Er opptaket til profesjonsstudiet i psykol reliabelt?

Kan vi stole på karaktergivningen ved eksamener i psykologi? Denne studien tar for seg stabiliteten i bedømmelsen ved sensurering.

TEKST
Svein Larsen
Bjørn Helge Johnsen
Ståle Pallesen

PUBLISERT 1. mars 2006

ABSTRACT:
Is the intake procedure to the professional school of psychology fair?
In Norway the system of grading on a scale ranging from 1.0 to 6.0 was substituted for the European/American grading system, ranging from A (Excellent) to F (Failure), as a function of a newly launched reform in higher education based on the Bologna Convention. The present study compares inter-rater reliability of grades given on introductory level exams in psychology before and after the new system was introduced. The data from the last exam before the introduction of the reform, show inter-rater reliabilities over the 15 commissions (two examiners) varying from .67 to .98. The data from the first exam after the inauguration of the reform show equally high intra-commission reliabilities varying from .46 to .93 over 13 commissions. The inter-examiner agreement was highest at the extreme ends of the grade scale.

EMNER reliabilitet karaktergivning sensurering

Innledning

Eksamener på lavere grad er den eneste formen for seleksjon til profesjonsstudiet i psykologi ved norske universiteter. Ved universitetet i Bergen (UiB), hvor det inntil høsten 2004 ikke var adgangsbegrensning til førsteårsstudiet, ble det høsten 2003 registrert noe i overkant av 1000 studenter på studiets første semester. Samtidig er det klart at tilgangen til profesjonsstudiet er sterkt begrenset. Det er bare 36 studenter som tas opp til dette studiet hvert semester ved UiB. Fordi karakterene fra førsteårsstudiet ligger til grunn for opptaket, er det meget viktig at sensureringen av disse eksamener er troverdig. Dominowski (2002) peker på at eksamenssystemer bør være rettferdige, pålitelige og forståelige for studentene. Han understreker også at eksamener bør samsvare med og reflektere kursenes målsettinger (Dominowski, 2002, s. 129).

Tidligere studier

Resultatene fra en meget tidlig studie på feltet som omhandler karakterfastsetting (Engvik, Kvale & Havik, 1970), tydet på at interbedømmer-reliabiliteten mellom tre uavhengige sensorer ved Universitetet i Oslo ikke var tilfredsstillende. I en mer nylig studie ga Raaheim (2000) syv sensorer oppgaven å sensurere femti besvarelser levert av førsteårsstudenter i psykologi. Han rapporterte at interbedømmer-korrelasjonene for disse syv eksaminatorene varierte mellom .55 og .91. I tillegg fant han at det var forskjeller i størrelsesorden 10 tiendedeler eller mer i 45 % av tilfellene. I 6.4 % av tilfellene var spriket mellom laud og stryk. Når han i tillegg spurte fire svært erfarne sensorer om å finne fram til de korrekte eller egentlige karakterene på de aktuelle besvarelsene, fant han at selv disse «supersensorene» ikke alltid var enige om hva den korrekte karakteren skulle være. Dette ledet Raaheim (2000) til å konkludere at sensorene sannsynligvis brukte ulike standarder, eller la vekt på ulike aspekter ved eksamensbesvarelsene under sensurarbeidet, noe han igjen antok kunne forklare den lave interbedømmer-reliabiliteten. En konsekvens av Raaheims studie er at den offentlige debatten ofte har fokusert på at vurderingssystemene er ustabile, og av og til kan man kanskje få inntrykk av at mange, særlig studenter, oppfatter karaktergivningen som å være mer eller mindre tilfeldig.

