AHELO mulighetsstudie

Oppsummering av erfaringene med å gjennomføre OECDs AHELO mulighetsstudie i Norge

Elisabeth Hovdhaugen
Vibeke Opheim
Jørgen Sjaastad
Rachel Sweetman

Rapport 11/2013

NIFU

AHELO mulighetsstudie

Oppsummering av erfaringene med å gjennomføre OECDs AHELO mulighetsstudie i Norge

Elisabeth Hovdhaugen
Vibeke Opheim
Jørgen Sjaastad
Rachel Sweetman

Rapport 11/2013

Rapport 11/2013

Utgitt av Nordisk institutt for studier av innovasjon, forskning og utdanning
Adresse PB 5183 Majorstuen, NO-0302 Oslo. Besøksadresse: Wergelandsveien 7, 0167 Oslo

Oppdragsgiver Kunnskapsdepartementet
Adresse Postboks 8119 Dep., N-0032 Oslo

ISBN 978-82-7218-903-6
ISSN 1892-2597 (online)

www.nifu.no

Forord

Denne rapporten er en dokumentasjon av gjennomføringen av AHELO mulighetsstudie (Assessment of Higher Education Learning Outcomes) i Norge. AHELO mulighetsstudie ble initiert av OECD i 2008/2009, og datainnsamlingen ble gjennomført i 2012. Hensikten med studien var å undersøke om det er mulig å måle studenters læringsutbytte på tvers av land og læresteder.

På oppdrag fra Kunnskapsdepartementet (KD) har NIFU hatt ansvar for koordinering og gjennomføring av AHELO mulighetsstudien i Norge, og også hatt ansvar for å dokumentere og oppsummere prosessen og resultatene fra studien. KD har formelt sett vært ansvarlig for AHELO mulighetsstudie i Norge, og hatt det overordnede ansvaret for samarbeid med OECD, samt kontakten med lærestedene.

Rapporten er skrevet av flere forskere i samarbeid. Elisabeth Hovdhaugen og Vibeke Opheim har skrevet de fire første kapitlene samt oppsummeringskapitlet, med hjelp av Rachel Sweetman som også har skrevet den engelske oppsummeringen. Jørgen Sjaastad har beskrevet analysene av testene som er brukt i kapittel 5 (5.2-5.4), mens resten av dette kapitlet er skrevet av Elisabeth Hovdhaugen. Tidligere utkast av rapporten har vært lest og kommentert av Kyrre Lekve og Per Olaf Aamodt.

Målgruppen for rapport er i første rekke oppdragsgiver, det vil si Kunnskapsdepartementet, men også andre som er interessert i internasjonale studier av læringsutbytte.

Oslo, mars 2013

Sveinung Skule
Direktør

Jannecke Wiers-Jenssen
Forskningsleder

Innhold

Sammendrag **7**

English summary **11**

1 Innledning **15**
1.1 Bakgrunn for AHELO 15
1.1.1 Ansvar og forankring av AHELO i OECD 16
1.2 Organisering av AHELO mulighetsstudie i Norge 17
1.2.1 Finansieringen av AHELO mulighetsstudie i Norge 18
1.2.2 OECDs finansieringsplan for AHELO 18
1.3 AHELO: et prosjekt med flere deler og faser 19

2 Beskrivelse av testinstrumentene **21**
2.1 Case-oppgaven 22
2.2 Multiple choice spørsmål 22
2.3 Måling av kontekstuelle forhold 23

3 Fase 1: Planlegging **25**
3.1 Oversettelse og tilpasning av testinstrumenter 25
3.1.1 Oversettelse av case-oppgaven (CLA) 25
3.1.2 Utvelgelse og oversettelse av multiple choice spørsmål 26
3.1.3 Oversettelse av kontekstinstrumentene 27
3.2 Pre-implementeringsfasen 27
3.3 Melding av prosjektet til Personvernombudet for forskning 28
3.4 Vurdering av planleggingsfasen (fase 1) 28

4 Fase 2: Gjennomføringsfasen **31**
4.1 Trekking av utvalg 31
4.2 Teknisk testing av systemene 32
4.3 Gjennomføring av testing 33
4.4 Skåringen av testen 34
4.5 Lærestedenes tilbakemeldinger på prosessen 35
4.6 Vurdering av gjennomføringen av AHELO mulighetsstudie 36

5 Analyser av data **39**
5.1 Analyser av hvem som har deltatt i AHELO 39
5.1.1 Studentenes innsats og testens relevans 41
5.1.2 Oppsummering 42
5.2 AHELO-oppgavene: psykometriske egenskaper 43
5.2.1 «Generic Skills Score variance explained» 43
5.2.2 «Generic Skills Assessment variable map» 44
5.3 AHELO-oppgavene: bruk i Norge 47
5.3.1 Et empirisk anker 48
5.3.2 «Generic Skills Assessment variable map» for MCQ 51
5.3.3 «Item fit» for MCQ 51
5.3.4 «EAP/PV Reliability estimates» 53
5.3.5 «Differential item functioning (DIF)» 53
5.4 Analyser av de norske data fra AHELO 55
5.4.1 MCQ-oppgavene 55
5.4.2 Hva slags MCQ-oppgaver gjorde de 115 norske studentene det best på? 57
5.4.3 «Lake-to-river» (CRT1) og «Catfish» (CRT2) 58
5.5 Oppsummering 59

6 Sammenfattende diskusjon **61**
6.1 Rekruttering av studenter – svarprosent 62
6.2 Hva var det mulig å få ut av data? 63
6.3 Utbytte av AHELO i Norge? 64

Referanser **67**

Vedlegg **69**
Vedlegg A: Oversettelser 69
Vedlegg B: Oversikt Cognitive Labs 71
Vedlegg C: CLA-dokumenter for oversettelse/tilpasning 73
Vedlegg D: Møter og kommunikasjon 75

Sammendrag

AHELO (Assessment of Higher Education Learning Outcomes) er et OECD-prosjekt som har som mål å undersøke om det er mulig å måle læringsutbytte i høyere utdanning på tvers av institusjoner og land. AHELO er en mulighetsstudie, og en mulighetsstudie skiller seg fra en pilot ved at den har som ambisjon å undersøke om det er *mulig* å gjøre det studien har som mål, det vil si å teste ut om dette er en type studie som egner seg for gjennomføring i større skala. En pilot derimot er en småskala test av en undersøkelse man har bestemt at man skal gjennomføre.

Denne rapporten dokumenterer prosessen med å gjennomføre AHELO mulighetsstudie i Norge. Hovedkonklusjonen er at det er teknisk mulig å oversette og tilpasse tester og gjennomføre elektronisk testing på læresteder. Derimot viste det seg svært vanskelig å rekruttere studenter til deltakelse, noe som resulterte i svært lav svarprosent.

Organisering av AHELO i Norge

Kunnskapsdepartementet har stått som ansvarlig for undersøkelsen i Norge, og er oppdragsgiver for rapporten. NIFU har fungert som nasjonal prosjektleder (National Project Manager, NPM) og hatt ansvaret for den praktiske gjennomføringen, samt analyse av data og rapportskriving. Kunnskapsdepartementet har hatt det overordnede ansvaret for samarbeid med OECD og med lærestedene, samt fattet alle formelle beslutninger om AHELO.

Enheten i AHELO er institusjon, ikke land, og i mulighetsstudien er det fem norske læresteder som deltar: NTNU, UMB, Universitetet i Stavanger, Høgskolen i Lillehammer og Høgskolen i Vestfold.

Om AHELO

AHELO er et prosjekt som består av flere deler, to moduler som skal teste ferdigheter i bestemte fag (Economics og Engineering) og en modul som skal undersøke generelle ferdigheter (Generic skills strand). Norge har kun deltatt i sistnevnte. For å måle generelle ferdigheter er to tester brukt, en case-oppgave som er en tilpasset versjon av den amerikanske testen Collegiate Learning Assessment (CLA) og flervalgsoppgaver (multiple choice) som er hentet fra en australsk oppgavedatabase. Med andre ord har man valgt å bruke to allerede eksisterende tester for å undersøke generelle ferdigheter i AHELO mulighetsstudie. I tillegg har prosjektet omfattet kontekstuelle spørreskjemaundersøkelser til deltakende studenter, vitenskapelige ansatte og institusjoner, som brukes på tvers av de tre delmodulene. Hensikten med å samle inn kontekstuelle data er å på sikt muliggjøre sammenligning på tvers av læresteder, ved at læresteder kan sammenligne seg med lignende læresteder. I mulighetsstudien vil slik sammenligning ikke være et tema, da hovedpoenget med studien er å se om det i det hele tatt er mulig å gjennomføre slike tester på tvers av land, kultur og institusjon.

Forberedelsesfasen

Første del av prosjektet dreide seg i stor grad om oversettelse og tilpasning av case-oppgaven, som det fra starten av var bestemt skulle brukes som testinstrument i Generic skills strand. Oversettelsen av CLA-oppgaven var omfattende, men det var satt av god tid til dette. Derimot kom ikke oversettelse og tilpasning av flervalgsoppgavene og de tre kontekstuelle spørreskjemaene i gang før endelig vedtak om at studien faktisk skulle gjennomføres ble tatt sommeren 2011. På grunn av finansielle vanskeligheter i prosjektet ble gjennomføringen utsatt med et år, og dermed ble siste del av forberedelsesfasen svært hektisk. Oversettelse av flervalgsoppgavene og spørreskjemaer ble derfor gjort under forholdsvis korte tidsfrister, men takket være gode rutiner for oversettelse fra AHELO konsortiets side ble de tilfredsstillende. Konklusjonen fra første del er at det er mulig å oversette og tilpasse oppgaver som er utviklet i andre land til norsk, men at dette for caseoppgavene var en forholdsvis kostnadskrevende prosess når omfanget på tekstene som skulle oversettes er stort. Småskalatesting av oversettelsen viste at studenter med ulik fagbakgrunn angrep oppgaven på ulike måter, noe som kan indikere at oppgaven ikke oppfattes likt av studenter med ulik fagbakgrunn.

Gjennomføringsfasen

Andre del av prosjektet var selve gjennomføringen av studien, inkludert utvalgstrekking av studenter og vitenskapelig ansatte, testing av de tekniske løsningene for gjennomføring og skåring av case-oppgavene etter gjennomføring. Den praktiske gjennomføringen av AHELO mulighetsstudie fungerte nesten helt uten tekniske problemer, mye takket være grundig testing av lærestedene før gjennomføringen startet. Men samtidig hadde Norge flere utfordringer i gjennomføringsfasen. Den største var å få studentene til å delta. Til tross for at lærestedene arbeidet for å rekruttere studenter på flere fronter, hadde undersøkelsen svært lav oppslutning. Totalt sett var det bare 115 av de 1500 uttrukne studentene som valgte å delta, noe som er så lavt at man ikke kan si noe substansielt om norske studenter eller om lærestedene basert på dataene. Dersom deltakelse i en fremtidig AHELO skal vurderes må det finnes en klar strategi for hvordan studenter skal rekrutteres til å delta i studien. Erfaringene fra mulighetsstudien tilsier at informasjonskampanjer om prosjektet og moderate insentiver for å delta ikke har vist seg å være tilstrekkelig til å få en akseptabel svarprosent. Hele prosjektet må derfor designes slik at studenter ønsker å delta, og dette må påvirke blant annet valg av utvalgsmetode, oppgaveutforming, testsituasjon og tidspunkt for testing.

Begrenset informasjon i data

Det er begrenset hvor mye det er mulig å få ut av dataene, siden Norge hadde få deltakere i studien. Men vi kan likevel si noe om hvordan testene har fungert i Norge. Både analyser gjort på de norske dataene og AHELO konsortiets analyser viser at case-oppgavene (CRT) fungerer mindre bra enn flervalgsoppgavene (multiple choice, MCQ) for norske studenter. Blant annet forklarer innsats på testen forholdsvis mye av variasjonen i de to case-oppgavene, mens det ikke i samme grad var tilfelle for flervalgsoppgavene. I tillegg gir ikke case-oppgavene noen klar tilbakemelding til lærestedene om hva studentene er gode eller mindre gode på, siden skårene på de tre dimensjonene oppgaven rettes etter sammenfaller i stor grad. Konsortiet har laget en samleskåre, som er basert på både case-oppgave og flervalgsoppgavene, og denne gir heller ikke mye informasjon til lærestedet.

Dersom man ønsker å gå videre med AHELO kan det ligge et potensial i å videreutvikle flervalgsoppgaver. Disse kan gi en mer differensiert tilbakemelding til både student og lærested, gitt at man tar utgangspunkt i et felles rammeverk, slik at man er enig om hvilken type ferdigheter testen er ment å måle, og bruker en utvalgsteknikk som gjør det mulig å gi tilbakemelding til både student og lærested.

Norge har lært mye gjennom å delta i AHELO mulighetsstudie, selv om svarprosenten var lav og data dermed ikke kan brukes på den måten det var ment å brukes. Det er flere årsaker til at svarprosenten ble lav. Det avhenger av en rekke faktorer som at siste semester i bachelorgraden ikke er et optimalt tidspunkt for å nå studenter på og at man valgte en metode for å trekke utvalg som ikke bidro til å gjøre rekrutteringsarbeidet lettere. I tillegg tar testen 2,5 timer å gjennomføre og den må gjennomføres på et datarom, noe som heller ikke gjorde rekrutteringen av deltakere lettere. Lærestedene la ned mye tid og energi i å rekruttere studenter, og alle lærestedene hadde ulike former for insentiver, men uten

at det hadde stor innvirkning på rekrutteringen. Dersom man ønsker å gå videre med AHELO bør rekruttering bygges inn i designet av studien fra starten av.

English summary

Introduction
AHELO (Assessment of Higher Education Learning Outcomes) is an OECD project investigating whether it is possible to measure learning outcomes in higher education, across a range of countries. AHELO has taken place as a feasibility study, running from 2010-2012, the aims of the feasibility study being: to test out specific approaches or tools to directly measure / test students' learning outcomes, and to investigate if such testing is feasible across countries and cultures.

This report documents the process of implementing the AHELO feasibility study in Norway and presents the results from analyses of the Norwegian data. The study was commissioned by the Norwegian Ministry of Education, which had overall responsibility for the AHELO feasibility study in Norway. The ministry delegated the implementation and analysis of data to NIFU as the selected National Project Manager (NPM) and NIFU was also responsible for preparing this report. The Ministry of Education remained in charge of overall cooperation with the OECD, of coordinating the participating Higher Education Institutions (HEIs), and had responsibility for all formal decisions regarding AHELO.

As a technical feasibility study, the AHELO project has shown it is possible to translate test instruments, set up and use secure, online test environments, and coordinate processes across a large number of sites; process and technical solutions that function well are in place and cooperation and communication was good throughout the project. However, the Norwegian project faced a major challenge in recruiting students to take part. The very low response rates limit the potential for the AHELO results to be used to draw firm conclusions or further our understanding of some of the study's broader aims, such as testing the validity and reliability of specific tests, over differing national and disciplinary contexts. Answering these broader questions about the potential for valid international, and cross-disciplinary comparisons of HE learning, and finding ways secure students' participation in such a study, remain as challenges for future work.

The AHELO test
The AHELO project was organized via three strands of testing, with two disciplinary-specific strands testing skills in Economics and Engineering, and a strand attempting to test general abilities (titled the 'Generic Skills Strand'). Norway only participated in the last strand. The generic skills strand was based on two tests. The first test was a set of Constructed Response Tasks (CRTs) based on an adapted version of the American Collegiate Learning Assessment (CLA). In CLA the test-taker is presented with a hypothetical situation where they are supposed to give a recommendation or present a viable explanation for a phenomenon, using a given set of digital documents. Their answer has to be presented as a form of essay.This test is intended to measure three distinct dimensions of ability, assessed in parallel within each task: analytic reasoning and evaluation, writing effectiveness, and problem solving. A second test instrument using multiple-choice questions (MCQs) was also used, sourced from Australian graduate student surveys. The project also included several contextual

surveys, to be completed by participating students, academic staff and institutions, respectively; these provided contextual information relevant to all three strands.

The preparation and planning stage

The first stage of the project largely involved the translation and adaptation of CRT items, which were originally planned to be used in the Generic Skills strand, and which were circulated early on to allow plenty of time for translations. The multiple choice tasks and the three contextual instruments (questionnaires for students, faculty and institutions) were only made available later, leaving a relatively short period for translation; this was due to a delay in the final confirmation that the project would go ahead, which was taken in July 2011. As the AHELO survey was to be conducted in spring 2012, this left only 6 months for the preparation and checking of these instruments. Thanks to good cooperation and good translation processes between the members of the consortium, the translation of the multiple choice tasks and context surveys were completed to a good standard. The conclusion that can be drawn from this first stage of the process is that it does seem to be possible to translate tasks developed in other countries for use in Norway, but that this can be a relatively time-consuming and expensive process if CRTs are involved, due to the extensive reference texts that require translation. It is not possible to conclude how well such tasks work in a Norwegian setting, as they have not been piloted in Norway (though the MCQs were piloted in Australia). However, small-scale testing suggested that students with different disciplinary backgrounds approached these CRTs in varying ways, which may indicate that they are not interpreted or completed consistently by students across disciplinary areas. It was hoped that the final AHELO data would shed more light on this issue, but the low response rates mean this cannot be resolved in regards to the Norwegian case at least.

The implementation stage

The second part of the project was the implementation of the study, including sampling students and academic staff and testing the technical solutions, the testing process and scoring of CRTs. Norway experienced several challenges in the implementation phase, the most significant of which involved getting students to participate. Despite the institutions working hard to recruit students, using a variety of approaches, very few responses were secured and the institutions' response rates ranged from 4.7 to 10 percent. Overall, only 115 of the 1,500 students selected in the original sample chose to participate; this gave a response rate so low that it is not possible to draw substantive conclusions about Norwegian students based on the data. In considering future participation in AHELO, it seems a clearer strategy will need to be identified to recruit more students.

The practical implementation of the AHELO test worked fine, almost without technical problems. This is partly due to institutions conducting thoroughly testing before the start of the test.

There were several other, less significant challenges. Norway was the first country to start testing, and initial delays in receiving login information for participants led to a small number of tests being postponed or cancelled. There were also challenges related to sampling: although the sampling instructions were extensive, the processes involved were complex and the instructions were not always sufficiently precise. Norway has very good quality records available about its student population (via national registers) which were used in sampling, although a number of adjustments had to be made to prepare the appropriate sample (for example to adjust the way study credit points are recorded). However, such adjustments may lead to major differences in the kinds of institutions selected in different countries, which may in turn have implications for the comparability of results across institutions. The scoring process was completed on time, and all CRTs were scored by two assessors (to ensure the scoring rubric was applied consistently). Overall the scoring process worked well, and it was also significantly less expensive than anticipated, primarily because there were only 115 students' responses to score.

Conclusions and implications for future studies

Norway has learned a lot from participating in the AHELO feasibility study, and the preparation and implementation processes showed that the technical capacity and solutions, and necessary structures for cooperation are in place for such a study. However, the Norwegian study was undermined by low response rates, resulting in additional work during testing and in data which is not robust to support all of the intended analyses, or to answer some of the central challenges related to testing of this kind.

Only limited conclusions and lessons can be drawn from the resulting AHELO data. However, some basic analyses investigating how these tests functioned in Norway have been possible. Both the national analysis of Norwegian data and the international AHELO consortium's overall analysis suggest the CRT items work less well than the multiple-choice tasks, in demonstrating unstable results across the items. In addition, the CRT items did not seem to perform well in terms of the three 'dimensions' measured, as students' results for the three skills were highly correlated; this limits the potential for the tests to offer clear feedback for institutions about specific areas or skills where their students performed better or worse. Hence, if AHELO is to continue it would be wise to prioritise MCQs over CRTs, as they are more cost effective to create, administer and score and they have the potential to give institutions feed-back on what students' are good at and what they need to work more on.

There does not seem to be one explanation for the low response rate, but rather it appears to have been caused by a number of factors: the last semester of their bachelor's degree is not an optimal time to approach students, the length of the test required a significant investment of time from students, and the sampling method used did not allow for a more targeted recruitment approach. The institutions put a lot of time and energy into recruiting students, and offered various forms of incentives to take part, but without any of these leading to any major improvement in recruitment.

The experience from this feasibility study suggests that more information about the project or larger incentives to participate will not be enough to significantly improve response rates. It seems likely that effective steps to motivate students to take part will need to be built into the overall design and approach in future studies, not simply be addressed by individual countries during implementation. Such changes might include using more targeted, stratified sampling (the selection of whole classes or course groups) which would make more intensive and structured recruitment and test scheduling possible, or devising a test system that can give students immediate feedback on their performance, thereby increasing engagement and motivation to take part.

1 Innledning

Denne rapporten er en dokumentasjon av gjennomføringen av AHELO mulighetsstudie (Assessment of Higher Education Learning Outcomes) i Norge. AHELO mulighetsstudie ble initiert av OECD i 2008/2009. Hensikten var å undersøke mulighetene for å måle studenters læringsutbytte på tvers av land og læresteder. På oppdrag fra Kunnskapsdepartementet (KD) har NIFU hatt ansvar for koordinering og gjennomføring av AHELO mulighetsstudien i Norge.

Rapporten omhandler de ulike delene av dette prosjektet – fra fasen med å oversette og tilpasse ulike testinstrumenter, til prosessen med å gjennomføre undersøkelsen, samt vurdering av utbytte av AHELO mulighetsstudie i Norge. Dette omfatter en gjennomgang og vurdering av hvordan prosessen med gjennomføringen av AHELO mulighetsstudie har fungert, både i det internasjonale prosjektet ledet av OECD og i Norge. I tillegg inneholder rapporten analyser og vurdering av data generert fra den norske delen av AHELO mulighetsstudie.

I neste del av dette kapitlet beskrives bakgrunnen for AHELO mulighetsstudie. Her beskrives den internasjonale organiseringen av prosjektet, forankringen i OECD, samt OECDs finansieringsplan for AHELO mulighetsstudie. Deretter beskrives organiseringen av prosjektet og de deltakende partene i den norske delen av AHELO mulighetsstudie. I siste del av dette kapitlet presenteres selve AHELO mulighetsstudien nærmere. Studien består av ulike deler og moduler.

1.1 Bakgrunn for AHELO

Opprinnelsen til AHELO var diskusjoner om utbytte av høyere utdanning på et møte for utdanningsministre innenfor OECD-området i juni 2006. Et tema her var hvilke indikatorer som var tilgjengelige som mål på utbytte av høyere utdanning, og ikke minst hvilke indikatorer som *ikke* fantes. Det ble konstatert at generelt var lite tilgjengelig data om studenters læringsutbytte i høyere utdanning, mens det var flere typer indikatorer på andre aspekter av høyere utdanning, som forskningsresultater eller utbytte av høyere utdanning i arbeidsmarkedet (lønn, mindre risiko for arbeidsledighet mm) (Opheim og Aamodt 2010).

Utdanningsministrene inviterte derfor OECD til å utrede hvordan dette ”informasjonshullet” kunne fylles (OECD 2007a). I løpet av 2007 arrangerte dermed OECD tre møter (såkalt ad-hoc ekspertmøter) der eksperter innenfor høyere utdanningsfeltet møttes og diskuterte målsettingene, hensiktene og mulighetene for et slikt prosjekt. De to første møtene hadde til hensikt å diskutere mulighetene for å utvikle internasjonalt sammenlignbare måleinstrumenter for læringsutbytte i høyere utdanning, mens det siste møtet diskuterte mulig design av en slik studie (OECD 2007a, 2007b, 2007c). Møtene medførte enighet om å gjennomføre en type pilotstudie – eller en *mulighetsstudie* for internasjonal måling av studenters læringsutbytte, men ekspertgruppen forordnet ikke en viss type undersøkelse, da de ikke kunne komme til enighet om dette (OECD 2007c). Derimot ble de enige om at en slik mulighetsstudie skulle ha to målsettinger:

1) Å fremskaffe støtte for eller verifisering av denne fremgangsmåten for å måle studentenes læringsutbytte på, på tvers av land og læresteder (*proof of concept*).

2) Å teste den praktiske gjennomføringen av en slik test i ulike land og ved ulike læresteder.

AHELO mulighetsstudie har med andre ord som mål å undersøke om det i det hele tatt lar seg gjøre å konstruere et internasjonalt mål på lærestedenes bidrag til studenters læringsutbytte, og gjennomføre en slik test på tvers av land, kultur og institusjon. Ytterligere et møte blant utdanningsministre fra ulike OECD-land i Tokyo i januar 2008 fortsatte diskusjonen om mulighetene for å måle læringsutbytte i høyere utdanning og det var enighet om viktigheten av å utvikle gode mål på studentenes læringsutbytte i høyere utdanning, samtidig som man understreket kompleksiteten og det problematiske knyttet til slike indikatorer. Det ble derfor anbefalt at en fremtidig AHELO mulighetsstudie ikke kunne ha som mål å undersøke alle aspekter ved studentenes læringsutbytte, men skulle ha som målsetting å ta hensyn til lærestedenes historiske, språklige og kulturelle kontekst, samt forskjellene mellom land både knyttet til utdanningenes innhold (*pensum/curricula*), studienes varighet og rekrutteringsmønstre (*enrolment rates*) (OECD 2008a). Dette ministermøte ryddet vei for videre arbeid med å etablere prosjektet, og IMHE (OECD Programme on Institutional Management in Higher Education) Governing Board tok det opp og bifalt opprettelsen av en Group of National Experts (GNE) som sammen med sekretariatet skulle overse prosessen med å gjennomføre AHELO, samt rapportere til IMHE (OECD 2008b).

1.1.1 Ansvar og forankring av AHELO i OECD

AHELO mulighetsstudie er forankret i OECDs program for styring og ledelse av høyere utdanningsinstitusjoner (IMHE) og styringen av mulighetsstudien har vært delt mellom EDCP (OECD Education Policy Committee) og IMHE Governing Board (Tremblay, Lalancette & Roseveare 2012: 79). Forankringen i IMHE underbygger det sterke fokuset på lærestedet i AHELO mulighetsstudie. Mens OECD vanligvis har fokus på det nasjonale nivået, understrekes det at enhetene i AHELO mulighetsstudie er utdanningsinstitusjoner og ikke land. Hensikten med AHELO er ikke å sammenligne studentenes læringsutbytte på tvers av land, men å fokusere på det enkelte læresteds bidrag til studentenes læringsutbytte. Samtidig innebærer dette fokuset på utdanningsinstitusjoner, og ikke land, en spenning i AHELO. Dette kommer ikke minst til syne i finansieringsplanene for AHELO. Den mest brukte modellen for finansiering av internasjonale prosjekter i OECD består i hovedsak av bidrag fra de deltakende landene. Fokuset på utdanningsinstitusjoner og ikke land, åpnet imidlertid for et annet syn på dette i AHELO, nemlig for større bidrag fra private aktører – aktører med interesse for og nytte av utviklingen av kvalitetsindikatorer for læringsutbytte i høyere utdanning.

IHME har i sin tur delegert ansvaret for beslutninger om metoder, prinsipper og timing i AHELO til GNE, Group of National Experts (Tremblay et al. 2012: 79). Medlemmene i GNE er nominert av respektive nasjonale delegasjoner til EDCP/IHME. I tillegg har OECD opprettet et AHELO-sekretariat, som tar seg av den daglige driften og fungerer som kontakt mellom deltakende land, kontraktører og andre konsulter som prosjektet bruker og gir tilbakemelding til IHME og EDCP. Sekretariatet skal også overvåke gjennomføringen og skrive sluttrapporter fra mulighetsstudien. AHELO-sekretariatet, sammen med GNE hadde ansvar for å utlyse konkurranse om å lede prosjektet, og dette ble gjort i 2009. Den praktiske ledelsen av prosjektet (management) ble gitt til et internasjonalt konsortium under ledelse ACER, Australian Council for Educational Research. Samtidig hadde man allerede fra OECDs side inngått avtale med CAE, Council for Aid to Education, om å bruke deres testinstrument i Generic skills strand, og dermed hadde man to kontrakter, en kontrakt med CAE for modul A (Generic skills) og en kontrakt med ACER-ledet konsortium for de andre modulene. I det ACER-ledete konsortiet inngikk også Educational Testing Services (ETS) fra USA, National Institute for Educational Policy Reserach (NIER) fra Japan, University of Florence i Italia, Centre for Higher Education Policy Studies (CHEPS) fra Nederland og Center for Postsecondary Research (CPR) fra USA. ETS hadde ansvaret for å utvikle instrumenter for Economics Strand, mens ACER sammen med NIER og University of Florence hadde ansvar for utviklingen av instrumenter for Engineering Strand. ACER, CHEPS og CPR hadde ansvar for å utvikle de tre kontekstuelle surveyene, og ACER hadde alene ansvar for styring og ledelse av hele prosjektet.

