

Diskrimineras kvinnliga lärare i kursvärderingar? Ett fältexperiment från Uppsala universitet

nr 1 2026 årgång 54

Internationell forskning visar att kursvärderingar ger kvinnliga universitetslärare något lägre omdömen, även när ämnesinriktning och kunskapsresultat kontrolleras för. En viktig fråga är om detta återspeglar diskriminering eller andra faktorer. Vi genomförde ett fältexperiment på grundkursen i nationalekonomi A under 2022, när en stor del av undervisningen var online. Mentorerna tilldelades slumpmässigt ett manligt eller kvinnligt namn, medan de faktiska lärarna inte kände till vilken könsidentitet som tillskrevs dem i kommunikationen. I kursutvärderingen ombads studenterna bedöma mentorernas hjälpsamhet, kunskap och svarstid. Resultaten visar inga tecken på att den kvinnliga mentorn bedömdes mer negativt än den manliga i någon av dimensionerna.

Studenters kursvärderingar har under lång tid spelat en central roll för hur universitet organiserar undervisning och hur lärares pedagogiska insatser bedöms (Kulik 2001; Benton and Cashin 2013). I praktiken utgör de även ett informellt meritunderlag: doktorander, postdoktorer och biträdande lektorer skickar i regel med sina undervisningsomdömen vid ansökningar till akademiska tjänster (Moore and Trahan 1998; Keng 2018). Mot den bakgrunden är det besvärande att en växande internationell forskning visar att kvinnliga lärare ofta får lägre omdömen än manliga, även när ämnesinriktning och observerbara kvalitetsfaktorer vägts in (Boring and Ottoboni 2016; Boring 2017; Mengel m fl 2019; Heffernan 2021; Ceci m fl 2023). Detta skulle kunna tyda på att kvinnliga lärare diskrimineras i kursvärderingar, något som stöds av att studentgrupper kan ha vissa sexistiska föreställningar redan när de kliver in i högre utbildning (t ex nationalekonomer i Chile; Paredes m fl 2023).

Större delen av den befintliga litteraturen bygger dock på observationsdata, ofta från stora kursvärderingsdatabaser med data från olika ämnen och universitet. Även om många studier använder sofistikerade metoder för att särskilja diskriminering från kvalitetsskillnader mellan discipliner och inriktningar, kvarstår två grundproblem: studenters deltagande i kurser är inte slumpmässigt, och forskaren kan sällan observera alla relevanta dimensioner av lärarens undervisningskvalitet. Detta är ett generellt och känt problem i all diskrimineringsforskning baserad på icke-experimentella data. Av detta skäl används i regel korrespondens- och auditsstudier (Bertrand and Duflo 2017), dvs experiment där en fiktiv person med ett tydligt gruppattribut (t ex namn) interagerar med subjekten. Ett känt exempel i detta sammanhang är MacNell m fl (2015), som lät lärare i en *online* kurs uppträda

OLA ANDERSSON, MALIN BACKMAN, NIKLAS BENGTSSON OCH PER ENGSTRÖM

Ola Andersson är professor i nationalekonomi vid Uppsala universitet. ola.andersson@nek.uu.se

Malin Backman är doktor i nationalekonomi och kvalificerad handläggare på Arbetsförmedlingen. annamalinheintz@gmail.com

Niklas Bengtsson är professor i nationalekonomi vid Uppsala universitet. niklas.bengtsson@nek.uu.se

Per Engström är professor i nationalekonomi vid Uppsala universitet. Per.Engstrom@nek.uu.se

under antingen kvinnligt eller manligt namn. De fann att den ”kvinnliga” läraren fick sämre omdömen, motsvarande ungefär en halv standardavvikelse.

En utmaning i korrespondensstudier med interaktioner över tid är att den person som ska gestalta en viss stereotyp tenderar att agera i linje med den fiktiva gruppstillhörigheten. Det finns alltså en risk att en anonym *onlinelärare* som ska agera ”som en person som heter Elin” antingen överdriver eller tonar ned könsaspekterna i kommunikationen. Kanske påverkas gestaltningen också av forskarnas förväntningar om vilket effekt som ska detekteras. Även detta problem är någorlunda generellt, och kallas för *experimenter-demand effect* på engelska. Det är ett potentiellt problem med MacNell m fl (2015) då lärarna i den studien själva visste vilket kön de skulle gestalta.

Vår studie replikerade metoden i MacNell m fl (2015), men med förbättrad design och betydligt större urval. Under perioden efter covidrestriktionerna, när undervisningen bedrevs i hybridform, kunde vi skapa en miljö där studenter kommunicerade med ”e-postmentorer” (detta var alltså innan artificiell intelligens slagit igenom), som bar antingen ett manligt eller kvinnligt namn. Det metodologiskt nya i vår studie är att de verkliga lärarna som besvarade frågorna inte visste vilket könstypiskt namn som skulle tillskrivas deras svar. Genom att administrera kommunikationen via en oberoende funktion kunde vi säkerställa att lärarna själva var ”blinda” för sin tilldelade könsidentitet. Detta är, såvitt vi vet, den första dubbelblinda studien av könsbias i kursvärderingar, och den första studien överhuvudtaget i Sverige.

