Carlton, S. T. (1961). *Factors in the judgment of writing quality*. NJ: Educational Testing Service.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25* (2), 155-185.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. Cambridge: MIT Press.

Evensen, L. S. (2009). Vurdering av skrivekompetanse, en kompleks utfordring [Assessment of writing proficiency, a complex challenge]. In O. K. Haugaløkken, L. S. Evensen, F. Hertzberg, & H. Otnes (Eds.), *Tekstvurdering som didaktisk utfordring* [*Writing assessment as didactic challenge*] (15-23). Oslo: Universitetsforlaget.

Evensen, L. S. (2012). Underveis mot et tolkningsfellesskap: lærerstemmer om elevtekster [Toward shared understanding: Teachers' voices about pupils' writings]. In S. Matre, D.K. Sjøhelle, & R. Solheim (Eds.), *Teorier om tekst i møte med skolens lese- og skrivepraksiser* [*Text theories in the context of the school's literacy praxises*] (151-160). Oslo: Universitetsforlaget.

Fasting, R., Thygesen, R., Berge, K. L., Evensen, L. S., & Vagle, W. (2009). National Assessment of Writing Proficiency among Norwegian Pupils in Compulsory Schools. *Scandinavian Journal of Educational Research, 53*, 617-637.

Glaser, B. G. & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Jersey: Aldine Transaction.

Godshalk, F. I., Swineford, F., and Coffman, W. E. (1966). *The Measurement of Writing Ability*. NY: College Entrance Examination Board.

Grauman, C. F. (1990). *Perspectival structure and dynamics in dialogue*. Retrieved September 15, 2013, from http://www.psychologie.uni-heidelberg.de/institutsberichte/SFB245/SFB021.pdf.

Habermas, J. (1992). *Moral consciousness and communicative action*. Cambridge, UK: Polity Press, (Discourse ethics: Notes on a program of philosophical justification, 43-115).

Hammersley, M. & Atkinson, P. (1995). *Ethnography. Principles in Practice*. London: Routledge.

Hamp-Lyons, L. (1994). Examining Expert Judgments of Task Difficulty on Essay Tests. *Journal of Second Language Writing, 3*, 49-68.

Hogarth, R. M. (1987). *Judgement and choice: The psychology of decision*. New York City: Wiley & Sons.

Huot, B. (2002). "(Re)articulating Writing Assessment for Teaching and Learning". *All USU Press Publications*. Book 137. Retrieved September 20, 2013, from http://digitalcommons.usu.edu/usupress_pubs/137

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130-144.

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12*, 108-128.

Knowledge promotion - Kunnskapsløftet (2007). Retrieved October 1, 2013, from http://www.udir.no/Stottemeny/English/Curriculum-in-English/_english/Knowledge-promotion---Kunnskapsloftet/

Leckie, G. & Baird, Jo-Anne (2011). Rater Effects on Essay Scoring: A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. *Journal of Educational Measurement*, *48*, 399-418.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*, 543-560.

Linell, P. (2001). *Approaching Dialogue. Talk, interaction and contexts in dialogical perspectives*. Amsterdam: John Benjamins.

Linell, P. (2009). *Rethinking Language, Mind, and World Dialogically*. Charlotte, NC: IAP, INC.

Long, D. L. & Bourg, T. (1996). Thinking aloud: Telling a story about a story. *Discourse Processes*, *21*, 329-339.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19*, 246-276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Maxwell, G. (2001). *Moderation of assessments in vocational education and training*. Brisbane, Queensland: Department of Employment and Training.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Moss, P. A. (1994). Can There Be Validity Without Reliability? *Educational Research, 23*, 5-12.

Neuendorf, K. A. (2002). *The content analysis guidebook*. Thousand Oaks, CA: Sage.

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing*, *7*, 143-164.

Purves, A. C. (1992). A Comparative Perspective on the Performance of Students in Written Composition. In A. C. Purves (red.). *The IEA study of written composition* (Vol. 2. 129-152). Oxford: Pergamon.

Potter, W.J. & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*, 258-284.

Rommetveit, R. (1992). Outlines of a dialogically based social-cognitive approach to human cognition and communication. In A. H. Wold (Ed.), *The Dialogical Alternative: Towards a Theroy of Language and Mind* (pp. 19-44). Oslo: Scandinavian University Press.

Sadler, R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education, 13,* 191-209.

Sadler, R. (2011). Academic freedom, achievement standards and professional identity. *Quality in Higher Education, 17*, 85-100.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment*. Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida (pp. 129-152). Cambridge: Cambridge University Press.

Skolinspektionen (2013). *Olikheterna är för stora. Omrättning av nationella prov i grundskolan och gymnasieskolan, 2013* [*The differencies are too big. Re-assessment of national tests in primary and secondary education, 2013*]. Retrieved September 28, 2013, from http://www.skolinspektionen.se/Documents/Rapporter/spara-2013/omrattning-nationella-prov-2013.pdf

Starch, D. & Elliott, E. C. (1912). Reliability of the Grading of High-school Work in English. *The School Review*, *20*, 442-457.

Stuhlmann, J., Daniel, C., Dellinger, A., Kenton, R. & Powers, T. (1999). A Generalizability Study of the Effects of Training on Teachers' Abilities to Rate Children's Writing Using a Rubric. *Reading Psychology, 20*, 107-127.

Suto, I. (2012). A Critical Review of Some Qualitative Research Methods Used to Explore Rater Cognition. *Educational Measurement: Issues and Practice*, *31*, 21-30.

Suto, I. & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal, 34*, 213-233.

Timperley, H. & Robinson, V. M. J. (2001). Achieving School Improvement through Challenging and Changing Teachers' Schema. *Journal of Educational Change*, *2*, 281-300.

Thygesen, R., Berge, K. L., Evensen, L. S., Fasting, R. B. (2007). *Sluttrapport: nasjonale prøver i skriving som grunnleggende ferdighet* [*Final report: National tests in writing as a key competency*]. Stavanger: Nasjonalt senter for leseopplæring og leseforsking, Universitetet i Stavanger.