En av oppgavene til konsortiet var å initiere Technical Advisory Group (TAG), en gruppe eksperter som skulle bistå AHELO GNE og konsortiet med råd med hensyn til utvikling av instrumenter og metode (OECD 2010b). Over tid tok TAG på seg mer og mer arbeid, blant annet ved å fungere som en ekspert gruppe for Generic Skills strand og den kontekstuelle dimensjonen, og for å få en mer fristilt posisjon ble TAG fra februar 2012 organisert under OECDs AHELO sekretariat, i stedet for under konsortiet (Tremblay et al 2012).

1.2 Organisering av AHELO mulighetsstudie i Norge

Oppdraget med praktisk å gjennomføre AHELO i Norge ble utlyst av Kunnskapsdepartementet (KD) som åpen anbudskonkurranse i 2009. NIFU (Nordisk institutt for studier av innovasjon, forskning og utdanning) fikk oppdraget, i samarbeid med Enhet for kvantitative utdanningsanalyser (EKVA) ved Institutt for lærerutdanning og skoleforskning (ILS) ved UiO. NIFU har ivaretatt funksjonen som nasjonal prosjektleder (National Project Manager, NPM) siden årsskiftet 2009/2010. EKVA/ILS har hatt ansvar for tilpassing og oversettelse av kartleggingsverktøyet til norske forhold, samt for poengsetting (skåring) av oppgavene[1].

Kunnskapsdepartementet har formelt sett vært ansvarlig for AHELO mulighetsstudie i Norge, og hatt det overordnede ansvaret for samarbeid med OECD, samt kontakten med lærestedene. KD har hatt ansvar for å organisere møter i prosjektgruppen, det vil si mellom KD, deltakende læresteder og nasjonal prosjektleder, organisere møter for AHELO styringsgruppe i departementet samt organisert informasjonsmøter om AHELO der representanter for Universitets- og høgskolerådet, studentorganisasjoner og ulike fagforeninger har deltatt (Konkurransegrunnlag 2009). Kunnskapsdepartementet har også fattet alle beslutninger knyttet til de finansielle sidene ved deltakelse i AHELO mulighetsstudie, inkludert deltakelse i gjennomføringsfasen (noe som medførte ytterligere finansiering).

Kunnskapsdepartementet har hatt ansvar for rekruttering av læresteder til å delta i studien. Fra 2009 var tre læresteder med: Norges teknisk-naturvitenskapelig universitet (NTNU), Universitetet for miljø og bioteknologi (UMB) og Høgskolen i Vestfold (HiVe). I 2011 ble også Universitetet i Stavanger (UiS) og Høgskolen på Lillehammer (HiL) med i studien. Dermed har totalt fem institusjoner i Norge deltatt i AHELO mulighetsstudien.

NIFUs ansvar har vært den praktiske gjennomføringen av studien: å tilpasse testinstrumenter og kvalitetssikre oversettelser, gjennomføre utvalgstrekkingen i samarbeid med lærestedene, følge opp støtte og drive opplæring av lærestedene i hvordan de tekniske systemene skal testes før gjennomføring og støtte og oppfølging under selve gjennomføringen samt rapportere til KD og OECD om prosessen og resultatene av gjennomføringen av AHELO mulighetsstudie i Norge (Konkurransegrunnlag 2009).

I følge konkurransegrunnlaget (2009) er lærestedenes ansvar i AHELO primært å gjennomføre selve testen på lærestedet. Lærestedenes ansvar er ytterligere spesifisert i et brev fra KD (Brev fra KD til læresteder i januar 2011), som sier at lærestedene har ansvar for følgende:

- *Etablering av en intern gjennomføringsorganisasjon*
- *Utpeking av prøverettere (skårere), som vil gjennomgå trening*
- *Tilrettelegging for og gjennomføring av selve testen, som vil omfatte ca 200 studenter i siste år av bachelorstudiet på hver institusjon. Disse skal gjennomføre en kontrollert 90 minutters test på elektronisk plattform*
- *Å sikre at tilstrekkelig antall studenter deltar, evt. ved å bruke enkle insentiver*

[1] Teamet ved NIFU ble fra prosjektoppstart ledet av forskningsleder Vibeke Opheim, med forskerne Elisabeth Hovdhaugen, Tine S. Prøitz og Per Olaf Aamodt som en del av prosjektteamet. Elisabeth Hovdhaugen tok over ansvaret som prosjektleder i mai 2011, og har vært prosjektleder ut prosjektperioden. I den forbindelse ble også AHELO-teamet ved NIFU utvidet til også å omfatte forsker Rachel Sweetman. Oversettelsen som EKVA/ILS gjorde ble gjennomført av forskerne Astrid Roe, Are Turmo og Inger Throndsen. For skåringen var professor Rolf Vegar Olsen ansvarlig kontaktperson ved EKVA/ILS og rådgiver Anna Eriksen var Lead scorer. Skåringen ble gjort av to vitenskapelig ansatte ved EKVA/ILS, som har tidligere erfaring fra skåring i PISA, og som fikk opplæring av Lead scorer.

- *Stille til rådighet datautstyr som testene legges inn på, samt tilrettelegging av elektronisk plattform. Testene og elektronisk plattform leveres av nasjonal prosjektleder i samarbeid med internasjonal prosjektledelse (ACER/CAE)*
- *Gjennomføring av en kort kontekstuell undersøkelse blant et antall ansatte på institusjonen*
- *Gjennomføre prøveretting, i tråd med prosjektets retningslinjer*
- *Overlevering av materiale til nasjonal prosjektleder*
- *Kort rapportering om gjennomføringen på institusjonen*

Dette dokumentet spesifiserer med andre ord hvilke oppgaver lærestedene kunne forvente at AHELO ville medføre, samtidig som dokumentet ikke inneholder informasjon om eventuelle direkte økonomiske kostnader (utover kostnader ved bruk av intern arbeidskraft) knyttet til gjennomføringen som lærestedene forventes å dekke. Dette kommer vi tilbake til i diskusjonen av gjennomføringen av AHELO mulighetsstudie i Norge i kapittel 4.

1.2.1 Finansieringen av AHELO mulighetsstudie i Norge

Den norske gjennomføringen av AHELO mulighetsstudie har primært vært finansiert med midler fra KD. KD har finansiert arbeidet fra NIFU, EKVA/ILS samt kostnader knyttet til internasjonal deltakelse. KDs arbeid med å rekruttere læresteder, arrangere møter med AHELO prosjektgruppen samt informasjonsmøter til eksterne grupper, har også være finansiert av KD. Men de deltakende lærestedene har også bidratt med finansieringen av AHELO mulighetsstudie i Norge, ved at de har dekket kostnader knyttet til gjennomføringen av selve testen på lærestedet.

Da kontrakten mellom NIFU og KD ble inngått i 2009 var det en del gjenværende usikkerhet i prosjektet. På det tidspunktet var lite kjent om hvordan gjennomføringen skulle gå til, siden det nyoppnevnte konsortiet fikk i oppdrag å utvikle plan for utvalg av studenter og faglig ansatte og kontekstuell survey. Dette gjorde at NIFU hadde en rekke forbehold i sitt tilbud, siden det var mange usikkerhetsmomenter i prosjektet med hensyn til fremdrift og gjennomføring. Den finansielle rammen for prosjektet har blitt utvidet underveis, grunnet forsinkelsen i prosjektet og merarbeid som ikke var tilstrekkelig spesifisert i den opprinnelige utlysningen. Blant annet var det ikke opprinnelig beregnet at NPM skulle være tilstede på alle møter i Paris og da det ble tilfelle måtte også den kostnaden inkluderes i prosjektet. I tillegg var det opprinnelig estimert et for lavt beløp for oversettelse av case-oppgaven, da den viste seg mye mer omfattende enn opprinnelig antatt. Ettersom materialet er rettighetsbelagt fikk man ikke anledning til å se utformingen av testen før man hadde sagt ja til å bli med i Generic Skills strand, det vil si man fikk først se hele testen i det man skulle velge ut de to test-oppgavene som skulle brukes og oversettelsen skulle begynne.

1.2.2 OECDs finansieringsplan for AHELO

Vanligvis finansieres OECD-prosjekter ved betaling fra medlemsland, blant annet finansieres både PISA og PIAAC på den måten. For AHELO hadde imidlertid OECD en annen plan, siden dette var en undersøkelse som primært kom læresteder, snarere enn land, til gode var det tenkt at det også kunne finnes andre som var interessert i å bidra til å finansiere AHELO. Dermed var den opprinnelige forretningsplanen for finansiering av AHELO basert på relativt beskjedne nasjonale bidrag, og at resten av finansiering skulle komme fra private aktører, fremfor alt gjennom ulike stiftelser eller veldedige organisasjoner (foundations) som har interesse i kvalitet i høyere utdanning (OECD 2008c). Imidlertid viste det seg allerede i 2008 at det var vanskelig å få inn midler fra private aktører. Dette ble delvis forklart med finanskrisen, som startet omtrent samtidig som AHELO ble lansert. I samband med finanskrisen i USA var det mange av de veldedige organisasjonene som vanligvis donerer penger til utdanningsforskning som mistet deler av sin kapital og dermed hadde de også mindre penger å dele ut. For å løse dette problemet ble det behov for å tenke i andre baner, og mange ulike initiativ ble prøvd ut. Men da heller ikke dette nådde frem ble det bedt om økte bidrag fra deltakerlandene og det ble åpning for at flere land, også land som ikke er med i OECD, kunne delta. Opprinnelig var Norge, Finland, Korea, Mexico, Kuwait og tre stater i USA de som hadde meldt interesse for å delta i AHELO Generic skills strand, men etter hvert ble også Colombia, Egypt og Slovakia med slik at man fikk totalt 9 land som deltok i Generic skills strand. Dette var nesten dobbelt så mange som foreslått i det opprinnelige for mulighetsstudien. De andre to modulene av AHELO (Economics og Engineering) ble

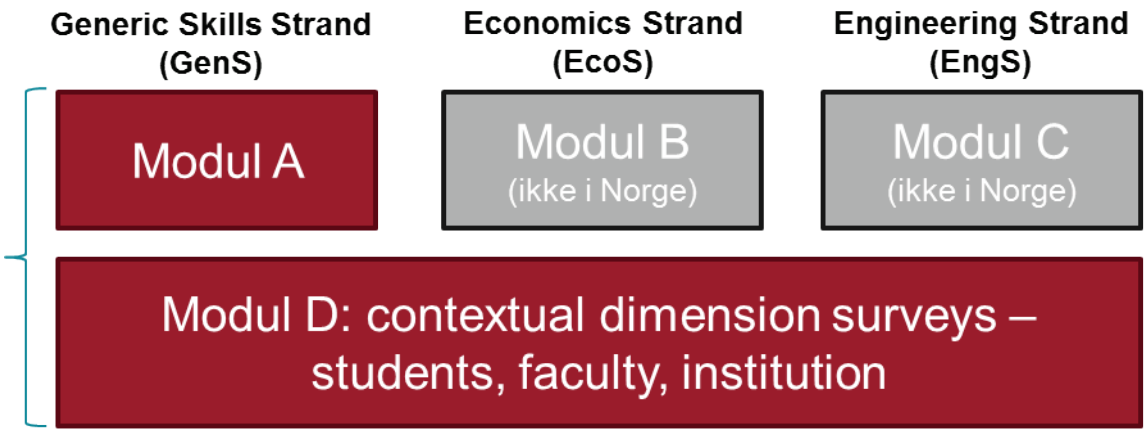
også utøket med flere land, og Mexico, Egypt og Slovakia deltok i alle tre modulene. I tillegg fikk man inn noe bidrag fra OECD-land som ikke deltok i prosjektet.

Utgangspunktet for AHELO-prosjektet var at omtrent halvparten av finansieringen skulle dekkes av private aktører, og resten av landene (OECD 2008c). Det endelige resultatet ble at 3 % av kostnadene i prosjektet ble dekket av OECD (oppstartmidler), 3,5 % ble dekket av bidrag fra land som ikke var med i studien, og 9,7 % ble dekket av bidrag fra private finansierer, i stor grad ulike stiftelser. Det resterende, 83,8 % ble dekket av nasjonale bidrag. Med andre ord ble AHELO i stor grad finansiert av bidrag fra deltakerlandene, slik som andre OECD prosjekter.

Det var flere negative konsekvenser av OECDs opprinnelig finansieringsplan for AHELO. For det første medførte den svært mye merarbeid for sekretariatet i OECD, som i liten grad ledet til bedre finansielle forutsetninger i prosjektet. For det andre ble også en belastning for deltakerlandene og GNE ved at videre finansiering av prosjektet ble et stadig tilbakevendende punkt på dagorden på alle GNE-møter. Vedvarende finansiell usikkerhet ga for det tredje også manglende oversikt over prosjektet som helhet og det gjorde i tillegg at det tok lang tid og mange forhandlinger for endelig avtale med konsortiet om gjennomføring av alle deler av AHELO kunne inngås. Den usikre finansielle situasjonen førte også til at prosjektet ble utsatt i et år, og beslutningen om utsettelse ble tatt forholdsvis sent i prosessen, i mai 2010. Prosjektet var opprinnelig tenkt i to faser, en forberedelsesfase og en gjennomføringsfase, men gjennomføringsfasen ble utsatt med et år, fra våren 2011 til våren 2012. Utsettelsen medførte at ytterligere merkostnader i prosjektet påløpte.

1.3 AHELO: et prosjekt med flere deler og faser

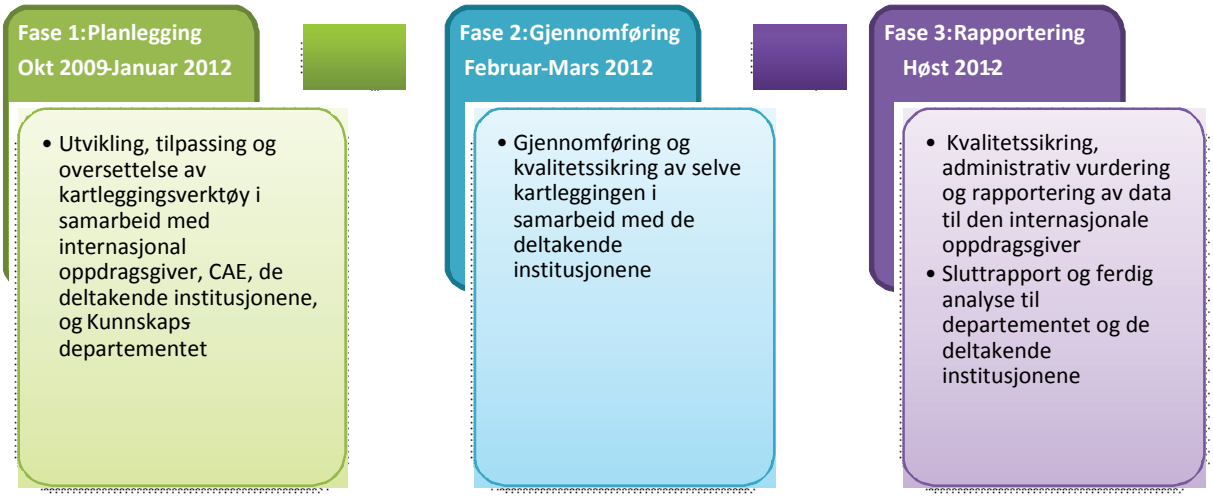
Allerede da AHELO ble lansert var det klart at AHELO mulighetsstudie skulle være et prosjekt med flere deler, eller «*strands*». Prosjektet skulle ha en *Generic skills strand* (modul A), og to disiplinere strands, en *Economics strand* (modul B) og en *Engineering strand* (modul C). I tillegg til de tre faglige delene skulle prosjektet ha en *kontekstuell dimensjon* for å sikre at informasjonen fra de faglige delene ble sett i sammenheng med kjennetegn ved den enkelte student, lærestedet og studieprogrammet studenten var en del av, lærestedets faglige profil, og en rekke andre kjennetegn ved lærestedet (modul D). Denne informasjonen skulle fremskaffes gjennom spørsmål til studenter, samt en egen spørreundersøkelse til et utvalg vitenskapelige ansatte ved hver enkelt institusjon, i tillegg til informasjon om lærestedet som helhet (størrelse, eieform, etc). Spørsmål om kontekstuelle forhold skulle alle studenter som deltok i undersøkelsen svare på, uavhengig av om de var med i Generic skills strand, Economics strand eller Engineering strand. Prosjektet moduler er presentert i figur 1.1. De røde feltene (modul A og D) illustrerer de delene av prosjektet der Norge deltok.



Figur 1.1: Hoveddeler i AHELO mulighetsstudie

I tillegg til modulene som fremkommer i figur 1.1 var det i den opprinnelige planen tenkt at man også skulle ha en «Value-added measurement strand», det vil si en modul som undersøker hva lærestedet bidrar med. Imidlertid ble det på et tidlig tidspunkt vedtatt at man ikke skulle undersøke dette i mulighetsstudien. For å kunne undersøke lærestedet bidrag må man ha minst to test-tidspunkt, siden man for å kunne si om de har lært noe mer gjennom å ta utdanning på lærestedet, må ha målt hva studentene kunne da de kom, og mulighetsstudien kun skulle ta utgangspunkt i ett test-tidspunkt. Derimot ble det helt i sluttfasen av prosjektperioden presentert en litteraturgjennomgang av ulike måter å måle lærestedets bidrag, som har vært brukt i grunnopplæringen og i høyere utdanning. Imidlertid konkluderer ikke litteraturgjennomgangen med at det finnes en god metode som kan brukes, snarere at alle modeller som skal måle lærestedets bidrag har svakheter ved at hvilken type modell som brukes og hvordan denne er spesifisert kan påvirke resultatet (OECD 2012b).

Prosjektet var opprinnelig delt inn i tre konkrete faser, slik det er illustrert i figur 1.2. Fase 1 består av utvikling, tilpassing og oversettelse av testverktøy til kartleggingen. Fase 2 består av selve gjennomføringen av kartleggingen ved de deltakende utdanningsinstitusjonene. Fase 3 består av kvalitetssikring av data, analyser og rapportering av resultater.



Figur 1.2: Opprinnelig prosjektplan for AHELO mulighetsstudie

Som figur 1.2 viser, utgjør Fase 1 den lengste fasen i tid. En stor del av AHELO mulighetsstudie har med andre ord vært knyttet til arbeidet med å oversette og tilpasse de ulike testinstrumentene og spørsmål knyttet til studentenes kontekstuelle forhold. Fase 1 ble imidlertid lengre enn opprinnelig planlagt, da gjennomføringen (fase 2) ble bestemt utsatt med ett år, fra 2011 til 2012. Utsettelsen kom som en følge av problemer med den internasjonale finansieringen av AHELO mulighetsstudie. Som beskrevet tidligere, gikk OECDs finansieringsplan ikke slik de hadde forventet. For å få tilstrekkelig finansiering til å kunne gjennomføre fase 2, ble derfor prosjektet utsatt med et år. Problemer med den internasjonale finansieringen medførte også at deler av fase 1, kom tett opptil fase 2. Dette kommer vi tilbake til i kapittel 3.

2 Beskrivelse av testinstrumentene

I dette kapitlet går vi igjennom de testinstrumentene som ble brukt i den delen av AHELO mulighetsstudien der Norge deltok: *Generic skill strand*.

I gjennomføringen av Generic skill strand ble det brukt to typer instrumenter for å måle læringsutbytte: case-oppgaver (constructed performance task) og et sett med flervalgsoppgaver (multiple choice). På ekspertmøtene som foregikk igangsettingen av mulighetsstudien ble det foreslått at man skulle bruke case-oppgaver som var basert på Collegiate Learning Assessment (CLA) utformet av Council for Aid to Education (CAE) (OECD 2007a, 2007b), mens det siste ekspertmøtet ikke anbefalte en viss test, da det ikke var enighet i ekspertgruppen om dette (OECD 2007c). Imidlertid ble CLA lagt frem før det første GNE-møtet som den anbefalte testen basert på ekspertmøtene, og bifalt som den testen som skulle brukes i Generic skills strand (OECD 2009a). Basert på anbefalinger fra ekspertmøtet var det også opprinnelig planlagt at mulighetsstudien skulle omfatte kortere testdel med flervalgsoppgaver, for å generere flere datapunkter (OECD 2007b). Imidlertid gikk man i en periode bort fra denne tanken, av økonomiske grunner. Men dette ble gjeninført etter ønske fra Technical Advisory Group (TAG) i forbindelse med at det ble vedtatt at gjennomføringsfasen skulle starte opp i juli 2011. Siden studien skulle gjennomføres våren 2012 hadde både konsortiet og landene kort tid til utvalg av flervalgsoppgaver og oversettelse. Arbeidet startet opp i august 2011. Konsortiet, ved ACER, valgte derfor å bruke oppgaver som var utformet i Australia, men som hadde et tilsnitt som også kunne fungere i andre land. Alle landene som var med i Generic skills strand fikk anledning til å uttale seg om et sett av oppgaver, og basert på den tilbakemeldingen valgte ACER ut et sett av kjerneoppgaver (core) som alle som tok testen skulle ta, samt fire sett med oppgaver som skulle rulleres (randomisert). Til sammen besto et testsett av 25 flervalgsoppgaver, 15 i kjernedelen og 10 i den rullerende delen.

En utfordring i arbeidet med testinstrumentene var at man i Generic skills strand ikke først hadde blitt enig om hva som egentlig ligger i begrepet «generic skills» og laget et felles rammeverk som sa hva man var ute etter å måle, men heller tatt utgangspunkt i at CLA oppgavene var sagt å måle kritisk tenking, problemløsing og skriveferdigheter, uten at det hadde blitt definert hva som lå i disse begrepene. Dette ble også kommentert av TAG ved flere tilfeller og derfor valgte konsortiet å utvikle et «Generic Skills Assessment Framework» samtidig som de satte sammen flervalgsoppgavene. Endelig versjon av rammeverket forelå høsten 2012 (OECD 2012c). Imidlertid var det ikke tid til å få en omforent internasjonal enighet om rammeverket, og det har derfor hatt liten videre betydning. Samtidig har rammeverket hatt den funksjonen at det har gjort det klart at man burte ha hatt et slikt rammeverk, som det var internasjonal enighet om, på plass før man bestemte seg for hvilke oppgaver man skulle velge ut. Dette er også påpekt både i konsortiets rapport til OECD og OECDs egen første delrapport fra prosjektet (OECD 2012c, Tremblay et al 2012).

2.1 Case-oppgaven

CLA er en case-oppgave der studenten blir konfrontert med en tenkt situasjon, der de med hjelp av informasjon som finnes i et digitalt dokumentbibliotek skal besvare en rekke oppgaver og gjøre vurderinger. De to oppgavene som er brukt i AHELO mulighetsstudien er «Catfish» og «Lake-to-river». I begge tilfeller tar det utgangspunkt i en tenkt situasjon, et case, der studenten skal gi råd til byens ordfører eller kommunestyre om hva som bør gjøres, basert på dokumenter. Catfish omhandler en deformert fisk som er funnet i et vassdrag nær en by. Studenten blir bedt om å vurdere argumenter for og imot ulike fortolkninger av hva som kan være årsaken til den deformerte fisken og skal til slutt gi sin egen vurdering av hvilken av de tre mulige årsakene som er den mest sannsynlige. Lake-to-river omhandler fordeler og ulemper ved å fjerne en eksisterende dam i kommunen. Studenten blir bedt om, basert på de tilgjengelige dokumentene, å komme med en oppsummering og en anbefalt strategi for hva kommunen bør gjøre med dammen. Felles for begge oppgavene er at studentene skal formulere et svar selv, det vil si dette er en form for langsvarsoppgave. Studenter som gjennomfører testen skal bare svare på en av de to oppgavene, og det er tilfeldig hvilken oppgave som kommer opp når de starter testen.

Testen rettes ved at en skårer vurderer svarene til studenten etter tre parametere: Analytisk resonnering og vurdering, problemløsning og argumenterende skriving, på en skala fra 1 til 6 der 1 er det laveste og 6 er det høyeste. Til grunn for vurderingen ligger en generell rettemanual, en «scoring rubric», som kan sies å være en form for generell sensorveiledning. I AHELO ble det bestemt at alle oppgaver skulle rettes av to skårere, slik at man kunne sammenligne de to skårene og derigjennom undersøke reliabiliteten mellom skårerne (interscorer reliabilitet).

2.2 Multiple choice spørsmål

Multiple choice, eller flervalgsoppgavene består av en rekke ulike typer spørsmål, som samlet er tenkt å vurdere ulike former for kritisk og analytisk tenkning og problemløsning. I følge et grunnlagsdokument, *OECD AHELO Feasibility Study Generic Skills MCQ description* (OECD ikke datert, mottatt september 2011), skal oppgavene som måler kritisk tenking eller analytisk tenking inneholde forståelse, analyse og evaluering av ideer og synspunkter, mens oppgavene som måler problemløsning inneholder prosessering og analyse av informasjon, og hvordan denne informasjonen siden kan brukes for å løse problemet.

Dokumentet spesifiserer hvilke aspekter av kritisk eller analytisk tenking den som tar testen skal vise.

In answering such items candidates may be required to:

- *comprehend and interpret arguments;*
- *recognise the way arguments and evidence relate to conclusions;*
- *analyse and assess dialectical strategies; and*
- *evaluate the strength and weakness of arguments.*

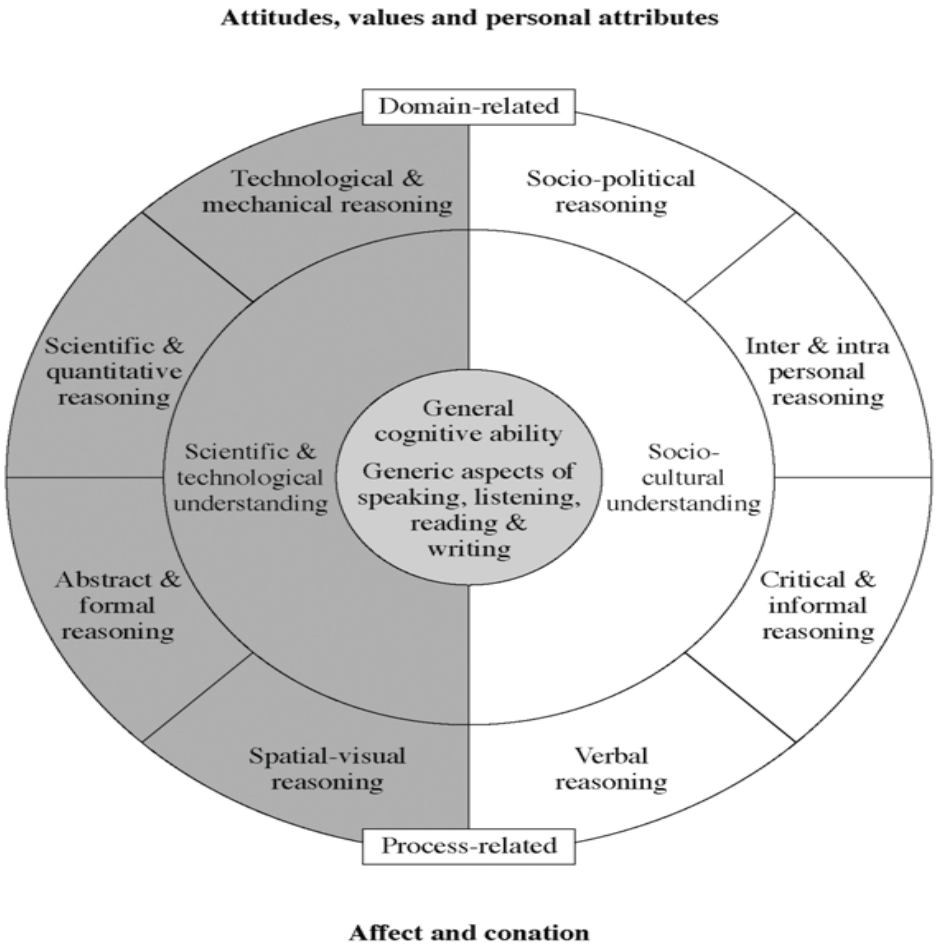
Videre spesifiserer også dokumentet hvilke aspekter av problemløsning den som tar testen skal demonstrere.