Efter kursens slut fick studenterna, som en del av den ordinarie kursutvärderingen, betygsätta mentorsinsatserna i termer av hjälpsamhet, kunskapsnivå och svarstid. Resultaten är tydliga: vi fann inga statistiskt eller praktiskt betydelsefulla skillnader mellan den kvinnliga och manliga mentorn i någon dimension. Konfidensintervallen är snäva och utesluter de effektstorlekar som rapporterats i tidigare studier. Vi fann inte heller några skillnader mellan hur kvinnliga och manliga studenter värderade lärarna. I slutet på denna artikel diskuterar vi varför vi finner effekter som skiljer sig från viss tidigare forskning. Detaljer återfinns i vår publicerade studie, Andersson m fl (2025).

1. Könsskillnader i kursvärderingar

Kursvärderingar har en lång tradition inom högre utbildning; den första studien kom redan 1927 (Remmers and Brandenburg 1927). Trots sin långa historia är de fortfarande kontroversiella inom akademien. En kritik är att kursvärderingar är lätta att manipulera och inte nödvändigtvis speglar undervisningens kvalitet, vilket gör dem olämpliga för beslut om t ex befordran och lön (Stroebe 2020). Även när kursvärderingar inte har någon direkt roll i formella beslut om befordran påverkar de indirekt karriärut-

vecklingen eftersom de styr matchningen mellan lärare och kurser (Kulik 2001). Kursvärderingar av låg kvalitet kan också i sig själva vara skadliga och nedslående för den som utvärderas (Bates 2015).

En relativt omfattande litteratur visar att kvinnor tenderar att få lägre poäng i kursvärderingar. Hamermesh och Parker (2005) rapporterar att kvinnliga lärare får något lägre betyg (men att denna skillnad samvarierar med lärarens attraktivitet). En nyare studie är Fan m fl (2019), som analyserar över 500 000 kursvärderingar och konstaterar att kvinnliga lärare får något lägre genomsnittliga poäng – även när studenterna ombeds utvärdera själva kursen och inte den specifika läraren. En annan färsk studie från Danmark, med data på över 100 000 studenter, finner ingen övergripande tendens, men visar att manliga studenter tenderar att ge män högre betyg och kvinnliga studenter tenderar att ge kvinnor högre betyg (Binderkrantz och Bisgaard 2024). Analyser av webbplatser av typen RateMyProfessor har gett blandade resultat. Tidigare studier fann inga skillnader mellan könen (Reid 2010; Stuber m fl 2009), medan nyare studier rapporterar lägre betyg för kvinnor (Rosen 2018; Boehmer och Wood 2017; Arceo-Gomez och Campos-Vazquez 2019). En studie av en studentdriven svensk webbplats konstaterade att kvinnliga lärare får sämre omdömen (Karlsson och Lundberg 2012).

Den deskriptiva skillnaden i läraromdömen som påvisats i observationsstudier är inte enorm – 0,1 av en standardavvikelse (Rosen 2018) – och uppstår inte i alla studier. Frågan är om skillnaden återspeglar diskriminering. I denna kontext skulle förekomsten av diskrimineringen vara särskilt intressant eftersom studenterna inte har några personliga intressen på spel när de svarar på anonyma kursutvärderingar. Eftersom kursvärderingsfrågor är retrospektiva och efterfrågar känd information, torde inte statistisk diskriminering vara en viktig mekanism. Det betyder att om utfallen återspeglar diskriminering är den av den smaksbaserade sorten (*taste-based discrimination*). En möjlig förklaring är att studenterna har sexistiska preferenser och vill bestraffa kvinnliga lärare även om det inte ger dem några förväntade personliga fördelar. Jämfört med studenter inom andra ämnesområden finns belägg för att ekonomistudenter hyser sexistiska uppfattningar redan när de börjar högre utbildning (Paredes m fl 2023). Sexism i anonyma nätforum, såsom Economics Job Market Rumors, är välkänd och dokumenterad i Wu (2018). Några liknande attityder har inte uppmätts i Sverige såvitt vi vet. Men om sådana uppfattningar påverkar bedömningen av prestationer kan anonyma studentutvärderingar skilja sig åt mellan manliga och kvinnliga lärare även när deras faktiska undervisning är identisk.

Observationsstudier kan inte helt avgöra vilka delar som är diskriminering och vad som reflekterar skillnader i de domäner som utvärderas. Två tidigare studier använder tentamensresultat för att kontrollera för den faktiska undervisningskvaliteten (t ex Boring 2017 och Mengel m fl 2019), vilket innebär att kvarstående skillnader mellan lärares kön tolkas som bias, alltså diskriminering. Det är dock oklart om betyg är bättre mått på

lärarkvalitet än kursvärderingar. Lärare kan kompensera för låg pedagogisk kvalitet genom att rätta generöst, eller, om de inte själva rättar eller konstruerar tentan, genom en *teaching to the test*-pedagogik. Det finns vissa belägg för att biträdande professorer rutinmässigt gör detta i takt med att *tenure*-beslutet närmar sig (Moore och Trahan 1998; Keng 2018). Man får hur som helst anta att de universitet som använder kursvärderingar vill mäta kvalitativa läraraspekter som inte direkt kan avläsas i studentgenomströmning och tentamensresultat (annars hade de ju inte använt sig av dem).

