Utdannings-
direktoratet

Sensorreliabilitet på skriftlig eksamen i videregående opplæring

Rapporten er en kartlegging og analyse av sensorreliabilitet ved skriftlig eksamen, altså i hvilken grad to eller flere sensorer er enige om vurderingen av en eksamensbesvarelse.

RAPPORT | SIST ENDRET: 03.03.2021

Tittel:
Sensorreliabilitet på skriftlig eksamen i videregående opplæring
Rapporten:
Last ned rapport i DUO (PDF)
Forfatter:
Julius K. Björnsson, ILS/EKVA-UIO og Gustaf B. Skar, Skrivesentret-NTNU
Utgiver:
Universitetet i Oslo
År:
2021

Bakgrunn for rapporten

Eksamensgruppa konkluderte at det trengs mer forskningsbasert kunnskap om eksamen.

Utdanningsdirektoratet har også fått i oppdrag å utrede sluttvurderingsordningene i programfagene, og undersøkelsen av sensorreliabiliteten på noen utvalgte skriftlige eksamener i videregående opplæring vil inngå som en del av denne utredningen.

Rapporten er starten på å bygge opp et mer solid kunnskapsgrunnlag om sensorreliabilitet.

Grunnlag for analysene

Analysene er basert på foreløpige karakterer fra de to sensorene som foretok ekstern sensurering. Dette er det beste estimatet vi kan få på sensorreliabilitet, siden det per i dag ikke er mulig å gjøre typen analyser på endelige eksamenskarakterer. Analysene er gjort med utgangspunkt i vurderinger av over 700 000 elevbesvarelser fra årene 2015–2019, i 40 utvalgte fag. I analysen er det brukt flere mål for å sikre en best mulig forståelse av sensorreliabilitet på eksamen.

Hovedfunn

- Sensorreliabiliteten på skriftlig eksamen i videregående opplæring varierer ganske mye.
- Sensorreliabiliteten er i noen fag stabilt høy over tid, mens den i andre fag er stabilt lav.
- Det er ikke statistisk grunnlag for å skille mellom seks nivåer av kompetanse.
- Det mulige tolkningsfellesskapet var tilsynelatende lite i mange fag.

Sensorreliabiliteten varierer ganske mye

I noen fag er sensorreliabiliteten så lav at vi ikke kan utelukke at eksamenskarakteren ikke bare gjenspeiler den kompetansen kandidatene har, men også vel så mye hvilke sensorer som har vurdert besvarelsen. Disse variasjonene har antakeligvis ulike årsaker i forskjellige fag. Dette må derfor utforskes nærmere for hvert fag, slik at passende tiltak kan iverksettes.

Stabilt høy reliabilitet i noen fag over tid, men lav i andre

Sensorreliabiliteten i matematikk og realfag er stabilt høy over tid, mens den er stabilt lav i norsk, engelsk og samfunnsfagene. Noen av de mindre fagene har ganske store variasjoner fra år til år. Dette er i noen tilfeller knyttet til at de har få kandidater.

Ikke grunnlag for å skille mellom seks nivåer av kompetanse

Funn i analysen kan tyde på at eksamen generelt sett er bedre på å skille mellom sensorers strenghet enn kandidaters kompetanse. Analysen viste videre at det ikke er statistisk grunnlag for å skille mellom seks nivåer av kompetanse. I gjennomsnitt klarte eksamen å skille tre nivåer av kompetanse presist nok, men her er det også forskjeller mellom fagene.

Det mulige tolkningsfellesskapet var tilsynelatende lite i mange fag

Analysen viste at det er få overlapp mellom sensorer, noe som gjør det vanskelig å sammenligne kandidater og sensorer på en god måte. For å kunne gjennomføre systematiske undersøkelser av

sensoratferd og koble disse undersøkelsene mot aspekter som eksempelvis sensorskolering, må det faktiske tolkningsfellesskapet utvides betraktelig. En måte å gjøre det på er å la et lite antall besvarelser være felles for alle sensorer i et fag.