In answering such items candidates may be required to:

- *use data to identify relationships, constraints, rules and consequences;*
- *make inferences;*
- *classify, generalize and hypothesize; and*
- *make decisions about everyday situations and problems.*

I forbindelse med at ACER tok frem forslag til flervalgsoppgaver som kunne brukes i AHELO mulighetsstudien foreslo de også et General Skills Assessment Framework som kunne brukes som et bakteppe, for å forklare hva flervalgsoppgavene var ment å undersøke. Rammeverket er anskueliggjort i figur 1.3. Slik som rammeverket indikerer skal oppgavene dekke ulike typer

ferdigheter, både mer verbalt eller språklig orienterte oppgaver og mer matematisk eller logiske oppgaver. Samlet er det tekt at disse oppgavene måler kritisk tenking, analytisk tenking og problemløsning.



Figur 2.1 Rammeverk for måling av generiske ferdigheter ved bruk av flervalgsoppgaver i AHELO mulighetsstudie (utviklet av ACER). Kilde: McCurry 2004

Spørsmålene er utviklet av ACER i Australia, inspirert av liknende spørsmål brukt i andre undersøkelser der. Felles for alle oppgavene var at de består av en lengre introduksjonstekst (stimulus), noen ganger med et bilde, en figur/tabell eller formler og mellom 2 og 5 flervalgsoppgaver som hører sammen med hver introduksjonstekst.

2.3 Måling av kontekstuelle forhold

I tillegg til å samle inn test-data var det også et ønske om å kontrollere for kontekstuelle forhold, det vil si forskjeller mellom land, institusjoner og studenter. Dette for å gjøre det mulig og meningsfylt å sammenligne resultater på tvers av land, kultur og høyere utdanningssystemer. Det ble derfor utviklet tre spørreskjemaer som skulle vurdere kontekstuelle forhold: et for studenter som de tok sammen med testen, et for vitenskapelig ansatte og et spørreskjema til institusjonen, som institusjonskoordinatoren skulle ta seg av.

Studentskjemaet inneholdt 23 spørsmål, og dekket temaer som studentens studievaner og tidligere prestasjoner, hvilke typer undervisning og tilbakemelding de får, og om studiet har bidratt til å hjelpe dem å utvikle ferdigheter de tror de har bruk for i arbeidslivet. I tillegg er det en rekke demografiske

spørsmål (for eksempel: kjønn, alder, fødeland, språk i hjemmet, heltids/deltidsstudent, foreldres utdanning og yrke).

Skjemaet til vitenskapelig ansatte inneholder 15 spørsmål, og dekket temaer som undervisningskvalitet og forskningskvalitet ved lærestedet, vitenskapelig ansattes tidsbruk og deres mening om hvor mye tid studentene bør bruke på ulike aktiviteter, samt hvor stor andel av studentene som har svake forkunnskaper. Dessuten er også en rekke demografiske spørsmål (for eksempel: kjønn, alder, fødeland, språk i hjemmet, heltids/deltidsstilling) stilt til de vitenskapelig ansatte.

Institusjonsskjemaet er det lengste skjemaet, med 49 spørsmål om alt fra antall studenter og antall ansatte på heltid og deltid, hvor stor andel av budsjettet som kommer fra offentlige kilder, kvinneandel blant studentene, andel internasjonale studenter, fullføringsandel til hvor stor andel av studentene som er rekruttert fra eget fylke.

3 Fase 1: Planlegging

Dette kapitlet omhandler det vi har kalt fase 1 i AHELO mulighetsstudie i Norge. Denne delen inneholdt planlegging, oversettelse og tilpasning av de ulike testinstrumentene til norske forhold (se figur 1.2 kapittel 1). Her presenteres og diskuteres arbeidet som ble gjort i denne delen av prosjektet. Kapitlet begynner med en gjennomgang av prosessen med å velge ut, oversette og tilpasse de to typene faglige testinstrumenter som ble brukt i den generiske delen (*Generic skills strand*). Deretter presenteres arbeidet med å utvikle, oversette og tilpasse spørsmålene til den kontekstuelle delen av undersøkelsen – både spørsmålene til studenter og til de faglig ansatte ved de deltakende lærestedene. Kapitlet beskriver også prosessen med å melde prosjektet til Personvernombudet for forskning (NSD). Til slutt i kapitlet vurderes styrker og svakheter ved denne fasen.

NIFU fikk oppdraget å være National Project Manager for Norge i oktober 2009, men arbeidet startet opp tidlig i 2010. Siden OECD allerede hadde tatt stilling til hvilken type test som skulle brukes i Generic skills strand, besto oppstarten av AHELO for Norge i å være med og velge ut oppgaver som skulle brukes i gjennomføringen av testen. I februar 2010 møttes representanter for landene som er med i Generic skills strand i New York for å diskutere hvilke av CLA-oppgavene CAE hadde som skulle brukes. Landene hadde på forhånd ble presentert med ni ulike oppgaver, der de skulle enes om to oppgaver som skulle brukes i AHELO mulighetsstudien (OECD 2010a). De to oppgavene landene kunne enes om bar navene «Catfish» og «Lake-to-river». Oppgavetypen er kortfattet beskrevet i kapittel 2.

Den første fasen av AHELO var primært karakterisert av oversettelse og tilpasning av de to CLA-oppgavene og andre tekster som skulle brukes i testingen, testing av hvordan de oversatte oppgavene fungerer, samt møter og oppdateringer med Kunnskapsdepartementet og lærestedene.

3.1 Oversettelse og tilpasning av testinstrumenter

3.1.1 Oversettelse av case-oppgaven (CLA)

Arbeidet med oversettelse og språklig tilpasning av de to case-oppgavene, omfattet også flere tilliggende dokumenter. Dette inkluderer tekst knyttet til utforming av vurderingsskjema, tekst til gjennomføringen av kognitive laboratorier (utdypes i neste avsnitt), tekst til introduksjon til case-oppgaven – en såkalt 'mini-testoppgave', tekst til instruksjon for sensurering av testene ('skåringen'), samt tekst til instruksjon av testpersonene (stu dentene) i den nettbaserte utgaven av testen, tekst til instruksjon for sensurering av testene ('skåringen'). Oversikt over alle dokumenter fra CLA som er blitt oversatt finnes i vedlegg C.

Dette arbeidet startet januar 2010 og ble fullført vinteren 2011, og oversettelsene av dokumentene var i hovedsak gjort av EKVA /ILS og gjennomgått av NIFU. Som en del av oversettelsesarbeidet, fikk prosjektgruppen besøk av oversettelsesansvarlig fra CAE juni 2010, professor Willy Solano-Flores.

Han deltok på et todagers seminar der hele prosjektgruppen, både fra NIFU og EKVA /ILS, gjennomgikk oversettelse av de ulike tekstene, samt plan for gjennomføring av de kognitive laboratoriene inkludert instruksjonstekster og plan for utvalg av studenter (antall, kjennetegn, kjønnsfordeling, fordeling etter fagfelt, incentiver for deltakelse, etc.).

Kvalitativ testing - Cognitive labs

I tillegg til oversettelsen og tilpasning av ulike tekster fra amerikansk engelsk til norsk, har fase 1 inkludert småskala kvalitativ testing av hvordan de to case-oppgavene fungerer blant et utvalg norske studenter. Dette er gjennomført i alle deltakerlandene og har hatt til hensikt å avklare eventuelle misforståelser i teksten eller annet som kan være til hinder for en komparativ gjennomføring av CLA-testen i de ulike deltakerlandene. Gjennomføringen av denne kvalitative testingen – såkalte «*cognitive labs*» – er gjort av EKVA/ILS med et lite utvalg studenter fra Universitetet i Oslo og Høgskolen i Oslo.

Den kvalitative testingen av oppgavene ble gjennomført ved at studenter gjennomgikk hele test-oppgaven sammen med en forsker fra EKVA/ILS og samtidig ble bedt om å resonnere/fortelle høyt hva de tenkte mens de leste gjennom de ulike deloppgavene og prøvde å løse dem. Hensikten med denne fremgangsmåten er å avdekke hvordan studentene oppfatter de ulike oppgavene og tekstene – om oppgavene blir forstått i tråd med hensiktene eller om det er noe i teksten som studentene misforstår. Til sammen 10 studenter gjennomførte den kvalitative testen av oppgavene (cognitive labs). Gjennomføringen viste ingen direkte problemer med studentenes forståelse av tekster og oppgaver, men det var en tendens til at studenter med en realfaglig bakgrunn angrep oppgaven på en litt annen måte enn studenter med annen faglig bakgrunn. Erfaringen fra gjennomføringen er oppsummert i vedlegg B. Siden dette var en kvalitativ undersøkelse basert på få intervjuer kan vi ikke trekke klare konklusjoner om at studenter med ulik fagbakgrunn har ulike måter å forholde seg til case-oppgavene på. Samtidig indikerer den kvalitative testen at det kan være disiplinære forskjeller i hvordan studentene forholder seg til denne typen oppgaver, og dermed ble det et mål å prøve å undersøke dette i videre da dataene fra mulighetsstudien foreligger.

3.1.2 Utvelgelse og oversettelse av multiple choice spørsmål

I den opprinnelige planen for AHELO skulle man i Generic skills strand dels bruke en caseoppgave i tillegg til multiple choice spørsmål, men på grunn av dårlig økonomi i prosjektet ble reduksjoner vedtatt, blant annet å bruke to forskjellige case-oppgaver i stedet for tre og ikke å utvikle flervalgsoppgaver som skulle være i tillegg til case-oppgavene. Imidlertid anbefalte Technical Advisory Group at man igjen burde ta inn flervalgsoppgaver som en del av Generic skills testen, og dette rådet valgte konsortiet ledet av ACER å følge. Grunnen til dette var at to tester ville generere flere datapunkter, og dermed mer informasjon. Derfor gikk man i gang med å sette sammen flervalgsoppgaver fra en eksisterende database i Australia med oppgaver som skal måle kritisk tenking, analytisk tenking og problemløsing sommeren/høsten 2011.

Deltakende land fikk i september 2011 tilsendt multiple choice oppgaver og et grunnlagsdokument, *OECD AHELO Feasibility Study Generic Skills MCQ description* (OECD ikke datert, mottatt september 2011), og fikk mulighet til å komme med tilbakemelding på de foreslåtte oppgavene. Alle som skulle se oppgavene måtte signere på at de hadde taushetsplikt og at de ikke kunne dele oppgavene med noen utenfor prosjektgruppen. Prosjektteamet ved NIFU samarbeidet om å gi konsortiet skriftlig tilbakemelding på oppgavene innen den gitte tidsfristen (en uke). Da vi fikk oversettelsen av oppgavene til gjennomsyn viste det seg at oppgavesettet var ytterligere utvidet med flere oppgaver som ikke hadde vært til vurdering. Imidlertid har ACER lang erfaring med å utvikle slike oppgaver og alle oppgavene er hentet fra en database av eksisterende oppgaver som er testet ut og validert i Australia.

Oversettelsen av multiple choice spørsmålene foregikk høsten 2011, og det var cApStAn, et oversettelsesbyrå som er en del av konsortiet, som i hovedsak gjorde oversettelsen, og denne ble deretter gjennomgått og kvalitetssikret av NIFU. I de tilfeller der multiple choice spørsmålene tok utgangspunkt i en litterær tekst, som for eksempel Ibsen eller Tolstoj, ble NIFU bedt om å finne en passende versjon, eller foreliggende oversettelse som teksten kunne hentes fra. Dette var ikke lett i alle tilfeller, og for å sikre likhet i oversettelse bør man i fremtiden gi mer opplysninger om hvor originalteksten er hentet fra, slik at en passende oversettelse kan velges. Oversettelsen ble gjort med

det cApStAn kaller en ATAV-prosess, hvilket er et Excel ark der ulike versjoner av oversettelsen holdes opp mot hverandre til man har funnet en oversettelse man kan enes om er god nok. Med andre ord foregår oversettelsesarbeidet i flere omganger, noe som er arbeidskrevende men som likevel forløp forholdsvis raskt, primært på grunn av korte tidsfrister. Dermed ble denne oversettelsen gjort under sterkt tidspress, siden arbeidet med å tilpasse multiple choice spørsmålene startet opp i september 2011.

Som prosess fungerte oversettelsen av multiple choice spørsmålene bra, ved at cApStAn var lydhørt for innspill og kommentarer fra det norske prosjektteamet. Samtidig var oversettelsen av multiple choice spørsmålene forholdsvis arbeidskrevende, og det var krevende å finne gode utgangstekster som oversettelsen kunne ta utgangspunkt i. For å sikre mer lik oversettelse av litterærere tekster i eventuelle fremtidige tester bør den som skal gjøre oversettelsen få informasjon om hvilken utgave av teksten som er opprinnelig brukt og dersom det er mulig og så informasjon om hvilken oversettelse som bør ligge til grunn. Slike oppgaver stiller dermed ganske store krav til informasjon om hva oppgaven er bygget på.

3.1.3 Oversettelse av kontekstinstrumentene

Ansvaret for utviklingen av den kontekstuelle dimensjonen ble lagt til det internasjonale konsortiet under ledelse av ACER. Dette arbeidet foregikk i løpet av 2011, først med å foreslå og å komme til enighet om passende spørsmål, og deretter med oversettelse av de tre kontekstinstrumentene.

Disse oversettelsene ble også gjort av cApStAn, etter samme modell som oversettelsen av multiple choice spørsmålene, det vil si ved bruk av en ATAV-prosess. Men oversettelsen var omfattende, siden den innehold nesten 90 spørsmål til de tre gruppene, og ikke alle måter å måle for eksempel fullføringsrate på var satt opp slik den pleier å måles i Norge. Men gjennom ATAV-prosessen kom vi til enighet om oversettelser som kunne fungere.

3.2 Pre-implementeringsfasen

CAE startet det de kaller 'pre-implementeringsfasen' av prosjektet vinteren 2011, som pågikk utover våren 2011. I denne fasen ble også ulike dokumenter oversatt fra amerikansk engelsk til norsk, blant annet skjermbilder og en manual, som skulle brukes av de som praktisk skulle gjennomføre studien («proctor manual»). Med andre ord brukte man fasen til å forberede gjennomføringen av testen, under antakelsen at man ville komme til å bruke CAEs tekniske plattform. Se vedlegg C for en oversikt over dokumenter tilsendt fra CAE til gjennomgang/oversettelse i denne fasen.

Oversettelsesarbeidet i denne fasen ble gjennomført av CAE ved hjelp fra eksterne oversettere i USA. Imidlertid var arbeidet av varierende kvalitet og innebar et betydelig arbeid i form av kvalitetssikring fra EKVA/ILS og NIFU. Manualen for gjennomføring av testen inneholdt en tekst som testansvarlig skulle lese opp for studentene som tok testen. Denne var dårlig oversatt, og NIFU så seg nødt til å søke råd hos KD om ny oversettelse skulle gjennomføres. KD ga klartegn til å bestille en ny oversettelse, og da av hele manualen for teknisk gjennomføring av testen («proctor manual»). Dette medførte at man måtte bestille en ny oversettelse av manualen fra et norsk oversettelsesbyrå siden den opprinnelige oversettelsen ikke holdt mål.

Få eller ingen av dokumentene som ble oversatt i denne fasen ble brukt i selve testingen. Grunnen til dette var at konsortiet ledet av ACER utviklet en egen teknisk plattform som skulle brukes høsten 2011, og innloggingsdelen fra den tekniske plattformen CAE brukte i sin test ble derfor ikke brukt i gjennomføringen av AHELO, kun databildene som inneholdt selve testen, og som tidligere hadde blitt oversatt av EKVA /ILS. Dermed viste det seg at mye av oversettelsene som ble gjort i denne fasen var unødvendige, siden de ikke ble brukt i gjennomføringsfasen. Samtidig var ikke dette noe som kunne ha vært forutsett da ny oversettelse ble bestilt, siden beslutningen om å bruke en annen teknisk plattform ble tatt flere måneder senere.

Denne fasen besto også i å teste ut den tekniske løsningen som CAE hadde utviklet, også det til ingen nytte siden heller ikke den ble brukt i gjennomføringsfasen. CAEs tekniske løsning besto, i tillegg til selve testsystemet (interface) også av en sikker nettleser som skulle lastes ned til maskinen for å sikre

at studentene ikke kunne bruke andre ressurser enn de som var gitt i oppgaven. Dermed var dette en type nettleser som «låste» maskinen for bruk av andre nettsider enn den som skulle brukes i testsituasjonen.

3.3 Melding av prosjektet til Personvernombudet for forskning

I følge norsk lov skal all datainnsamling som omfatter informasjon om enkeltindivider meldes inn til Personvernombudet for forskning. AHELO mulighetsstudie omfattes av dette regelverket. Imidlertid er det sider ved prosjektet som har komplisert dette. Ved registrering av prosjekt til Personvernombudet skal all informasjon om undersøkelsen foreligge, inklusive utvalgsplan og spørreskjema/ testinstrumenter. I et internasjonalt prosjekt av typen AHELO ligger ansvaret for å utvikle spørreskjema/testinstrumenter utenfor den nasjonale prosjektlederens kontroll, og det nasjonale prosjektteamet har dermed liten påvirkning på når utvalgsplan, spørreskjema og testinstrumenter foreligger. Dette gjorde det ikke mulig å melde inn prosjektet før forholdsvis sent i prosessen, først høsten 2011, og endelig vedtak om godkjenning fra Personvernombudet for forskning kom 11. juni 2012.

3.4 Vurdering av planleggingsfasen (fase 1)

Som beskrevet i dette kapitlet har den første fasen av AHELO mulighetsstudie egentlig bestått av flere deler. I hovedsak kan man skille mellom to deler – en første del som primært har omhandlet oversettelse og tilpasning av case-oppgaven, og en andre del som har omfattet oversettelse og tilpasning av flervalgsoppgaver (multiple choice), og oversettelse og tilpasning av kontekstuelle spørreskjemaer til studenter, faglig ansatte og læresteder.

Når det gjelder tidsbruk skiller de to delene seg veldig fra hverandre. Den første delen hadde en forholdsvis romslig tidsplan og dertil grundig prosess med oversettelse og tilpasning av case-oppgavene. Andre del hadde forholdsvis kort tid til oversettelse av flervalgsoppgaver og spørreskjemaer. Samtidig hadde konsortiet etablert gode rutiner for oversettelse og tilpasning som ble brukt i den andre delen og dette bidro til å gjøre oversettelsesarbeidet av flervalgsoppgaver og spørreskjemaer enklere. Siden dette var en mulighetsstudie var man ikke primært opptatt av å perfeksjonere oversettelsene og dermed fungerte det å gjøre oversettelsene av flervalgsoppgaver og spørreskjemaer forholdsvis raskt.

Selv om oversettelsen av spørreskjema til institusjoner gikk forholdsvis bra viste gjennomføringen at det ikke var mulig for lærestedene lett å finne de indikatorer som var brukt. Med andre ord hadde man, sannsynligvis på grunn av liten tid, ikke gjort godt nok arbeid med å finne indikatorer på lærestedsnivå som passer i alle land, og man hentet i begrenset grad inspirasjon fra andre studier, som for eksempel EUs U-Map/U-Multirank, til å definere hvordan indikatorer skulle rapporteres. Tilbakemeldingen fra lærestedene var dermed at de ikke klarte å svare på alle spørsmål i institusjonssurveyen, fordi de ikke har data rapportert på den måten det blir spurt om. Her finnes det med andre ord et forbedringspotensial dersom AHELO skal gjennomføres i fullskala. Kontekstinformasjon, særlig om lærestedet, er svært viktig for å kunne gjøre sammenligninger på tvers av lærested.

At de ulike oppgavene/delene i fase 1 skiller seg fra hverandre med hensyn til prosess og tidsbruk må ses i sammenheng med den internasjonale prosjektledelsen og ansvarsdelingen i de ulike delene av prosjektet. Mens CAE har hatt ansvaret for case-oppgavene, har konsortiet ledet av ACER hatt ansvaret for multiple choice-oppgavene og de tre kontekstinstrumentene. Konsortiet ledet av ACER kom i gang med sin del av arbeidet på et betydelig senere tidspunkt enn CAE på grunn av med OECDs finansieringssituasjon (se kapittel 1). Dette var også knyttet til det faktum at OECD inngikk kontrakt med CAE om å bruke deres testinstrument allerede før konsortiet ledet av ACER fikk i oppdrag å gjennomføre de andre delene av AHELO. Sett i etterpåklokskapens lys var det å binde seg til et allerede etablert testinstrument før prosjektet egentlig hadde startet opp ikke særlig klokt av OECD, og det er i dag bred enighet om at man burte ha startet ut med et felles rammeverk for «generic skills» og deretter utarbeidet tester som kunne brukes for å måle det man ønsket å måle.

For Norges del (i likhet med de andre deltakende landene) har de to kontraktspartnerne i Generic skills strand, og tidspunktene da deres arbeid i prosjektet begynte gjort at ulike deler av prosjektet har fått totalt ulikt tempo. Den korte tiden for oversettelser av de delene konsortiet ledet av ACER hadde ansvar for (flervalgsoppgaver og kontektsspørreskjemaer) kunne ha blitt et problem dersom ikke oversettelsene ble gode nok. Men siden man hadde etablert gode rutiner for oversettelse innad i konsortiet, gjennom cApStAns ATAV-prosess, ble ikke dette noe stort problem.

I fremtiden bør det settes av tilstrekkelig tid til alle oversettelser, for å sikre kvaliteten i arbeidet. Samtidig fungerte de etablerte ATAV-prosessene, som også brukes i andre store OECD prosjekter (som PISA), godt som et oversiktlig instrument for å gjøre oversettelser. Oversettelse og tilpasning av instrumenter er svært viktig for at dataene som kommer ut av testen skal bli sammenlignbare.

Fase 1 av prosjektet var, med unntak av oversettelser, primært karakterisert av møter, både på internasjonalt plan og nasjonalt plan (oversikt over møter NPM har deltatt i finnes i vedlegg D). Samarbeidet mellom de deltakende parter i Norge har fungert godt, KD har innkalt til jevnlige prosjektmøter og det har bidratt til god kommunikasjon i prosjektet. NPM har blitt inkludert i GNE-møtene i OECD, noe som har vært bra for å få oversikt over prosjektet og for å sikre gode samarbeidsrelasjoner med andre land som har deltatt i Generic skills strand. Norge var et av de første landene som oppnevnte nasjonal prosjektleder, og selv om det var separate møter for alle lands nasjonale prosjektledere i samband med GNE-møtene, var det likevel verdifullt å være tilstede på begge to.

Samarbeidet med CAE har stort sett fungert godt, særlig i forhold til oversettelsen av case-oppgavene. Her hadde man en profesjonell oversettelse og en svært omfattende prosess med kvalitetssikring av den oversettelsen. Imidlertid var ikke oversettelsen av skjermbilder og manualer like god, og den åpnet heller ikke for kvalitetssikring fra norsk side. Konsekvensen av dette ar at noen av oversettelsene måtte gjøres på nytt, i Norge, og dette medførte merkostnader.

Den største utfordringen i fase 1 av prosjektet lå i det internasjonale samarbeidet, ved at prosjektet ble endret og utsatt fordi OECD manglet finansiering. Utsettelsen medførte at prosjektet nesten gikk i dvale og man kunne dermed ha risikert å miste læresteder som man allerede hadde rekruttert. Norge gjorde ikke dette, men flere andre land opplevde at læresteder som hadde sagt seg interessert i å delta valgte å ikke være med da studien ble gjennomført.

4 Fase 2: Gjennomføringsfasen

I dette kapitlet presenteres selve gjennomføringen av AHELO mulighetsstudie i Norge. Implementeringen av AHELO begynte høsten 2011, og i denne fasen av studien inngikk trekking av utvalg studenter og vitenskapelig ansatte, teknisk testing av systemene som skulle brukes og gjennomføringen av selve testingen, samt gjennomføring av survey (kontekstinstrumentene). Til slutt i kapitlet diskuteres ulike sider ved gjennomføringen av AHELO mulighetsstudie.

4.1 Trekking av utvalg

Den første oppgaven relatert til gjennomføringen av AHELO var å trekke utvalg (*sampling*). Utvalgsprosessen besto først av å lage lister over de to populasjonene som skulle undersøkes: studenter og vitenskapelig ansatte, deretter å trekke ut et tilfeldig utvalg fra hver av de to listene.

Etter et møte i prosjektgruppen der både læresteder og KD var til stede i begynnelsen av desember 2011 ble det enighet om å bruke Felles studentsystem, FS, for å plukke ut en målpopulasjon av studenter for AHELO.. Etter noen runder med avgrensninger av hvem dataene skulle omfatte, var dataene ferdige for avlevering fra FS 10. januar 2012, og selve utvalgstrekkingen ble gjort 12. januar 2012.

Kriteriene for hvordan utvalget skulle avgrenses var i utgangspunktet satt av konsortiet, og beskrevet i dokumentet *Sampling manual* (OECD 2011a). I følge denne manualen skulle bachelorstudenter i sitt tredje studieår trekkes ut. Dette kan virke som en enkel definisjon, men viste seg å kreve en rekke tilpasninger for å få avgrenset og bestemt populasjonen av studenter. I utgangspunktet skulle kun bachelorstudenter være med i utvalget, men siden det i Norge finnes en del integrerte masterprogrammer, som indirekte består av et sammensatt bachelor- og masterprogram, for eksempel sivilingeniør/master i teknologi, ble også slike programmer inkludert. Argumentet for også å inkludere integrerte masterprogrammer var at studentene formelt sett kunne ha tatt en bachelor i ingeniørfag for så å bygge på med en master for å bli sivilingeniør. Skille mellom bachelor- og masterprogram for denne gruppen studenter er dermed uklart.[2]

Å isolere gruppen «studenter i sitt 3. studieår» - underforstått 'studenter som nærmer seg fullført bachelorstudium' viste seg å by på ytterligere utfordringer, ettersom en betydelig andel av studentene i Norge er forsinket i studiene (eventuelt har hatt avbrudd fra studiene eller ikke fulgt normert studieprogresjon gjennom hele studieløpet). Dermed kan studentene være i sitt tredje studieår, men likevel ikke være nær fullføring, fordi de ikke har tilstrekkelig med poeng. Som en tilpasning til dette,

[2] NPM fikk medhold i konsortiet for også å la studenter som gikk på integrerte masterprogrammer (slik som sivilingeniørutdanning) være en del av målpopulasjonen i Norge. Ett tilleggs-argument for å inkludere studenter på integrerte masterutdanninger i populasjonen var at dersom det ikke ble gjort ville over halvparten (55 prosent) av studentene i målgruppen (tredjeårsstudenter) ved NTNU være utenfor og det samme ville ha vært tilfelle for 39 prosent av studentene i målgruppen ved UMB.

valgte vi å definere studentpopulasjonen som samtlige studenter med fullført minst 100 studiepoeng men ikke mer enn 150 studiepoeng innenfor programmet de gikk på per 15.8.2011. En slik definisjon omfattet også studenter på integrerte masterprogrammer, siden de som har mellom 100 og 150 studiepoeng ved studiestart kan antas å være i sitt tredje studieår.

I utvalgsmanualen var det ikke lagt opp til at det var en kobling mellom student og vitenskapelig ansatt, selv om informasjonen fra de vitenskapelige ansatte var sagt å skulle fungere som kontekstinformasjon til studien. Dette ble påpekt av NPM til konsortiet allerede tidlig i prosessen, men siden man fra konsortiets side, med tilslutning fra OECD, hadde valgt et opplegg der det ikke skulle være noen kobling mellom student og vitenskapelig ansatt var det ingenting man kunne gjøre med det. Vi var derfor klar over allerede da uttrekket ble gjort at svarene fra de vitenskapelig ansatte sannsynligvis skulle bidra med lite informasjon inn mot undersøkelsen. Det er nytteløst å spørre vitenskapelig ansatte generelt om nivået på deres innkommende studenter, dersom man ikke kan koble det til studentene på det aktuelle faget eller studieprogrammet. Data fra surveyen til vitenskapelig ansatte er heller ikke inkludert i datafilen fra studien, slik at dette nok er et forhold som gjelder på tvers av land.