En lite snårigare konceptuell poäng är att även om vi hade ett objektivt mått på lärarkvaliteten i klassrummet så behöver inte studentutvärderingar nödvändigtvis spegla irrelevant information. Kursvärderingar kan vara informativa om undervisningens externaliteter, som inte återspeglas i tentamensresultat. Exempelvis kan lärare skapa negativa eller positiva externaliteter för studiemiljön eller arbetsbördan för övriga lärarkollegor, doktorander eller kursadministratörer (en typisk kursvärderingsfråga är om läraren behandlade alla studenter likvärdigt oavsett kön, etnicitet, etc). En annan konceptuell poäng är att institutioner kan ha ett egenintresse av att anställa inspirerande, underhållande eller provocerande lärare även vid givna lärandeutfall, om sådana lärare ökar konsumtionsvärdet av högre utbildning och därmed ämnets attraktion. Detta är relaterat till vad som i litteraturen kallas för ”förebildseffekter” (*role model effects*), dvs att studenter blir genuint inspirerade av att läraren är lik en själv. Sådana förebildseffekter brukar framhållas som kraftfulla styrmedel för att nå specifika studentgrupper, men är på sätt och vis en slags kunddiskriminering.

En ideal kontroll i observationsstudier vore ett exakt mått på en professors värdeskapande för studenterna minus de externaliteter som åläggs administration, kollegor och framtida studenter. Sådana kontroller existerar inte, delvis eftersom uppfattningarna skiljer sig åt om vad som är värdeskapande och vad som är kostnader (eller borde vara det). Observationsstudier kan därför underskatta eller överskatta studenters bias mot kvinnliga lärare om dessa abstrakta kostnader och nyttor inte hålls konstanta. Fördelen med korrespondensexperiment är att förväntade professionella kvaliteter inte behöver observeras, eftersom de är identiska mellan behandlings- och kontrollgrupperna genom designen.

2. Fältexperimentet

Institutionen för nationalekonomi vid Uppsala universitet ger varje termin en introduktionskurs i nationalekonomi till ca 250 studenter. Kursen består av föreläsningar och övningar där problemuppgifter löses. Institutionen använder studentmentorer för att leda övningarna. Dessa mentorer är vanligtvis masterstudenter som hjälper studenterna att lösa uppgifter i mindre grupper (”mentorskapet” avser enbart vägledning vid lösning av problemuppgifter och omfattar inte studie- eller karriärvägledning). Under våren 2021, då pandemirestriktionerna hindrade full närvaro på campus, genom-

fördes en del av mentorskapet *online* via e-post. Denna andrahandslösning blev en framgång; studenterna var aktiva med att skicka frågor per mejl och lämnade uppmuntrande omdömen i kursutvärderingarna.

Den påtvingade övergången till textbaserat mentorskap blev så framgångsrik att vi fick idén att återinföra e-postmentorskap även efter det att covidrestriktionerna inte längre begränsade campusundervisningen. Därför meddelade vi våren 2022 studenterna att de kunde mejla vilka frågor som helst om kursinnehållet till den generiska adressen mailmentor@nek.uu.se och att två av institutionens doktorander skulle svara på frågorna. I praktiken bestod e-postmentorerna av endast en doktorand (Malin Backman, kvinna) samt kursens två huvudlärare (Per Engström våren 2022 och Ola Andersson hösten 2022, båda män).

Vi benämner härnäst de personer som faktiskt svarade på mejlen som ”lärare”. För att se till att dessa lärare inte visste om huruvida deras signatur var kvinnlig eller manlig var det endast en administratör som hade tillgång till den generiska e-postadressen på institutionen. Administratören vidarebefordrade studenternas frågor till läraren, som sedan skickade svaret till administratören, som i sin tur slumpmässigt signerade det med antingen Anton (man) eller Elin (kvinna) – två vanliga svenska förnamn. Vi använde parvis blockrandomisering, vilket innebar att mentors namn randomiserades inom varje par av mejl som skickades från mentoradressen.

I de fall studenten hade uppföljningsfrågor randomiserades signaturerna även på dessa. Det experimentella upplägget innebar att en student kunde exponeras för en lång mejltråd med svar från både manliga och kvinnliga mentorer. Den uppfattade könstillhörigheten hos läraren kunde endast utläsas ur namnet i signaturen, i enlighet med tidigare korrespondens- och auditstudier (Bertrand och Mullainathan 2004). Både Elin och Anton är svenska namn som ger en tydlig könssignal. Enligt Statistiska centralbyrån heter ca 37 000 kvinnor Elin som tilltalsnamn (endast 14 män) och 28 500 män heter Anton (endast 11 kvinnor). För ökad trovärdighet hänvisade namnen till verkliga doktorander vid institutionen. Båda gav sitt samtycke till att deras namn användes.