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

Wegerif, R. & Mercer, N. (1997). A Dialogical Framework for Investigating Talk. In R. Wegerif and P. Scrimshaw (Eds.), *Computers and Talk in the Primary Classroom* (pp. 49-65). Clevedon: Multilingual Matters.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*, 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*, 263-87.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145-178.

Wenger, E. (1998). *Communities of Practice. Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

Wheel of writing (2012). Retrieved January 3, 2014, from http://www.skrivesenteret.no/uploads/files/WritingWheel_Static_(rev_03_2012)_2.pdf

Williamson, M. (1994). The Worship of Efficiency: Untangling Theoretical and Practical Considerations in Writing Assessment. *Assessing Writing, 1*, 147-174.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*, 150-173.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers´ judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*, 59-75.

Yancey, K. B. (1999). Looking Back as We Look Forward: Historicizing Writing Assessment. *College Composition and Communication*, *50*, 483-503.

Artikkel 2

Is not included due to copyright

Artikkel 3

Rater strategies for reaching agreement on pupil text quality

Abstract

Novice members of a Norwegian national rater panel tasked with assessing Year 8 pupils' written texts were studied during three successive preparation sessions (2011-2012). The purpose was to investigate how the raters successfully make use of different decision-making strategies in an assessment situation where pre-set criteria and standards give a rather strict framework. The data sources were the raters' pair assessment dialogues. The analysis shows that the raters use a 'shared standards strategy', but when reaching agreement on text quality they also seem to make very good use of assessment strategies related to their work as writing teachers. Moreover, asymmetries in knowledge and participation among raters contribute to creating an image of writing assessment as a highly fragile hermeneutic practice. It is suggested that future rater preparation would gain from being attentive to the internalized assessment practices teachers bring to the fore when working as raters.

Introduction

Within a writing assessment system raters are the people who turn performance into outcome through their judgments. However, history has shown that this is an extraordinarily challenging task (e.g. McNamara, 1996; Purves, 1992; Skolinspektionen, 2012; Thygesen, Berge, Evensen, & Fasting, 2007). Various studies have shown that numerous variables influence this 'transliteration'. Some are related to the marking method (Barkaoui, 2011a; Knoch, 2007) and the quality of the prompts (Weigle, 1999) and the texts (Kuiken and Vedder, 2014), while others focus on the rater(s), such as reader proficiency (Thygesen, Berge, Evensen and Fasting, 2007), shared practice (Borgström and Ledin, 2014), background (Leckie & Baird, 2011; Lim, 2011; Wiseman, 2012), values (Baker, 2010) and training effect (Brown, Glasswell & Harland, 2004; Eckes, 2008; Stuhlmann, Daniel, Dellinger, Kenton, & Powers, 1999).

While the number of studies showing the relationships between these different variables and scoring is growing, we still know little about raters' decision-making processes (but see Crisp, 2010; Cumming, Kantor, & Powers, 2002; Sakyi, 2000; Weigle, 1994). What we in particular know little about is how raters learn to rate – progressing from being experienced teachers to becoming proficient raters – and what claims they make when deciding about text

quality. Raters are expected to balance their use of different types of criteria (Sadler, 1985). There are different traditions for dealing with this (Wyatt-Smith and Klenowski, 2013). What can be viewed as the classical, technical, positivistic tradition emphasizes the necessity of a rater's 'subordination' to *explicit criteria* in order to create reliable scoring. In fact, any interference of 'alien' criteria will be used to explain why a test score turned out to be unreliable. Another more hermeneutic-oriented tradition does not believe that a set of explicit criteria is sufficient for taking into consideration all aspects of what counts as text quality. Within this latter tradition *tacit* and *taken-for-granted criteria* therefore emerge as central dimensions in a rater's professional judgment practice. Without neglecting the importance of defined standards, this tradition will stress that "reliability is not everything" (Sadler, 2009, p. 176).

The present paper presents an empirical study on how raters balance between different types of judgment strategies with appurtenant criteria in preparation for the Norwegian learning supportive tests in writing as a basic skill. Analyses reported in a previous study using the same dataset indicated that despite extensive preparation raters only slightly changed their rater behaviour towards working more in alliance with shared normative assessment resources (Jølle, 2014). This finding was both a surprise and an invitation to further investigation. When raters were recruited to the project a key criterion to ensure text competence within the group of raters was that they were all experienced writing teachers. However, this also means that the raters bring more or less idiosyncratic scoring practices from their different classroom contexts into this new context. An important question emerges: How do their previous decision-making processes ally with new demands when exposed to defined standards and pre-set criteria?[49]

Contextual framing

By organizing the Norwegian writing assessment as sample-based, and not as assessing whole age cohorts, it has become feasible for an assessment panel (the National Assessment Panel, NAP) consisting of approximately 90 teachers recruited nationwide to carry out the writing assessments. In preparing the teachers for the job, that is, to turn the teachers into competent raters, they have met for three days on two occasions each year since 2010. The test they are

[49] Sadler (1985) offers a clarification of the differences between standards and criteria.

to rate is to be given annually from 2014. There are several reasons for the considerable amount of time that is being spent on preparing the raters.

First, the Norwegian curriculum from 2006 (the Knowledge Promotion Reform), where writing is introduced as a key competency, encouraged Norwegian writing researchers to develop a theoretical writing construct that forms the basis of how to understand writing. A model, called the Wheel of Writing, displays the basic concepts in this construct, where acts and purposes of writing are central elements (Wheel of Writing, 2012). As this theoretical model was new to the raters, they needed time to familiarize themselves with the concept. Second, it has been important to move the conceptual understanding of writing from exclusively belonging to Norwegian language instruction (L1) to an understanding of writing as a key competency, which makes it relevant in every subject for a number of purposes (DeSeCo., 2005). This turn has been difficult for teachers and raters to adapt to. What is meant by proficient writing in science has for instance not been a frequent topic of discussion among writing instructors and raters. As a response to this fundamental change, the raters within the assessment panel represent the full range of school subjects, making extensive discussions about subject-specific writing necessary. Third, an attempt to carry out a national writing test was made in 2005 but then without an assessment panel to score the texts. This led to such doubtful scoring that the test was put on hold. Lack of both rating proficiency and shared assessment culture were possible reasons that were given for the unreliable scoring (Fasting, Thygesen, Berge, Evensen & Vagle, 2009). Hence, by now giving the raters the possibility to discuss and rate together within the assessment panel for a period of time before the 2014 re-launch, the intention has been to improve the reliability of the scoring.

