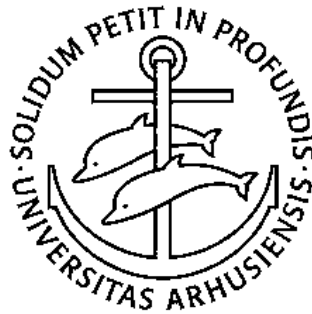


To prøvevalideringsprojekter fra AU

Lektor Lotte O'Neill

Center for Sundhedsvidenskabelige Uddannelser (CESU), AU.



Hvorfor validitet?

- Interesse
- Eksempler på anvendt validitetsteori
- Prøvevaliditet er fundamental
- Er vi kritiske nok i DK?

Teori og retningslinjer

- Kane, M.T. (2006). Validation. (In R.L. Brennan, (Ed), Educational Measurement (pp. 17-64). Westport: ACE/Praeger).
- Brennan, R.L. (2001). Generalizability theory. New York: Springer.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington, DC: AERA.

Projekter

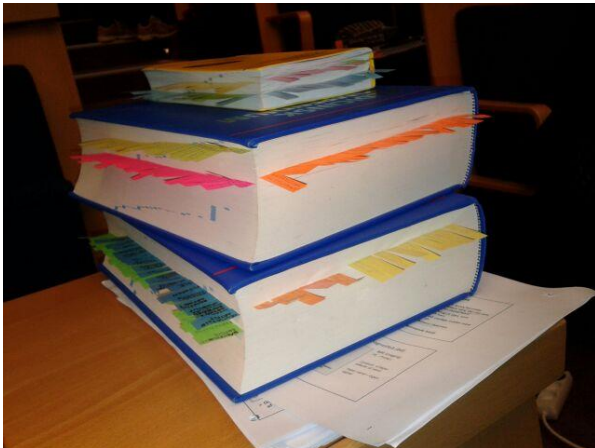
1) En multiple choice eksamen i medicinsk viden med åben bog og fri netadgang



2) En pilottest af nye optagelsessamtaler til medicinstudiet



1) MCQ eksamenerne



- 4 store eks. på ka. (95 ECTS)
- Tester medicinsk/klinisk viden
- MCQ format: One-Best-Answer
- 80 items på 2 timer
- Alle hjælpemidler tilladt

Faktuel viden

('Knows')

Hvad er den hyppigste årsag til trillinge og andre multi-graviditeter?

- A. Baggrundsstråling
- B. Stress
- C. B-vitamin mangel
- D. Kunstig befrugtning

Anvendt viden

('Knows how')

Kurt på 67 år med mangeårig diabetes har aktuelt været til sin årlige status undersøgelse. Du har i den forbindelse konstateret, at han har mikroalbuminuri. Han har således en albumin/creatinin ratio på henholdsvis 35 mg/g og 37 mg/g på to forskellige dage. Hans BT er 120/70 og hans blodsukker er ligeledes rigtig fint med et Hba1c på 6,7.

Hvilken behandling vil du anbefale Kurt?

- A. Fortsatte uændret med kontrol da HbA1c er under 7,0 % og BT < 130/80 mmHG
- B. Henvise til nefrologisk afdeling
- C. Opstarte ACE hammer behandling

Moderne validitetsteori

(Kane 2006)

Validity	Steps
Scoring	from observed performance to observed score
Generalizability	from observed score to universe score
Ekstrapolation	from the universe score to the level of skill
Decision	from conclusion about level of skill to decisions taken