Uttrekket av vitenskapelig ansatte viste seg vesentlig mer komplisert å gjøre enn uttrekket av studenter. For det første var spesifikasjonene i *Sampling manual* (OECD 2011a) svært spesifisert for vitenskapelig ansatte, om hvem som skulle inkluderes og ikke, og for det andre er dette ikke registre som lærestedene har lett tilgjengelige. Dermed tok det en del tid før NIFU fikk lister over vitenskapelig ansatte fra alle de fem deltakende lærestedene. På samme måte som for studentene forelå det en spesifikasjon for hvilke vitenskapelig ansatte som skulle regnes med i populasjonen, men siden registrene på de fem lærestedene ikke er helt like var det ikke alltid mulig å gjøre helt like avgrensninger og det ble derfor foretatt noen tilpasninger. Men alle lærestedene tok utgangspunkt i vitenskapelig ansatte som tilhørte de instituttene som har ansvar for programmene studentene var trukket ut fra. En utfordring her er at vitenskapelig ansatte hører hjemme på et institutt, mens studenter er tilordnet studieprogrammer. Dermed vil det kunne være problemer med å matche studenter på et gitt studieprogram med de vitenskapelige ansatte som underviser akkurat de studentene. Dessuten leverte alle lærestedene svært varierende informasjon om de vitenskapelig ansatte, noen læresteder inkluderte bare informasjon om institutt og fakultet personen tilhørte, mens andre også inkluderte stilling og andre variabler. Hva som ble inkludert i filen avhengte i stor grad av hva som fantes i lærestedenes registre, men siden det likevel ikke var noen kobling til studentene spilte det ikke veldig stor rolle. Utvalget ble trukket 29. mars 2012 og godkjent av konsortiet 2. april 2012.

4.2 Teknisk testing av systemene

Før undersøkelsen skulle gjennomføres ble det gjort en teknisk testing av testplattformen. Den tekniske testingen har foregått i to faser, først som en del av pre-implementerings-fasen av case-oppgaven (CLA), og siden som en forberedelse til gjennomføringen av studien. Testingen av den tekniske løsningen for CLA-oppgaven besto i at NPM fikk tilsendt innloggingsinformasjon sommeren 2011 og kunne gå inn for å se at det fungerte. CLA-oppgavens tekniske løsning var konstruert slik at den som tar testen laster ned en sikker nettleser, som låser datamaskinen slik at man bare kunne være på testen, det vil si ikke bruke internett generelt. Imidlertid kom ACER, i samarbeid med resten av konsortiet frem til at de heller ville bruke en annen plattform som omfattet hele testen, slik at studenter som skulle ta testen bare trengte å logge inn et sted. Den nye plattformen, som kom i løpet av høsten 2011, inneholdt ikke en sikker nettleser, men krevde i stedet at datamaskinene som skulle brukes for testen ble satt opp slik at de bare kunne gå til en gitt nettside, testsiden, og at alle andre sider på internett var blokkert. Den tekniske testingen gikk dermed i stor grad ut på å undersøke hvor mye arbeid som fulgte med dette, og hvordan plattformen forholdt seg til lærestedenes brannmurer og andre tekniske ting koblet til testoppsettet.

Selve den tekniske testingen, som forberedelse til gjennomføring av testen, ble ikke gjort før 25. januar 2012, fordi det viste seg komplisert å få integrert hovedplattformen for undersøkelsen med den plattformen som CLA-oppgaven var i. Opprinnelig skulle denne testingen har foregått for jul 2011, eller tidlig i januar 2012. Dermed fikk lærestedene forholdsvis kort tid til å gjennomføre testene, men de erfarte relativt få problemer. Derimot var det en del av lærestedene som støtte på problemer da testen

skulle gjennomføres, blant annet at det tok svært lang tid å forberede maskinene som skulle brukes ved gjennomføringen og at dette ikke kan gjøres lenge på forhånd – siden maskinene da ikke kan brukes på internett i det hele tatt.

En annen utfordring ved de tekniske testene var at lærestedene ikke fikk anledning til å se testen studentene skulle ta. Alt materiale som brukes i testen, både case-oppgaven og multiple choice spørsmålene er rettighetsbelagt og eies av CAE respektive ACER. Nasjonal prosjektleder og resten av prosjektteamet har måtte signere på taushetserklæringer for å kunne se testene, og det var ikke lagt opp til at lærestedsansvarlig eller testansvarlig skulle signere på slike taushetserklæringer. Dermed ga de tekniske testene bare tilgang til nettsiden der testen var plassert, men ikke anledning til å se igjennom hva testen ville bestå i. Dette ledet til mye frustrasjon blant de lærestedsansvarlige, siden de mente at mer informasjon om hva testen besto i nok hadde gjort dem bedre rustet til å informere studentene om hva testen gikk ut på.

4.3 Gjennomføring av testing

Norge var først ute med å gjennomføre AHELO, selv om oppstart ble noe forskjøvet på grunn av at vi ikke hadde mottatt innloggingsdetaljer i tide. Alle lærestedene gjennomførte testingen i løpet av februar og mars måned, det vil si før påske. Tabell 4.1. X viser hvor mange studenter som tok testen ved vært lærested, når lærestedene gjennomførte testene, hvor lang de ulike lærestedenes testperioder var og hvilket insentiv de brukte for å rekruttere studenter til å delta.

Tabell 4.1 Antall besvarelser fra deltakende institusjoner, testperiode og insentiv

	Tatt testen	Testperiode	Insentiv	Kontaktmetode
UMB	14	To dager, 8.-9. mars	Utlodding av Ipad2	Epost, sms
NTNU	30	Uke 10, 5.-9. mars	Gavekort kr 250	Epost, sms
HiL	29	Uke 6, 6.-10. februar	Utlodding av Ipad2 + 5 konsertbilletter	Epost, sms
UiS	25	Flere sesjoner i løpet av februar og mars	Kinobillett kr 200	Epost, telefon, sms
HiVe	17	Flere sesjoner i løpet av mars måned	Utlodding av gave-kort kr 500 x 10	Epost, sms

Lærestedene valgte litt forskjellige løsninger for hvordan de gjennomførte undersøkelsen. Universitetet i Stavanger og Høgskolen i Vestfold valgte å kjøre flere test-sesjoner i uken i flere uker, mens Høgskolen i Lillehammer og NTNU konsentrerte testingen til en gitt uke, med flere sesjoner per dag. UMB valgte å bare ha to testdager, men med fire ulike sesjoner de to dagene. Høgskolen i Lillehammer hadde opprinnelig planlagt å ha to uker med testing men fordi vi ikke fikk innloggingsdetaljer i tide ble oppstart utsatt og de måtte avlyse en rekke test-sesjoner som var satt opp tidlig i februar.

Selv om lærestedene valgte ulike former for insentiver og ulike lange testperioder er det ikke noe som indikerer at en løsning har fungert bedre enn en annen – ved alle lærested var det få studenter som ønsket å delta i AHELO. Svarprosenten varierte fra 5 % til 10 % og gjennomsnittlig svarprosent er 7,67 %. Med en slik svak svarprosent kan ikke dataene gi noe tilbakemelding til lærestedene på lærestedsnivå. Lærdommen er at det er svært vanskelig å rekruttere studenter til å delta i en undersøkelse som AHELO. Det er ikke bare Norge som har svak svarprosent, også Finland og USA erfarte at det var vanskelig å rekruttere studenter til å delta, og det var også enkelte land i de andre modulene som slet med å rekruttere deltakere. Det tar omtrent 2,5 timer å ta testen i Generic skills strand og dette har sannsynligvis vært en medvirkende årsak til at så få studenter har ønsket å delta. Kostnadene i form av tid sammenlignet med hva studenten får ut av det (som i mulighetsstudien var relativt moderate insentiver) oppfattes dermed som for høye. Dessuten viste det seg at mange av studentene som ble trukket ut til å delta var i en fase av studiet der de skriver individuelle bacheloroppgaver, noe som gjorde at de sjelden var på campus og det dermed også være en ekstra kostnad knyttet til å møte opp i test-lokalet. Dillman (2000) argumenterer for at deltakelse i

spørreskjemaundersøkelser kan ses i lys av bytte-teori, ved at folk vurderer kostnader og belønninger knyttet til å delta i studien. Selvsagt må den oppfattede belønningen for å delta være større enn kostnaden, eller respondenten må i hvert fall ha tillit til at belønningen på sikt vil være større enn kostnaden ved å delta. Da testen ble utformet og man vedtok å bruke en forholdsvis lang test, gjorde man nok ikke like grundig vurdering av hvilke implikasjoner dette ville kunne ha for den opplevde kostandene ved å delta for studenter.

Den svake svarprosenten kommer ikke av at lærestedene ikke arbeidet hardt for å rekruttere studenter til å delta i AHELO, de lærestedsansvarlige gjorde mye for å nå ut til studentene som var trukket ut. Alle lærestedene brukte epost for å informere de som var trukket ut til å delta i undersøkelsen og oppfordre dem til å delta, og flere steder sendte de ut purre-eposter for å få flere til å melde seg på. NIFU har bistått lærestedene med å utforme en powerpoint-presentasjon om AHELO, som har vært brukt til å informere lærestedets ledelse, studentorganisasjoner og studenter. Alle lærestedene laget nyhetssaker om AHELO som lå på nettsidene, på læringsplattformer og/eller elektroniske informasjonstavler, og en av lærestedene laget også postere/flyer for å informere om undersøkelsen. I tillegg var det flere av lærestedene som gikk ut i klasser eller kurs for å prøve å rekruttere deltakere, men dette har sannsynligvis ikke hatt veldig stor effekt siden det kun er få studenter på hvert kurs som var trukket til å delta. Ved et lærested ansatte de en studentassistent til å ringe alle studenter som hadde fått eposten for å minne dem på at de skulle melde seg på. Studenter som ikke møtte til avtalt tid ble også oppringt ved flere av de andre lærestedene, men også det ga lite uttelling på svarprosenten.

UMB var det lærestedet som brukte minst ressurser på rekruttering og testing. For å rekruttere sendte de ut en epost og de hadde oppslag på nettet og elektroniske informasjonstavler, og de gjennomførte testing bare to dager. Da NPM spurte om det var mulig at de kunne tenke seg å bruke ytterligere tid på å få flere til å ta testen var tilbakemeldingen at de allerede hadde trukket premien og dermed ikke ville gjennomføre flere tester. I tillegg ble det fastslått at kostnaden for å få inn flere responser bare ville øke kostnaden for deltakelse i AHELO for lærestedet, og det ønsket ikke lærestedet.

Teknisk fungerte gjennomføringen ganske godt, det var kun få rapporter om problemer med å gjennomføre testen. Noen få hadde innloggingsproblemer, men det var ofte knyttet til oppsettet på maskinen og kunne løses ved å bytte til en annen maskin. UMB opplevde at en student ble kastet ut av testen underveis, og derfor mangler all informasjon fra spørreskjema for denne testtakeren. Det samme gjelder for fem studenter ved UiS, som fikk fullført testen men ikke levert surveyen. Institusjonsansvarlig ved UiS antar at problemene ved UiS er knyttet til en teknisk skrivefeil i en URL, som ikke var mulig å oppdage ved hjelp av testrutinene på forhånd. Men med unntak av noen få problemer fungerte den tekniske gjennomføringen av testen forbausende bra, fremfor alt med tanke på at den tekniske testingen hadde foregått kun kort tid før gjennomføringen startet og lærestedene ikke hadde hatt tilgang på selve testen, og dermed ikke kunnet sjekke at hele systemet fungerte som det skulle.

4.4 Skåringen av testen

Skåringen av testen ble gjennomført av EKVA /ILS, med rådgiver Anna Eriksen somhovedansvarlig (lead scorer). Skåringen ble gjennomført i perioden 1.-25. mai 2012, av Lead Scorer sammen med to andre skårere. Siden det var kjent da skåringen skulle starte at det var forholds få besvarelser gikk man bort fra den opprinnelige planen om å ansette og lære opp masterstudenter til å gjøre skåringen og brukte heller to ansatte på EKVA som har tidligere erfaring med å vurdere og å skåre PISA-oppgaver. De to skårerne fikk opplæring etter samme mønster som Lead scorer treningen i Paris hadde hatt, dvs ved gjennomgang av eksempeloppgaver. Siden det var få besvarelser i Norge måtte noen av de eksiterende oppgavene brukes som mønsteroppgaver for å finne de ulike skåringsnivåene (1-6). Begge skårere fikk opplæring i begge oppgaver og alle oppgaver ble skåret to ganger. Skåringen ble gjort på papir siden det var enklest, selv om det også finnes et elektronisk system for skåringen. Alle resultatene ble i stedet lagt inn i det elektroniske systemet på samme dag da skåringen var ferdig. Lead scorer gikk igjennom og godkjente alle skårene som ble lagt inn i datasystemet, og over halvparten av besvarelsene ble også kontrollert enkeltvis. En av oppgavene til Lead scorer er å finne avvik på mer enn et ett skåringsnivå, og da se til at den oppgaven blir skåret på ny. Imidlertid

viste det seg at Lead scorer ikke kunne sende oppgaven inn i køen igjen siden en skårer bare kan skåre en oppgave en gang, og begge skårere hadde satt en skåre på oppgaven. Man måtte derfor heller velge «rescore», som innebar at Lead scorer satte ny skår på besvarelser som i første omgang ble sendt tilbake i systemet for ny vurdering. Med andre ord ser det ut til at man egentlig hadde behov for flere skårere, for å ta seg av tilfeller der det var avvik mellom skåren som de to skårerne hadde gitt. Imidlertid fremgikk det ikke av instruksene hvor mange skårere som trengs, og problemet ble løst ved at Lead scorer da skåret oppgaven en tredje gang og overprøvde en av skårene.

Lead scorer kan ikke, basert på de få norske besvarelsene, si noe om nivået på de norske studentene, men kommer med en refleksjon om hvordan studentenes svar fremstår. I den ene case-oppgaven er studentene bedt om å komme med argumenter som støtter en av tre ulike tolkinger av hvorfor det finnes en deformert fisk i dammen og en del studentene velger å besvare denne oppgaven med forholdsvis korte svar, siden det er totalt seks korte spørsmål som spør om argumenter før respektive mot de ulike tolkingene. Med andre ord kan det se ut til at måten oppgaven er formulert på kan bidra til å influere hvordan studenten velger å besvare oppgaven, gjennom at studentene gir korte svare på kort formulerte oppgaver, selv om instruksen sier at svaret «bør beskrive alle detaljer som er nødvendige for å begrunne ditt standpunkt». Med andre ord kan dette delvis henge sammen med hvordan oppgaven har blitt oversatt, eller at litt ulike oversettelse gir ulike signaler for hvordan svaret bør skrives.

4.5 Lærestedenes tilbakemeldinger på prosessen

Lærestedene var primært involvert i fase 2 av AHELO, selv om det også i den første fasen var oppdateringsmøter mellom KD, NIFU og lærestedene. I gjennomføringsfasen derimot hadde NPM utstrakt kontakt med lærestedene, både med hensyn til trekking av utvalg, for å få sammenstilt hvilke programmer studentene skulle komme fra og få lister over vitenskapelig ansatte, i forhold til informasjon om undersøkelsen og som støtte i gjennomføringen av undersøkelsen. Tilbakemeldingen fra lærestedene på den nasjonale prosjektorganisasjonen er at det stort sett har fungert godt, at NPM har vært tilgjengelig for å svare på spørsmål og at informasjons-utvekslingen gjennom møter og epost har vært god. Imidlertid har det noen ganger blitt litt vel mye informasjon, da veldig lange eposter fra konsortiet med mye informasjon har blitt videresendt til institusjonsansvarlig. Med andre ord finnes det et forbedringspotensial med hensyn til informasjonsflyt mellom NPM og institusjonene som bør være fokus i en eventuell ny runde med AHELO

Den store bøygen i AHELO var rekruttering av studenter. Institusjonsansvarlige jobbet mye for å nå ut til studentene, men med dårlig resultat. I tillegg viste seg rekrutteringsarbeidet mye mer ressurskrevende enn opprinnelig antatt, blant annet fordi institusjonsansvarlig måtte prøve å nå studenter gjennom ulike kanaler og gjennom personlig å informere på kurs. Generelt hadde alle institusjonene i Norge ønsket seg en annen type utvalg, der kurs eller programmer ble samplet. Det ville ha gjort informasjonsarbeidet enklere og gjort det lettere å nå mange av de uttrukne studentene personlig.

En klar tilbakemelding fra flere av lærestedene er at kostnadene for lærestedet knyttet til å delta i AHELO var vesentlig høyere enn de opprinnelig hadde antatt, og i tillegg ble det uttrykket frustrasjon over at de økonomiske sidene ved prosjektet ikke var klarer da institusjonene sa ja til å være med i AHELO. Denne frustrasjonen er primært knyttet til at lærestedene ble bedt om å dekke kostnaden for skåring av besvarelser. Fordi svarprosenten ble lav ble kostnaden for skåringen ikke så høy som først antatt, men det var tydelig både i prosessen og i tilbakemeldingen fra lærestedene at det at kostnaden for skåringen ble veltet over på lærestedene ble oppfattet som negativt, og som et dårlig signal fra KD. I tillegg kan det virke som om det for lærestedene er lettere å dekke kostnader som ikke er direkte synlige, som å oppnevne en institusjonsansvarlig eller få IT-personale til å forberede datamaskiner og nettverk for testing, enn å dekke direkte kostnader i form av en faktura for skåring av oppgaver. En institusjon var også direkte klar på at de ikke ønsket å bruke mye tid og energi på å rekruttere flere studenter til deltakelse, siden det bare ville før til merkostnader for institusjonen – både direkte og indirekte.

AHELO gjennomføres som en eksamenssituasjon, og tilbakemelding fra testtakere til lærestedsansvarlige viser at dette ikke har bidratt positivt til å øke deltakelsen. Samtidig er AHELO ment å være en test som måler læringsutbytte, og dermed må den foregå i en kontrollert setting, i en form for eksamenssituasjon. Dermed er det ikke noen annen måte å få gjennomført en test av den typen man ser for seg å bruke i AHELO uten at det framstår som en eksamenssituasjon. Dette er reell utfordring med hensyn til rekruttering av studenter og dersom en eventuell fullskala AHELO skal ha en fremtid er dette noe man må finne en løsning på.

Lærestedene ble også bedt om å gi en vurdering av om de kunne ønske å være med i en ny runde av AHELO. Selv om det viste seg vanskelig å rekruttere studenter til å delta i testen var ikke de fem lærestedene negative til ny deltakelse, men de hadde forbehold om at måten utvalget trekkes på måtte endres for at de skulle være interessert i å være med videre. Slik de oppfattet det bidro måten utvalget var trukket på (enkel tilfeldig trekking) til at det var mye vanskeligere å rekruttere studenter, enn det nok ville ha vært dersom studentene var trukket ut fra kurs eller klasser. Alle lærestedene ga uttrykk for at de var med i AHELO fordi de var interessert i problematikken kring læringsutbytte og at de så på AHELO som et utviklingsarbeid.

4.6 Vurdering av gjennomføringen av AHELO mulighetsstudie

Gjennomføringen av AHELO mulighetsstudie (fase 2, jamfør figur 1.2), kan deles inn i flere deloppgaver og prosesser. I denne fasen av studien inngikk trekking av utvalg studenter og vitenskapelig ansatte, teknisk testing av systemene som skulle brukes og gjennomføringen av selve testingen, samt gjennomføring av AHELO testen.

Norge hadde flere utfordringer i forbindelse med trekking av utvalg. Utfordringene var først og fremst knyttet til manualene laget av konsortiet. Til tross for omfattende instrukser, viste det seg at disse ikke var godt nok tilpasset norske forhold. Norge har en rekke studieprogrammer som er integrerte masterprogrammer. I utgangspunktet skulle AHELO mulighetsstudie kun omfatte studenter i bachelorprogrammer, noe som ville utelate studenter ved integrerte masterprogrammer. Dette var problematisk for Norge, ettersom dette ville utelate en betydelig andel av studentene ved enkelte av lærestedene. Etter diskusjoner med konsortiet vant Norge frem med sitt syn og fikk inkludert studenter ved integrerte masterprogrammer i sitt utvalg. Dette illustrerer likevel utfordringer man står overfor i internasjonale undersøkelser der det er avgjørende at sammenligningsgrunnlaget er likt på tvers av land og institusjoner. Trolig bør en videre utvikling av AHELO omfatte en grundigere diskusjon og tydeligere avklaring av hvilke grupper av studenter som skal inkluderes i denne type undersøkelser.

Norge hadde også en utfordring med utdanninger som er organisert på annen måte enn treårige bachelorstudier eller femårige masterstudier. Dette gjelder lærerutdanningen, som er en fireårig utdanning. Det viste seg også at ved kun å ta utgangspunkt i studenter som var i begynnelsen av tredje studieår, inkluderte man en stor gruppe studenter med lavt antall fullførte studiepoeng. Dette kan ha flere årsaker – studenter kan ha hatt avbrudd i studiene eller av andre grunner ikke fulgt normal studieprogresjon. En måte å løse denne type problematikk i en fremtidig undersøkelse, kan være å definere studentpopulasjonen med utgangspunkt i antall fullførte eller oppnådde studiepoeng, uavhengig av studieprogram eller antall år studenten har vært registrert som student.

Når det gjelder selve gjennomføringen av AHELO mulighetsstudie, fungerte den tekniske gjennomføringen tilnærmet uproblematisk ved samtlige læresteder. Imidlertid bød gjennomføringen på ulike andre utfordringer. Disse kan deles inn i to hovedgrupper; tidsrommet for gjennomføringen og utfordringer ved å få studenter til å gjennomføre AHELO testen.

Norge var først ut med gjennomføringen av selve testen. Dette hadde sammenheng med et ønske blant lærestedene om å gjennomføre testen før påskeferien, for å unngå konflikt med eksamensavvikling som starter opp i ukene etter påske. Det internasjonale konsortiet hadde satt tidligste oppstartdato til 1. februar 2012 (OECD 2011b), og den norske delen av studien var ment å starte opp denne datoen. Dessverre viste det seg at de tekniske testsystemene ikke var ferdige før helt i slutten av januar noe som medførte at testoppstart ble forskjøvet til 6. februar. Et lærested måtte dermed utsette oppstart til uken etter og mistet dermed sannsynligvis noen respondenter som hadde meldt seg på testing de første dagene i februar.

Den største utfordringen for gjennomføringen av AHELO testen i Norge må likevel si å handle om vanskeligheter ved å få studenter til å delta. Alle lærestedene valgte å rekruttere studenter etter en modell der de uttrukne studentene fikk tilsendt en epost og der de kunne velge når de skulle melde seg på å ta testen. Alle læresteder la til rette for flere testtilfeller, og noen hadde svært mange testtilfeller spredt ut over flere uker. Det var dermed mange muligheter for studentene til å delta i AHELO mulighetsstudie, men deltakelsen ble likevel svært lav. Gjennomsnittlig svarprosent var 7,7 %, med laveste svarprosent på 4,7 % og høyeste på 10 %.

Utfordringen med å få studentene til å delta i en 2,5 timer lang test synes i større grad å handle om studentenes manglende ønsker eller insentiver til å delta enn å handle om tilrettelegging eller informasjon fra lærestedenes side. Lærestedene synes å ha jobbet hardt for å rekruttere studenter, men ingen av rekrutteringsstrategiene har vært vellykket med hensyn til å oppnå tilstrekkelig deltakelse fra studentene. Alle lærestedene hadde en form for insentiver til studentene, men verken utlodding av premie eller betaling i form av gavekort til alle som hadde gjennomført testen ser ut til å ha generert høyere deltakelse.

I prosessen med gjennomføringen av AHELO mulighetsstudie kom uklarheter med hensyn til fordeling av kostnader mellom lærestedene og KD til syne. I avtalen mellom det enkelte lærested og KD kom det klart frem at lærestedene skulle dekke kostnader til gjennomføring og testretting (skåring). At dette ville medføre direkte økonomiske kostnader og ikke kun interne kostnader i form av arbeidstid, kom derimot ikke klart frem. Dette ble et diskusjonstema i forbindelse med gjennomføringsfasen. De direkte økonomiske kostnadene omhandlet insentiver til studentene og betaling av ekstern skåring. Vi går ikke inn i detaljene, men ser det som nødvendig å påpeke at uenigheten om finansiering av denne delen av undersøkelsen kan ha hatt negative konsekvenser for gjennomføringen av AHELO mulighetsstudie i Norge. Ved at lærestedene opplevde det som en uventet økonomisk belastning å skulle dekke kostnader til insentiver til studentene, kan det hevdes at dette bidro til å holde nivået på insentiver på et relativt lavt nivå. Nå bør det likevel sies at nivået på insentiver trolig ville ha måttet øke til et betydelig høyere nivå for ha noen effekt på deltakelsen blant studenter. To forhold er verdt å nevne her: Variasjonen i størrelsen på insentiver mellom lærestedene synes ikke å ha hatt noen betydning for deltakelsen. Det kan tyde på at variasjonen ikke var stor nok og at betydelig større insentiver må til for å få studenter til å delta i en undersøkelse av et slikt omfang og varighet. Et annet forhold som kan nevnes er størrelsen på insentiver, samt andre sider ved gjennomføringsfasen i andre typer undersøkelser av tilsvarende omfang og varighet. Et nærliggende eksempel er OECDs undersøkelse om voksnes læring: PIAAC (Programme for the International Assessment of Adult Competencies). I denne undersøkelsen var innsamlingsfasen både betydelig lenger (mer enn ett semester), tidspunkt og sted for gjennomføring av undersøkelsen var mer fleksibel (hjemme hos deltakerne, hvor intervjueren hadde med pc), i tillegg ble deltakerne tilbudt et høyere beløp som kompensasjon for tiden det tok å delta. Det er ikke gitt å overføre erfaringer fra en type undersøkelse til en annen, det er mange elementer som vil variere og ikke så lett lar seg sammenligne. I en eventuell videreføring av AHELO kan det likevel være nyttig å undersøke hva som er gjort i andre undersøkelser.

I lys av erfaringer fra gjennomføringen av AHELO mulighetsstudie i Norge er en vurdering at det ved en eventuell fremtidig undersøkelse må fokuseres mer på hvordan studenter rekrutteres til å delta. Her synes det å være to hovedstrategier; enten å øke størrelsen på insentivene til studentene eller å gjøre testen som en obligatorisk del av studiet eller på annen måte gi studenten noe igjen for å ta testen, for eksempel i form av individuell tilbakemelding. Slik AHELO mulighetsstudie har vært organisert, med matrisesampling, har det ikke vært mulig å gi slik tilbakemelding. OECDs første rapport fra prosjektet (Tremblay et al. 2012) åpner for at man i fremtiden skal vurdere om testen også kan gi individuell tilbakemelding til de studentene som tar testen. En slik skåre vil gi studentene informasjon om hva hun/han er god til, og hva hun/han bør jobbe mer med. Samtidig er det mulig at et slikt system vil kreve andre typer tester og muligens en test som tar enda mer enn 2,5 timer, hvilket sannsynligvis ikke vil bidra til å bedre sjansene for å rekruttere mange studenter til å delta. Dette er forhold som en eventuell videreføring av denne type test vil måtte diskutere nærmere. Et annet spørsmål er om målgruppen som OECD har satt, studenter som er i sitt siste år av bachelor, er en gruppe som det er realistisk å kunne nå. I Norge bød dette på utfordringer, siden tilbakemeldingen fra mange av de uttrukne studentene var at de i siste semester på bachelor var opptatt med å skrive bacheloroppgave, og derfor ikke var tilstede på lærestedet. Dette var sannsynligvis både et problem relatert til det å nå

studentene med informasjon om at de var trukket ut til å delta i undersøkelsen, og noe som gjorde at de var mindre villige til å delta, siden de ikke vanligvis var på campus.

5 Analyser av data

Dette kapittelet vil se på ulike måter å analysere dataene fra AHELO. Fordi det var få studenter som deltok i AHELO er det begrenset hvilken type analyser som kan gjøres, men det er likevel noen analyser som kan gjennomføres. Vi ser på i hvilken grad studentene som deltok i AHELO skiller seg vesentlig fra utvalget og populasjonen, vi presenterer og tolker noen av de psykometriske testene av hvor godt de to typene oppgaver har fungert som konsortiet har gjort, samt se hvor godt testene har fungert for norske studenter, med utgangspunkt i ulike psykometriske egenskaper ved testene (det vil si hvor godt testoppgavene har fungert for norske studenter generelt og for ulike grupper spesielt). Til slutt presenteres en oversikt over hvordan de norske studentene som har deltatt i AHELO har gjort det på testene, i gjennomsnitt, for å gi et bilde av hvilken type informasjon man kan forvente å få fra en slik test.

5.1 Analyser av hvem som har deltatt i AHELO

Norge fikk som nevnt svært lav deltakelse i AHELO, kun 115 av 1500 uttrukne studenter ved de fem lærestedene deltok i undersøkelsen. Alle lærestedene slet med å få studentene til å delta i testen, men det var noe variasjon i hvor mange som deltok per lærested. Lavest antall hadde UMB og HiVe, og felles for dem var at de begge startet gjennomføringen forholdsvis sent med testingen i mars måned. Samtidig er det ikke så store forskjeller i antall svar mellom lærestedene at det er grunnlag for å si at det kun var tidspunktet som spilte inn.