The members of the panel are introduced to, and discuss, the theoretical writing construct, as well as the implications of using a primary trait model; they participate in plenary and group discussions both on criteria for assessment and what to expect from pupils' texts at different year levels; and they rate pilot assessments. The raters score according to a five-level scale ranging from M1 (much lower than what is expected from most pupils at the actual year level), through M3 (as expected), up to M5 (much better than what is expected from most pupils at the actual year level) on each of the following six text domains: Communication, Content, Composition, Use of language, Spelling and Punctuation.

The raters are asked to use two normative documents developed for the purpose. The *Guideline with its benchmark texts* is a quite extensive document that has two main sections.

The first presents general information about the writing test, such as specification of what is to be assessed (writing as a key competency and not an L1 test) and key features of the primary trait model and the six domains to be assessed. Section two presents the current assignment and information about which act of writing (for instance "to describe" or "to persuade") the pupils' texts are written according to. Specification of competence aims in relevant curricula depending on the assignment is also provided. Finally, verbal descriptors of expectations are presented for each text domain followed by benchmark texts. The second normative document is the *Rubric with its support questions,* which is the actual scoring sheet. The rubric contains the different text domains, the scale (where the raters are to score from M1 to M5 in each domain), and selected "support questions" for each domain (these questions are a sort of extract of the descriptors found in the *Guideline*). The raters are advised to use both documents when deciding which text features to illuminate and how to judge them according to the normative five-level scale.

One last contextual feature has to be mentioned. An essential part of the test's design is pair assessment. Each text is assessed by at least two raters working together. The assumption behind the assessment dialogue is that it will boost learning and shared understanding. The ambition is to create a community of proficient raters who talk about and treat texts in relatively similar ways.

Purpose of this study

The raters have been placed together to turn pupil writing performance into outcome. The pair assessments have a specific purpose. The assessment dialogue is as such '*institutional discourse'* (Linell, 2001, p. 240) where contextual factors give the talk a stricter structure than in everyday conversation. Linell proposes that the analytical concept 'communicative project' (2001, 2009) should be used when studying such dialogic sequences, defining them as 'comprehensive units of meaningful action' (2001, p. 233). The larger global project for the assessment dialogue is to agree upon what counts as hallmarks of pupil text quality, and more specifically to score pupil texts. However, nested within this global project we will find numerous smaller ones where phases of the larger project are carried out. The concept of communicative projects will be used to analyse how the raters structure their workload.

Even though the communicative project is a useful analytical concept, it is not sufficient if we are to reach an understanding of how raters jointly make decisions about text

quality. The question is which ways of doing this are repeatedly in use when raters talk their way towards agreement? Linell offers a second concept for this purpose, defining 'a specific (...) way, or method, of going about solving (trying to solve) the problem' as a 'communicative strategy' (Linell, 2001, p. 227). Linell gives several examples: The politician's tendency to reformulate the journalist's question to enable her to give the preferred answer, the gossip telling her own side of a story to lure the interlocutor into telling the juicy parts of a story and the doctor asking for the patient's perspective first as a 'self-assessment' before giving her own professional version (Linell, 2001, p. 228-229). In all these cases, the actors have strategically (even though often unconsciously) chosen how to behave to get the job done with favourable results.

Strategies occur wherever communicative projects take place, including where raters undertake their pair assessments. The questions to ask are which strategies are used and if knowledge about them can contribute to better understanding about how raters, in their joint effort to assess pupils' writings, reach decisions about pupil text quality. In this study these two concepts (*communicative strategies* and *communicative projects*) will be front and centre as they offer an analytical approach to investigating which methods the raters use to carry out the different writing assessments within the assessment panel's preparatory stages. The main question to discuss is:

Which communicative strategies are successfully employed to reach agreement about pupil text quality, and how are these to be understood in the raters' efforts to establish a shared understanding of writing assessment?

Rater development

A writing assessment's trustworthiness has long been an issue, and many attempts have been made to adjust different variables to increase an assessment's reliability. Still, recognizing that writing assessments have reliability challenges, this paper seeks to deal with the matter from a different angle; the rating processes are explored without concerns about the outcome. The focus is on *what* the raters talk about and *how* they talk when assessing pupils' texts, and not the scores itself. However, it is held that more knowledge about the judging process will make us better prepared to deal with reliability related challenges. For instance, is unreliable scoring mainly due to rater incompetence, or to genuinely different views about what is relevant

(Broad, 2000; Moss, 1994; Petruzzi, 2008)? How we treat this seems to be important for future actions. So far it looks like this challenge has mainly been treated as a 'rater quality problem' (Behizadeh & Engelhard Jr., 2011; Huot, 2002; Jefferey, 2009). However, in a study of reader training experiences in two large-scale portfolio assessments, Colombini and McBride (2012) find evidence that disagreement is not a danger to consensus, but rather a necessary step towards shared norms/standards. Several writing assessment moderation studies underline in similar ways how different teacher educational backgrounds and career experiences necessarily lead to a difference of opinions about what to value in writing, but that this is no impediment for achieving common understandings of pupil text quality through moderation (Reid, 2007; Wyatt-Smith, Klenowski & Gunn, 2010). To understand the development of rater proficiency in these terms is to state that rater agreement and shared practice is a goal not a requirement.

Participants

The Norwegian assessment panel consists of 90 raters, where half assess Year 5 pupils' texts and half Year 8 pupils' texts. In this study 33 raters from the latter group participated.[50] Snowball sampling was used to recruit experienced teachers to the panel, and the main criterion was that geographically and socially distributed teachers had shown an interest in and commitment to writing and assessment of writing at their local school level. But, as emphasized earlier under 'Contextual framing', experience gained from the former attempt to carry out a national writing test suggests that there is no one-to-one relationship between rater commitment and a test's success. That is, the raters are to be considered novices in this particular writing assessment context.