Karakterskalaen

I den norske universitetstradisjonen har det i mange år vært brukt en numerisk 51-steg-skala ved karakterfastsettingen. Under dette systemet var karakteren 1.0 best, mens karakteren 6.0 var dårligst. Skalaen hadde latinske benevnelser for de ulike karakterene (Se Tabell 1). Karakterer mellom 1.0 og 1.5 (Laudabilis Praeceteris) ble sjelden brukt i samfunnsvitenskapelige fag (inklusive psykologi). Samtidig fikk kandidater som ble bedømt til «immaturus» (som best oversettes med «umoden» og som innebærer stryk) heller ikke oppgitt den eksakte karakteren (se Tabell 1). Karakterer på individuelle eksamener ble gitt med en desimal, slik at eksamenskommisjonene på tiendedelen måtte angi hvilken karakter de hadde gitt. Fordi det endelige vitnemålet skulle uttrykke den samlede karakteren (gjennomsnitt av alle karakterer oppnådd under profesjonsstudiet), inneholdt denne karakteren hundredeler. På denne måten kunne en kandidat oppnå Laud («være laudabel» eller «prisverdig») med karakteren 2.50, mens en annen kandidat kunne få karakteren Haud illaudabilis (neppe uprisverdig) med karakteren 2.51. Dette skillet ble ofte opplevd som svært dramatisk. Det fortelles historier om at dette skillet mellom «den laudable» og den «ikke laudable» kandidaten i enkelte fag kunne få svært alvorlige følger for karriereutviklingen. Tidligere var det også slik at man måtte ha laudabel karakter fra grunnfag for å komme inn på profesjonsstudiet i psykologi.

Tabell 1. Den tradisjonelle og nye karakterskalaen

Den tradisjonelle skalaen (i bruk til og med våren 2003)

Laudabilis Praeceteris:	1.00 – 1.50
Laudabilis:	1.51 – 2.50
Haud Illaudabilis	2.51 – 3.25
Non Contemnendus	3.26 – 4.00
Immaturus	4.01 – 6.00
Den nye skalaen (i bruk fra og med høsten 2003)	
A: Fremragende	Fremragende prestasjon som klart utmerker seg. Kandidaten viser svært god vurderingsevne og stor grad av selvstendighet.
B: Meget god	Meget god prestasjon. Kandidaten viser meget god vurderingsevne og selvstendighet.
C: God	Jevnt god prestasjon som er tilfredsstillende på de fleste områder. Kandidaten viser god vurderingsevne og selvstendighet på de viktigste områdene.
D: Nokså god	En akseptabel prestasjon med noen vesentlige mangler. Kandidaten viser en viss grad av vurderingsevne og selvstendighet.
E: Tilstrekkelig	Prestasjonen tilfredsstiller minimumskravene, men heller ikke mer. Kandidaten viser liten vurderingsevne og selvstendighet.
F: Ikke bestått	Prestasjon som ikke tilfredsstiller de faglige minimumskravene. Kandidaten viser både manglende vurderingsevne og selvstendighet.

I mange år ble det brukt tre sensorer på grunnfagseksamen i psykologi. Ettersom denne ordningen var kostbar, gikk man over til et to-sensorsystem. I våre dager er selv dette systemet under angrep, og ved noen læresteder har man i noen fag innført et system hvor det bare er *en* sensor, ofte faglæreren selv, selv om en-sensor-systemet ikke er innført for introduksjonsstudier i psykologi.

På førsteårsstudiet i psykologi vil vanligvis hver enkelt kommisjon vurdere mellom 20 og 40 besvarelser, og ved Det psykologiske fakultet i Bergen, hvor vi inntil høsten 2004 ikke har hatt adgangsbegrensning, har antallet kommisjoner vært rundt 20. Ved UiB er hver enkelt kommisjon satt sammen av to sensorer.

Nytt karaktersystem

Det nye karaktersystemet ble gjort gjeldende fra og med høsten 2003 ved UiB. Dette nye systemet utgjør en av tilpasningene til Bologna-konvensjonen, og representerer derfor et forsøk på å tilpasse eksamenskarakterene til en internasjonal standard. Bologna-erklæringen omfatter seks hovedpunkter ifølge Nicolaysen (2004), og han skriver at «Gradssystema skal vere samanliknbare, vere jamnlike med omsyn til tid og innsats og vere lett skjønelege for alle; i hovudsak skal alle land bruke eit system som byggjer på to nivå, nemleg lågare og høgare grad; eit system for studiepoeng skal gjere det lett å jamføre og også utveksle delar av eller heile studietilbod; utveksling av lærarar og studentar skal bli lettare, og eit europeisk sams system for å vurdere kvalitet i utdanninga skal etablerast» (s. 9).