Men selv om det er få norske besvarelser er det mulig å se på hvem det er som har svart og sammenligne dem med utvalget som helhet. Risikoen når man får svært lav svarprosent er at man har fått et utvalg som avviker svært mye fra utvalget som helhet. I filen med informasjon om hele utvalget finnes variabler som kjønn, program studenten går på og karakterpoeng ved opptak. Dessverre er variabelen karakterpoeng ved opptak kun tilgjengelig for noen få av studenter ved HiVe, men ved alle de andre fire lærestedene finnes det karakterinformasjon for mellom 70 og 88 prosent av studentene i utvalget.

Sammenligning av kjønnsfordeling og fordeling på bachelor- og masterprogrammer mellom de i utvalget som har tatt testen og de som ikke har tatt testen viser at det ikke var stort avvik i gjennomsnitt. Både blant de som har tatt testen og de som ikke har tatt testen er det en overvekt av kvinner og det er i begge grupper omtrent en av fem som går på et integrert masterprogram.

Tabell 5.1: Kjønnsfordeling blant de som har deltatt og de som ikke har deltatt i testen

	Mann	Kvinne
ikke tatt testen	43,0	57,0
tatt testen	43,5	56,5

Dersom vi derimot ser på kjønnsfordelingen på det enkelte lærested er det noe større variasjon, men ved alle læresteder unntatt UMB er forholdet mellom kvinner og menn likt, selv om andelen blant de som har tatt testen og de som ikke har tatt testen er forskjellig. Ved NTNU er det litt større andel kvinner enn menn i populasjonen og utvalget, mens det er motsatt ved Høgskolen i Lillehammer, Høgskolen i Vestfold og Universitetet i Stavanger, og blant de som har tatt testen er det tilsvarende forhold mellom kvinner og menn. Ved UMB er det derimot litt større andel kvinner enn menn i populasjonen og i utvalget, mens det er 8 menn og 6 kvinner som har fullført AHELO-testen, og dermed er størrelsesforholdet omvendt for de som har tatt testen. Men det er mye å forvente at størrelsesforholdene skal være like når antallet som har tatt testen er så lite som 14 personer.

Tabell 5.2: Fordeling på programtype blant de som har deltatt og de som ikke har deltatt i testen

	Bachelor-program	Integr. Master-program
ikke tatt testen	78,1	21,9
tatt testen	80,0	20,0

Tilsvarende finner vi større variasjon mellom andel studenter på bachelorprogrammer respektive integrerte masterprogrammer på institusjonsnivå. Da utvalget ble trukket ble det kontrollert at fordeling i utvalget var i overensstemmelse med populasjonen. Blant de deltakende lærestedene er det UMB, NTNU og UiS som har studenter på integrerte masterprogrammer. Ved UMB og UiS er det færre masterstudenter som har tatt testen enn andelen deres tilsvarer av utvalget, mens det ved NTNU er en noe større andel studenter på integrerte mastere enn andelen deres skulle tilsi som har deltatt i AHELO-testen. Men i gjennomsnitt er det en av fem studenter både i utvalget og blant de som har tatt testen som går på et integrert masterprogram.

Det er heller ikke noen store avvik i gjennomsnittskarakter mellom studenter som har tatt testen og de som ikke har tatt testen, ved de fire lærestedene som har gode data for gjennomsnittskarakter. Dette er forbausende siden en kunne anta at det stort sett ville være studenter med svært gode karakterer, de som vet at de er flinke, som frivillig hadde valgt å ta testen.

Tabell 5.3: Gjennomsnittskarakter og standardavvik, etter lærested og deltakelse i testen

	Har tatt testen			Har ikke tatt testen			Alle		
	Gj.snitt	Antall	Std.avvik	Gj.snitt	Antall	Std.avvik	Gj.snitt	Antall	Std.avvik
UMB	45,32	10	5,03	46,19	208	5,77	46,15	218	5,73
NTNU	51,40	29	6,42	52,27	237	10,29	52,17	266	9,94
HiL	39,00	18	6,72	39,41	204	5,68	39,38	222	5,75
UiS	43,93	18	5,20	42,79	211	5,79	42,88	229	5,75
HiVe*		2		40,62	19	6,61	40,50	21	6,34
Alle	45,65	77	7,70	45,32	879	8,72	45,34	956	8,64

*kun 2 av studentene som har tatt testen er registrert med gjennomsnittskarakter, og disse vises derfor ikke.

Dermed kan vi konkludere at de som har deltatt i testen ikke ser ut til å være veldig forskjellige fra de som ikke deltok, i hvert fall ikke med tanke på kjønn, programtype og karakter.

Dersom vi i stedet ser på hvilke faggrupper som er representert, ser vi at utvalget er skjevt med tanke på andelene de ulike gruppene utgjør av populasjonen. Her er det ikke mulig å bryte tallene ned på lærested på grunn av få responser, og det er viktig å huske på at siden vi ikke har et tilfeldig utvalg læresteder har vi heller ikke et tilfeldig utvalg på fagområdenivå. Imidlertid ble fagfordelingen på det opprinnelige utvalget (300 per lærested) sammenlignet med fagfordelingen generelt i populasjonen, og da var det ikke noen store avvik. Med andre ord var det samsvar mellom fagsammensetningen i utvalget ved et lærested og fagsammensetningen i populasjonen ved det lærestedet. Derimot viser tabell 5.4 at det i gjennomsnitt er det er en større andel av studenter i humaniora, samfunnsvitenskap, teknologi og helsefag som har tatt testen og en mindre andel studenter i pedagogiske fag, enn deres andel utgjør av populasjonen.

Tabell 5.4: Fagsammensetning blant de i utvalget som har tatt testen og de som ikke har tatt testen

	PED	HF	SV/ØK/JUS	MN	TEK	AGR	HELSE	SERVICE	N=100%
Har ikke tatt testen	11 %	4 %	26 %	10 %	21 %	3 %	14 %	10 %	1385
Har tatt testen	3 %	8 %	30 %	9 %	23 %	2 %	17 %	9 %	115

5.1.1 Studentenes innsats og testens relevans

I spørreskjema til studentene var det spørsmål om hvordan stor innsats de la i å ta testen og i hvilken grad de oppfattet testen som relevant for studiet de var i gang med og for sitt fremtidige yrkesliv. I spørsmålsstillingen var yrkesliv formulert som «fremtidig faglig praksis». Frekvensene for dette er vist i tabell 5.5 og 5.6. Tabell 5.5. viser over 63 prosent av studentene som deltok i testen sier at de la stor innsats i testen, det vil si at de faktisk anstrengte seg for å klare oppgavene. Dette stemmer godt overens med hvordan norske elever som deltar i PISA svarer, der det også er en klar majoritet som sier at de gjorde en god innsats eller sitt beste på prøven (Kjernsli & Roe 2010: 228). Imidlertid er ikke spørsmålene stilt på samme måte, i PISA brukes utsagn som elevene kan si seg enige eller uenige i, mens det her er brukt en skala fra liten/ingen til svært mye innsats.

Tabell 5.5: Hvor stor innsats studentene la i testen

	Antall	Andel
Liten/ingen innsats	1	0,9
Litt	39	35,8
Mye	60	55,0
Svært mye	9	8,3
	109	100,0

Tabell 5.6 viser i hvilken grad studentene mener at AHELO-testen i den utforming den har hatt i mulighetsstudien er relevant for graden de er i med å ta og for yrkeslivet. Generelt mener studentene at testen har forholdsvis lite relevans, nesten en av fem mener at den ikke er relevant i det hele tatt, verken for graden eller for yrkeslivet. Kun 4 prosent mener at den har ganske mye relevans for graden og 9 prosent mener den har relevans for yrkeslivet. Med andre ord ser ikke studentene hvordan testen er relevant verken for den graden de er i gang med og heller ikke for yrkeslivet og dette er noe som også kan være problematisk, gitt at det ikke vil være noe som bidrar til å gjøre rekrutteringen til å delta i studien lettere.

Tabell 5.6: Testens relevans for graden og for yrkeslivet (fremtidig faglig praksis)

	For graden		For yrkeslivet	
	Antall	Andel	Antall	Andel
Ikke i det hele tatt	21	19,3	19	17,4
Svært lite	58	53,2	47	43,1
Litt	26	23,9	33	30,3
Ganske mye	4	3,7	9	8,3
Svært mye	0	0,0	1	0,9
	109	100,0	109	100,0

Konsortiet har, basert på de to deltestene som AHELO består av (CRT og MCQ), beregnet en skåre for alle deltakere. Fordi det er få deltakere i den norske mulighetsstudien skal vi ikke vise gjennomsnittsskårene, men vi skal nå se nærmere på i hvilken grad det er korrelasjon mellom studentenes skåre på testen og ulike variabler. Hensikten med dette er å prøve å si noe om hva som bidrar til å forklare skåren studentene oppnår, eller i hvilken grad opptakskarakter, innsats på testen, testens relevans for graden og for yrkeslivet hver for seg bidrar til å forklare deler av skåren. Tabell 5.7 viser at det er forholdsvis sterk korrelasjon mellom opptakskarakter og skåre på testen, 0,54, og denne er signifikant. At testskåre korrelerer med studentenes generelle karakternivå ved opptak til høyere utdanning, er ikke uventet med tanke på at AHELO er utformet som en test for å måle studentenes kunnskapsnivå og ferdigheter innen bestemte områder. En korrelasjon på 0,54 er relativt sterk, men uttrykker samtidig en del variasjon mellom de to prestasjonsmålene. Vi finner også signifikante korrelasjoner mellom skåre og innsats på testen, og skåre og relevans av testen for graden. Derimot er korrelasjonen mellom testens relevans for yrkeslivet og skåren studentene oppnår ikke signifikant, men vi ser at det er en svak sammenheng.

Tabell 5.7: Korrelasjon mellom testskåre og opptakskarakter, innsats på testen, relevans for graden og for yrkeslivet

	Korrelasjon	sig	N
Opptakskarakter	0,543	0,000	77
Innsats på testen	0,224	0,019	109
Testens relevans for graden	0,221	0,021	109
Testens relevans for yrkesliv	0,177	0,064	109

5.1.2 Oppsummering

Denne oversikten viser at i gjennomsnitt avviker ikke de få studentene som har tatt AHELO-testen fra utvalget eller populasjonen, men at det blir større variasjoner dersom vi bryter det ned på institusjon. Det som derimot er gledelig er at det er liten forskjell i gjennomsnittskarakter mellom de som har tatt testen og de som ikke har tatt testen, dersom vi sammenligner basert på lærested. Dette tilsier at testen i hvert fall ikke bare har rekruttert de flinkeste studentene. Samtidig er antallet som har tatt testen så lite at det aldri kan være representativt for det enkelte lærested, og siden det ikke er et tilfeldig utvalg av læresteder er det heller ikke representativt for Norge.

Spørreskjemadata viser at studentene har lagt forholdsvis mye innsats i å ta testen, samtidig som de i liten grad synes at testen er relevant for graden de er i gang med eller for fremtidig yrkesliv. Vi finner videre at det er en forholdsvis sterk korrelasjon mellom studentens opptakskarakter og hvordan studentene skårer på testen, og en noe svakere korrelasjon mellom studentens innsats på testen eller oppfatning av testens relevans for graden og skåren studenten oppnår.

5.2 AHELO-oppgavene: psykometriske egenskaper

Vi skal først se på noen av de psykometriske egenskapene som konsortiet beskriver i sine analyser, for å se hvordan de ulike deltestene i AHELO har fungert. Generic skills strand tok utgangspunkt i to typer tester: en case-oppgavene som kom i to varianter (CRT1 og CRT2), samt et sett med multiple choice-spørsmål (heretter omtalt som MCQ), som besto av en kjernedel som alle testtakere tok og en roterende del (fire ulike varianter). I denne delen forholder vi oss til hele settet av MCQs mens vi i analysene av de norske dataene kun forholder oss til kjernedelen av MCQ (14 oppgaver).

Vi starter ut med å se nærmere på hva konsortiets analyser av hvordan studentens innsats på testen henger sammen med de tre testene vi kan utskille, de to caseoppgavene og settet med flervalgsoppgavar (MCQ) viser.

5.2.1 «Generic Skills Score variance explained»

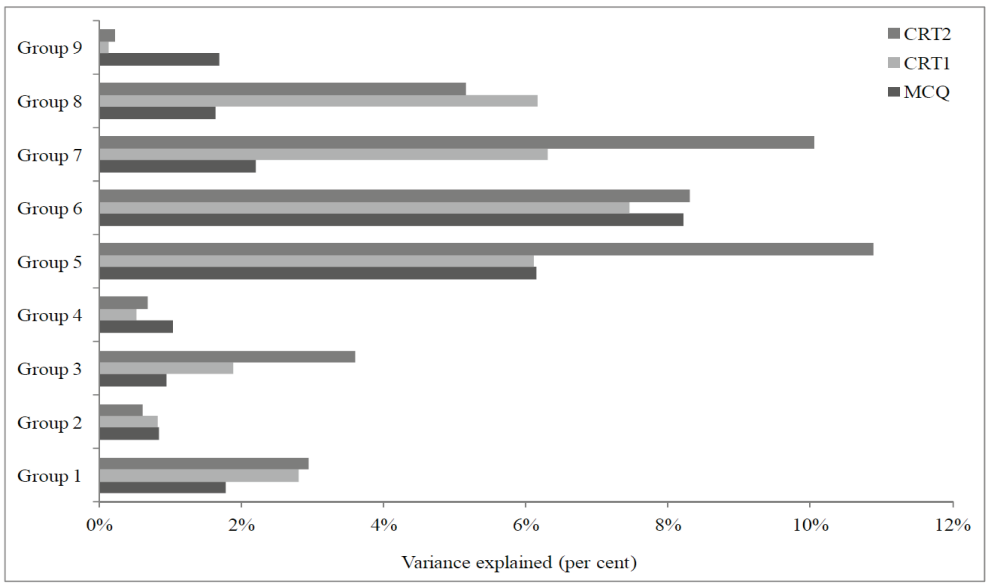


Figure 6: Generic Skills Score variance explained by group by effort by task type (n=10657)

Figur 5.1: Forklart varians etter studenters innsats på testen. Kilde: AHELO consortium 2012:81

I slutten av spørreskjemaet fikk studentene spørsmål om egen innsats i besvarelsen. Ved hjelp av flernivåanalyser kan man derfor estimere hvor mye av variansen som er forklart av studentenes innsats. Dette er gitt i figur 5.1 (AHELO consortium 2012:81). Med «grupper» menes her land, og for hvert land har vi tre stolper som representerer de tre testene CRT1, CRT2 og MCQ. Norge er gruppe 7. Med «varians forklart av studentenes innsats» mener vi hvor mye av variasjonen i studentenes resultater som skyldes deres innsats i testbesvarelsen og ikke hvor mye de faktisk kan. Hypotetisk sett, om oppgaven bestod i at studenten skulle skrive sitt eget navn, ville dette kun avhenge av om studenten ønsket å gjøre dette. I et slikt tilfelle ville 100 % av variasjonen i resultatene være forklart av innsats.

Derfor: Jo lengre stolpene i figuren er, desto mer avhenger resultatene av innsats fremfor kunnskap for studentene i dette landet. Vi ønsker korte stolper fordi variasjonen i resultatene i størst mulig grad bør avhenge av det vi ønsker å måle, nemlig «generic skills» og ikke innsats. Vi merker oss blant annet at omtrent 10 % av variasjonen i Norge på CRT2 skyldes innsats. Videre ser vi at i begge CRT-

oppgavene er det klart høyere forklart varians basert på innsats for testtakerne i Norge, sammenlignet med MCQ. Med andre ord er MCQ mindre sårbart for innsatsvilje hos testtakerne.

5.2.2 «Generic Skills Assessment variable map»

Konsortiet har også foretatt «Item response modelling» for hele testen (AHELO consortium 2012:83-89) De presenterer tre figurer, en for hele testen sett under ett, en figur som kombinerer CRT1 med MCQ, og en figur som kombinerer CRT2 og MCQ. Disse figurene gir en grafisk illustrasjon av det som er en av Raschmodellens viktigste egenskaper: At personer og oppgaver kan plasseres på samme skala (Bond og Fox 2001). En slik figur inneholder mye viktig informasjon om testen.

En grunnide i Raschteori er at man ønsker å måle et «latent trait» - en egenskap ved et menneske som ikke er direkte observerbart. Det er for dette «noe» at vi ønsker å utvikle en skala. Man kan ha mye av egenskapen, eller man kan ha lite av egenskapen. Personers plassering på denne skalaen skal kun avhenge av hvor mye de har av denne egenskapen. Kryssene på venstre side av figur 5.2 representerer grupper av personer (omtrent 50 personer i hver gruppe i dette tilfellet). Et kryss langt ned i figuren representerer en gruppe personer som har lite av denne egenskapen, mens et kryss langt oppe i figuren representerer noen som har mye av denne egenskapen.

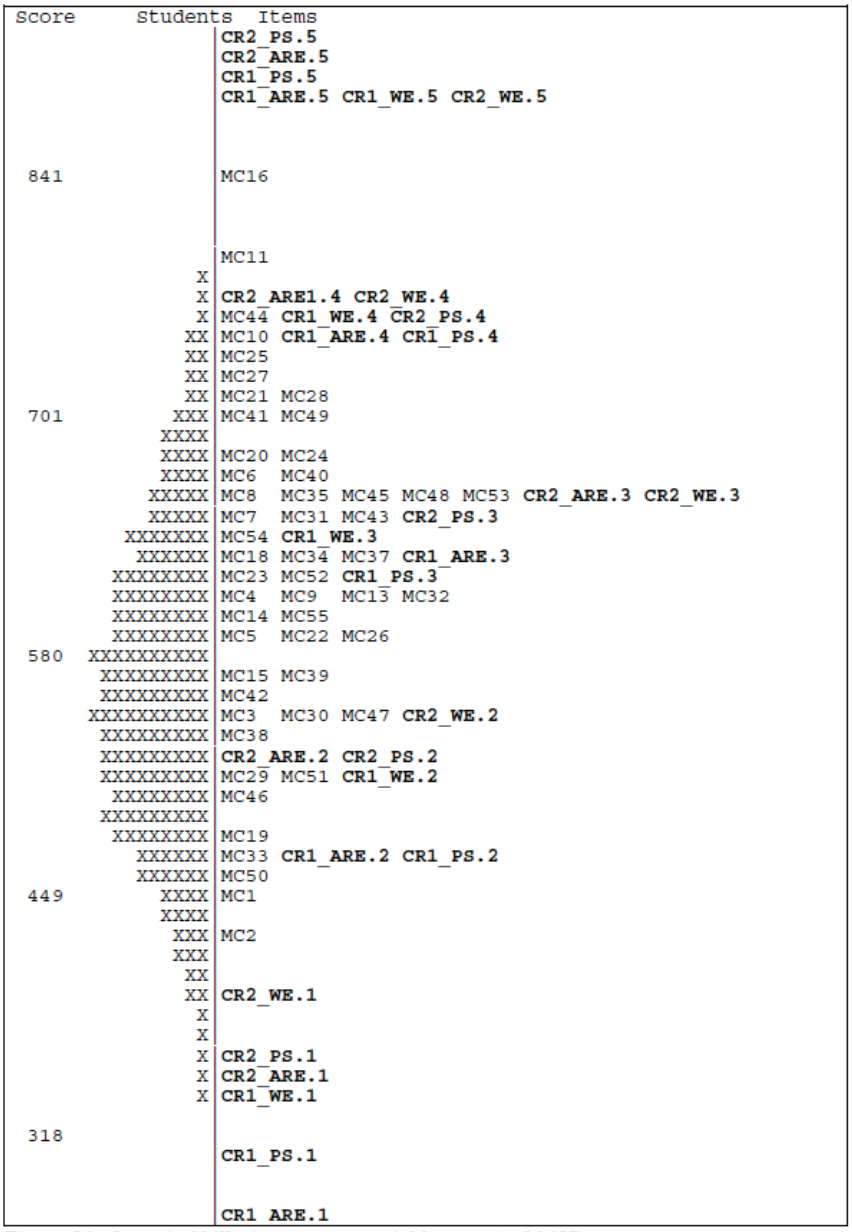


Figure 10: Generic Skills Assessment variable map (n=10657)

Figur 5.2: Variabelkart for Generic skills test. Kilde: AHELO consortium 2012:84

Som vi ser av fordelingen i figur 5.2 har de fleste fått en skår mellom 449 og 701. Tallene i skalaen er ikke den faktiske skåren til personene på testen, det er gjort noen omregninger fra den opprinnelige skåren på testen slik at personenes dyktighet får et tall som ligner det man kjenner fra PISA og TIMSS. Man kunne også gjort omregninger slik at skalaen fikk et gjennomsnitt på 0, hvor de fleste personene blir plassert mellom -4 og +4. Dette er gjort i figur 5.3 og hvordan omregningen gjøres er opp til den som gjennomfører testen. Konsortiet har valt å lage en skala med 500 som gjennomsnitt og 100 som standardavvik. Hensikten er uansett ivaretatt: At man kan vurdere personene i forhold til hverandre og i forhold til oppgavene.

Oppgavenes plassering på skalaen har en bestemt betydning: Oppgavens vanskelighetsgrad (plassering på skalaen) er «så mye av egenskapen en person må ha for å ha 50 % sannsynlighet for å klare oppgaven». Eksempelvis vil personene som er representert med det øverste krysset i figur 5.2 ha *litt under* 50 % sjanse for å besvare MC11 riktig, og *litt over* 50 % sjanse for å besvare MC44 korrekt. Oppgaver langt ned på skalaen, for eksempel oppgave MC2, vil disse besvare korrekt i

praktisk talt alle tilfeller. Desto større avstand det er fra krysset og *opp* til en oppgave, desto mindre er sjansen for korrekt respons, og desto større avstand det er fra krysset og *ned* til en oppgave, desto større er sjansen for korrekt respons. Det at sannsynligheten minsker og øker på denne måten, er grunnen til at vi kaller modellen for «probabilistisk» (probability = sannsynlighet). Det at en person ligger høyere opp enn en oppgave, betyr altså at det ikke er garantert at personen klarer oppgaven, men at det er mer sannsynlig.

Merk at vi i figuren finner oppgaver som heter «CR1_ARE.1», «CR_ARE.2», osv. Ettersom man for hver at de tre områdene (ARE, PS og WE) på de to deltestene (CRT1 og CRT2) kan få fra 0 til 5 poeng, kan man anse hvert nytt steg opp på skalaen, for eksempel fra 0 til 1 poeng eller fra 3 til 4 poeng, som en slags oppgave i seg selv med en bestemt vanskelighetsgrad: Det krever en bestemt dyktighet for å gå fra å ha størst sjanse til å få 3 poeng til å ha størst sjanse til å få 4 poeng.

Verdien av å se fordelingen av personenes dyktighet og oppgavenes vanskelighetsgrad på denne måten, er blant annet at man kan avgjøre hvorvidt testen er velegnet for denne populasjonen: Man ønsker at fordelingen av personer og oppgaver skal se så like ut som mulig. Hvorfor? Fordi man i det ferdighetssjiktet man har mange personer (for figur 5.2: i området rundt 580 poeng) trenger mange oppgaver med passende vanskelighetsgrad for å rangere disse personene med presisjon. Om alle oppgavene hadde vært altfor enkle, ville man ikke klart å skille mellom personene, fordi de fleste hadde fått full skår. Tilsvarende hadde de fleste fått null poeng om alle oppgavene hadde vært altfor vanskelige. Skal man rangere tre studenter i 10. klasse etter hvem som er flinkest i matematikk, vil det ikke være til hjelp å gi dem enkle addisjonsoppgaver – alle tre studentene vil svare riktig. Tilsvarende vil alle svare feil om de får avanserte integrasjonsoppgaver. Derfor er det viktig at man har flest oppgaver i det sjiktet man har flest personer.

Det er fortsatt viktig å ha *noen* enkle og *noen* krevende oppgaver for å skille mellom de som henholdsvis har lite og mye av egenskapen man måler. I lys av dette, ser testen ut til å ha en fin vanskelighetsgrad. Figur 5.3 viser at CRT2 er litt vanskelig («bølgen» av oppgaver ligger litt høyere på skalaen enn bølgen av personer), men dette er likevel på et akseptabelt nivå, siden forskjellen ikke er veldig stor.

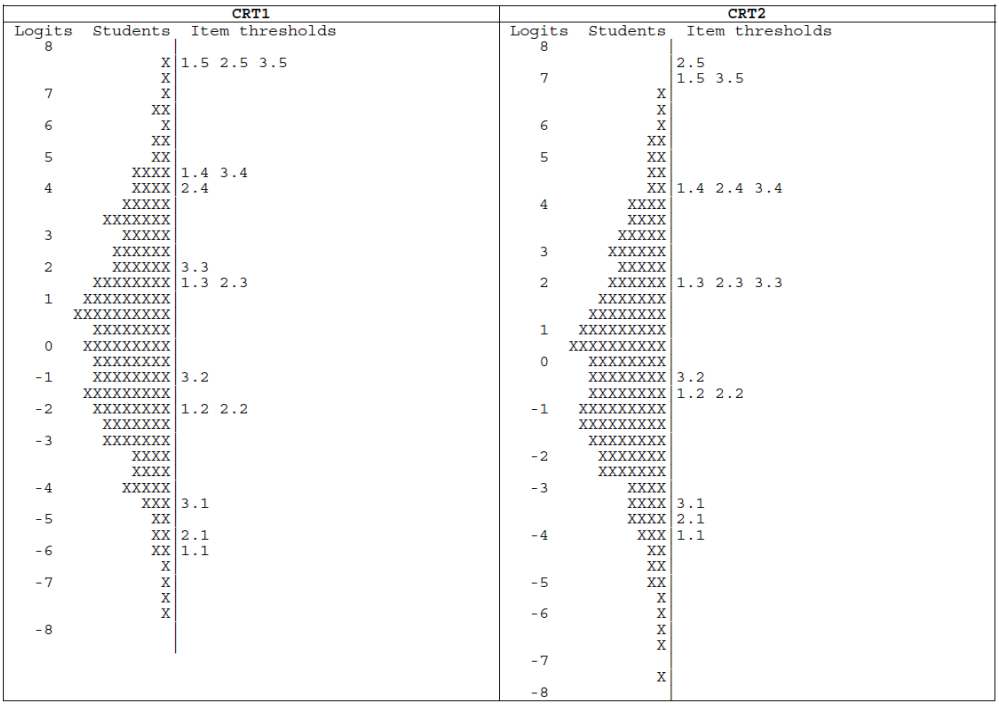


Figure 14: Generic Skills Assessment CRT1 and CRT2 variable maps with score category thresholds (n=10657)

Figur 5.3: Separate variabelkart for CRT1 og CRT2. Kilde: AHELO consortium 2012:89

Det mest problematiske vi finner i disse figurene, er knyttet til noen av skårkategoriene på CRT-oppgavene. I figur 5.2 ser vi at de letteste oppgavene kanskje er for lette (nesten alle får minst ett poeng på CRT-oppgavene) og de vanskeligste er for vanskelige (praktisk talt ingen får 5 poeng på verken ARE, PS eller WE på CRT1 eller CRT2). For å se nærmere på CRT1 og CRT2, har man i figur 5.3 tatt for seg de to testene hver for seg, med CRT1 i venstre del og CRT2 i høyre del. Mens studentene i figur 5.2 var rangert etter skår på hele AHELO-testen, har man her rangert dem kun basert på resultater fra den av de to CRT-testene de har besvart. Her ser man at det er noen som faktisk har mulighet til å skåre i skalaens ytterpunkter. Men fortsatt ser det ut til å være en utfordring at svært mange respondenter må rangeres ved hjelp av relativt få trinn på skalaen. I CRT2 (høyre del) ser vi for eksempel at det store antallet studenter med dyktighet fra 0 1.8 «logits»[3] ikke skilles av noen trinn.

Skalaen for CRT-oppgavene er for unyansert i det sjiktet hvor de fleste studentene befinner seg. Siden nesten alle studenter får 2, 3 eller 4 poeng så skiller oppgavene i liten grad mellom dem. Dette kan også tyde på at kravene for å få høyeste poengsum er for høye. En skala som denne kan fungere godt dersom det er stor variasjon i ferdighetsnivå mellom de som testes (for eksempel om vi hadde testet grunnskoleelever sammen med universitetsstudenter) men siden alle testpersoner er på universitets- og høgskolenivå er deres ferdighetsnivå for likt til at spennet i denne skalaen er hensiktsmessig. Man hadde heller hatt behov for større differensiering i sjiktet fra 2 til 4 poeng.