Data collection and analysis

The study was conducted over three successive National Assessment Panel meetings, in June 2011, November 2011 and April 2012. During this period 26 assessment dialogues were tape-recorded, where each recording documents the assessment of one to three Year 8 texts. Most of the assessments are carried out in pairs, but for practical reasons some assessments are completed by a group of three raters. It is common in such situations to prepare raters in

[50] Three raters asked to be excused from the study and this, combined with an effort to discourage pairs from developing their own assessment practices within the panel by continuously changing the pair composition, limited the participants to 28 raters.

terms of giving them instructions on how to express their decision-making behaviour (e.g. Cumming, Kantor and Powers, 2002; Huot, 1993; Lumley, 2005; Wolfe, Kao and Ranney, 1998). Such an 'as-if' situation was avoided in this study since it was the pair/group assessments *per se* that were of interest (see for instance Barkaoui (2011b) for a presentation of potential bias effects when asking the informants to 'think aloud'). Hence, data collection bias was limited to possible effects from the raters' awareness of being tape-recorded. Furthermore, the data corpus also consists of observations in the form of field notes and document analysis (mainly the assessment resources).

Through reiterated readings of the transcribed tape-recordings all episodes where the raters made judgments about text quality were classified. While recognizing that this often is a highly sensitive part of a study, it is worth noting that the institutional character of the raters' discourse made the different smaller communicative projects 'manifest' (Potter & Levine-Donnerstein, 1999), making peer coding unnecessary. In other words, since the raters followed the structure found in the *Rubric with its support questions*, the analysis in this first phase was limited to isolating the raters' dialogue related to each support question as single communicative projects. The following excerpt is a good illustration of this:

November 2011. Raters R12 and R13

R13 'Use of language'. 'Are words and expressions varied in an appropriate way?'

R12 Yes, isn't that just-

R13 Isn't that okay?

R12 Yes, I think so.

Here R13 starts by citing a support question from the *Rubric* related to the text domain 'Use of language'. In the subsequent three short turns R12 and R13 agree that the text is 'okay' with respect to the question, meaning that it is as expected (i.e. according to the standards), and the raters are now ready to move on to the next support question. Hence, within the coding scheme the above excerpt has been coded as one communicative project.

Strategies to reach agreement on text quality

The analyses of the assessment dialogues carried out in this study rest on some basic theoretical principles about the dialogue (cf. Linell, 2001; Markova & Linell, 1996). Any

contribution to dialogue is seen as simultaneously an answer to what just has been said (a response) and deciding for what to be said next (an initiative). Hence, the first theoretical principle is that a contribution becomes meaningful within a *sequential chain* of utterances. An utterance is not meaningful in itself, but only as a response to other contributions. This stresses a dialogue's social dimension, which leads us to the second principle; a dialogue is a *joint construction*. Further, just as a single contribution in a dialogue is simultaneously a response and an initiative, is the (institutional) dialogue both responsive to former dialogues and guiding for future dialogues. This points to a third theoretical principle about the dialogue; it is nested within a larger *communicative genre* (a routinized way of going about) that support the interlocutors in the ongoing dialogue. Linell writes: "Communicative genres are thus originally interactionally developed, then historically sedimented, often institutionally congealed, and finally interactionally reconstructed in situ (2001, p. 239). Together these basic principles help us to understand how meaning is constructed in and through dialogue. It is also these dialogical principles this study rests on when investigating a specific group of raters' writing assessment practice as they, through dialogue, talk their way toward agreement on text quality.

In general, the raters' work within the National Assessment Panel in its preparatory stages can be categorized as being cooperative, where they for the most part agree with their rating partner without much discussion (cf. Jølle, 2014). But even though it is correct that they are supposed to agree, unconditional acceptance is problematic in this context. Valid judgments depend on an open professional discussion among the raters in the pair assessments. When the assessment dialogues are too consensus oriented we can think of different reasons for this: People in general seem to consider face-to-face disagreements to be unpleasant, and in this particular case they are even supposed to agree. It is likely that this is a strong incitement for the raters to at least appear to approve what the other rater is saying. Furthermore, the raters, as pointed out above, are novices in this context, and as such it is likely that they doubt their own competence. When feeling uncertain or when lacking a strong counter-argument in a discussion it might be just as easy to agree with the other. Even though they share an interest in the teaching and the assessment of writing, their assessments have previously been related to specific pupils in their specific writing developmental stages. Assessing the text as a product, and not as a text under process, contributes to making the task alien to the raters (Solheim & Matre, 2014).

However, although the raters often and quickly reach agreement, they also have lively debates when assessing texts. In both cases, whether consensus or alterity characterizes the assessments, the raters seem to use a variety of methods to reach a decision about text quality. When analysing the transcripts, three main strategies emerge as successful and recursive. In what follows, a short discussion of each of these assessment strategies will be presented.

Formative assessment strategy

There is reason to assume that Norwegian teachers, especially in the lower years, have tended to adjust their feedback to the pupils according to what they have seen as the pupils' individual potential.[51] This means that the teachers have not necessarily based their comments on standardized expectations, but rather on goals considered to be realistic for the pupils. It is likely that this follows from a tradition where marks not are used for pupils in the lower years, a tradition that has allowed teachers to work independently of norms of expectations. However, following the Curriculum reform from 2006 (Knowledge Promotion), Norwegian teachers are now instructed to assess pupils' performances according to Year specific competence aims. The aims are not to be lowered for any pupil, but not every pupil is meant to fulfil all criteria for each of the aims.

At any rate, whether the teachers lower the bar to get the pupils over it or accept that not every pupil is supposed to get over all bars every time, the distinction between performer (pupil) and performance (e.g. pupil text) tends to be quite blurry (cf. Elliot, 2013). Feedback requires contextual knowledge in order to be successful (Hattie & Timperley, 2007); how different pupils develop competence in different situations is for the teachers to know. As such, the performance is closely linked to the performer. However, whether the teacher sees the performance as an expression of pupil competence or as an expression of a pupil's personal character makes a difference. To prevent the latter from happening, the texts that the raters assess in the sample-based writing test are blinded. The message is clear: The raters are supposed to assess the text in relation to given pre-set criteria only.