I hovedsak skal alle land som slutter seg til konvensjonen, bruke et system som bygger på to nivåer (lavere og høyere grad) og et system for studiepoeng (ECTS). Hensikten med denne internasjonaliseringen er å standardisere det europeiske utdannings- og gradssystemet, noe som vil gjøre studentutvekslinger, lærerutvekslinger og undervisningssamarbeid på tvers av landegrensene lettere. Som Tabell 1 viser, er de nye karakterene bredere enn de gamle (bare seks kategorier i forhold til det gamle systemets 51 (eller mer nøyaktig 31, siden karakterer mellom 4.1 og 6.0 alle ble gitt karakteren «Immaturus»)). I tillegg ser vi fra Tabell 1 at hver karakter (A–F) etterfølges av en kvalitativ beskrivelse av hvilke minimumskrav som skal stilles til den spesifikke bokstavkarakteren. Det er ikke til å undres over at noen av sensorene blir forvirret under det nye systemet, slik det har framstått i media i løpet av året 2004 (Weldeghebriel, 2004). Det er heller ikke rart at mange arbeidsgivere både her hjemme og i utlandet er usikre på hva de nye karakterene betyr (Aftenposten, 2004).

Problemstilling

Vi ønsket å studere interbedømmer-reliabiliteten for lavere grads eksamener i psykologi (førstesemester-eksamener), først under det gamle systemet med tallkarakterer, og dernest etter at det nye systemet med bokstavkarakterer var innført. I motsetning til Raaheim (2000) ville vi studere reliabilitetsproblemet innenfor en økologisk valid situasjon. Dette vil si at vi ønsket å se på hvordan stabiliteten i bedømmelsen var ved reell sensurering. I henhold til Raaheims (2000) tidligere funn ville vi forvente at karaktersettingen var meget ustabil og følgelig ikke reliabel, og dermed heller ikke valid.

Metode

Studie 1: Alle sensorer fra den siste sensuren (våren 2003) under det gamle systemet med tallkarakterer på kurset «PS101: Innføring i psykologi» oppga hvilke karakterer de hadde gitt individuelt (før de hadde konferert med den andre sensoren). De oppga også hvilken karakter som ble den endelige for den enkelte kandidaten i kommisjonen. På eksamen skulle kandidatene besvare tre av fire essayoppgaver. I alt 15 kommisjoner (to sensorer i hver kommisjon), og totalt 512 kandidater var omfattet av denne første delen av studien. Stryk-karakterer (karakterer fra 4.1 og ned til 6.0) blir ikke gitt annen karakter enn «Immaturus». Disse ble derfor kodet som 4.10.

Studie 2: I denne studien fikk vi data fra begge sensorene i 13 av i alt 18 kommisjoner på kurset «PSYK100: Innføring i psykologi» vedrørende kandidatens (totalt 508) bokstavkarakter (etter den nye skalaen). Vi vet ikke hvorfor noen av sensorene (n = 5) ikke sendte inn sine uavhengige bedømmelser selv etter purring. Karakterene ble kodet på følgende måte: «A» ble gitt tallverdien «1», «B» ble gitt tallverdien «2» og så videre. Ved avviklingen av denne første eksamen i «PSYK100» under kvalitetsreformen (høsten 2003) var det også innført en flervalgsprøve som del av eksamen. Flervalgsprøven skal svare for en tredel av eksamensresultatet, mens to essayoppgaver (to av tre essayoppgaver skulle besvares) skal svare for en tredel hver. For å kunne sammenlikne data fra studie 1 og 2 oppga hver sensor i de 13 kommisjonene hvilken karakter de hadde gitt individuelt (før de hadde konferert med den andre sensoren) på de to essayoppgavene samlet.

Statistikk. Alle karakterer ble lagt inn i og analysert med statistikkprogrammet SPSS, versjon 13.0 (SPSS, Inc, 2003). Samsvaret (interbedømmer-reliabiliteten) ble kalkulert og uttrykt som intraklasse-korrelasjoner.

Resultater

Studie 1: Gjennomsnittlig karakter for studenter som besto eksamen, var 2.69 (SD = .43). Totalt strøk 33.4 % av kandidatene. Intraklasse-korrelasjonene i de 15 kommisjonene varierte fra .67 til .98 (se Tabell 2). En ANOVA med kommisjoner som mellomgruppefaktor og de to sensorers karakterer som repeterte målinger i hver kommisjon viste ingen statistisk signifikant effekt av kommisjon (F = 1,614, p = .70). Den gjennomsnittlige intraklasse-korrelasjonen var .89. For de 341 kandidatene som bestod var det bare for 13 (3.8 %) at spriket mellom de to sensorene var ti tiendedeler (en hel karakter) eller mer. For en av de 512 kandidatene (0.2 %) gikk spriket mellom sensorene mellom laud og stryk. I de 186 tilfellene der en sensor hadde stryk, hadde den andre det også i 137 tilfeller (73.7 %). Den gjennomsnittlige karakterdiskrepansen (uttrykt i absoluttverdi) var .22 (SD = .28). Tabell 3 viser en oversikt over karakterdiskrepanser mellom sensorene for ulike karaktergrupper.