5.3 AHELO-oppgavene: bruk i Norge

I utgangspunktet skulle Generic skills strand kun bruke case-oppgavene basert på CLA (heretter omtalt som CRT1 og CRT2), men etter anbefaling fra TAG og som foreslått i July Interim Feasibility Study Report (AHELO consortium 2011) valgte man også å ha med multiple choice spørsmål (heretter omtalt som MCQ), som skulle fungere som et empirisk anker.

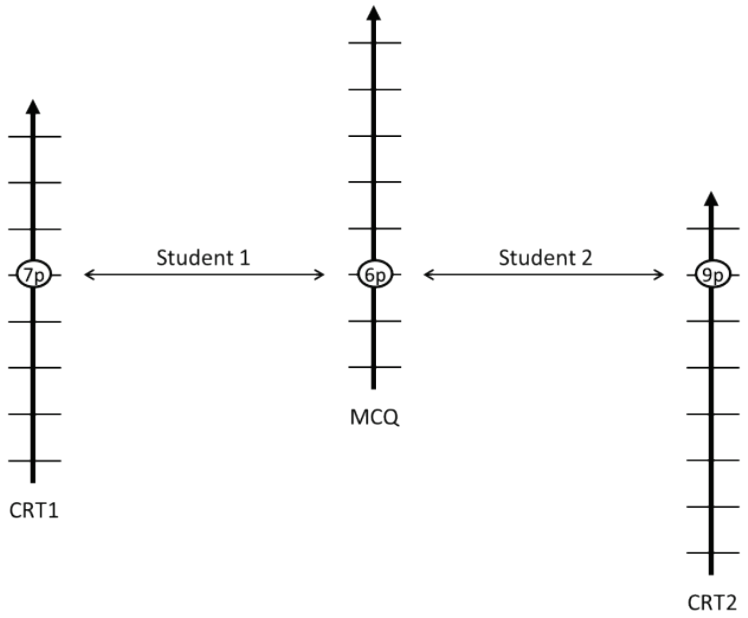
[3] Logits er en vanlig benevning for enheter på slike skalaer og er en forkortelse av ordet «log-units». Dette har å gjøre med omregningen fra faktisk skår på testen til skalaen vi bruker, hvor logaritmeverdier spiller en avgjørende rolle

5.3.1 Et empirisk anker

Behovet for et såkalt «empirisk anker» er et hovedpoeng i utvikling av testen. Grunnen til dette er at den eneste deltesten alle deltakere tar er de første 15 MCQ-oppgavene. CRT1 og CRT2 distribueres til ulike studenter, og disse kombineres med fire ulike delsett med 10 tilleggsoppgaver (MCQ). Spørsmålet er derfor hvordan kan man sammenligne en student som besvarer CRT1 med en student som besvarer CRT2, når man ikke vet hvordan vanskelighetsgraden på de to oppgavesettene forholder seg til hverandre. De antas å være like, men dette må kontrolleres. Det kan, hypotetisk sett, være vanskeligere å skåre høyt på CRT1 enn på CRT2, og dermed blir det feil å sammenligne poengskår fra CRT1 med poengskår fra CRT2. Dette løser man ved å legge inn noen ekstra spørsmål som *alle* studentene besvarer. I vårt tilfelle ble det lagt til en samling flervalgsoppgaver (MCQ-testen). Svært forenklet kan man si at logikken er som følger:

Vi ser på CRT1, CRT2 og MCQ som tre tester som alle måler det samme. Disse har ulik vanskelighetsgrad som vi ikke kjenner til. Det vil si: Vi vet ikke hva Student 1 som skårer 7 poeng (tilfeldig valgt skår) på CRT1 ville ha skåret om han hadde besvart CRT2.

Siden alle studentene også besvarer MCQ, vet vi hva Student 1 skårer på denne testen: 6 poeng (tilfeldig valgt skår). Det vi da kan gjøre, er å lete blant studentene som har tatt CRT2, og spørre: De som får 6 poeng på MCQ, hva får de på CRT2? Blant disse finner vi Student 2, som får 6 poeng på MCQ og 9 poeng på CRT2. Da kan vi argumentere for at «det å skåre 7 poeng på CRT1 tilsvarer å skåre 9 poeng på CRT2». Hvorfor? Fordi Student 1 og Student 2 som er like flinke – noe vi antar siden begge skårer 6 poeng på MCQ – skårer henholdsvis 7 poeng på CRT1 og 9 poeng på CRT2. Dette er illustrert i 5.4.



Figur 5.4: Kalibrering av tre ulike skalaer tilhørende tre tester CRT1, MCQ og CRT2. Forenklet eksempel til illustrasjon.

På denne måten får man kalibrert de ulike skalaene mot hverandre, og man kan vurdere vanskelighetsgraden til oppgaver i CRT1 i forhold til oppgaver i CRT2. Det avgjørende punktet her er at de tre skalaene virkelig måler det samme konstruktet: Hvis CRT1 og CRT2 måler «generic skills», mens MCQ måler noe annet enn dette, gir ikke kalibreringen mening. Dette gjelder også dersom de ulike oppgavesettene måler ulike former for generic skills. Dette punktet er så viktig at vi illustrerer det med et eksempel:

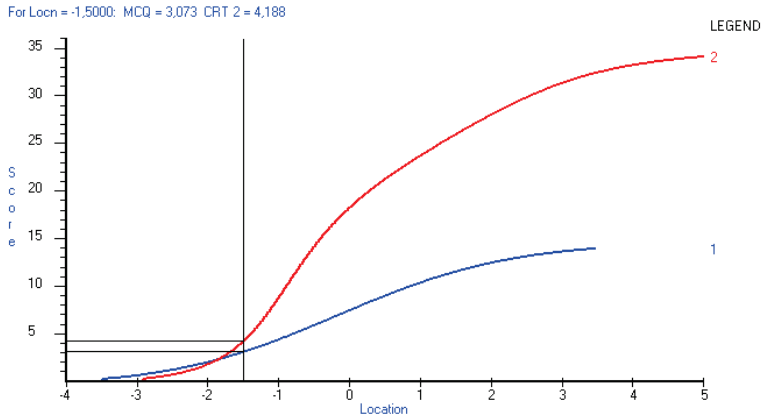
Si at CRT1 og CRT2 måler «generic skills» og MCQ måler «mathematics skills». Da kan man godt konkludere med at de som skårer 7 poeng på CRT 1 i snitt skårer 6 poeng på MCQ. Videre kan man godt konkludere med at de som skårer 9 poeng på CRT2 også skårer 6 poeng på MCQ. Problemet er

at vi ut fra dette *ikke* kan konkludere med at 7 poeng på CRT1 tilsvarer 9 poeng på CRT2. Hvorfor ikke? Fordi det vi har funnet ut, er at «folk som er like gode i matematikk skårer 7 poeng på CRT1 og 9 poeng på CRT2». Vi har ikke fått bekreftet av CRT1 og CRT2 faktisk måler det samme og dermed ikke hvordan de to skalaene forholder seg til hverandre. Det vi trenger å vite, er hva folk som er like gode i «generic skills» skårer på de CRT1 og CRT2. *Derfor er det avgjørende at CRT1, CRT2 og MCQ måler det samme om resultatene fra disse skal kunne sammenlignes*.

De norske resultatene tyder på at denne sammenkoblingen er svært usikker. Hvorfor? Fordi det ser ut til at MCQ måler noe annet ved de norske studentene enn det CRT1 og CRT2 gjør, noe som blir begrunnet empirisk i neste avsnitt. Dermed blir det også mer usikkert at CRT1 og CRT2 måler det samme. Dette har vi ingen mulighet til å undersøke, ettersom ingen studenter har besvart både CRT1 og CRT2. Uansett må vi konkludere med at skalaen som er utviklet, har stor usikkerhet når den er anvendt på norske studenter.

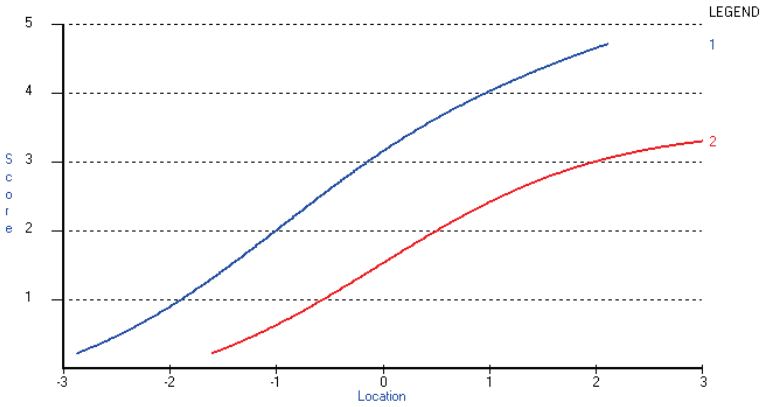
Påstanden om at MCQ måler noe annet enn CRT-oppgavene begrunner vi her empirisk ved å vise til en såkalt «Principal Component Analysis» med CRT2 som eksempel (resultatene for CRT1 er tilsvarende): Vi antar at MCQ og CRT2 sammen utgjør én test. Denne testen rangerer personene og gir dem en dyktighet. Ut fra deres dyktighet beregnes deres forventede skår på hver oppgave. Ved å trekke denne verdien fra det de faktisk skårer på testen, får man for hver person og hver oppgave en restverdi (et «residual»). Ved å analysere disse residualene, finner vi i vårt tilfelle en sammenheng mellom residualene som er knyttet til oppgavene fra CRT2. Det viser seg at det er noe med disse oppgavene som gjør de systematisk avviker fra MCQ-oppgavene med tanke på residualene. Oppgavene fra CRT2 har altså noe felles som MCQ-oppgavene ikke har.

Effekten av dette viser seg om man beholder estimatene på personenes dyktighet, men behandler MCQ og CRT2 som to separate tester. For disse to testene kan vi tegne opp en kurve som viser forholdet mellom dyktighet og observert skår. I figur 5.5 er begge kurvene tegnet inn, med MCQ merket «1» og CRT2 merket «2».



Figur 5.5: Forholdet mellom dyktighet («Location») og observert skår («Score») på MCQ (test 1, blå kurve) og CRT2 (test 2, rød kurve). Antall poeng på de to testene er avlest for personer med dyktighet -1.5 logit.

I figur 5.5 ser vi at de to kurvene oppfører seg svært forskjellig. På MCQ får studentene jevnt flere poeng ved økende dyktighet. På CRT2 får studentene først ikke får særlig uttelling for økende dyktighet (fra -3 til -1.5 logit), før en endring i dyktighet brått gir stor uttelling i form av mange poeng (fra -1.5 og oppover). At CRT2 belønner studentene på denne måten er ikke noe problem i seg selv, men dette betyr at det er noe ved denne testen som skiller seg fra MCQ. Til sammenligning viser vi i figur 5.6 hva som skjer hvis vi deler MCQ i to basert på en tilsvarende residualanalyse. Her er det ingen indikasjon på at de to delene måler noe ulikt. Til tross for at del 1 er lettere enn del 2 (studenter med en gitt dyktighet får høyere skår på del 1 enn del 2), ser vi at de to testene oppfører seg likt med tanke på hva en økning i dyktighet betyr for observert skår.



Figur 5.6: Forholdet mellom dyktighet («Location») og observert skår («Score») på MCQ del 1 og MCQ del 2. De to delene er gitt av hvilke oppgaver i MCQ som skiller seg mest fra hverandre i en Principal Component Analysis.

En viktig empirisk indikasjon på at de to deltestene MCQ og CRT2 måler noe forskjellig, finner vi ved å ta for oss hver persons skår på MCQ og CRT2. Om testene måler det samme, skal en students skår på MCQ kunne si oss noe om studentens skår på CRT2. Siden hver person har et eget estimat på målesikkerhet (personer med dyktighet på et nivå hvor det også er mange oppgaver, har større målesikkerhet enn personer på et nivå med få oppgaver), kan vi basert på skåren på MCQ finne et intervall for skår på CRT2 som personen med 99 % sikkerhet skal ligge innenfor. I vårt tilfelle skårer 9 % av studentene utenfor dette intervallet, noe man kan forvente at kun omtrent 1 % skal gjøre om testene måler det samme. Det er altså to ulike former for «dyktighet» som skal til for å skåre på de to testene.

Vi finner også andre indikasjoner på forskjeller mellom MCQ og CRT2. Korrelasjonen mellom de to deltestene er 0.34. Dette er mindre enn det man kunne forvente om testene hadde målt det samme. Videre, når man betrakter MCQ og CRT2 sammen som en test, viser det seg at det er de tre delene som utgjør CRT2 som passer dårligst når disse testes mot Raschmodellen. Dette omtales som dårlig «item fit»; et begrep som blir presentert i kapittel 5.3.3.

Tilsvarende resultater får vi ved sammenligning av CRT1 og MCQ. Konklusjonen her blir at MCQ ser ut til å måle noe annet enn CRT1 og CRT2, og av denne grunn bør MCQ ikke brukes som «empirisk anker» for de andre testene. Derfor er det fortsatt uklarhet i hvordan CRT1 og CRT2 forholder seg til hverandre. Videre betyr det at skalaen som baseres på personers skår på MCQ og enten CRT1 eller CRT2, ikke gir god mening. Det blir som å summere to ulike mål for å beskrive noe. Dette er sammenlignbart med å legge sammen høyde i centimeter og vekt i kilo. Hvis jeg gjør dette, får jeg svaret «255». De to delmålene har opplagt en sammenheng, men ut fra tallet «255» er det vanskelig å si noe sikkert om meg. Jeg kan enten være 165 cm høy og veie 90 kilo, eller 185 cm høy og veie 70 kilo, eller noe midt i mellom.

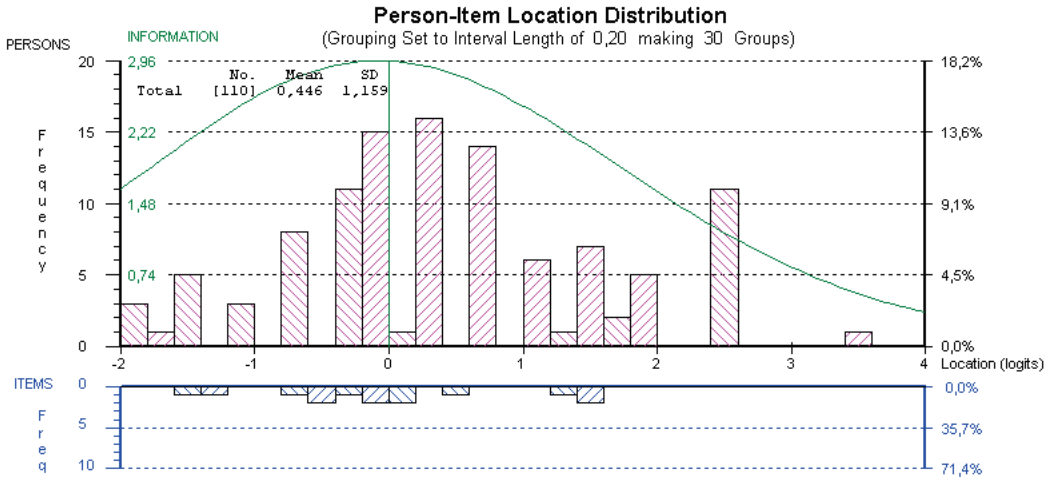
Denne øvelsen viser at testene muligens kan fungere godt hver for seg, men å sammenligne studenter som har tatt MCQ og CRT1 eller CRT2 er problematisk. Derfor skal vi nå gå videre for å se nærmere på hvordan MCQ alene har fungert blant norske studenter. Denne delen velges ut fordi den i det internasjonale tilfellet «captured 90 per cent of the data in 25 per cent of the testing time» (AHELO consortium 2012:68). Det ser med andre ord ut til å være svært kostnadsbesparende å velge MCQ – forutsatt at den fungerer godt og måler noe vi ønsker å måle. I det følgende vil vi undersøke om den fungerer godt, mens spørsmålet om vi måler noe vi ønsker å måle blir diskutert senere.

Før psykometriske egenskaper for MCQ blir presentert i det påfølgende, er det grunn til å trekke frem noe viktig vi allerede vet om MCQ: Den første indikasjonen på at MCQ har fungert fint, ble gitt i figur 5.6. Om man grupperer de oppgavene i MCQ som tilsynelatende er mest forskjellige (basert på en analyse av residualene), viser det seg at disse to delene fortsatt oppfører seg svært likt. I dette tilfellet er det ingen personer som skårer utenfor det 99 %-konfidensintervallet i skår på del 2 som er estimert basert på personenes skår på del 1. Det er heller ingen som skårer utenfor et 95 %-konfidensintervall.

Dette er en indikasjon på at de to delene måler samme egenskap, som igjen er et argument for at MCQ som helhet måler én underliggende egenskap.

5.3.2 «Generic Skills Assessment variable map» for MCQ

Skalaen vi nå presenterer, er annerledes enn den vi har sett på tidligere. Beregningen av studentenes dyktighet og oppgavenes vanskelighetsgrad er ikke er ikke knyttet til studentenes besvarelser av CRT1 eller CRT2. Skalaen er derfor kalibrert med et annet utgangspunkt, og man kan derfor ikke sammenligne dyktighet eller vanskelighetsgrad mellom de foregående og de påfølgende skalaene. I 5.7 er personer og oppgaver plassert på MCQ-skalaen slik programvaren RUMM2030 fremstiller det. Fremstillingen er noe annerledes enn i konsortiets rapport hvor programvaren ConQuest er anvendt (AHELO consortium 2012:84). I forhold til denne har man i RUMM2030 «vippet skalaen ned» så den ligger horisontalt, med økende dyktighet (personer) og vanskelighetsgrad (oppgaver) fra venstre til høyre. De rosa søylene representerer personer og de blå søylene representerer oppgaver.



Figur 5.7: MCQ variable map, også kalt Person – Item map. Rosa søyler representerer personer, blå søyler representerer oppgaver. Stigende dyktighet/vanskelighetsgrad fra venstre til høyre. Den grønne kurven er informasjonskurven.

Som vi ser av figur 5.7, er det en litt dårlig match mellom personene og oppgavene. Det største problemet er at ingen oppgaver har vanskelighetsgrad over 1.6 logit, mens en vesentlig andel av personene (omtrent 18 %) har dyktighet over dette. Med andre ord skulle man gjerne hatt flere oppgaver av høyere vanskelighetsgrad for å gi presise estimat av disse personenes dyktighet.

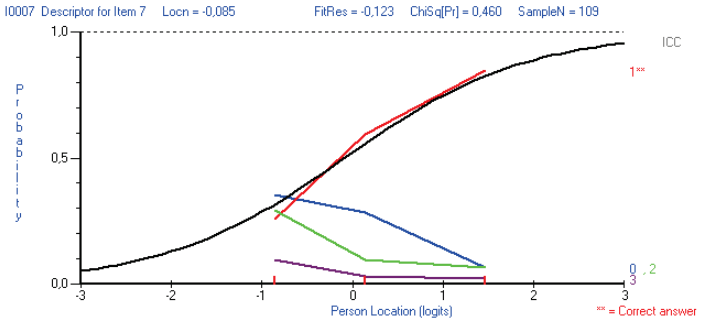
Den grønne kurven i figuren illustrerer dette. Det er den såkalte informasjonskurven, som har en topp hvor man også har flest oppgaver. Man får mest informasjon om en person gjennom oppgaver i passende vanskelighetsgrad, og informasjonskurven viser altså i hvilket område av skalaen man har mest informasjon. Kurven er lav i siktet fra 2 logit og oppover ettersom det ikke er noen oppgaver her. All informasjon man får om personene i dette sjiktet, er gjennom oppgaver som er lette for dem i større eller mindre grad. Vi ser blant annet at den ene søylen stikker over kurven. Dette er en visuell måte å kommunisere resultatet på: Vi har for lite informasjon om de flinkeste studentene, og trenger flere vanskelige oppgaver.

5.3.3 «Item fit» for MCQ

Raschmodellen er, som nevnt, en probabilistisk modell. Med utgangspunkt i en persons mengde av en egenskap, gir modellen sannsynligheten for at personen klarer en oppgave knyttet til denne egenskapen. Dermed kan vi tegne opp en graf som viser «sannsynligheten for at en person klarer en oppgave» på y-aksen, gitt «hvor mye av egenskapen (hvor mange logits) personen har» på x-aksen. I figur 5.8 er Raschmodellens estimat gitt av den heltrukne linjen («Item Characteristic Curve», ICC). For å se hvor godt oppgavene passer modellen, noe som kalles «item fit», ser man hvordan de

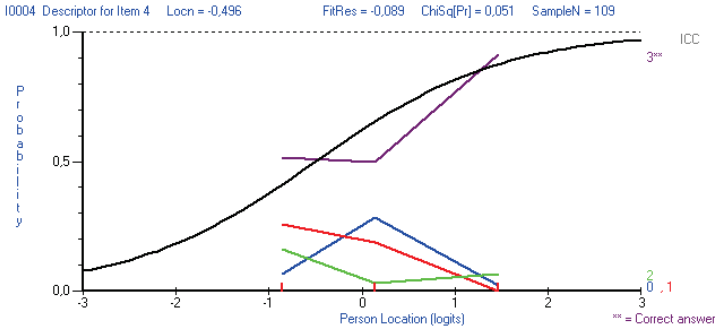
observerte verdiene for personer med ulike grader av dyktighet plasserer seg i forhold til ICC-en. God «item fit» har man om de observerte verdiene til det korrekte svaralternativet ligger tett til den heltrukne linjen.

De fleste oppgavene i MCQ ser ut til å fungere tilfredsstillende. For å illustrere dette, ser vi i 5.8 ICC-en til det som er oppgave 7. De observerte verdiene viser at oppgave 7 passer godt til Raschmodellen: Linjen som viser de observerte verdiene for det korrekte svaralternativet (rød linje) ligger tett til modellens ICC. De observerte verdiene for de alternativene som er feil («distraktorene») viser at sannsynligheten for å svare feil synker med økende dyktighet.



Figur 5.8: ICC for oppgave 7 i MCQ. Oppgaven ser ut til å fungere godt. Sannsynligheten for korrekt besvarelse (rød linje) ligger nært modellens estimat (svart kurve). Sannsynligheten for å velge en av distraktorene (svaralternativene som er feil: blå, grønn og lilla linje) synker med økende dyktighet.

To oppgaver ser ikke ut til å fungere like godt som de tolv øvrige for de norske studentene: Oppgave 4 og oppgave 8. ICC-en til oppgave 4 er gitt i 5.9 og ICC-en til oppgave 8 er gitt i 5.10.



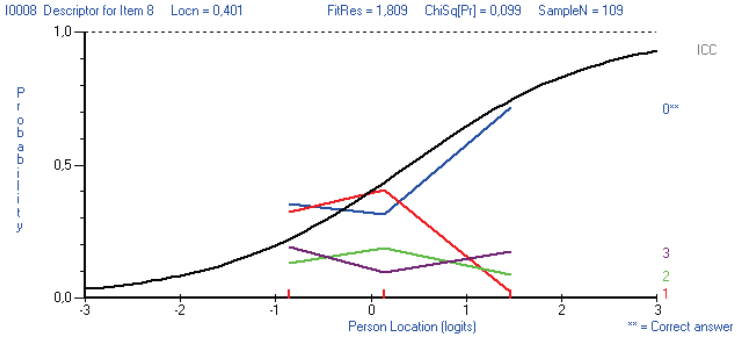
Figur 5.9: ICC for oppgave 4. Oppgaven fungerer ikke helt som den skal, da sannsynligheten for korrekt besvarelse (lilla linje) synker fra -0.9 logit til 0.1 logit.

Figur 5.9 viser hva som er problemet med oppgave 4: Flere med dyktighetsgrad -0.9 logit svarte riktig enn de med dyktighet 0.1 logit. Dette avviket fra modellen fører til at den lilla linjen (de observerte verdiene) «bommer» på den svarte kurven (modellens estimat). Her fins det ingen fasit for hva som er akseptert avvik, men dette vurderes her som problematisk. Man bør se nærmere på oppgaven og undersøke hva som kan være årsaken til avviket.

En mulig årsak til avviket ser ut til å være knyttet til en av distraktorene: Den blå kurven stiger betraktelig fra -0.9 logit til 0.1 logit. Dette er et galt svaralternativ som lokker til seg flere middels flinke studenter (gjennomsnitt 0.1 logit) enn svake studenter (gjennomsnitt -0.9 logit). Slike distraktorer bør man se nærmere på, ettersom man ønsker oppgaver hvor dyktighet gir seg utslag i lavere sannsynlighet for svar som er feil.

Figur 5.10 viser at vi har den tilsvarende situasjonen for oppgave 8: Flere med dyktighetsgrad på -0.9 logit svarte riktig enn de med dyktighet 0.1 logit. Også her er det tydelig at en av distraktorene kan være opphav til noe av avviket fra modellens estimat. Svaralternativ 1 (den røde distraktoren) er mer sannsynlig enn det korrekte svaralternativ 0 (den blå linjen) for personer med 0.1 logit dyktighet. Dette

er ikke et problem i seg selv, da det simpelthen kan indikere at oppgaven er vanskelig. Men det er et problem når situasjonen er motsatt for personer med -0.9 logit i dyktighet. Disse har større sannsynlighet for å svare riktig enn å svare alternativ 1. I denne oppgaven blir man dermed «straffet» for å være middels flink fremfor å være blant de svakeste studentene. Det er en uønsket egenskap ved en oppgave, og man bør derfor se nærmere på svaralternativene.



Figur 5.10: ICC for oppgave 8. Oppgaven fungerer ikke som den skal, da sannsynligheten for korrekt besvarelse (blå linje) synker fra -0.9 logit til 0.1 logit. En av distraktorene (rød linje) ser ut til å være en medvirkende årsak til avviket.

5.3.4 «EAP/PV Reliability estimates»

Testens evne til å rangere personer etter grad av dyktighet er gitt i rapporten av «EAP/PV Reliability estimates», som er programvaren ConQuest sin måte å rapportere dette på. RUMM2030 (programvaren vi bruker) har et tilsvarende estimat som kalles Person Separation Index (PSI). I dette navnet ligger også nøkkelen til å forstå hva EAP/PV-reliabilitet forteller oss: Hvor presist klarer testen å skille (derav «separate») personer etter ferdighetsnivå?

For tilfellet med de norske studentenes besvarelser på MCQ-oppgavene oppgir denne programvaren en PSI = 0.65. RUMM2030 markerer dette som «reasonable», men det vurderes her som noe svakt.

En årsak til denne svake PSI-en har vi sett i det foregående. I kapittel 5.3.2 så vi at vi hadde få oppgaver som ga oss informasjon om de flinkeste studentene. En konsekvens av dette er at testen har større usikkerhet når den skal rangerer de flinke personene etter ferdighetsnivå. Dette gir lavere PSI enn ønskelig. Det er grunn til å tro at man vil kunne øke denne PSI-verdien ved å inkludere flere vanskelige oppgaver.

5.3.5 «Differential item functioning (DIF)»

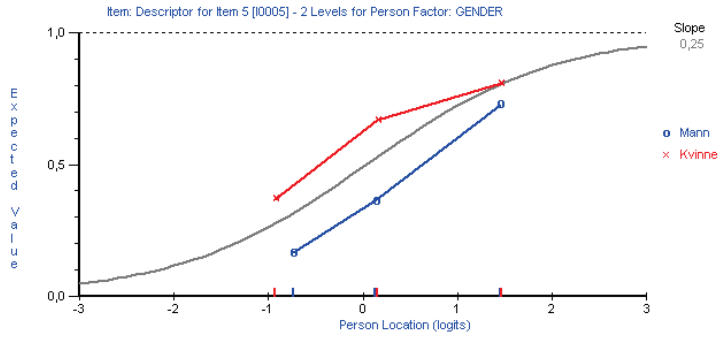
Vi skal til slutt se om like flinke studenter fra ulike grupper skårer likt på oppgavene i MCQ, det vil si om *oppgavene fungerer likt* for ulike grupper. I internasjonale undersøkelser er det for eksempel vanlig å gruppere etter kjønn, fagområde og nasjonalitet. Særlig er det viktig å undersøke «DIF for nasjonalitet» i undersøkelser som AHELO: Fungerer oppgavene forskjellig i ulike land?