[51] Preliminary findings from the research project *Developing national standards for the assessment of writing. A tool for teaching and learning* (http://norm.skrivesenteret.no/wp-content/uploads/2012/02/-Prosjektskisse.pdf) confirm such an understanding (personal communication with the project leader, Prof. Synnøve Matre, September 21, 2014).

Yet, in the dialogues there are numerous examples where the raters find it hard not to take into account the writer who tends to manifest himself/herself in the text. Below is an example where raters R3 and R4 are about to conclude on the text domain 'Use of language'. Earlier in the assessment of this particular text the raters have expressed frustration about where to 'penalize' for the writer's extensive use of brackets:

November 2011. Raters R3 and R4:

R3 Yes, but it was here we were supposed to penalize [for the brackets.

R4 [Yes, yes, true.

R3 So, perhaps it's a bit silly to-

R4 To?

R3 To penalize here, or if we penalize her here then it will be in a domain where she actually is very good.

R4 Yes, and I think that this is a girl who rather should be given guidance about, eh-.

R3 'Your writing is a bit too colloquial and perhaps too colloquial with all these brackets'.

R4 Yes. If she is going to reach the highest level she has to be given guidance about what it takes. 'Because you master so many other things so well'.

R3 Yes.

R4 So, she has a high level of mastery after all.

The raters see a 'model pupil' behind the text, and she has their sympathy. In this example they even construct an imagined dialogue with the girl where they instruct her on how to improve while they at the same time decide not to mark the text down even though they admit it has flaws. It is worth mentioning that this also applies when antipathy arises as well, and we find examples where the raters read 'lazy pupil' out of the texts. The raters' use of language is then typically hostile, for example: 'I don't understand what *he's* babbling on about!', '*He* doesn't understand anything!' and '*He*'s totally lost!'. Such language is here seen as indicative of a more severe assessment practice than the one undertaken in the 'model pupil' texts, but what is important in this context is that they are equally successful as strategies to make decisions about text quality.

Whereas the raters' tendency to 'judge' the writer (pupil) rather than the text is a highly questionable assessment strategy, their tendency to 'look ahead', to judge the text in

relation to a conception of a writing development continuum, is in most cases desirable. Teachers are experts in finding such signs of development, and utterances like 'she is on her way' and 'there are places in the text where he shows that he knows how to do it' flourish in the data material: The raters have brought their formative assessment classroom practice with them to the assessment panel.

Strategy of using personal standards

One common strategy found amongst the raters is to refer to their standards obtained through their individual classroom praxis. Working with pupils' texts over years has given the teachers an idea of what to expect from pupils at different age levels. In the following excerpt we see how R3 and R19 discuss a support question related to a text's spelling. They agree that the text has few errors, but R19 thinks this is 'as expected' since the pupil has a simple vocabulary. This leads R3 to compare the present text with texts written by pupils in her Year-7 class, and she finds the spelling in the present text 'way above average'. R19 accepts this as a valid argument and they assess accordingly:

November 2011. Raters R3 and R19

R3 But he doesn't have many [errors.

R19 [I agree.

R19 Mhm. All in all, I think this is as expected. He doesn't use many of those complicated words, really.

R3 In my Year-7 class he would have been way above average.

R19 Yes?

R3 Yes. In my experience the spelling here is better than what can be expected.

R19 Then let's mark that as a plus.[52]

The raters do not always verbalize their personal standards in this way. Quite often they seem to agree without reasoning, as the next excerpt exemplifies:

[52] "Plus" is a response alternative to a support question in the *Rubric*, meaning that the raters find the current characteristic to be better than what can be expected.

November 2011. Raters R12 and R13

R13 It's a girl.

R12 Who writes directly to a reader. So, I think we should mark this as M3, yes. Of course, I agree with you there.

R13 Agree?

R12 Great, great, great, great, great. So, then it's M3 on both Communication and Content.

R13 Yes, I would say so.

R12 Yes.

The use of (tacit) personal standards as a strategy is both expected and problematic. Expected because it is an expression of the competence the raters already possess prior to their participation in the assessment panel. In fact, it was this competence that made them candidates for the panel in the first place. However, the raters' use of classroom assessment standards is also problematic because earlier studies have shown that Norwegian teachers do not share views on what to expect from pupils at different Year levels (Fasting, Thygesen, Berge, Evensen & Vagle, 2009). An extensive use of their personal standards as an assessment strategy would challenge the development of a shared assessment culture. In addition, when this praxis is tacit, the decision-making strategies remain unknown. Hence, regardless of whether the verbalized or the tacit personal standards are in play, such praxis is a major contributor in making large-scale writing assessments quite blurry; the first in terms of being potentially unreliable and the other in terms of being hard to validate.

Strategy of using shared standards

I have already pointed out that the *Rubric with its support questions* structures the assessment dialogue into smaller communicative projects (see 'Data collection and analysis' above). The *Rubric* and the *Guideline with its benchmark texts* are the most important assessment resources that contribute to setting national standards for what to expect from pupils' texts. The *Rubric*'s support questions work normatively both in terms of sharpening the focus and in terms of the questions' scoring alternatives (-/0/+). The general information in the *Guideline* is a reminder to the raters about key contextual features, like the scale and the primary trait model, while the second section in the *Guideline* presents standard descriptors and

benchmark texts that exemplify these standards. Descriptors and benchmark texts are held to be key elements when assessing according to given standards (Sadler, 1987).

In the following excerpt we see how R21 wants to seek advice in the *Guideline* when the raters are unsure about a punctuation issue:

June 2011. Raters R8, R21, R23 and R24[53]

R23 But the question is [how serious such errors in the use of commas are?

R8 [She is on her-

(…)

R23 Yes.

R8 But she is on her way.

(…)

R21 But what did the *Guideline* say about this?

R24 She might have-

R23 It says that it is to be expected that the pupil knows how to use commas both with co-ordinated and dependent clauses.