För att mäta förekomsten av bias i kursvärderingen införde vi tre förregistrerade frågor i kursutvärderingen, en för varje mentors kön. Studenterna ombads specifikt att betygsätta Elin och Anton på en femgradig skala vad gäller hjälpsamhet, kunskap och svarstid. Eftersom de olika frågorna hade samma förväntade tecken förregistrerade vi ett samlat utfall för maximal statistisk styrka, definierat som det ovägda medelvärdet av de tre dimensionerna hjälpsamhet, snabbhet och kunskap.

Upplägget gjorde att undervisningsmiljön torde upplevas som helt naturligt ur studenternas perspektiv. Däremot innebar den exakta formuleringen av kursutvärderingsfrågorna en liten avvikelse från institutionens normala praxis. Den vedertagna formuleringen av studentutvärderingar vid Uppsala universitet är komponent/funktion, inte person/egenskap (t ex ”Var föreläsningarna väl förberedda?” i stället för ”Var Ola kunnig?”). Vi

avvek något från dessa riktlinjer för att bättre kunna anknyta till befintlig litteratur. MacNell m fl (2015) använder nio ytterligare dimensioner, men vi kunde inte inkludera så många frågor om mentorer utan att äventyra den naturliga känslan i kursutvärderingen.

3. Skattningsmetodik

Även om behandlingen är randomiserad förregistrerade vi flera estimeringsmetoder för att öka precisionen. Eftersom samma student ombads betygsätta både den kvinnliga och den manliga mentorn är en estimator med studentfasta effekter tekniskt möjlig. Att använda samma students skillnad i betyg mellan kvinnlig och manlig mentor löser problemet att allmänt missnöjda studenter annars kan påverka estimatet på ett sätt som kraftigt minskar precisionen (Uttl och Violo (2021) visar att ett fåtal sådana studenter kan ha påverkat slutsatserna i MacNell m fl (2015)). Även om *fixed-effects*-metoden är mer effektiv vid givna svarsfrekvenser kräver den att studenterna svarar på frågor om båda mentorerna.

Vi förväntade oss att ca 200 studenter skulle delta i kursutvärderingarna då experimentet genomfördes över två terminer. Däremot räknade vi med att bara en delmängd av studenterna skulle interagera med mentorerna och därmed betygsätta dem. Eftersom studenter kunde skicka flera frågor, och varje fråga då behandlades med ett slumpmässigt val av kvinnlig eller manlig mentor, var det vid planeringsstadiet oklart hur många som skulle exponeras för båda könen. Med tanke på dessa osäkerheter beskrev den förregistrerade forskningsplanen tre angreppssätt: en enkel jämförelse av medelvärden, regressionsjusterade estimat med kontroller för studentens ålder, kön och utbildningsprogram, samt en *fixed-effects*-analys. Vårt förregistrerade åtagande var att den regressionsjusterade modellen med kontroller skulle vara vår föredragna metod om färre än 80 studenter betygsatte båda mentorerna; annars skulle vi använda *fixed-effects*-estimering.

4. Resultat

Data och balanstest

E-postmentorerna blev populära: totalt besvarades 765 mejl över två terminer från 203 unika studentadresser, ca 42 procent av alla inskrivna studenter. Av dessa 203 var 116 kvinnor och 87 män. Formuleringarna i studenternas frågor tyder dock på att studenter ofta kontaktade mentorn i små grupper, så den faktiska räckvidden är troligen större än 203 individer. 153 unika adresser hade minst en interaktion med den uppfattade kvinnliga mentorn Elin, och 148 med den uppfattade manliga mentorn Anton; 98 studenter hade minst en interaktion med båda versionerna.

Fördelningen av inkomna mejl var jämnt spridd över de båda terminerna då experimentet pågick. Totalt besvarades 383 mejl av ”Anton” och 382

mejl av ”Elin”. I vår publicerade artikel (Andersson m fl 2025) presenterar vi detaljerade balanstest som visar att behandlings- och kontrollgrupperna är mycket väl balanserade samt att mejlens egenskaper (ämne, längd, tidpunkt m m) inte skiljer sig signifikant mellan de två mentorerna – vilket är precis vad man förväntar sig vid en lyckad randomisering.

Ett centralt resultat är att kvinnliga studenter – givet att de kontaktade någon mentor överhuvudtaget – inte var vare sig mer eller mindre benägna att vända sig till den kvinnliga mentorn ”Elin” än till den manliga mentorn ”Anton” (och vice versa för manliga studenter). Genomsnittlig svarstid och medellängd på svaren var i princip identiska mellan mentorerna. Vi kan heller inte finna någon korrelation mellan den undervisande lärarens kön och signatörens kön. Sammantaget innebär detta att kvaliteten på den skriftliga undervisningen, mätt med dessa indikatorer, statistiskt sett var likvärdig oavsett om mejlen besvarades av ”Anton” eller ”Elin”.