Det at «oppgaven fungerer forskjellig» henviser til situasjoner hvor forhold ved en gruppe gjør at de har mindre eller større sannsynlighet til å besvare oppgaven korrekt – forhold som ikke har med deres ferdighetsnivå å gjøre. Et eksempel på dette kan være matematikkoppgaver som er knyttet til skiløping. Elever i Norge har større sannsynlighet enn elever i Spania til å besvare slike oppgaver korrekt. Dette har de ikke fordi de nødvendigvis er flinkere i matematikk, men fordi de har mer erfaring med skiløping. Da blir det enklere å visualisere situasjonen, det blir lettere å gjøre anslag basert på erfaring, de kan ha tenkt gjennom problemstillingen før, osv. Oppgaver som «viser DIF for nasjonalitet» er uegnet når man senere skal sammenligne prestasjoner i ulike land.

En måte å undersøke DIF på, er å inspisere besvarelsene til personer fra ulike grupper som testen for øvrig vurderer til å ha samme dyktighet: Har de fortsatt den samme sannsynligheten for å besvare oppgaven riktig? Med andre ord skal gutter og jenter, nordmenn og spanjoler, ingeniør- og historiestudenter – om de alle er på samme ferdighetsnivå – ha like stor sannsynlighet til å klare en oppgave. Hvis ikke, sier man at oppgaven også «måler kjønn», «måler nasjonalitet» eller «måler

fagområde». Muligheten til å undersøke DIF avhenger av bakgrunnsinformasjonen man har om personene som tar testen.

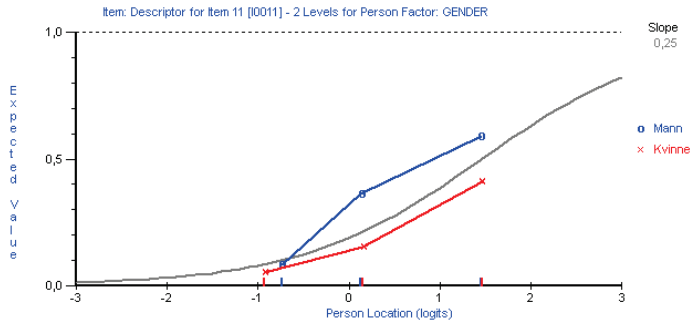
I vårt tilfelle har vi blant annet informasjon om kjønn og fagområde. To oppgaver pekte seg ut i så måte. Figur 5.11 tar for seg oppgave 5, mens figur 5.12 og figur 5.13 tar for seg oppgave 11. I alle tre figurene er forskjellene som fremkommer signifikante.



Figur 5.11: Oppgave 5 viser DIF for kjønn. Kvinner har høyere forventet verdi enn menn med samme ferdighetsnivå.

I figur 5.11 ser vi at oppgave 5 viser DIF for kjønn. Noe med denne oppgaven gjør at kvinner (rød linje) skårer høyere enn menn (blå linje) som er på samme ferdighetsnivå. I et slikt tilfelle bør man undersøke oppgaven og svaralternativene for se om man kan forstå hvorfor. I tilfellet for oppgave 5 er det ikke opplagt hvorfor kvinner skal gjøre det bedre enn menn. Oppgaven handler om å oppfatte mønsteret i en tabell og se for seg hvordan de fire påfølgende kolonnene vil se ut.

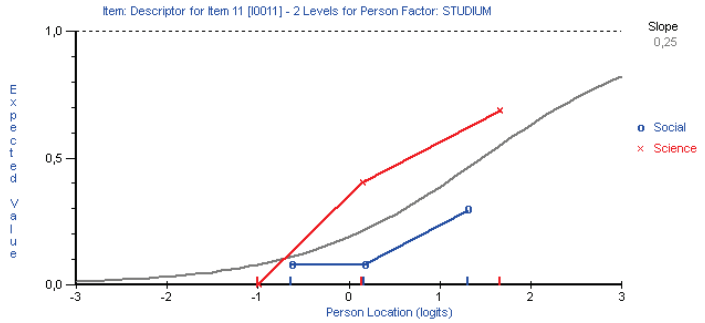
Når man skal undersøke oppgaver for DIF, må man spørre: «Er det noe i tillegg til kompetansen man er ute etter å måle, som hadde hjulpet meg å besvare dette?» Her er det mulig at man bruker ekstra tid på oppgaven på grunn av det visuelle oppsettet av svaralternativene. Tett ved siden av benevningen til svaralternativene (A, B, C og D) står det et ord som er tilsynelatende likt for alle alternativene. Blikket kan derfor falle direkte på de ulike alternativene som står beskrevet et stykke ytterligere til høyre. Ordene man da hopper over er imidlertid ikke like. En bokstav gjør dem forskjellige fra hverandre, og det måtte man være oppmerksom på for at de ulike svaralternativene skal gi mening. En hypotese for utslaget av DIF i oppgave 5, er derfor knyttet til oppsettet av svaralternativene. Dette kan muligens favorisere de som leser grundig. Det hadde vært interessant å teste oppgave 5 med en annen layout på svaralternativene, og se om oppgaven til en viss grad også «måler lesing» – som kvinner presterer bedre i enn menn i Norge.



Figur 5.12: Oppgave 11 viser DIF for kjønn. Menn har høyere forventet verdi enn kvinner med samme ferdighetsnivå.

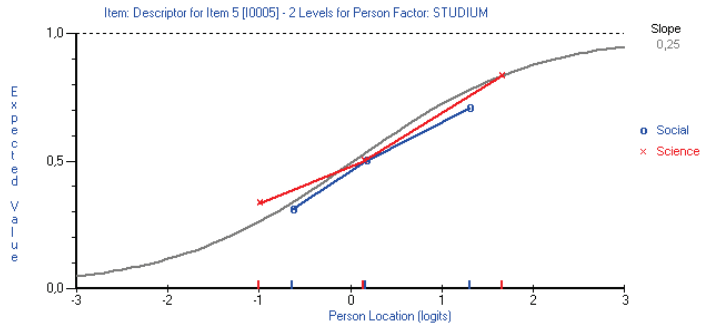
Figur 5.12 viser også DIF for kjønn. Denne gangen gjelder det oppgave 11 og forskjellen er i menns favør (blå linje). Igjen må man se nøyere på oppgaven. Det er grunn til å tro at man i dette tilfellet ikke har å gjøre med en kjønnseffekt, men en effekt av fagområde. Hvorfor? På grunn av den store overvekten menn på studier innen realfag og ingeniørfag. Oppgave 11 handler om tolkning av funksjoner. Dette trener studenter innen realfag og ingeniørfag vesentlig mer på enn studenter i for eksempel sosialvitenskap, økonomi og administrasjon og jus. Om denne hypotesen er sann, må det

også være DIF for fagområde på oppgave 11 i favør av «science». I figur 5.13 ser vi at dette er tilfelle – og forskjellen er *større* enn DIF for kjønn. Dermed er det god grunn til å tro at hele forskjellen man finner i menns favør på oppgave 11 er inkludert i forskjellen man finner i favør av «science». Men igjen: For å være sikker på dette, må man gå tilbake til oppgaven og granske denne.



Figur 5.13: Oppgave 11 viser DIF for fagområde. Studenter i realfag/ingeniørfag («science») har høyere forventet verdi enn studenter i sosialvitenskap/øk.ad/jus («social»). Dette kan være bakenforliggende årsak til DIF for kjønn påvist i figur 5.12.

Hvis det er så at DIF for kjønn på oppgave 11 lar seg forklare av DIF for fagområde, kan det også hende at dette er tilfelle for oppgave 5, som vi så på i figur 5.11. Kanskje er det slik at denne oppgaven favoriserer fagområder hvor kvinner er overrepresentert? Figur 5.14 tyder på at dette ikke er tilfelle. Her finner vi ingen overbevisende DIF for fagområde. Dermed står vi litt mindre rustet til å identifisere hva det er med oppgave 5 som gjør at vi finner DIF enn det vi gjør i tilfellet med oppgave 11.



Figur 5.14: Oppgave 5 viser *ikke* DIF for fagområde. Personene med samme ferdighetsnivå har omtrent samme forventede verdi, uavhengig av fagområdet de tilhører.

5.4 Analyser av de norske data fra AHELO

Selv om de norske dataene inneholder svar fra kun 115 studenter, skal vi her se på i hvilken grad studentene har klart de ulike oppgavene. Vi tar utgangspunkt i kjernespørsmålene i MCQ-oppgavene, totalt 14 spørsmål (ett av de 15 opprinnelige spørsmålene er slettet) som alle testtakere besvarte, og de to CRT-oppgavene, der omtrent halvparten av testtakerne har besvart en av oppgavene.

5.4.1 MCQ-oppgavene

Her presenteres svarfordelingen til de norske studentene på MCQ-oppgavene i AHELO. Ettersom det er 115 respondenter i utvalget, er prosenttallet som oppgis omtrent det samme som det faktiske antall studenter som svarer dette. Svaralternativet som er uthevet i grå rute er korrekt svaralternativ.

Ettersom oppgavene i AHELO holdes hemmelig, har vi her ikke mulighet til å beskrive oppgavene i detalj. I det følgende har vi derfor forsøkt å beskrive hva slags ferdigheter oppgavene måler.

Oppgave 1-3: «Leadership»

Disse oppgavene er relatert til ledelse i organisasjoner og bedrifter.. Spørsmålene er lite verbale og handler mest om konseptuell tenkning.

Ulike ledelsesstiler blir presentert. I oppgave 1 skal disse linkes til andre politiske og ledelsesrelaterte ord, i oppgave 2 skal en av de presenterte ledelsesstilene utdypes, og i oppgave 3 skal man gjøre en vurdering av en av de presenterte ledelsesstilene i lys av interpersonelle forhold.

Tabell 5.8: Svarandeler oppgavesett «Leadership»

	A	B	C	D	Svar mangler
Oppgave 1	11 %	**80 %**	4 %	1 %	4 %
Oppgave 2	12 %	**64 %**	13 %	6 %	4 %
Oppgave 3	0 %	**70 %**	17 %	9 %	4 %

Oppgave 4-6: «Changing the crop»

Disse oppgavene skal ta for seg en praktisk og prosedural aktivitet. Spørsmålene krever logiske resonnement vedrørende fysiske konsepter.

Oppgaven tar utgangspunkt i systematikken i hvordan en bonde planter en åker, og dette blir så fremstilt i en tabell. I oppgave 4, 5 og 6 skal man på bakgrunn av systematikken i tabellen gjøre beregninger og forutse utviklingen i tiden fremover.

Tabell 5.9: Svarandeler oppgavesett «Changing the crop»

	A	B	C	D	Svar mangler
Oppgave 4	10 %	12 %	8 %	**65 %**	4 %
Oppgave 5	10 %	**54 %**	20 %	10 %	6 %
Oppgave 6	28 %	19 %	**29 %**	18 %	6 %

Oppgave 7-9: «Globalisation»

Disse oppgavene skal ta for seg forståelse for internasjonale forhold. Spørsmålene er lite verbale og handler om konseptuell og dialektisk tenkning.

En generell vurdering av globaliseringsfenomenet blir presentert. I oppgave 7, 8 og 9 presenteres tre påstander, og studentene skal avgjøre om påstandene er et argument *for* den generelle vurderingen, et argument som er *mot*, et argument som kan gå *begge veier*, eller et argument som er *irrelevant*.

Tabell 5.10: Svarandeler oppgavesett «Globalisation»

	A	B	C	D	Svar mangler
Oppgave 7	20 %	**58 %**	13 %	4 %	4 %
Oppgave 8	**48 %**	21 %	12 %	15 %	4 %
Oppgave 9	**63 %**	17 %	10 %	6 %	5 %

Oppgave 10-13: «Fish eggs»

Disse oppgavene skal ta for seg vitenskapelig tenkning rundt et biologisk spørsmål. Spørsmålene handler om å tolke et eksperiment, lese grafer og tolke datamateriale.

Et eksperiment med fiskeegg blir presentert, og resultatet av eksperimentet blir visualisert med en graf. I oppgave 10 og 11 skal verbal forståelse for grafen komme til uttrykk. I oppgave 13 beskrives en endring i eksperimentet, og studentene skal velge hvilken av fire grafer som beskriver hvordan resultatene nå ser ut. Oppgave 12 har utgått fordi den ikke fungerte godt i flere land.

Tabell 5.11: Svarandeler oppgavesett «Fish eggs»

	A	B	C	D	Svar mangler
Oppgave 10	4 %	5 %	57 %	**28 %**	5 %
Oppgave 11	**29 %**	6 %	6 %	53 %	6 %
Oppgave 13	**51 %**	17 %	14 %	10 %	8 %

Oppgave 14-15: «Rain»

Disse oppgavene skal ta for seg et konkret, fysisk fenomen. Spørsmålene forutsetter prosessering av informasjon og logiske resonnement. Oppgaven er verbal, men ikke verbalt kompleks.

Et mønster for når det regner blir presentert verbalt. I oppgave 14 og 15 presenteres to hypotetiske situasjoner, og ved hjelp av værmønsteret skal man avgjøre situasjonenes varighet.

Tabell 5.12: Svarandeler oppgavesett «Rain»

	A	B	C	D	Svar mangler
Oppgave 14	10 %	10 %	17 %	**54 %**	9 %
Oppgave 15	3 %	6 %	**73 %**	9 %	10 %

5.4.2 Hva slags MCQ-oppgaver gjorde de 115 norske studentene det best på?

Vanskelighetsgraden på de 14 oppgavene presentert over er ikke kjent. Likevel, for å få en ide om hva slags oppgaver studentene gjorde det best på, rangeres her spørsmålene etter hvor mange studenter som besvarte spørsmålene korrekt. I kolonnen til høyre er NIFUs kortfattede oppsummering av hva slags ferdigheter spørsmålene måler.

Igjen må det understrekes at tolkningene av tabellen som følger må være svært usikre. Om mange av de 115 respondentene klarte oppgaver vedrørende «politiske konsepter og argumentasjon» og få av disse klarte oppgaver vedrørende «tolkning av grafer», kan dette henge sammen med at oppgavene vedrørende tolkning av grafer faktisk er vanskeligere enn oppgavene gitt for politiske konsepter og argumentasjon.

Tabell 5.13: Svarandeler, etter oppgave og ferdighet som måles

Oppgave	% korrekt	Oppgavesett	Ferdigheter som måles
1	80 %	Leadership	Politiske og ledelsesrelaterte konsepter
15	73 %	Rain	Logisk resonnement, system presentert verbalt
3	70 %	Leadership	Politiske og ledelsesrelaterte konsepter
4	65 %	Changing the crop	Logisk resonnement, system presentert ved tabell
2	64 %	Leadership	Politiske og ledelsesrelaterte konsepter
9	63 %	Globalisation	Politiske konsepter og argumentasjon
7	58 %	Globalisation	Politiske konsepter og argumentasjon
5	54 %	Changing the crop	Logisk resonnement, system presentert ved tabell
14	54 %	Rain	Logisk resonnement, system presentert verbalt
13	51 %	Fish eggs	Tolkning av grafer
8	48 %	Globalisation	Politiske konsepter og argumentasjon
6	29 %	Changing the crop	Logisk resonnement, system presentert ved tabell
11	29 %	Fish eggs	Tolkning av grafer
10	28 %	Fish eggs	Tolkning av grafer

Tabellen viser at de 115 norske studentene svarte best på oppgaver vedrørende politiske og ledelsesrelaterte konsepter og svakest på oppgaver vedrørende tolkning av grafer. Her kan det være

tilfelle at oppgavene vedrørende tolkning av grafer var vanskeligere formulert enn oppgavene vedrørende politiske og ledelsesrelaterte konsepter. Likevel kan det hende at dette gjenspeiler de faktiske egenskapene ved de 115 studentene: At disse er relativt sett er har større kompetanse innen politiske og ledelsesrelaterte konsepter enn innen tolkning av grafer.

5.4.3 «Lake-to-river» (CRT1) og «Catfish» (CRT2)

De åpne CLA-oppgavene ble skåret av to uavhengige personer etter de tre dimensjonene som vi her kaller «logisk tenkning», «problemløsning» og «skriveferdighet». Tabellen som følger viser hvor mange prosent av de som besvarte de to CLA-oppgavene som fikk en bestemt gjennomsnittskår på de ulike dimensjonene. Ettersom det er 55 og 60 personer i de to gruppene, tilsvarer én person nesten 2 %. Merk at den nederste linjen i tabellen ikke gir et prosenttall, men viser gjennomsnittlig antall poeng for denne gruppen på denne dimensjonen.

Tabell 5.14: Gjennomsnittsskår på de to CRT-oppgavene (case-oppgavene)

	Lake-to-river			**Catfish**		
Poeng	**Logisk tenkning**	**Problem-løsning**	**Skrive-ferdighet**	**Logisk tenkning**	**Problem-løsning**	**Skrive-ferdighet**
0,0	0 %	0 %	0 %	2 %	0 %	2 %
0,5	0 %	0 %	0 %	0 %	0 %	0 %
1,0	3 %	3 %	7 %	4 %	4 %	2 %
1,5	2 %	5 %	3 %	0 %	2 %	6 %
2,0	10 %	3 %	12 %	18 %	11 %	13 %
2,5	7 %	13 %	10 %	0 %	6 %	9 %
3,0	23 %	28 %	28 %	33 %	31 %	29 %
3,5	15 %	7 %	12 %	11 %	10 %	9 %
4,0	20 %	20 %	18 %	22 %	24 %	24 %
4,5	13 %	16 %	5 %	6 %	7 %	6 %
5,0	7 %	3 %	5 %	4 %	4 %	2 %
5,5	0 %	0 %	0 %	0 %	2 %	0 %
6,0	0 %	0 %	0 %	2 %	0 %	0 %
Gjennomsnittlig antall poeng	**3,38**	**3,31**	**3,07**	**3,17**	**3,21**	**3,05**

Denne tabellen viser oss ulike ting. For det første ser vi at studentene som besvarte «Lake-to-river» fikk noe høyere gjennomsnittskår enn studentene som besvarte «Catfish». Dette kan enten bety at førstnevnte gruppe er flinkere enn sistnevnte, eller at det var lettere å gjøre det godt på førstnevnte oppgave enn sistnevnte. Den mest sannsynlige tolkningen her er at «Catfish»-oppgaven er noe vanskeligere enn «Lake-to-river».

Videre ser vi at forholdet mellom skår på logisk tenkning, problemløsning og skriveferdighet er ganske lik mellom de to CLA-oppgavene: Det er skriveferdigheter de 115 norske studentene skårer svakest på. På «Lake-to-river»-oppgaven skårer studentene omtrent 0,27 poeng svakere på denne dimensjonen enn de to andre, og på «Catfish»-oppgaven skårer de omtrent 0,14 poeng svakere.

Til tross for at de norske studentene skårer noe svakere på skriveferdighet enn på logisk tenkning og problemløsning, ser det ut til at skåren varierer lite mellom disse tre dimensjonene. Dette kan tyde på at dimensjonene henger sammen på en eller annen måte – muligens knyttet til det som her omtales som «generic skills». Ved å se hvordan disse dimensjonene korrelerer, kan man se om denne sammenhengen gjelder på individnivå. For «Lake-to-river» korrelerer dimensjonene slik (Pearson-korrelasjon, signifikante på 0,01-nivå):

- Logisk tenkning – Problemløsning: 0,90
- Logisk tenkning – Skriveferdighet: 0,76
- Problemløsning – Skriveferdighet: 0,82

For «Catfish» korrelerer dimensjonene slik:

- Logisk tenkning – Problemløsning: 0,92
- Logisk tenkning – Skriveferdighet: 0,83
- Problemløsning – Skriveferdighet: 0,87

For begge CLA-oppgavene ser vi at det er en svært høy korrelasjon mellom dimensjonene. Dette vil si at om man vet en students skår på dimensjonen for logisk tenkning, kan man med stor sannsynlighet gjette seg fram til hva studenten skårer på problemløsning og skriveferdighet. Det er altså en tett empirisk sammenheng mellom dimensjonene slik de er operasjonalisert og vurdert i AHELO.

Dette betyr i sin tur at CLA-oppgavene gir lite informasjon til lærestedet om hva studentene må arbeide mer med – det de får av tilbakemelding om sine studenter kan sammenfattes med at «gode studenter gjør det godt på alle dimensjoner» og «svake studenter skårer lavt på alle dimensjoner». Dette er sannsynligvis noe som lærestedet kunne ha gjettet seg frem til, basert på karakterinformasjon om studentene som de får ved opptak. Analysene av korrelasjon mellom skåre og opptakskarakter tidligere i kapittelet viser også at AHELO-testen fanger opp en del av det samme som studentenes opptakskarakter (0,54). Samtidig er det ikke uventet å finne korrelasjon mellom to prestasjonsmål. En korrelasjon på 0,54 er relativt sterk, men uttrykker samtidig en del variasjon mellom de to prestasjonsmålene.

5.5 Oppsummering

Analysene presentert i dette kapitlet er basert på det samlede antallet studenter som deltok i den norske delen av AHELO mulighetsstudie. Siden dette kun utgjør 115 studenter, er det begrenset hvor mye vi kan si basert på dataene. Selv om vi har sett på norske studenters svarmønstre på MCQ-oppgavene er det ikke grunnlag for å trekke noen klare konklusjoner basert på dataene vi har. Analysene viser likevel at dersom man går videre med MCQ-oppgaver og gir samme oppgaver til alle studenter, vil man basert på dataene sannsynligvis kunne gi tilbakemelding til den enkelte om hvilke av ferdighetene som testen måler som studenten er god på, samtidig som man kan gi tilbakemelding til lærestedene om hvordan studentenes kompetanseprofil ser ut – hvilke ferdigheter de er gode på og hvilke ferdigheter de kan videreutvikle.

Sammenligningen av de som har deltatt i AHELO mulighetsstudie med den delen av utvalget som ikke tok testen og med populasjonen, viser at studentene som deltok i testen ikke skiller seg svært mye fra utvalget av deltakere eller populasjonen av tredjeårsstudenter, verken med hensyn til kjønnsfordeling eller karaktergjennomsnitt.

Gjennomgangen av noen av konsortiets analyser av dataene viser at en forholdsvis stor del av variasjonen i hvordan de norske studentene gjør det på CRT-oppgavene forklares av innsats, det vil si av andre forhold enn det testen er ment å måle. MCQ-oppgavene er mindre sårbare for testtakerens innsatsvilje. I tillegg kan det se ut som om MCQ-testen ikke måler det samme som CRT-testene. MCQ kan altså ikke brukes som empirisk anker for de to variantene av CRT. Dermed kan vi heller ikke si om CRT1 og CRT2 måler det samme eller hvordan deres vanskelighetsgrader forholder seg til hverandre.

De psykometriske analysene av de norske dataene viser at de tre dimensjonene som måles i CRT-oppgavene har svært høy korrelasjon. Dermed gir de lite informasjon til lærestedet om hva de bør jobbe videre med: studentenes skriveferdigheter, evne til problemløsning eller kritisk tenking. Videre viser de psykometriske analysene av MCQ at disse fungerer forholdsvis godt, selv om noen av oppgavene må utvikles videre og at oppgavesettet sannsynligvis må utvides med flere vanskelige oppgaver. I så fall bør man jobbe videre med de spesifikke underdimensjonene, slik at man får muligheten til å informere lærestedene om hvor deres studenter har størst utviklingspotensial.

Gjennomgangen av psykometriske egenskaper ved testen handler dypest sett om *validitet*. Man bruker psykometri som et trinn i prosessen med å validere en test. Validitet handler ikke om testen i seg selv, men slutningene man trekker fra resultatene av testen. Har man grunnlag for å si det man sier? Vil det man sier på grunnlag av testen kunne føre til endring i ønskede retninger?

Derfor begrenses ikke diskusjonen om validering av AHELO seg til en gjennomgang av psykometriske egenskaper. Man må også spørre seg om man antar at resultatene av denne undersøkelsen kan føre til endringer i ønskede retninger. En forutsetning for dette er allerede nevnt: Utdanningsinstitusjonene bør få tilbakemeldinger hvor det er presisert hvilke områder studentene gjør det godt på og hvilke områder hvor forbedringspotensialet er størst. Tilbakemeldingen kan dermed ikke komme som en samleskår, men som et mer differensiert mål. Videre er det en forutsetning at man faktisk ønsker å se ting endres på grunn av disse resultatene, og om det er troverdig at dette skjer. Skal ingeniørutdanningen endre innhold om elevene ikke kan nok om det politiske systemet? Og er det troverdig at en matematikkprofessor vil endre undervisningspraksis om hun får vite at hennes studenter ikke kommuniserer godt nok skriftlig? Disse spørsmålene må besvares før man kan snakke om at man har et valid instrument som er hensiktsmessig å utvikle videre.

6 Sammenfattende diskusjon

AHELO mulighetsstudie hadde som målsetting å få undersøkt to forhold. Det ene var å undersøke mulighetene for å kunne konstruere et internasjonalt mål på lærestedenes bidrag til studenters læringsutbytte (*proof of concept*), Det andre var å teste den praktiske gjennomføringen av en slik test i ulike land og ved ulike læresteder (Opheim og Aamodt 2010, Coates og Richardson 2011). I dette kapitlet oppsummeres og diskuteres erfaringene fra gjennomføringen av AHELO mulighetsstudie i Norge.

De praktiske delene av prosessen med å gjennomføre AHELO mulighetsstudie har omfattet mange forskjellige arbeidsprosesser:

- oversettelse og tilpasning av instrumenter, skjermtekster, skåringsrubrikk og surveyer
- rekruttering av læresteder
- trekking av utvalg og rekruttering av studenter
- trekking av utvalg og rekruttering av vitenskapelig ansatte
- praktisk gjennomføring av testingen på lærestedene (inklusive forberedelser)
- testretting/skåring av resultater
- informasjonsflyt mellom alle parter i prosjektet (KD, NIFU, institusjoner, OECD/konsortium)
- kostnadsmessig gjennomføring av AHELO mulighetsstudie i Norge

Resultatet av AHELO mulighetsstudie er at stort sett alle disse prosessene, med unntak av rekruttering av studenter til deltakelse i mulighetsstudien, har fungert godt i Norge. Oversettelse og tilpasning av testinstrumenter, skåringsrubrikker og annen tilhørende tekst har vært en relativt tidkrevende prosess. Dette må ses i sammenheng med valg av test og testinstrument (se kapittel 2). Erfaringene er likevel at relativt komplekse og omfattende testinstrumenter lar seg oversette og tilpasse til en norsk kontekst og gjøres forståelige for testens målgruppe – studentene. Dette gjelder også skåringsrubrikkene, til tross for utfordringer knyttet til formuleringer og ord som ikke enkelt lar seg oversette fra engelsk til norsk. Omfanget av oversettelse var ikke kjent ved prosjektstart, og dermed var heller ikke kostnadene for å oversette og tilpasse testinstrumenter, skåringsrubrikker og annen tilhørende tekst lagt på et tilstrekkelig nivå fra starten av.

Trekking av utvalg av studenter og vitenskapelig ansatte fungerte også godt, og Norge har stort sett gode registre som slike data kan hentes fra. Lærestedenes register over vitenskapelig ansatte er det fortsatt noen utfordringer knyttet til – de er ikke likt strukturert, men alle lærestedene klarte å hente ut informasjonen som var etterspurt, etter de gitte kriteriene, om enn i noe forskjellig form.

Den praktiske gjennomføringen forløp nesten helt uten tekniske problemer. Dette må ses i sammenheng med at lærestedene hadde gjort grundig testing av systemet på forhånd. Lærestedenes vurderinger av prosessen er også i stor grad positive, til tross for forsinkelser i oppstarten av testingen og lav deltakelse fra de uttrukne studentene. Derimot ble det i løpet av gjennomføringsfasen tydelig at

lærestedene og KD hadde ulik oppfatning av kostnadsfordelingen mellom partene. Flere av lærestedene mente departementet burte ha dekket utgiftene til skåring av oppgaver og insentiver til studentene, og at kostnadsnivået for lærestedets deltakelse i prosjektet fremgikk ikke klart da lærestedene sa ja til å være med i AHELO mulighetsstudie. Departementet på sin side, deler ikke denne oppfatningen, men mener avtalen med lærestedene beskriver kostnadsansvar og -fordeling tydelig. Denne type uenigheter om kostnadsfordeling mellom ulike parter er et tema man bør være observant på i en eventuell videreføring av AHELO. Trolig er dette særlig viktig i undersøkelser der flere parter er involvert, mange elementer er uavklart ved starten av prosjektsamarbeidet og hvor ansvar for kostnader knyttet til ulike deler av prosjektet er avtalt på en slik måte at det åpner for ulike oppfatninger.