R8 Yes, that's right.

R23 So, perhaps she is a bit under then?

R21 Yeah.

Rater R8 would clearly like to award the pupil for being 'on her way', but by relying on the *Guideline*'s descriptors they all agree that the pupil's use of commas in the text in question is 'a bit under'. Considering that the main purpose for establishing the assessment panel was to create shared assessment practice, this example of a standardization strategy must be evaluated as exemplary. The fact that this is a joint effort is noteworthy: R23 raises an important question that R8 wishes to answer by using a competing *formative assessment strategy*. Then a third rater, R21, offers another strategy that is quickly given the necessary support. In this way both the dialogue, involving both acknowledgement and manners of alterity, and the use of shared standards contribute to making the assessment transparent, and thereby verifiable and possible to replicate.

[53] In rare occasions, and with irregular intervals, the researchers that lead the assessment panel merged two pairs and made groups of four raters in order to get the raters to share their understanding among several.

However, one needs to guard against being too faithful. It is argued that when assessing writing there is a need for a critical rater who is able to see both the rationality and the irrationality when it is reflected in the different assessment situations. Rater R3 shows this ability in the following short excerpt:

November 2011. Raters R3 and R4

R3 So, again I think it turns out to be a bit silly-

R4 Yes, but-

R3 -that the support question awards him for doing something that is not positive.

Here R3 questions the support question in the *Rubric* and uses her position as a member of the assessment panel to override a pre-set criteria. In doing this she avoids what Sadler so wisely has warned against, that pre-set criteria are likely to make even the most competent rater lose sight of qualities in the text at hand (Sadler, 2009. See also Borgström and Ledin, 2014, about raters' suffering from 'technical knockout', i.e. the pre-set criteria force raters to give scores that contradict with their professional judgment).

Asymmetries of knowledge and participation

The raters have joined an assessment panel where the social dimension – both formal and informal – is supposed to stimulate learning and shared understanding and practice. Prior to joining the panel the raters are held to be a fairly homogenous group; all of them are experienced teachers with special interest in writing and the assessment of writing. The pair assessments should therefore be expected to be relatively symmetrical with raters having equal rights and obligations to contribute (cf. Linell, 2001, p. 258). However, wherever meaning is negotiated, a level of asymmetry between the interlocutors is inevitable. Lack of equality when meaning is constructed in interaction accentuates tension and competition (Linell, 2001; Matusov, 1996). This is also evident in the assessment dialogues; the raters must *agree upon* (negotiate meaning about) pupil text quality, a situation where the raters inevitably want their colleagues to be influenced by their own views on text quality.

This asymmetry appears to be due to a rater's potentially advantageous position in the panel, a position that can be obtained in many ways. During the assessment panel's plenary discussions, in the pair assessments and during other social gatherings, the raters have rich

opportunities to present themselves as particularly knowledgeable. It could be that they show an in-depth understanding of the writing construct and/or the assessment resources, or they might present themselves as particularly interested in the assessment of writing from their work as teachers. However, when such a position is successfully established it becomes clear that *who* utters something about text quality is in many cases just as important as the soundness of *what* is said. For instance, in the excerpt below raters R4 and R3 both make it clear that they find the rich profusion of brackets in the pupil text as communicatively destructive, even though other aspects of the text's communication are very good. Rater R4 has therefore suggested that they should mark this 'as expected', i.e. M3. But rater R3 disagrees:

November 2011. Raters R4 and R3

R3 I believe we should mark it M4.

R4 Yes?

R3 Then we do not let [the-, 'Communication' -, although it is quite right-

R4 [Yes

R3 -that it disturbs the communication.

R4 Yes, you're right, it does.

R3 Yes, it does.

Notice how R3 cleverly mentions the counterargument; that she recognizes that the brackets in the text disturb the communication, but notice also that she does so without presenting any new argument for her suggested mark (M4, i.e. better than what can be expected). She, as the speaker, is the argument. The interesting result is that the raters reach agreement where one rater's advantageous position has made them both act against their professional judgment as they both agreed that the brackets disturb the communication.

The way asymmetry between raters plays a role in the writing assessment is sometimes quite transparent. The next two excerpts are presented side by side to illustrate this. Excerpt A is the same excerpt as presented above as an example of 'Strategy of using personal standards'. Here rater R3 finds support for a judgment from her classroom praxis when she states that '[i]n my Year-7 class he would have been way above average'. Even though rater R19 first thinks the text is 'average' (i.e. as expected), she lets R3's personal praxis argument

win. Now, look at excerpt B. The same rater, R3, is now assessing together with another rater, R4. They are about to conclude their scoring of a given text's handwriting when R4 argues that the pupils' texts in general show a lack of coherent handwriting, a feature she finds 'incredibly conspicuous' and in contrast to what she in her personal praxis emphasizes ('All my pupils (…) had coherent handwriting'). R3 counters this argument by referring to the *Guideline* ('according to the Guideline they do not expect that either') which is fair enough; the *Guideline* is an important normative assessment resource. However, the raters have been asked to be critical of the assessment resources during the panel's preparatory stages. That is also what R4 insists on doing ('No, let's put a question mark behind this'), but R3 rejects such a strategy by simply stating that '[t]his is the situation in the Kingdom of Norway' and she soon just ends the discussion.

November 2011. R3 and R19 about to finish their assessment of a pupil text's spelling	**November 2011. R3 and R4 about to finish their assessment of a pupil text's handwriting**
R3 But he doesn't have many [errors. R19 [I agree. (…) R19 Mhm. All in all, I think this is as expected. He doesn't use many of those complicated words, really. R3 In my Year-7 class he would have been way above average. R19 Yes? R3 Yes. In my experience the spelling here is better than what can be expected. R19 Then let's mark that as a plus.	R4 Yes. But then they don't master it. R3 No, but- R4 That's why I'm- R3 - according to the Guideline they do not- R4 Yes. R3 - expect that either. R4 No, let's put a question mark behind, because-. Does not master coherent handwriting, and I find that incredibly conspicuous, so many who don't have coherent handwriting. R3 This is the situation in the Kingdom of Norway. R4 Not everywhere. All my pupils who finished Year 7 last year had coherent handwriting. R3 Really? R4 Yes. Absolutely everyone. They were really good – yes. R3 Then we're finished with the text.