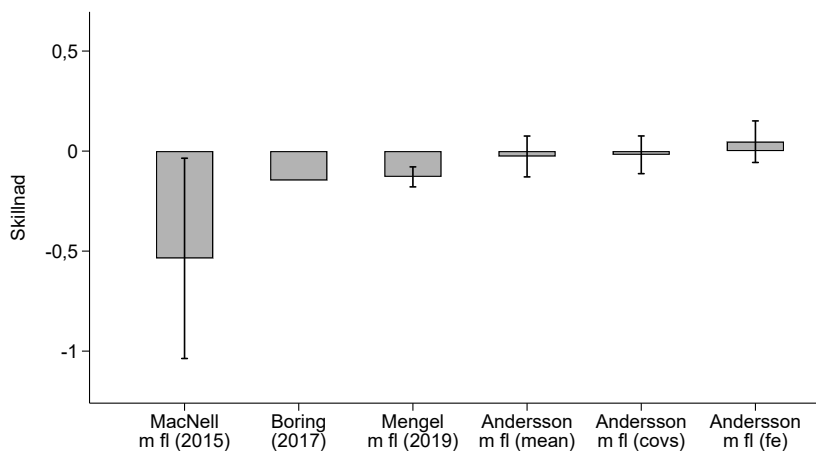
Sammanlagt svarade 253 studenter på kursutvärderingen: 136 kvinnor (54 procent) och 117 män (46 procent). Av dessa svarade 109 på minst en fråga om någon mentor, och 78 studenter betygsatte både den kvinnliga och den manliga mentorn (vilket innebär att den regressionsjusterade modellen med kontroller blev vår föredragna estimeringsmodell).

De studenter som betygsatte mentorerna gav i genomsnitt höga betyg, ca 4,62 på en femgradig skala. Det kan bero på att i praktiken svarade erfarna lärare på studentfrågorna, vilket förmodligen skiljer sig från hur en sådan funktion normalt skulle sättas upp med lärarassistenter. Vi övervägde att medvetet sänka kvaliteten på mentorssvaren inför den andra rundan, men övergav idén av både praktiska och etiska skäl.

Förregistrerad analys

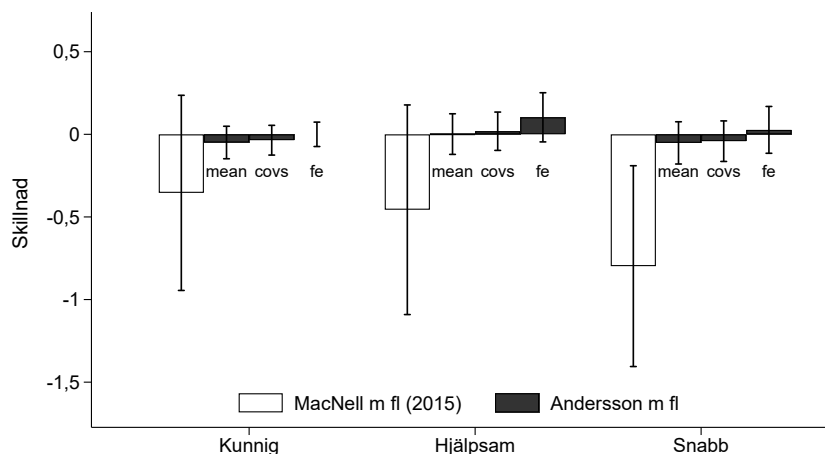
Vårt huvudresultat är att kursvärderingarna inte uppvisar någon bias mot den kvinnliga mentorn. Elin och Anton fick i slutändan nästan identiska kursvärderingspoäng. Vårt huvudresultat, som vi kontrasterar mot befintliga studier, finns i figur 1. Figuren reproducerar de studier som ligger närmast vår i metodik (dvs, de adresserar selektions- och kvalitetsskillnader på något sätt). Det sammanvägda betyget för MacNell m fl (2015) är beräknat av oss med originaldata och baseras på de tre utvärderingsdimensioner som analyseras i denna artikel. För Boring (2017) används ett viktat medelvärde av effektstorlekarna för kvinnliga och manliga studenter samt för vår- och hösttermin (tabell 2 i originalartikeln), viktat efter antalet observationer och omräknat från författarens fyrgradiga till en femgradig betygsskala. Effektstorleken för Mengel m fl (2019) avser *Instructor-related*-betyget i tabell 5 i Mengel m fl (2019) och kommer från den modell som exkluderar interaktionstermen mellan kvinnlig student och kvinnlig lärare (resultat tillhandahållna av författarna). Andersson m fl avser uppskattningarna i denna studie, där *mean* är den enkla ojusterade medelvärdeskillnaden, *covs* är vår föredragna modell med kovariatkontroller och *fe* är *fixed-effects*-modellen med kurs- och termin-*fixed effects*. Samtliga effektstorlekar uttrycks som

Figur 1
Kvinnlig–manlig
skillnad i betyg:
Jämförelse av effekt-
storlekar med tidigare
litteratur



Källa: Andersson m fl (2025).