Tabell 2. Intraklasse-korrelasjonene ved PS101 eksamen våren 2003 (n = 512).

Kommisjon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Intraklasse-korrelasjon	.91	.87	.67	.90	.91	.97	.88	.74	.87	.98	.86	.91	.98	.95	.84

Tabell 3. Gjennomsnittlige diskrepans mellom de to sensorene (i absoluttverdi) for ulike karaktergrupper ved PS101 eksamen våren 200 (standardavvik i parentes).

Endelig karakter	? = 2.2	2.3-2.5	2.6-3.2	3.3-4.0	4.1-6.0

	(n = 46)	(n = 109)	(n = 144)	(n = 42)	(n = 170)
Gjennomsnittlig diskrepans	.14 (.10)	.20 (.22)	.29 (.28)	.54 (.39)	.09 (.23)

Studie 2: Gjennomsnittskarakteren for dem som bestod eksamen var 2.96 (SD = 1.16). Totalt strøk 25 % av kandidatene. Intraklasse-korrelasjonene i de 13 kommisjonene der vi fikk data fra begge sensorer, varierte fra .46 til .93 (se Tabell 4). Den gjennomsnittlige intraklasse-korrelasjonen var .83. Intraklasse-korrelasjoner skiller seg fra vanlige korrelasjoner (Pearsons) ved at intraklasse-korrelasjonen også påvirkes av forskjeller i nivået mellom bedømmerne. Derfor blir det mer korrekt å bruke denne korrelasjonen for å kontrollere samsvaret i nivået på bedømmingene, ikke bare om karakterforskjellene er identiske hos de to sensorene. En ANOVA med kommisjoner som mellomgruppefaktor og de to sensorers karakterer som repeterte målinger i hver kommisjon viste ingen statistisk signifikant effekt av kommisjon (F = 1,529, p = .110). En enveis variansanalyse med kommisjoner som uavhengig og persentilskåre som avhengig variabel viste heller ingen effekt av kommisjoner (F = 1,465, p = .137).

Tabell 4. Intraklasse-korrelasjonene ved PSYK100 eksamen høsten 2003 (n = 508).

Kommisjon	1	2	3	4	5	6	7	8	9	10	11	12	13
Intraklasse-korrelasjon	.80	.88	.84	.92	.93	.90	.85	.86	.46	.90	.74	.93	.90

For kandidatene som bestod eksamen (n = 372) i de aktuelle kommisjonene, var det i 163 tilfeller (43.8 %) ingen diskrepans i essayvurderingen mellom de to sensorene, i 164 tilfeller (44.1 %) var diskrepansen en karakter, i 36 tilfeller (9.7 %) to karakterer og i 9 tilfeller (2.4 %) tre karakterer. Diskrepanser på fire eller fem karakterer forekom ikke i materialet. Følgelig hadde ingen kandidater fått stryk (F) av en sensor og laud (A eller B) av en annen. I de 146 tilfeller der minst en av sensorene hadde gitt stryk (F) hadde den andre også gjort dette i 68.5 % av tilfellene. Den gjennomsnittlige karakterdiskrepansen (uttrykt i absoluttverdi) var .61 (*SD* = .74).

Tabell 5 viser en oversikt over karakterdiskrepanser mellom sensorene samt gjennomsnittlig persentilskåre på flervalgsprøven for de ulike karaktergruppene. Intraklasse-korrelasjonen mellom persentil-skåren på flervalgsprøven og sensor 1 og 2s vurdering av essayene var i begge tilfeller –.61. Intraklasse-korrelasjonene mellom persentilskåren på flervalgsprøven og endelig karakter var også –.61. Denne korrelasjonen er negativ fordi en høgere persentilskåre uttrykker en bedre besvarelse, mens en lavere tallskåre på bokstavkarakterene, uttrykker det samme.

