HÍ | HEILBRIGÐISVÍSINDASTOFNUN

Próffræðilegir eiginleikar matstækja
– og af hverju þeir skipta máli

Próffræðistofa Heilbrigðisvísindastofnunar

2024

Efnisyfirlit

Orðskýringar 3

Af hverju skipta próffræðilegir eiginleikar máli? 5

Áreiðanleiki 6

Aðferðir við að meta áreiðanleika 7

Innri áreiðanleiki 7

Cronbach‘s alpha 7

Ordinal alpha 9

Kuder-Richardson (KR-20) 9

Omega 9

Aðrir stuðlar 10

Endurprófunaráreiðanleiki 10

Áreiðanleiki matsmanna 11

Helmingunaráreiðanleiki 12

Áreiðanleiki hliðstæðra matstækja 12

Réttmæti 12

Aðferðir við að meta réttmæti 13

Vísbendingar byggðar á svarferli 14

Vísbendingar byggðar á innri formgerð 15

Vísbendingar byggðar á tengslum 16

Samantekt 18

Heimildir 19

Orðskýringar

- Dreifing (eða dreifni, e. variance): Öll tölulegu gildi sem breyta tekur, og hversu oft þau koma fyrir
- Formgerð (e. internal structure): Úr þáttagreiningu (e. factor analysis) – þetta er fjöldi þátta / vídda og innbyrðis tengsl þátta / vídda og atriða
- Fylgni vs. sammæli (e. correlation vs. agreement): Fylgnistuðlar eru hannaðir til að meta samband tveggja fyrirbæra. Fylgni milli X og Y getur verið há án þess að sammæli sé hátt. Sammæli metur samræmi í tveimur mælingum *á sama fyrirbæri*. Ef sammæli er hátt er fylgni örugglega há
- Háreist dreifing (e. leptokurtic): Dreifing sem hefur háan topp, eða hærri en normaldreifing
- Hlutbundið fyrirbæri (e. objective): Sjáanlegt og / eða mælanlegt með beinum hætti. Dæmi: Fjöldi, stærð, þyngd, hitastig
- Hlutur vs. þáttur: (e. component vs. factor): Hlutur er orð yfir það sem meginhlutagreining (e. principal component analysis) skilar okkur. Hlutur er í praxís hliðstæður þætti í þáttagreiningu, en theorían er ekki sú sama. Hlutur er samantektarbreyta, fall af atriðum matstækis – hentug lýsing á tengslum margra breyta sem þó þurfa ekki að eiga saman fræðilega. Þáttur er aftur á móti orsakabreyta – undirliggjandi fyrirbæri sem við göngum úr frá því að skýri tengsl margra breyta sem allar eiga eitthvað sameiginlegt, þ.e. þáttinn. Dæmi um hlut: Félagsstaða, sem er fall af kyni, aldri, menntun, innkomu, búsetu, fjölskylduhag o.s.frv. Dæmi um þátt: Þunglyndi, sem torsakar depurð, leiða, áhugaleysi, vanvirkni o.s.frv.
- Hugsmíð (e. construct): Annað orð yfir (óhlutbundið) fyrirbæri sem er skilgreint með fræðilegum hætti
- Klassíska raungildiskenningin (e. classical test theory): Einn stærsti kenningarammi próffræðinnar, helst eignuð Lord & Novick á 7.áratug síðustu aldar
- Lágreist dreifing (e. leptokurtic): Dreifing sem hefur lágan eða flatan topp, flatari en normaldreifing
- Magnbinding (e. quantification): Að ákvarða magn einhvers, oft með því að gefa því tölugildi Dæmi: Heildarskor á spurningalista sem metur einkenni þunglyndis er tilraun til magnbindingar á þunglyndi
- Mæld breyta (e. measured variable): Má einnig kalla vísi. Fyrirbæri eða eiginleiki sem „vitnar um" eitthvað annað. Dæmi: Depurðaratriði á spurningalista sem metur þunglyndi er mæld breyta (þunglyndið sjálf er undirliggjandi breyta)
- Mæling vs. mat (e. measurement vs. assessment): Mælingu má skilgreina sem magnbindingu á mælanlegu *additívu* fyrirbæri (Michell, 1997). Mat má hins vegar líta á sem flokkun eða röðun á fyrirbæri sem er ekki magnbindanlegt eða additívt (Michell, 2012). Þar sem matstæki og spurningalistar meta mun oftar eiginleika (e. qualities, t.d. félagslyndi) heldur en magnbundnar stærðir (e. quantities, t.d. ummál handleggs) er réttara að tala um að þau *meti* eitthvað heldur en *mæli* (sbr. matstæki, ekki mælitæki)
- Óhlutbundið fyrirbæri (e. subjective): Ósýnilegt og / eða ekki mælanlegt með beinum hætti
- *Polychoric* fylgni: Fylgnistuðull sem ætlaður er fyrir mat á tengslum tveggja undirliggjandi breyta með tveimur mældum breytum. Gengið er út frá því að hinar undirliggjandi breytur séu normallaga, en mældar á raðkvarða
- Raunvís gögn (e. empirical data): Gögn sem fást með athugunum (e. observations) eða tilraunum (e. experiments). Dæmi: Tölulegar niðurstöður próffræðilegra rannsókna á matstækjum eru raunvís gögn sem upplýsa um áreiðanleika og réttmæti