I et stort internasjonalt prosjekt som AHELO er informasjonsflyt viktig, og dette har fungert godt i Norge. KD har, i egenskap av leder av prosjektgruppen, sett til at alle parter har vært informert om utviklingen i prosjektet, og NIFU har hatt ansvar for å videreformidle informasjon og instrukser til lærestedene med hensyn til gjennomføringen av studien. Organiseringsmodellen man valgte i Norge vil dermed kunne være en god modell også for en fremtidig AHELO. Informasjonen fra OECD og konsortiet har også fungert tilfredsstillende, selv om utfordringer oppsto da de to testinstrumentene fra ulike miljøer skulle integreres. Alle land som deltok i Generic skills strand måtte forholde seg til testinstrumenter fra to ulike miljøer (CAE og ACER), og koordineringen av implementeringen av disse to instrumentene ledde til noen mindre utfordringer knyttet til informasjonsutveksling i den internasjonale delen av prosjektet.

Kostnadsmessig ble gjennomføringen av AHELO mulighetsstudie dyrere enn først antatt, også den norske delen av prosjektet. Dette skyldes flere forhold. Utsettelser i prosjektet, samt mer omfattende oversettelsesarbeid enn antatt, medførte økte kostnader for gjennomføringen av arbeidet i Norge (se kapittel 3 for mer om dette). I tillegg ble kostnaden til OECD for å være med i prosjektet, høyere for Norge enn først antatt. Dette har sammenheng med at OECD måtte endre den opprinnelige finansieringsplanen, da omfanget av finansiering fra private aktører ble betydelig lavere enn forutsatt. Dette medførte økte bidrag fra de deltakende (og også noen ikke-deltakende) landene. Uten å foreta en detaljert sammenligning av kostnader knyttet til ulike internasjonale undersøkelser, bør det likevel nevnes at AHELO mulighetsstudie har hatt et relativt moderat kostnadsnivå sammenlignet med andre OECD-prosjekt som kan være rimelig å sammenligne med, eksempelvis PISA eller PIAAC.

6.1 Rekruttering av studenter – svarprosent

Den absolutt største utfordringen i Norge var svarprosent, det å få studenter til å ønske å delta i AHELO mulighetsstudie. Det var kun 115 av de 1500 uttrukne studentene som deltok, og dermed var svarprosent på under 10 prosent. Her ligger sannsynligvis den aller største utfordringen for en fullskala AHELO – for at AHELO skal ha en fremtid må en tilstrekkelig andel av studentene gjennomføre testen ved de enkelte lærestedene.

Det er vanskelig å fastslå én grunn til hvorfor svarprosenten ble så lav. Det viste seg blant annet at det siste semesteret i bachelorgraden ikke er et optimalt tidspunkt for å rekruttere norske laveregradsstudenter, og utvalgsmetoden som skulle brukes gjorde at det kun var få studenter på hvert program som var trukket ut til å delta. En utvalgsmetode som tar utgangspunkt i kurs eller grupper kan virke positivt på rekrutteringen, men likevel må nok testen endres for å være som noe som norske studenter ønsker å bruke tid på.

AHELO-testen er relativt omfattende, den tar omtrent 2,5 timer å gjennomføre og utfordringen ligger her i studentenes manglende ønsker eller insentiver til å delta snarere enn manglende tilrettelegging eller informasjon fra lærestedenes side. Lærestedene synes å ha jobbet hardt for å rekruttere studenter, med mye informasjon og flere test-tilfeller, men ingen av rekrutteringsstrategiene har vært vellykket med hensyn til å oppnå tilstrekkelig deltakelse fra studentene. Det ser heller ikke ut til at de insentiver lærestedene tilbød var nok, verken utlodding av premie eller betaling i form av gavekort til alle som hadde gjennomført testen ser ut til å ha generert høyere deltakelse.

I tillegg gjorde ikke måten utvalget som skulle delta i AHELO rekrutteringen lettere, siden det var pålagt å bruke et enkelt tilfeldig utvalg. Det er mulig at det hadde vært lettere å rekruttere studenter til

å delta dersom man hadde brukt en form for stratifisert utvalg, og dermed kunne ha inkludert alle studenter som tar et visst kurs. Men siden dette ikke er testet ut vet vi ikke om det ville ha fungert bedre. Det som derimot er klart er at oppgaven med å nå ut med informasjon til de uttrukne studentene blir enklere dersom de er samlet i et kurs. Da får også lærestedene mulighet til å nå studentene direkte, i undervisningssituasjonen, og ikke bare gjennom e-post, sms og ulike former for informasjonsoppslag.

Det finnes flere mulige løsninger på problemet: en vil være å sette av tid til dette i timeplanen, men siden AHELO ikke er inkludert i noen studieprogrammer i Norge kan det ikke gjøres obligatorisk uten at det også gir studentene studiepoeng eller virker tellende i graden de skal oppnå. Uansett kan man stille spørsmål om det er en god løsning å gi studenter studiepoeng for en test som de ikke kan forberede seg til og som ikke inkluderer noe pensum. En annen mulig måte er å gi studentene svært store insentiver. I AHELO mulighetsstudie kunne det virke som om insentiver i størrelsesorden kr 250 for noe som tar drøyt 2 timer ikke var nok til å motivere studentene til deltakelse. Store insentiver for å delta vil dermed fort bli dyrt, fremfor alt dersom det er mange studenter som skal delta i AHELO. I tillegg kommer også diskusjonen om hvem som har ansvar for å betale for insentivet – Kunnskapsdepartementet som er de som fatter beslutningen om Norge skal være med i en fremtidig AHELO eller ikke, eller lærestedet som er de som får tilgang på data om seg selv (gitt at studentdeltakelsen blir stor nok). En tredje mulighet, som kanskje kan virke motiverende for studenter, er om de kan få vite hvordan de har gjort det på testen, det vil si om den gir dem noen individuell tilbakemelding om deres prestasjon. Her er en mulighet at studentene som tar testen får tilbakemelding om sin kompetanseprofil, som kan gi dem en ide om hva de bør arbeide videre med. Det er ikke sikkert at en slik individuell tilbakemelding virker positivt på rekrutteringen, men det hadde i nok i hvert fall gjort det lettere å markedsføre testen overfor studenter.

Men uansett hvilke strategier som brukes for å rekruttere flere studenter til å delta må man også vurdere utvalgsmetodene i lys av rekruttering av studenter. Dersom studentene samples i grupper, slik at man kan ha målrettet markedsføring eller rekruttering overfor de gruppene som er plukket ut til å delta i AHELO så vil dette sannsynligvis gjøre rekrutteringen noe lettere. I AHELO mulighetsstudie var alle studentene plukket ut ved hjelp at enkelt tilfeldig trekking og dette gjorde ikke rekrutteringen lettere, fordi det bare var et fåtall i hver gruppe som var trukket ut.

Hovedpunktet med hensyn til rekruttering av studenter til å delta i AHELO er at dette må gjøres på en helt annen måte i en fullskala AHELO for å kunne fungere, og for at man skal sikre god nok rekruttering av testtakere. I AHELO mulighetsstudien var målet for svarprosent egentlig 75 prosent svar, noe svært få land oppnådde. De land som oppnådde det brukte enten sensus (det vil si de testet alle) eller de gjorde deltakelsen i AHELO mulighetsstudie obligatorisk.

6.2 Hva var det mulig å få ut av data?

Dataene fra AHELO mulighetsstudie har mye mindre verdi enn opprinnelig antatt, siden svarprosenten er så lav at det ikke er mulig å si noe på lærestedsnivå. AHELO har aldri hatt som ambisjon å kunne si noe om norsk høyere utdanning, siden studien er laget for å måle læringsutbytte på institusjonsnivå og ikke på nasjonal nivå og den heller ikke er basert på et tilfeldig utvalg av læresteder. Derimot hadde vi håpet å få data som kunne gi lærestedene noe tilbakemelding, men det er i grunn ikke mulig, siden det er så få besvarelser per lærested. Vi har likevel gjort analyser på hele materialet for å se hvordan svarfordelingene ser ut. Analyser av de MCQ-oppgavene som alle studenter har tatt viser stor variasjon i hvor stor andel som klarer en oppgave, men fordi det er så få studenter i utvalget kan vi ikke konkludere på basis av dette. Analyser av CRT-oppgavene indikerer at den ene oppgaven er noe vanskeligere enn den andre oppgaven, noe som også bekreftes av de analysene som konsortiet har gjort av data fra alle land.

Småskalatesten av case-oppgaven, «cognitive labs» indikerte at studenter med ulik fagbakgrunn angrep oppgaven på forskjellig måte. Dette var derfor noe vi egentlig ønsket å undersøke videre, men fordi det er så få deltakere i AHELO mulighetsstudie var det ikke mulig å forfølge den problemstillingen videre. Derimot er dette noe som også taler imot videre bruk av den typen CRT-oppgave som CLA har

utviklet, siden mesteparten av de oppgavene har en naturvitenskapelig eller kvantitativ vinkling og at dette kan gjør at oppgavene ikke måler studenter med ulik fagbakgrunn på samme måte.

Men det er mulig å gjøre psykometriske analyser av hvordan de to ulike oppgavetypene har fungert på de norske studentene som deltok i AHELO, for å se nærmere på hvordan oppgavene har fungert, om de har klart å måle det man var ute etter. De psykometriske testene, både de som konsortiet har gjort og de NIFU har gjort at CRT-oppgavene fungerer mindre bra enn MCQ-oppgavene, for norske studenter. CRT-oppgavene gir heller ingen klar tilbakemelding til lærestedene om hva studentene er gode eller mindre gode på, siden de tre skårene som gis til de tre dimensjonene oppgaven rettes etter (kritisk tenking, problemløsing og skriveferdigheter) sammenfaller i stor grad. Derimot kan det ligge et potensiale i å videreutvikle MCQ-oppgavene, men det krever da forholdsvis omfattende videre arbeid. Kun den delen av MCQ-oppgavene som alle studenter har tatt slik den er i dag er ikke en fullgod test av studentenes læringsutbytte, dels fordi det er for få oppgaver i settet slik det fremstår i dag men fremfor alt fordi det mangler et rammeverk som sier hvilken type læringsutbytte som oppgavene er ment å dekke. Antallet oppgaver i kjernedelen er lavt, fordi matrisesampling har blitt brukt. Dermed besvarte ikke alle studenter alle oppgaver. Korrelasjonsanalyser viser en forholdsvis sterk sammenheng mellom opptakskarakter og skåren studenten oppnår på testen, både på MCQ-oppgavene, og på samleskåren som konsortiet har regnet ut basert på både case-oppgave og flervalgsoppgave (0,54). At studentenes testskåre korrelerer med deres generelle karakternivå ved opptak til høyere utdanning, er ikke uventet med tanke på at AHELO er utformet som en test for å måle studentenes ferdigheter innen bestemte områder. En korrelasjon på 0,54 er relativt sterk, men uttrykker samtidig en del variasjon mellom de to prestasjonsmålene.

6.3 Utbytte av AHELO i Norge?

Hvilket utbytte har vi hatt av å delta i AHELO mulighetsstudie? Til tross for lav svarprosent og dermed begrensede muligheter til å gjøre analyser av data, har deltakelsen gitt nyttige erfaringer.

AHELO mulighetsstudie har bidratt til et økt fokus på læringsutbytte i høyere utdanning, og er et første forsøk på direkte å måle læringsutbytte på tvers av institusjon, land og kultur. Mulighetsstudien har vist at det i det store og hele er praktisk og teknisk mulig å gjennomføre en slik studie. Derimot viste det seg vanskelig å rekruttere studenter til å delta i studien, og på grunn av lav svarprosent er det dermed vanskelig å si hvordan testene har fungert i Norge og om det er mulig å gjøre valide sammenligninger mellom læresteder basert på en slik test. Det er ikke mulig å analysere data på lærestedsnivå på grunn av få besvarelser, og det begrenser informasjonen vi kan få ut av studien.

En studie som AHELO er helt avhengig av studentenes oppslutning, studentene må ta testen for at lærestedene skal få den informasjonen ut av data som de ønsker. Det blir dermed viktig at testen utformes på en måte som gjør den attraktiv for studenter å delta i. Erfaringene fra mulighetsstudien i Norge tilsier at moderate insentiver ikke er nok til å rekruttere studenter til deltakelse. De deltakende lærestedene arbeidet mye for å spre informasjonen om AHELO, men møtte lite interesse. Blant annet viste det seg at studenter i sitt siste studieår på bachelornivå sannsynligvis er en gruppe som er særlig vanskelig å nå, fordi de i mange tilfeller arbeider med individuelle oppgaver. Videre er det mulig at andre samplingsmetoder eller andre måter å utforme testen på kan gi en høyere svarprosent. Disse forslagene til hvordan svarprosenten kan økes bør bygges inn i designen av en eventuell fremtidig AHELO.

En kommentar fra lærestedene er at de, som en bieffekt av AHELO, har fått teste hvordan det ville være å organisere en elektronisk eksamen på lærestedet, blant annet hva som kreves av forberedelser, tekniske løsninger og lignende. Med tanke på at flere læresteder har brukt ressurser på å teste ut ulike løsninger for elektronisk eksamen (se for eksempel UiA 2011), har AHELO mulighetsstudie fungert som en måte å teste ut en slik løsning i praksis. Siden den tekniske gjennomføringen av AHELO har fungert forholdsvis bra er dette en lærdom de deltakende lærestedene kan ta med seg videre i arbeidet med elektronisk eksamen.

En annen lærdom er at det er mange utfordringer knyttet til å bruke case-oppgaver (CRT). Det er kostnadskrevende å oversette og tilpasse oppgavene, og disse er i tillegg kostnadskrevende å skåre. Siden analysene av dataene viser at de tre dimensjonene som case-oppgavene skåres etter i stor

grad sammenfaller, gir de heller ikke nok informasjon tilbake til lærestedet til at prosjektgruppen ved NIFU kan anbefale å gå videre med en test basert på case-oppgaver.

Dersom Norge skal vurdere å være med i en videre utvikling av AHELO bør en slik test trolig være basert på MCQ-oppgaver. En test av typen som AHELO ønsker å være, krever imidlertid et rammeverk, og et av hovedproblemene med Generic skills strand er at man ikke først ble enig om et slikt rammeverk. Dette er også kommentert av konsortiet og sekretariatet som noe som må gjøres dersom AHELO skal kunne gjennomføres i fullskala (OECD 2012a, Tremblay et al 2012). Prosjektgruppen ved NIFU støtter dette fullt ut. I tillegg må det instrumentet man skal bruke være pilotert, og man må sikre at det inneholder nok vanskelige og enkle oppgaver, slik at man klarer å skille både gode og mindre gode studenter. Analysene her har særlig pekt på behovet for å få inkludert flere vanskelige oppgaver. I tillegg har vi påpekt i analysene av de norske dataene at noen av oppgavene er problematiske. Disse bør man derfor prøve å endre dersom de samme oppgavene skal brukes igjen. Om man skal gå videre med en test av generelle ferdigheter bør man strebe etter å rendyrke de ulike fasettene av generelle ferdigheter som oppgavene som ble brukt i mulighetsstudien bygger på. Slik kan lærestedet få informasjon om hva de bør jobbe med å utvikle hos sine studenter, og studentene kan få tilbakemelding om sin egen kompetanse.

Men hovedutfordringen for AHELO mulighetsstudie i Norge har vært rekruttering av studenter til deltakelse i studien. Dette er ikke et unikt problem for Norge, flere andre land har også svak svarprosent, noe som tilsier at rekruttering bør bygges inn i designet av studien fra starten av, dersom man ønsker å gå videre med AHELO. Dette innebærer at man må arbeide på flere fronter for å gjøre studentdeltakelse i studien mer attraktiv for å sikre en tilstrekkelig god svarprosent.

Referanser

Bond, Trevor G. & Christine M. Fox (2001): *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, N.J: Erlbaum

Coates, Hamish & Sarah Richardson (2011): An international assessment of bachelor degree graduates' learning outcomes, *Higher Education and Management Policy* 23 (3): 51-69

Kjernsli, Marit & Astrid Roe (red) (2010): *På rett spor. Norske elevers kompetanse i lesing, matematikk og naturfag på PISA 2009.* Oslo: Universitetsforlaget

Opheim, Vibeke & Per Olaf Aamodt (2010): AHELO mulighetsstudie: Bakgrunn, innhold og målsettinger. *Norsk Pedagogisk Tidsskrift* 94(6), 440-449.

Tremblay, Karine, Diane Lalancette & Deborah Roseveare (2012): *AHELO Feasibility Study Report, Volume 1: Design and implementation.* OECD: Paris

UiA (2011): *Anbefalinger til UiA sitt videre arbeid med digital eksamen.* Prosjektrapport fra prosjektet «Digital eksamen» Universitetet i Agder, høsten 2011. Tilgjengelig på nettet: http://www.uia.no/portaler/aktuelt/nyhetsarkivet/ny_rapport_om_digital_eksamen

Dokumenter fra OECD

OECD (2007a): Assessing higher education learning outcomes. Summary of a first meeting of experts. EDU(2007)8. Paris/Washington: OECD.

OECD (2007b): Assessing higher education learning outcomes. Summary of the second meeting of experts. EDU(2007)9. Paris: OECD.

OECD (2007c): Assessing higher education learning outcomes. Summary of the third meeting of experts. EDU(2007)14. Paris/Seoul: OECD.

OECD (2008a): *Informal meeting of OECD Education Ministers, Chair's summary*. Web page: http://www.oecd.org/document/45/0,3343,en_2649_39263238_39903213_1_1_1_1,00.html.

OECD (2008b): *Roadmap for the OECD Assessment of Higher Education Learning Outcomes (AHELO) Feasibility Study.* EDU/IMHE/GB(2008)7. Paris: OECD.

OECD (2008c): *OECD Feasibility Study for an Assessment of Higher Education Learning Outcomes (AHELO): Progress report and work plan for 2009-10.* EDU/IMHE/AHELO/GNE(2008)2/FINAL. Paris: OECD.

OECD (2009a): *Feasibility study for the Assessment of Higher Education Learning Outcomes (AHELO): Progress report*. EDU/IMHE/AHELO/GNE(2009)1/FINAL Paris: OECD.

OECD (2009b): *Planning AHELO activities at the national level*. GNE(2009)12. Paris: OECD.

OECD (2010a): *Progress report on Generic skills strand*. EDU/IMHE/AHELO/GNE(2010)2. Paris: OECD.

OECD (2010b): AHELO feasibility study technical advisory group – composition and terms of reference. EDU/IMHE/AHELO/GNE(2010)19. Paris: OECD.

OECD (2011a): *Sampling manual*. October 2011. EDU/IMHE/AHELO/GNE(2011)21/ANN3. Paris: OECD.

OECD (2011b): *National management manual*. November 2011. EDU/IMHE/AHELO/GNE(2011)21 /ANN6. Paris: OECD.

OECD (2012a): *Advanced draft of the feasibility report*. AHELO consortium (ACER) EDU/IMHE/AHELO/GNE(2012)25 Paris: OECD.

OECD (2012b): *Literature review on value added measurement*. EDU/IMHE/AHELO/GNE(2012)27 Paris: OECD.

OECD (2012c): *Generic skills assessment framework*. EDU/IMHE/AHELO/GNE(2012)36 Paris: OECD.

OECD (ikke datert, mottatt september 2011): *OECD AHELO Feasibility Study Generic Skills MCQ description.* Unpublished document.

AHELO consortium (2011): *July Interim Feasibility Report*. Unpublished document.

AHELO consortium (2012): *AHELO Feasibility Study Report*. Unpublished document.

Konkurransegrunnlag, brev og møtereferater

Konkurransegrunnlag: National project manager AHELO feasibilty study in Norway, utlyst av KD juli 2009

Referat fra møte i AHELO prosjektgruppe. 20. september 2011

Brev fra KD til deltakende læresteder i AHELO, 25. oktober 2011

Brev fra Høgskolen i Vestfold til KD, 6. desember 2011

Vedlegg

Vedlegg A: Oversettelser

Oversettelse	Oversettere
Oppgaveteksten og dokumentbiblioteket for de to scenario-oppgavene.	EKVA, UiO
Tekst på databildene som ledsaget testtakeren inn i testen utformet av CLA, samt instrukser til testleder (proctor manual)	Ekstern oversetter i USA*
Tekst på databildene som ledsaget testtakeren gjennom i testsystemet utformet av konsortiet (oppstart, multiple choice spørsmål og kontekstspørsmål for studenter).	cApStAn/BranTra/ NIFU
Multiple choice oppgavene, både ledetekster og spørsmål.	cApStAn/BranTra/ NIFU
Spørreskjema kontekstspørsmål, studenter.	cApStAn/BranTra/ NIFU
Spørreskjema kontekstspørsmål, vitenskapelig ansatte.	cApStAn/BranTra/ NIFU
Spørreskjema kontekstspørsmål, institusjoner.	cApStAn/BranTra/ NIFU

* Fordi kvaliteten på denne oversettelsen var svært dårlig ble vi nødt til å gjøre en ny oversettelse av proctor manual. Denne oversettelsen ble gjort av Allegro, et oversettelsesfirma i Bergen. De gjorde en god jobb, men dette medførte en ekstrakostnad i prosjektet som Kunnskapsdepartementet dekket.

Vedlegg B: Oversikt Cognitive Labs

Summary – Cognitive Labs in Norway

Participating students:

Student	Background	Institution	Sex	Performance task	Time
1	Teacher education	University	Female	Lake to river	1 h 30 min
2	ICT	University	Male	Lake to river	1 h 15 min
3	Pedagogic	University	Male	Lake to river	1 h 15 min
4	Pedagogic	University	Female	Lake to river	1 h
5	Medicine	University	Female	Lake to river	1 h
6	Teacher education	University College	Female	Catfish	1 h 15 min
7	Teacher education	University College	Female	Catfish	1 h 10 min
8	Biochemistry	University	Female	Catfish	1 h
9	Microbiology	University	Male	Catfish	1 h 30 min
10	Natural Sciences	University	Male	Catfish	1 h 10 min

Vedlegg C: CLA-dokumenter for oversettelse/tilpasning

Dokumenter oversatt/tilpasset fra US engelsk til norsk som en del av fase 1:

- Performance Tasks: Catfish and Lake to River (GS.33-GS.34)
- Scoring Rubric (GS.35)
- Cognitive Lab (mini pilots) Materials (GS.37)
- Mini Performance Task for tuning purposes (GS.38)
- Scoring Handbook Charts (GS.39)
- Internet Interface Instructions (student interface) (GS.42)

Dokumenter som skal gjennomgås og/eller oversettes/tilpasses fra US engelsk til norsk som en del av pre-implementeringsfasen:

GS.49_AHELO_PT_Instructions_and_Questions_TESTLANGUAGE
GS.50_AHELO-P-LR-INTL-Document Library_ TESTLANGUAGE
GS.51_Lake-to-River-Document3_English
GS.52_Lake-to-River-Document7_English
GS.53_AHELO-P-CA-INTL-Document Library_ TESTLANGUAGE
GS.54_Catfish-Document3-map_English
GS.55_Catfish-Document7-Chart1_English
GS.56_Catfish-Document7-Chart2_English
GS.57_AHELO_Displays_UI_ TESTLANGUAGE
GS.58_Stopsign_English
GS.59_Student_StartTest_OECD_ TESTLANGUAGE
GS.60_ Student_StartTest_OECD_English_Reference
GS.61_Student_StartTest_Interface_Screenshots
GS.62_Proctor_Interface_OECD_ TESTLANGUAGE
GS.63_Proctor_Interface_OECD_English_Reference
GS.64_Proctor_Interface_Screenshots
GS.65_Human_Scorer_OECD_ TESTLANGUAGE
GS.66_ Human_Scorer_OECD_English_Reference

Vedlegg D: Møter og kommunikasjon

I tillegg til kommunikasjon via epost, har det vært flere møter mellom de ulike partene i AHELO-prosjektet. Nedenfor oppsummeres denne aktiviteten. Vi har her ikke tatt med informasjonsmøter og lignende som KD har tatt initiativ til, ettersom dette har vært rettet mot eksterne parter (kommunikasjon utad) og ikke som en del av AHELO-prosjektet.

Møtetype/tema	Tidspunkt	Arrangør/sted	Deltakere
GNE-møte nr. 3	18-19 nov. 2009	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Bjørnulf Stokvik, Vibeke Opheim
CAE AHELO Generic Strand Meeting	15-18 feb. 2010	CAE/New York, USA	CAE, NPM og eksperter på oversettelse fra deltakerlandene innen Del A: Generic strand. Fra Norge: Are Turmo, Vibeke Opheim
GNE-møte nr. 4	15-16 mars 2010	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Vibeke Opheim
CAE AHELO Generic Strand Meeting	17 mars 2010	OECD/Paris, Frankrike	CAE, representanter fra deltakerlandene innen Del A: Generic skill strand. Fra Norge: Astrid Roe
CAE Representative Site Visit	29-30 juni 2010	NIFU ILS/Oslo	CAE oversettelsesansvarlig: Willy Solano-Flores. NPM og eksperter på oversettelse fra Norge: Astrid Roe, Are Turmo, Inger Throndsen, Elisabeth Hovdhaugen, Tine S. Prøitz, Per Olaf Aamodt, Vibeke Opheim
CAE AHELO Generic Strand Telephone conference	27 sept. 2010	CAE/ skype (telefonmøte)	CAE, Finland, Norge. Fra Norge: Inger Throndsen, Vibeke Opheim
AHELO prosjektgruppemøte	18 okt. 2010	KD/Oslo	KD, representanter fra deltagende norske læresteder, ILS og NIFU.
GNE-møte nr. 5	25-26 okt. 2010	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Olve Sørensen, Vibeke Opheim
OECD AHELO National Project Manager Meeting	27-28 okt. 2010	ACER/Paris, Frankrike	ACER, NPM fra deltakerlandene innen Del A, B og C. Fra Norge: Elisabeth Hovdhaugen (27. oktober)
AHELO update December 2010: Telephone conference	3 des. 2010	ACER/skype (telefonmøte)	Hamish Coates, ACER, NPM fra Kuwait, Egypt, Finland og Norge. Robert Keeley, CAE (NYC), Scott Elliot, CAE (San Fran), Dianne Lalancette, OECD. Fra Norge: Vibeke Opheim
CAE AHELO Generic Strand Telephone conference	2 feb. 2011	CAE/ skype (telefonmøte)	Robert Keeley, CAE (NYC), Scott Elliot, CAE (San Fran). Fra Norge: Vibeke Opheim
AHELO prosjektgruppemøte	15 mars 2011	KD/Oslo	KD, representanter fra deltagende norske læresteder og NIFU
GNE-møte nr. 6	28-29 mars 2011	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Olve Sørensen, Vibeke Opheim, Elisabeth Hovdhaugen
OECD AHELO National Project Manager Meeting	29-30 mars 2011	OECD/Paris, Frankrike	ACER, CAE, NPM fra deltakerlandene. Fra Norge: Vibeke Opheim, Elisabeth Hovdhaugen
GNE-møte nr. 7	01 juli 2011	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Olve Sørensen, Elisabeth Hovdhaugen

AHELO prosjektgruppemøte	20 sept. 2011	NIFU/Oslo	KD, representanter fra deltagende norske læresteder, ILS og NIFU.
OECD AHELO National Project Manager Meeting	23-25 nov. 2011	OECD/Paris, Frankrike	ACER, CAE, NPM fra deltakerlandene. Fra Norge: Elisabeth Hovdhaugen, Rachel Sweetman, Anna Eriksen (lead scorer)
GNE-møte nr. 8	28-29 nov.	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Olve Sørensen
AHELO prosjektgruppemøte	1 des. 2011	KD/Oslo	KD, representanter fra deltagende norske læresteder, ILS og NIFU.
AHELO prosjektgruppemøte	25 jan. 2012	KD/Oslo	KD, representanter fra deltagende norske læresteder, ILS og NIFU.
OECD AHELO National Project Manager Meeting	16-17 mars 2012	Paris, Frankrike	ACER, NPM fra deltakerlandene. Fra Norge: Elisabeth Hovdhaugen
GNE-møte nr. 9	19-20 mars 2012	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Olve Sørensen
GNE-møte nr. 10	17-18 okt. 2012	OECD/Paris, Frankrike	GNE-representanter fra de deltakende landene. Fra Norge: Jan Levy, Elisabeth Hovdhaugen

Nordisk institutt for studier av
innovasjon, forskning og utdanning

Nordic Instititute for Studies in
Innovation, Research and Education

www.nifu.no