Tabell 5. Gjennomsnittlige diskrepans mellom de to sensorene (i absoluttverdi) og gjennomsnittlig persentilskåre på flervalgsprøven for ulike karaktergrupper ved PS101 eksamen høsten 2003 (standardavvik i parentes).

Endelig karakter	A (n = 43)	B (n = 86)	C (n = 113)	D (n = 98)	E (n = 32)	F (n = 134)
Gjennomsnittlig diskrepans	.51 (.51)	.70 (.74)	.70 (.72)	.78 (.84)	.81 (.74)	.33 (.62)
Gjennomsnittlig persentilskåre på flervalgsprøven	93.50 (6.6)	77.95 (14.2)	61.23 (18.3)	41.17 (19.4)	28.90 (15.7)	24.66 (20.3)

Diskusjon

Gjennomsnittskarakteren før og etter kvalitetsreformen kan vanskelig sammenliknes direkte da karakterskalaene er forskjellige. Strykprosenten ser imidlertid ut til å være lavere etter kvalitetsreformen. Det er tenkelig at denne forbedringen i strykprosenten reflekterer at skalaen er utvidet og at det derfor er en bredere «latitude of acceptance» etter innføringen av det nye systemet når det gjelder hva som vurderes som bestått. Det er også tenkelig at denne forskjellen er helt tilfeldig. I tillegg kan det være at vanskelighetsgraden på eksamen kan ha vært ulik i de to tilfellene. Etter innføringen av det nye systemet var den gjennomsnittlige intraklasse-korrelasjonen .83, mot .89 før kvalitetsreformen. Denne forskjellen oppfatter vi som marginal. Reliabilitetsindeksene indikerer en høy grad av konsensus mellom sensorene, og er ikke svært forskjelllige fra de funnene Raaheim (2000) rapporterte. På den annen side mener vi at resultatene ikke kan gi grunnlag for en påstand om at sensorer bruker ulike standarder og vektlegger ulike aspekter ved sin bedømming, slik Raaheim (2000) gjorde. Dette skyldes at det for de som oppnådde A i vår studie, ikke fantes noen tilfeller hvor den ene av sensorene ga dårligere enn B. For et par av kommisjonene var imidlertid intraklasse-korrelasjonene noe lave også i det foreliggende materialet.

> «Samlet sett mener vi våre data viser at karakterfastsetting på lavere grad psykologi ved Universitetet i Bergen skjer på en betryggende måte»

Før kvalitetsreformen var det bare i 3.8 % av tilfellene at karakterforskjellen mellom de to sensorene var i størrelsesorden ti tiendedeler eller mer. Denne frekvensen er påfallende lavere enn den tilsvarende (45 %) som Raaheim (2000) rapporterte. Han viste at sprik mellom laud og stryk oppstod i 6.5 % av tilfellene, mens våre data viser at dette kun oppstod i 0.2 % av tilfellene før kvalitetsreformen og i ingen av tilfellene etter kvalitetsreformen. Hvorfor disse funnene er så vidt forskjellige, er vanskelig å avgjøre, men vi tilskriver dette det forhold at vår studie, i motsetning til Raaheims (2000) er gjort i en reell og naturalistisk kontekst. I de sjeldne tilfeller et slik sprik har oppstått,

kan det tenkes at den aktuelle kandidaten er blitt bedømt som å ha levert to delprøver med svært sterk laud, og en delprøve hvor den ene sensoren har gitt ståkarakter (på 3-tallet), mens den andre har strøket den tredje delprøven og således hele besvarelsen, da man må besvare alle delprøver tilfredsstillende for å bestå. Dermed behøver ikke slike store diskrepanser å være alarmerende eller urovekkende. Diskrepansen mellom karakterer etter kvalitetsreformen kan vanskelig sammenliknes med diskrepansen før kvalitetsreformen, men det er i alle fall betryggende å registrere at diskpanser i størrelsesorden laudstryk ikke forekom.

Det er også et markant funn i undersøkelsen at den største enigheten mellom sensorene finnes der hvor de mest ekstreme karakterene er gitt. Når kandidaten får A eller stryker, er sensorene i mindre tvil enn de er ved fastsettingen av andre karakterer. Uttrykt på en positiv måte kan vi si at sensorene er mest enige med hverandre når de gir karakteren A, mens de er nest mest enige når karakteren F blir gitt. Dette kan kanskje best forklares som en konsekvens av skalaene: Når den ene sensoren gir høyeste eller laveste karakter, er det bare mulighet for uenighet hos den andre sensoren i *en* retning på skalaen, noe som minsker sannsynligheten for uenighet i begge endene av skalaen. Likevel er dette funnet særlig viktig, da det i dag er slik at en må ha en overvekt av A fra eksamener i lavere grad psykologi for å komme inn på profesjonsstudiet.