Figur 2
Kvinnlig–manlig
skillnad i betyg: Jäm-
förelse med MacNell
m fl (2015) – effekt-
storlekar över olika
dimensioner



Källa: Andersson m fl (2025).

skillnaden i betyg (på femgradig skala) när mentorn eller läraren uppfattats som kvinnlig jämfört med manlig, där negativa värden innebär lägre betyg för uppfattat kvinnlig mentor eller lärare.

Avsaknaden av bias drivs inte av motverkande effekter mellan könen eller av skillnader mellan de olika kvalitetsdimensionerna – se tabell 1. De uppskattade effekterna är både statistiskt insignifikanta och nära noll för såväl manliga som kvinnliga studenter. Nollresultatet är också konsekvent över samtliga utvärderade dimensioner (hjälpksamhet, kunskap och svars-tid) – se figur 2.

Vår publicerade artikel (Andersson m fl 2025) innehåller också resultat med ojusterade t-tester samt *fixed-effects*-modeller. Dessa modeller bekrä-

	Samlat	Hjälpsam	Kunnig	Snabb
Alla studenter				
Kvinnlig mentor	- 0,0184 (0,0478)	0,0188 (0,0588)	- 0,0359 (0,0459)	- 0,0408 (0,0622)
Antal observationer	187	186	182	179
Antal studenter	109	108	106	105
Medelbetyg (manlig mentor)	4,621	4,645	4,687	4,514
Std avvikelse betyg	0,689	0,737	0,702	0,810
Kvinnliga studenter				
Kvinnlig mentor	0,00318 (0,0516)	0,0277 (0,0621)	- 0,0350 (0,0409)	0,0168 (0,0752)
Antal observationer	123	122	119	117
Antal studenter	72	71	69	69
Medelbetyg (manlig mentor)	4,738	4,770	4,782	4,641
Std avvikelse betyg	0,486	0,494	0,472	0,636
Manliga studenter				
Kvinnlig mentor	- 0,0404 (0,0958)	- 0,0030 (0,125)	- 0,0192 (0,101)	- 0,112 (0,0973)
Antal observationer	64	64	63	62
Antal studenter	37	37	37	36
Medelbetyg (manlig mentor)	4,396	4,406	4,508	4,274
Std avvikelse betyg	0,892	0,980	0,894	1,024

Tabell 1
Huvudeffekter –
skillnad kvinnlig
mentor jämfört med
manlig

Anm: Varje cell visar resultat från separata regressioner med kontroller för studentens kön, ålder och program. Standardfel klustrade på studentnivå inom parentes. ”Samlat” är ovägt medelvärde av de tre dimensionerna hjälpsam, kunnig och snabb (skala 1–5).

Källa: Andersson m fl (2025).

tar huvudresultatet (noll). Som förväntat är modellerna med kontroller och fasta effekter de mest precisa.

Koncentrationen av betygen kring 4–5 innebär att standardavvikelsen för lärarbetygen är lägre än vi antog i *power*-analysen, vilket gör att konfidensintervallen för den uppskattade behandlingseffekten är snävare än förväntat. Punktuppskattningen för det samlade betyget (*covs* i figur 1) är - 0,0184, vilket motsvarar 0,4 procent av medelbetyget eller 2,7 procent av en standardavvikelse. På 95-procentsnivån är den nedre gränsen i konfidensintervallet -0,112, ca 2,4 procent av medelbetyget eller 15,9 procent av en standardavvikelse – se Andersson m fl (2025). Vi kan alltså förkasta även mycket små effektstorlekar.

Sammanfattningsvis beror vår oförmåga att förkasta nollhypotesen inte på oprecisa uppskattningar i förhållande till tidigare studier – tvärtom är våra konfidensintervall så snäva att de inte överlappar de tidigare studiernas punktskattningar.

5. Ytterligare analys

I den publicerade analysen (Andersson m fl 2025) undersöker vi flera kompletterande mekanismer kring eventuell könsbias i studenters interaktion med e-postmentorerna.

Först testar vi om studenterna förväntar sig mer hjälp av en kvinnlig mentor och därför skickar fler uppföljningsfrågor till henne – något som i observationsstudier utan randomisering skulle kunna dölja en underliggande betygsbias genom att kvinnor tvingas arbeta hårdare för samma betyg. Resultaten visar dock ingen skillnad: varken frekvensen av uppföljningsmejl eller andelen rena nya frågor påverkades av om föregående svar signerats som Elin eller Anton. Detta gäller oavsett studentens eget kön. Vi undersöker också artighetsnivån i uppföljningsmejlen (både med ChatGPT och två mänskliga bedömare) och finner att studenterna var mycket artiga överlag, men att artighetsgraden var exakt densamma oberoende av mentorns uppfattade kön.