- Samdreifnifylki (e. covariance matrix): Samdreifni er dreifing sem safn mældra breyta / atriða deila hver með annarri. Fylki er stærðfræðileg framsetning á tengslum af því tagi. Samdreifnifylki koma við sögu í þáttagreiningu
- Samfelldur svarkvarði (e. continuous response scale): Strangt til tekið kvarði sem getur tekið öll möguleg gildi á einhverju talnabili (t.d. 0 til 10 þar sem breyta getur verið 0,0004, 5,182, 9,9999, o.s.frv.). Einu svarkvarðarnir af þessu tagi eru VAS kvarðar (e. visual analog scale), en mun oftar eru matstæki og spurningalistar með atriði sem mæld eru á raðkvarða (e. ordinal scale) sem taka bara nokkur heiltölugildi
- Samleitni (e. convergence) vs. sundurgreining (e. divergence): Samleitni er samræmi í niðurstöðum skyldra matstækja, oft metin með fylgniútreikningi. Sundurleitni er hlutfallslega lítið samræmi í niðurstöðum tveggja óskyldra matstækja. Hvort tveggja gefur vísbendingar um réttmæti. Dæmi: Við búumst við hárri fylgni milli X1 sem metur almennan kvíða og X2 sem metur áhyggjur. Við búumst við hlutfallslega lægri fylgni milli X1 og Z sem metur líkamsmynd
- Samstofna atriði (e. congeneric): Atriði sem meta sama fyrirbærið en tengjast því ekki endilega öll jafn sterkt. Dæmi: Atriði sem meta kjarnaeinkenni þunglyndis og atriði sem meta jaðareinkenni þunglyndis eru samstofna, en inntak þeirra tengist þunglyndi missterkt
- Skor (e. score): Annað orð yfir stig eða niðurstöðu á atriði / matstæki. Dæmi: Heildarskor eða summuskor á spurningalista (koma við sögu í umfjölluninni að neðan og eru notuð sitt á hvað – þó heildarskor þurfi ekki endilega að vera reiknuð með samlagningu)
- Strjáll svarkvarði (e. discrete response scale): Kvarði sem tekur aðeins fá gildi. Þetta gildir um megnið af svarkvörðum sem notaðir eru í spurningalistum almennt – Likertkvarðar og aðrir raðkvarðar eru strjálir (…en oft meðhöndlaðir eins og samfelldir!)
- Svarskekkjur (e. response bias): Kerfisbundin tilhneiging til þess að svara atriðum á grunni einhvers annars en inntaks þeirra (Paulhaus, 1991). Dæmi: Jáhneigð (e. acquiescence) er tilhneiging til að samþykkja / vera sammála – velja já – án tillits til inntaks. Miðjusvörun er tilhneiging til að velja (hlutlausan) miðjusvarkost óháð inntaki. Ýkt svörun er tilhneiging til að velja sterkustu svarkostina óháð inntaki
- Tvíkosta svarkvarði (e. dichotomous / binary response scale): Kvarði sem tekur bara tvö gildi. Dæmi: Já / nei, rétt / rangt
- Undirliggjandi breyta (e. latent variable): Annað orð yfir (óhlutbundið) fyrirbæri sem er skilgreint með fræðilegum hætti og talið „orsaka" þau gildi sem vísar taka. Dæmi: Þunglyndi (undirliggjandi breyta) er talið orsaka depurð (vísir / mæld breyta)
- Vigtun – unit weight vs. optimal weight: Unit weight vísar til þess þegar atriði á matstæki vega öll jafn þungt í heildarskori. Optimal weight er vigtun þar sem atriði vega misþungt í heildarskori eftir því hve sterk tengsl þau hafa við fyrirbærið sem á að meta
- Villuliður (e. error term): Frávik í mati (skekkja)
- Vídd (e. dimension): Annað orð yfir fyrirbæri / hugsmíð / þátt
- Vísir (e. indicator): Annað orð yfir mælda breytu. Fyrirbæri eða eiginleiki sem „vitnar um" eitthvað annað. Dæmi: Depurðaratriði á spurningalista sem metur þunglyndi er talinn vera vísir að þunglyndi
- Þáttabygging (e. factor structure): Annað orð yfir formgerð. Talað eru um einvíða (e. unidimensional) og fjölvíða (e. multidimensional) þáttabyggingu
- Þáttahleðsla (e. factor loading): Stærð sem lýsir styrk tengsla mældrar breytu (atriðis) og undirliggjandi breytu / þáttar
- Þáttur (e. factor): Orð yfir það sem þáttagreining skilar okkur. Orsakabreyta – undirliggjandi fyrirbæri sem við göngum úr frá því að skýri tengsl margra breyta