Thus, in excerpt A, rater R3 rules out her colleague's argument by referring to what the pupils she has in her classroom normally achieve; *R3 finds the personal standard strategy valid*. In excerpt B, rater R4 is the one to refer to what her pupils achieve 'back home' when she states that the given pupil's handwriting is poor. But R3 now refutes this as a sound argument; *R3 finds the personal standard strategy invalid*. It is hard to understand this shift without underlining how a rater's position, or status, always has to be accounted for when

negotiations take place. One and the same rater might use his/her position to 'win' by using one particular strategy while (s)he in other situations uses the same position to argue against other raters' use of the same strategy. The two excerpts exemplify disharmonious use of assessment strategies. Hence, what we see is that both the raters' inconsistent use of different decision-making strategies and the raters' different participation in the pair/group assessments, have consequences for the scoring.

Discussion

We have seen how the raters apply different assessment strategies when making judgments about text quality; sometimes they look for a text's potential and assess thereafter (formative strategy), other times they find support in their respective standards developed over the years as teachers and yet other times they rely on the shared standards when making decisions. With the latter, the raters seem to choose assessment strategy quite explicitly (e.g. 'But what did the *Guideline* say about this?', '[A]ccording to the *Guideline* they do not expect that either'), while they seem to be more unaware of their use of the two other strategies. This is perhaps not surprising as the first two strategies are cultivated through years of teacher experience and as such a part of the raters' "doxa" (Bourdieu, 1977), or what would be the dominant ideology, and thereby "non-ideological" (Fairclough, 1995). Hence, we can understand the different strategies as belonging to different "communicative genres" (Linell, 2001): The raters' use of formative assessment strategy and personal standard strategy belong to what we could call a dominant *writing teacher genre*. But the assessment panel has been established to bring about change; the teachers are to become raters. To make change happen, a new strategy is made available (using shared standards) that belongs to a different communicative genre – we could call it *the writing assessment genre*.

The *writing teacher genre* is understood as a communicative genre whereby specific values and practices are appreciated. In the Norwegian context this would for instance be assessment for learning, individual feedback, adapted education and contextual affordances. *The writing assessment genre*, on the other hand, has other key features with terms like summative assessment, standards, objectivity and reliability. In between, of course, there are many common features, or 'boundary objects' (Star & Griesemer, 1989), such as validity and fairness. However, an important matter to resolve seems to be how we would like the relationship between these two major assessment strategies to develop.

It might be fruitful to not treat the raters' work in the assessment panel context as either belonging to the *writing teacher genre* (and therefore potentially being viewed as low-grade practice?) or belonging to the *writing assessment genre* (and therefore being viewed as good practice?), but rather treat this as a mixed blessing. The raters have been recruited to the assessment panel due to their competence as experienced writing teachers. Their know-how has been deemed to be worth something in this new context. It should not come as a surprise, then, that the raters sometimes allow themselves to question the standards and to rather rely on their previous assessment practice. As such, they view their own judgment practice as 'cumulative', where 'the influence of previous judgements can flow into and colour each new or subsequent judgement' (Wyatt-Smith & Klenowski, 2013, p. 37. Cf. Linell, 2001, about the concept of 'double dialogicality').

However, the analyses show how aspects of the raters' assessment practice are indicative of inconsistent assessments. First, the unsystematic use of different communicative strategies when making judgements on text quality makes the assessments hard to reproduce. It is not easy to find any pattern for when the raters use the different judgement strategies. Instead, it seems like the raters in this preparatory stage of the panel's existence consciously/unconsciously try different strategies to see what 'works', that is, what leads to a judgment. Second, as exposed in the analyses, the social design of the assessment has advantages but also challenges. When raters are forced to articulate decision-making arguments, they might be adjusted and even contradicted by the co-rater(s). Several voices are likely to lead to better validated judgements (Jølle, 2014; Matusov, 1996). But such beneficial outcome depends on mutual trust and relative symmetry between the raters.

The different strategies which this study has displayed as being in play when assessing writing can be, but are not necessarily contradictory. A rater's personal standard might be aligned with the shared standards, but it might also deviate from them. The formative assessment strategy might add a dimension to a shared standard strategy without changing the outcome, but, on the other hand, it might lead to change. What this study has revealed is that different decision-making strategies exist even within an assessment panel where rubrics and guidelines, together with lectures and practice, are given much space and attention. An important implication to draw from this would be that future rater preparation would benefit from giving more attention to the internalized assessment practices the teachers bring to the fore, also when they are working as raters. To jettison all but one assessment strategy to

streamline the scoring does not seem to be a good idea when considering what it takes to carry out valid writing assessments. On the other hand, due to the present study's design with in-depth analyses of a limited data set, more research is needed on the topic 'rater proficiency' to enrich our knowledge about raters' decision-making processes.

Conclusion

A key question was raised early in this paper: Are possible inconsistencies in scoring an expression of rater incompetence or rater disagreement? If treated as the first, rater incompetence, it makes sense to strive for greater reliability, i.e. removing any 'noise' that leads to dissimilar assessment practice. But treated as rater disagreement, the situation changes. Then it is not first and foremost a matter of a lack of shared assessment practice (although this is certainly part of the issue), but rather about rater *values* and/or rater *choices* within a professional judgement practice. How this issue is treated is obviously important.

The raters turn pupil writing performance into outcome, evidently an extremely fragile hermeneutic practice. In the specific context presented in this paper the raters are sitting in pairs or small groups trying to agree on what counts as hallmarks for text quality, what is it in the different texts that shows (or reveals a lack of) evidence of these hallmarks, and what are different achievements worth on a given scale? This is, of course, extremely challenging. Rater R27 puts words to this wisdom in the following long-drawn sigh during one of her rating sessions: '*Just sitting here looking at these texts without well-defined borders. We are not able to-, we are constantly moving one step back or one step forward, you know*.' Her frustration is an expression of why we should be cautious when working with writing assessment: Cautious in being meticulous when designing a writing test, cautious in being patient when developing a test and cautious in being prudent when interpreting the outcome.