När det gäller extern validitet diskuterar vi två potentiella begränsningar. En fråga är om det spelar roll vilken kön den faktiska läraren hade. En hypotes är att tonaliteten och språkbruket i de mejl som undertecknades av en mentor av motsatt kön än den faktiska läraren skulle kunna framstå som apart och udda. För att undersöka detta använder vi att den faktiska (osynliga) lärarens kön varierade mellan terminerna: under höstterminen svarade den kvinnliga doktoranden på ca 80 procent av mejlen, medan andelen var lägre på våren. Trots detta finner vi inga statistiskt säkerställda skillnader i betygsbias mellan terminerna, vilket talar för att resultatet är robust även för olika kombinationer av verklig lärare och uppfattat kön.

En annan fråga om extern validitet handlar om att våra mentorer fick genomgående mycket höga betyg och att vi, till skillnad från vissa tidigare studier (t ex MacNell m fl 2015), saknar starkt missnöjda studenter i datamaterialet. Även denna fråga kan adresseras i viss mån, då olika studentgrupper är olika nöjda överlag. I vår publicerade artikel presenterar vi resultat där vi jämför programstudenter (med höga antagningspoäng och mycket höga mentorbetyg) med fristående kursstudenter (lägre gymnasiebetyg och klart lägre nöjdhet med mentorerna). Även här finner vi ingen interaktionseffekt: nollresultatet består även bland den grupp som var minst nöjd med mentorerna.

6. Slutsatser

I denna studie randomiserade vi manliga och kvinnliga namn på e-postkorrespondens med lärare och undersökte därefter hur studenterna interagerade med och betygsatte dessa lärare i kursutvärderingar. Tack vare randomiseringen hölls den faktiska undervisningsprestationen i genomsnitt identisk mellan könen, vilket innebär att eventuella skillnader i betyg kan tolkas som ren studentbias.

Vi finner inga belägg för att kursutvärderingar missgynnar kvinnliga

lärare. Tvärtom utesluter vårt 95-procentiga konfidensintervall, baserat på det sammanvägda betygsnittet, de effektstorlekar som tidigare rapporterats i litteraturen. Till skillnad från de flesta föregående studier var vårt experiment dubbelblindt och genomfördes som ett randomiserat naturligt fältexperiment. Vi bekräftar också att kvaliteten på mejlsvaren var balanserad mellan de uppfattade könen och att kvinnliga lärare inte mötte högre krav eller en mer krävande undervisningsmiljö i detta sammanhang.

Resultaten är alltså mycket positiva då de tyder på att kursvärderingar kan användas som ett mått på undervisningsprestation utan att riskera systematisk diskriminering av kvinnor. Detta innebär dock inte att vi kan avfärda alla slutsatser från tidigare forskning. Sverige 2025 är ett betydligt mer jämställt samhälle än den amerikanska södern var när MacNell m fl (2015) genomförde sin studie för femton år sedan. De europeiska studier som funnit negativ bias mot kvinnliga lärare (Boring 2017; Mengel m fl 2019) har dessutom genomförts i fysiska klassrum, där "genusdoseringen" – dvs exponeringen för könssignaler via röst, kroppsspråk och utseende – är avsevärt större än i en renodlad digital miljö. Ju fler icke-observerbara könade signaler som finns närvarande, desto större utrymme finns också för bias att uppstå. Å andra sidan uppstår också fler icke-mätbara kvalitetskillnader – vilket ju är skälet till att just *online*experiment framställts som särskilt värdefulla (Ceci m fl 2023). Den samlade kunskapen om var diskriminering äger rum ger inga särskilda skäl att anta att diskriminering är starkare i fysiska miljöer jämfört med digitala.

Slutligen är det värt att påminna om att det finns evidens för att kvinnor inom akademien fortsatt möter nackdelar i flera andra sammanhang, exempelvis vid publicering, anslagstilldelning och befordran, samt att hemarbete och obetalt arbete ofta fördelas ojämnt mellan könen. Det är tänkbart att sådana faktorer indirekt kan påverka kvinnors förutsättningar och prestationer i undervisningssituationer. Våra resultat utesluter därmed inte att könsskillnader i traditionella kursutvärderingar – där de förekommer – delvis kan spegla verkliga kvalitetsskillnader som har sitt ursprung i bredare samhälleliga eller strukturella förhållanden.

Mer forskning behövs därför, gärna med liknande randomiserade fältexperiment i olika länder, undervisningsformer och ämnen, för att vi ska få en mer fullständig bild av när och varför könsskillnader i kursutvärderingar uppstår – och när de inte gör det.

Andersson, O, M Backman, N Bengtsson och P Engström (2025), "Are Economics Students Biased against Female Teachers? Evidence from a Randomized, Double-blind Natural Field Experiment", *Journal of Economic Behavior & Organization*, vol 228, s 1–15.
Arceo-Gomez, E O och R M Campos-Vazquez (2019), "Gender Stereotypes: The Case of MisProfesores.com in Mexico", *Economics of Education Review*, vol 72, s 55–65.