Vi vil også argumentere for at nettopp ved å ha to sensorer som skal komme frem til en konsensus, vil eventuelle idiosynkratiske forhold ved sensorene kunne bli korrigert. Det er også slik at studentenes rettssikkerhet er ivaretatt ved at de har adgang til å klage på vurderingen som er gjort av deres eksamensbesvarelse, og med dette få en annen kommisjon til å gjøre en ny og uhildet vurdering. Innføringen av flervalgsprøve som en deleksamen etter kvalitetsreformen viser at den ser ut til å skille mellom dem som skriver relativt gode og dem som skriver relativt dårlige essays, da det er en rimelig sterk negativ intraklasse-korrelasjon (–.61) mellom essayvurderingene og persentilskårene på flervalgsprøvene. At korrelasjonen er mindre mellom resultatet på flervalgsprøven og på essayoppgavene kan kanskje skyldes at de to formene ikke måler nøyaktig de samme ferdighetene.

Innføringen av det nye karaktersystemet ser ikke ut til å ha gjort karakterfastsettingen mindre pålitelig. Likevel kan det fortsatt være grunnlag for å overvåke situasjonen når det gjelder karakterfastsetting, både i psykologi og i andre fag hvor studentene er avhengig av karakterer for videre studieforløp. Et viktig bidrag i så måte kan være å undersøke interkommisjons-reliabilitet (dvs. hvorvidt ulike kommisjoner som vurderer samme besvarelse, kommer fram til samme konsensusresultat). Samlet sett mener vi våre data viser at karakterfastsetting på lavere grad psykologi ved Universitetet i Bergen skjer på en betryggende måte. Med referanse til spørsmålet stilt i artikkeltittelen kan vi konkludere med at reliabiliteten i bedømmelsen innen kommisjonene er høy. Det er likevel problematisk at data ikke gir grunnlag for å si noe om hvordan de enkelte oppgavene ville blitt bedømt i forskjellige kommisjoner; om nivået på bedømmelsene er valid. Til dette formålet ville en trenge data hvor samme besvarelser var blitt vurdert av ulike sensorer og av ulike kommisjoner.

Svein Larsen

Institutt for samfunnspsykologi

Christiesgt. 12

5015 Bergen

Tlf 55 58 86 28

E-post svein.larsen@psysp.uib.no

Teksten sto på trykk første gang i Tidsskrift for Norsk psykologforening, Vol 43, nummer 3, 2006, side 221-225

TEKST

Svein Larsen, Institutt for samfunnspsykologi, Universitetet i Bergen

Bjørn Helge Johnsen, Institutt for samfunnspsykologi, Universitetet i Bergen

Ståle Pallesen, Institutt for samfunnspsykologi, Universitetet i Bergen

KONTAKT: staale.pallesen@psysp.uib.no

Referanser

Aftenposten (2004). Problemer i utlandet. Aftenposten. Nedlastet 14. juli 2004, fra http://www.aftenposten.no

Dominowski, R. L. (2002). Teaching undergraduates. Mahawa, NJ: Laurence Erlbaum Associates.

Engvik, H., Kvale, S., & Havik, O. E. (1970). Rater reliability of essay and oral examinations. Scandinavian Journal of Educational Research, 14, 195–220.

Nicolaysen, B. K. (2004). Kvalitetsreformasjonen. Dag og tid, 14. august, s. 9.

Raaheim, A. (2000). En studie av interbedømmer reliabilitet ved eksamen på psykologi grunnfag). Tidsskrift for Norsk Psykologforening, 37, 203–213.

SPSS (2000). SPSS for Windows (versjon 12.0) [computer program]. Chicago, IL: SPSS, Inc.

Universitets- og høgskolerådet. (2002) Nasjonal karakterskala – generelle, kvalitative beskrivelser, http://www.uhr.no/utvalg/studie/nasjonalkarakterskala.htm

Weldeghebriel, L. H. (2000). Nye karakterer kan koste studenter jobben. Aftenposten. Nedlastet 14. juli 2004, fra http://www.aftenposten.no