Moderne validitetsteori

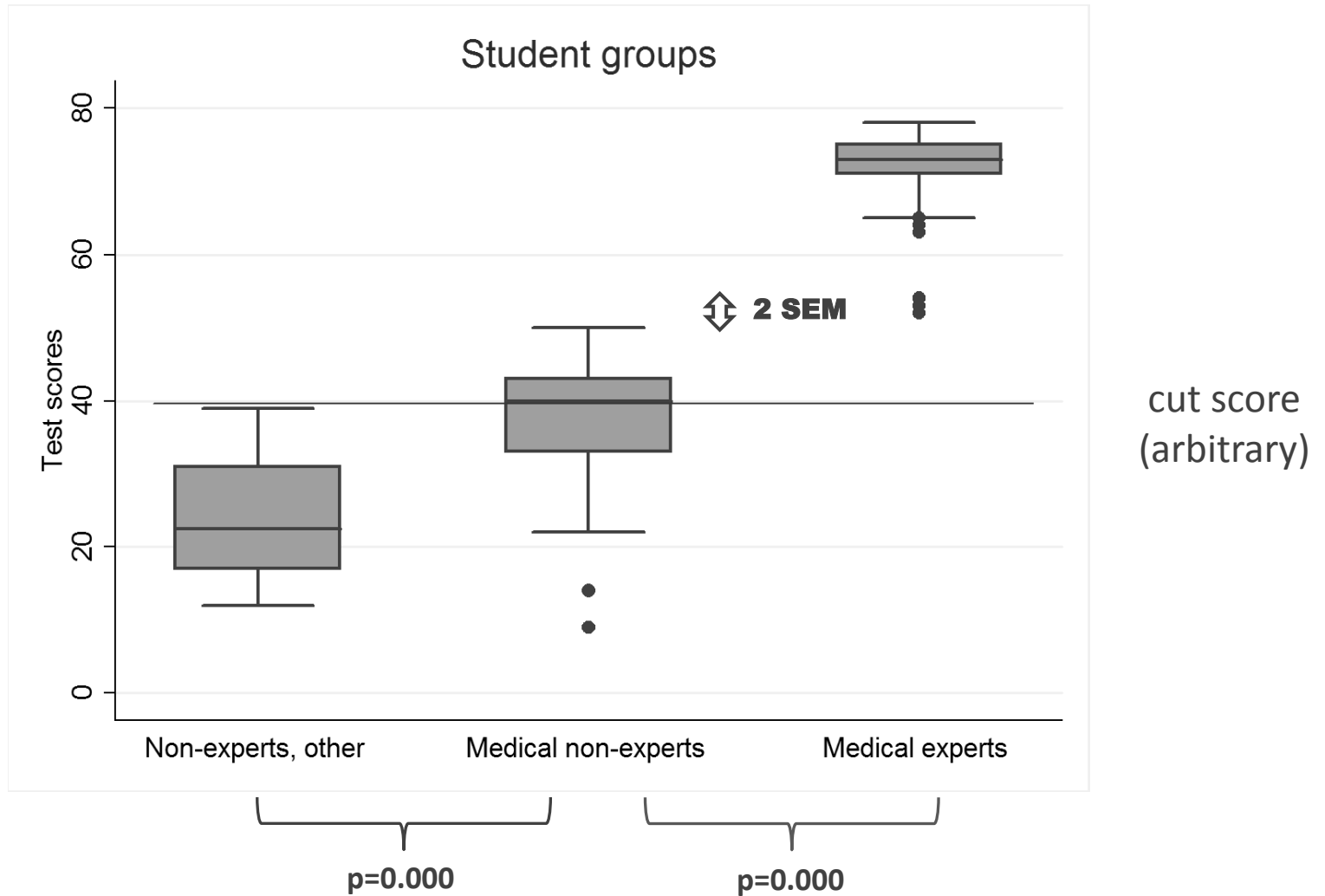
(Kane 2006)

Validitet	<u>ANTAGELSER</u> for mcq-prøverne på medicin AU
Scoring	<ul style="list-style-type: none">• De elektroniske besvarelser registreres korrekt• Svarnøglen er korrekt• Svarnøglen anvendes korrekt og ensartet
Generaliserbarhed	<ul style="list-style-type: none">• Spørgsmålene i eksamenssættet er dækkende for faget• Antallet af spørgsmål er stort nok (stikprøvestørrelse)
Ekstrapolering	<ul style="list-style-type: none">• Vi kan ekstrapolere fra test-scorerne til ekspertiseniveauet• Vi tester ikke andre 'konkurrerende' kompetencer end medicinsk viden (f.eks. opslags-evner i en åben-bog test)
Konsekvenser	<ul style="list-style-type: none">• Studerende med ringe kompetencer består sandsynligvis ikke testen

Kan alle bare bestå en MCQ prøve med åben bog og netadgang?