Bates, L (2015), "Female Academics Face Huge Sexist Bias – No Wonder there Are so Few of them", *The Guardian*, 14 februari 2015.

Benton, S L och W E Cashin (2013), "Student Ratings of Instruction in College and University Courses", i Paulsen, M B (red), *Higher Education: Handbook of Theory and Research*, vol 29, Springer, Dordrecht.

Bertrand, M och E Duflo (2017), "Field Ex-

REFERENSER

- periments on Discrimination”, i Banerjee, A V och E Duflo (red), *Handbook of Economic Field Experiments*, vol 1, North-Holland Elsevier, Amsterdam.
- Bertrand, M och S Mullainathan (2004), ”Are Emily and Greg more Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination”, *American Economic Review*, vol 94, s 991–1013.
- Binderkrantz, A S och M Bisgaard (2024), ”A Gender Affinity Effect: The Role of Gender in Teaching Evaluations at a Danish University”, *Higher Education*, vol 87, s 591–610.
- Boehmer, D M och W C Wood (2017), ”Student vs. Faculty Perspectives on Quality Instruction: Gender Bias, ’Hotness,’ and ’Easiness’ in Evaluating Teaching”, *Journal of Education for Business*, vol 92, s 173–178.
- Boring, A (2017), ”Gender Biases in Student Evaluations of Teaching”, *Journal of Public Economics*, vol 145, s 27–41.
- Boring, A och K Ottoboni (2016), ”Student Evaluations of Teaching (mostly) Do not Measure Teaching Effectiveness”, *ScienceOpen Research*.
- Ceci, S J, S Kahn och W M Williams (2023), ”Exploring Gender Bias in Six Key Domains of Academic Science: An Adversarial Collaboration”, *Psychological Science in the Public Interest*, 15291006231163179.
- Fan, Y m fl (2019), ”Gender and Cultural Biases in Student Evaluations: Why Representation Matters”, *PLoS ONE*, vol 14, e0209749.
- Hamermesh, D S och A Parker (2005), ”Beauty in the Classroom: Instructors’ Pulchritude and Putative Pedagogical Productivity”, *Economics of Education Review*, vol 24, s 369–376.
- Heffernan, T (2021), ”Sexism, Racism, Prejudice, and Bias: A Literature Review and Synthesis of Research Surrounding Student Evaluations of Courses and Teaching”, *Assessment & Evaluation in Higher Education*, vol 47, s 144–154.
- Karlsson, M och E Lundberg (2012), ”I betraktarens ögon – betydelsen av kön och ålder för studenters läraromdömen”, *Högere utbildning*, vol 2, s 19–32.
- Keng, S-H (2018), ”Tenure System and Its Impact on Grading Leniency, Teaching Effectiveness and Student Effort”, *Empirical Economics*, vol 55, s 1207–1227.
- Kulik, J A (2001), ”Student Ratings: Validity, Utility, and Controversy”, *New Directions for Institutional Research*, vol 109, s 9–25.
- MacNell, L, A Driscoll och A N Hunt (2015), ”What’s in a Name: Exposing Gender Bias in Student Ratings of Teaching”, *Innovative Higher Education*, vol 40, s 291–303.
- Mengel, F, J Sauermaann och U Zöllitz (2019), ”Gender Bias in Teaching Evaluations”, *Journal of the European Economic Association*, vol 17, s 535–566.
- Moore, M och R Trahan (1998), ”Tenure Status and Grading Practices”, *Sociological Perspectives*, vol 41, s 775–781.
- Paredes, V M, D Paserman och F J Pino (2023), ”Does Economics Make you Sexist?”, *Review of Economics and Statistics*, vol 107, s 1247–1259.
- Reid, L D (2010), ”The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessors.com”, *Journal of Diversity in Higher Education*, vol 3, s 137.
- Remmers, H H och G C Brandenburg (1927), ”Experimental Data on the Purdue Rating Scale for Instructors”, *Educational Administration and Supervision*, vol 13, s 399–406.
- Rosen, A S (2018), ”Correlations, Trends and Potential Biases among Publicly Accessible Web-based Student Evaluations of Teaching”, *Assessment & Evaluation in Higher Education*, vol 43, s 31–44.
- Stroebe, W (2020), ”Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis”, *Basic and Applied Social Psychology*, vol 42, s 276–294.
- Stuber, J M, A Watson, A Carle och K Staggs (2009), ”Gender Expectations and On-line Evaluations of Teaching: Evidence from RateMyProfessors.com”, *Teaching in Higher Education*, vol 14, s 387–399.
- Uttl, B och V C Violo (2021), ”Small Samples, Unreasonable Generalizations, and Outliers: Gender Bias in Student Evaluation of Teaching or Three Unhappy Students?”, *ScienceOpen Research*.
- Wu, A H (2018), ”Gendered Language on the Economics Job Market Rumors Forum”, *AEA Papers and Proceedings*, vol 108, s 175–179.