References

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing, 15*, 133-153.

Barkaoui, K, (2011a). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*, 279-293.

Barkaoui, K. (2011b). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. Language Testing, 28 (1), 51-75.

Behizadeh, N. & Engelhard Jr., G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*, 189-211.

Borgström, E. & Ledin, P. (2014). Bedömarvariation. Balansen mellan teknisk och hermeneutisk rationalitet vid bedömning av skrivprov. *Språk & Stil*, *24*, 133-165. [Rater variance. The balance between technical and hermeneutic rationality when assessing writing].

Bourdieu, P. (1977). *Outline of a Theory of Practice*. Cambridge: Cambridge University Press.

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English*, 35 (2), 213-260.

Brown, G., Glasswill, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9,* 105-121.

Colombini, C. B. & McBride, M. (2012). "Storming and norming": Exploring the value of group development models in addressing conflict in communal writing assessment. *Assessing Writing, 17*, 191-207.

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*, 1-21.

Cumming, A., Kantor, R. and Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67-96.

DeSeCo. (2005). *The Definition and Selection of Key Competencies*. Retrieved April 18, 2014, from http://www.oecd.org/pisa/35070367.pdf.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155-185.

Elliot, V. (2013). Empathetic projections and affect reactions in examiners of 'A' level English and History. *Assessment in Education: Principles, Policy & Practice, 20*, 266-280.

Fairclough, N. (1995). *Critical Discourse Analysis*. Boston: Addison Wesley.

Fasting, R., Thygesen, R., Berge, K. L., Evensen, L. S., & Vagle, W. (2009). National assessment of writing proficiency among Norwegian pupils in compulsory schools. *Scandinavian Journal of Educational Research, 53*, 617-637.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Huot, B.A. (1993). The influence of holistic scoring procedures on reading and rating pupil essays. In M. M. Williamson and B. A. Huot (eds), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Creskill, NJ: Hampton Press.

Huot, B. (2002). "(Re)articulating writing assessment for teaching and learning". *All USU Press Publications*. Book 137. Retrieved May 20, 2014, from http://digitalcommons.usu.edu/-usupress_pubs/137

Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing, 14*, 3-24.

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic appriach for styding the development of rater competence. *Assessing Writing, 20*, 37-52.

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12*, 108-128.

Kuiken, F. and Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329-348.

Leckie, G. & Baird, Jo-Anne (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, *48*, 399-418.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*, 543-560.

Linell, P. (2001). *Approaching Dialogue. Talk, interaction and contexts in dialogical perspectives*. Amsterdam: John Benjamins.

Linell, P. (2009). *Rethinking Language, Mind, and World Dialogically*. Charlotte, NC: IAP, INC.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York: Peter Lang.

Markova, I. & Linell, P. (1996). Coding elementary contributions to dialogue: Individual acts versus dialogical interactions. *Journal for the Theory of Social Behaviour, 26*, 353-373.

Matusov, E. (1996). Intersubjectivity without Agreement. *Mind, Culture and Activity, 3*, 25-45.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Moss, P. A. (1994). Can there be validity Without reliability? *Educational Research, 23*, 5-12.

Petruzzi, A. (2008). Articulating a hermeneutic theory of writing assessment. *Assessing Writing*, 13 (3), 219-242.

Potter, W. J. & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*, 258-284.

Purves, A. C. (1992). A comparative perspective on the performance of pupils in written composition. In A. C. Purves (red.). *The IEA study of written composition* (Vol. 2. 129-152). Oxford: Pergamon.

Reid, L. (2007). Teachers talking about writing assessment: valuable professional learning? *Improving Schools, 10*, 132-149.

Sadler, R. (1985). The origins and functions of evaluative criteria. *Educational Theory, 35*(3), 285-297.

Sadler, R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13,* 191-209.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*, 159-179.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment*. Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida (pp. 129-152). Cambridge: Cambridge University Press.

Skolinspektionen (2013). *Olikheterna är för stora. Omrättning av nationella prov i grundskolan och gymnasieskolan, 2013* [*The differences are too big. Re-assessment of national tests in primary and secondary education, 2013*]. Retrieved June 28, 2014, from http://www.skolinspektionen.se/Documents/Rapporter/spara-2013/omrattning-nationella-prov-2013.pdf

Solheim, R. & Matre, S. (2014). Forventninger om skrivekompetanse. Perspektiver på skriving, skriveopplæring og vurdering i 'Normprosjektet'. *Viden om læsning, 15*, 76-89. [Expectations about writing skills. Perspectives on writing, the teaching of writing and assessment within the 'Norms project']

Star, S. L. & Griesemer, J. R. (1989). Institutional ecology, `translations` and boundary objects: Amateurs and professionals in Berkeley's Museum of vertebrate zoology, 1907-1939. *Social Studies of Science*, 19 (3), 387-420.

Stuhlmann, J., Daniel, C., Dellinger, A., Kenton, R. & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology, 20*, 107-127.

Thygesen, R., Berge, K. L., Evensen, L. S., Fasting, R. B. (2007). *Sluttrapport: nasjonale prøver i skriving som grunnleggende ferdighet* [*Final report: National tests in writing as a key competency*]. Stavanger: Nasjonalt senter for leseopplæring og leseforsking, Universitetet i Stavanger

Weigle, S. C. (1994). Weigle, S. C. (1994). *Effects of training on raters of ESL compositions: Quantitative and qualitative approaches*. Unpublished PhD dissertation, University of California, Los Angeles.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*, 145-178.

Wheel of writing (2012). Retrieved August 3, 2014, from http://www.skrivesenteret.no/-uploads/files/WritingWheel_Static_(rev_03_2012)_2.pdf

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*, 150-173.

Wolfe, E. W., Kao, C. and Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication*, 15, 465-492.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers´ judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*, 59-75.

Wyatt-Smith, C. & Klenowksi, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgment practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 35-52.