- Design
 - ‘Known-groups comparison’
- Deltagere
 - 178 ‘experter’ (eksaminander juni 2013)
 - 71 ikke-experter (3. års medicin/uni studerende)
- Materialer
 - Samme test indhold
 - Same ipads & testfaciliteter
- Analyser
 - Levene’s test, quadratic transformation, ANOVA, Tukey’s post hoc test.

Resultater



Konklusioner

- Robusthed mod konkurrerende kompetencer (opslagsevner)
 - 1,5 min/item var ikke nok
- De items som testede faktisk viden (16/79)
 - Hjalp ikke vores ikke-eksperter
 - Skelnede ikke mellem eksperterne
- Hovedtruslen mod validiteten var en uforsvarlig beståscore, men
 - ..forkerte beslutninger (B/IB) var alligevel usandsynlige for 98.3% af eksaminanderne vi undersøgte

2) De nye optagelsessamtaler



- ‘Multiple Mini Interviews’ (MMI) pilot
- 8 adskilte stationer med
- 8 uafhængige opgaver
- 4 test domæner
 - Kommunikation
 - Empati
 - Robusthed
 - Samarbejde
- 1-2 bedømmere/station
- 8 minutter/station

Baggrund

- Nye kvote 2 på medicin fra f16
 - Optagelsessamtaler (MMI)
- Pilot-testet af MMI 13. nov. 2015
- Deltager-feedback fra
 - 26 bach. stud. med'ere
 - 31 regionale læger
 - 4 observatører (CESU)
- Forløb
 - Minimal bedømmer instruktion
 - 2 rul af 2 identiske spor med 8 interview stationer af 8 min
 - Skejby Sygehus gæstekantine

Validitetsantagelser – scoringen

- Scoreark/ankre er egnede
 - Stor aftagerinvolvering i udviklingen (domæner & indhold)
 - Bedømmer survey
- Scoreark/ankre anvendes som anvist
 - Observatør survey & interview
- Scoringen er fri for bedømmer bias
 - $G=0,85$ (8 stationer med 2 bedømmer/st.)
 - $G=0,74$ (8 stationer med 1 bedømmer/st.)

Validitetsantagelser – generaliserbarhed

- Stikprøven af stationer er stor nok
 - G 'på dagen'=0,85
- Stikprøven af stationer er repræsentativ
 - $G=0,50$ for 8 stationer med alternativt prøveindhold
 - $G=0,75$ ville kræve 25 interview stationer!

Validitetsantagelser - ekstrapolering

- True/universe scoren er associeret med effektmålet
 - Scorerne på stationer med samme domæne blev undersøgt for korrelation (ES: lille-moderat)
- Der er ingen systematiske fejl som underminerer ekstrapoleringen
 - MMI scorer og adgangsgivende karakterer, $r=-0,15$, $p=0,452$.

Validitetsantagelser - konsekvenser

- Testens implikationer er rimelige/forsvarlige
 - Deltagernes oplevelse af acceptabilitet
 - Præstationer vs. køn og alder: ingen sammenhænge

Begrænsninger

- De simulerede ansøgers repræsentativitet
- Statistiske begrænsninger
 - 'Restriction of range'
 - Ret lille stikprøve

Fremadrettet

- Validerings-studier f16
 - Hvordan scorer bedømmerne (kvalitative interviews)
 - Hvad er det, vi kan sige at vi bedømmer? (CFA)
 - Konkurrerende faktorer (karakterer, køn, alder)
- Senere (2016-2020)
 - Prospektive kohortestudier (frafald, progression, præstationer)

Rapport

http://cesu.au.dk/fileadmin/www.medu.au.dk/centret/MMI_evalueringsrapport.pdf

Kommentarer/spørgsmål?