Meiningin með eftirfarandi umfjöllun er að skýra hugtökin áreiðanleiki og réttmæti með sem einföldustum hætti, fjalla um þær aðferðir sem helst eru notaðar við mat á þeim og takmarkanir þeirra aðferða þegar við á. Okkar von er að efnið veki lesendur til meðvitundar um flækjustig og blæbrigði við mat á próffræðilegum eiginleikum. Við hvetjum rannsakendur á Heilbrigðisvísindasviði til að nýta sér umfjöllunina að vild – athugið bara að vísa til heimilda þegar við á.

Próffræðilegir eiginleikar (e. psychometric properties) matstækja[1] veita vísbendingar um gæði þeirra og réttmæti til ákveðinna nota, í hversdagslegum skilningi þess orð. Þessir eiginleikar – áreiðanleiki og réttmæti – skipta máli í öllum þeim tilvikum þar sem magnbinda (e. quantify) á fyrirbæri sem ekki er hægt að mæla með beinum hætti, eða sem er í eðli sínu ómælanlegt. Mæling (e. measurement) á slíku fyrirbæri – sem kannski væri réttara að kalla mat (e. assessment) – felur í sér tilraun til magnbindingar með svonefndum vísum (e. indicators) sem taldir eru vitna um fyrirbærið. Í próffræði (e. psychometrics) er óhlutbundið fyrirbæri af þessu tagi nefnt hugsmíð (e. construct) eða undirliggjandi breyta (e. latent variable) og það sem við notum til að meta það nefnist mældar breytur (e. measured variables). Dæmi: PHQ-9 spurningalistinn er tilraun til magnbindingar á þunglyndi. Á honum er þunglyndi hin undirliggjandi breyta. Atriði listans eru taldir vera vísar fyrir þunglyndi – þau eiga að tilgreina einkenni sem orsakast af og vitna um „magn“ þess. Athuganir á áreiðanleika og réttmæti PHQ upplýsa okkur um trúverðugleika þessa – þ.e. að atriðin nái sannarlega utan um þunglyndi og endurspegli raunverulegan breytileika í því.

Hugtökin áreiðanleiki og réttmæti eru hluti af stöðluðu orðfæri í rannsóknum, en reynslan hefur sýnt að orðin eru stundum skrifuð og sögð án þess að vera fyllilega skilin. Á næstu síðum munum við setja fram skilgreiningar á hvoru um sig og greina frá þeim aðferðum sem helst koma við sögu við mat á þessum eiginleikum. En fyrst skulum við átta okkur betur á því af hverju þetta skiptir í alvöru máli.

Af hverju skipta próffræðilegir eiginleikar máli?

Þó flestir rannsakendur átti sig á því að próffræðilegir eiginleikar matstækja útheimti lágmarks umfjöllun í fræðigreinum eru þeir sjaldan aðal áhugamál fræðafólks (annarra en þeirra sem skrifa þennan texta!). Það er eðlilegt – læknar, hjúkrunarfræðingar, lyfjafræðingar, matvæla- og næringarfræðingar, tannlæknar og sálfræðingar hafa sennilega meiri áhuga á lífi og heilsu en áreiðanleika og réttmæti. En tilfellið er að alltaf þegar matstæki er notað verður að vera ljóst að það a) *skili stöðugri og nákvæmri niðurstöðu*, og að b) *niðurstaðan sé magnbinding á fyrirbærinu sem ætlunin er að meta og engu öðru*. Notkun matstækja og túlkun niðurstaðna verður þar að auki að vera studd af kenningalegum bakgrunni og í samræmi við ætlað notagildi. Þegar áreiðanleiki og réttmæti matstækja er ekki fullnægjanlegt getur það leitt til þess að upplýsingar sem með þeim fást verða afvegaleiðandi eða beinlínis rangar. Slíkt hefur afleiðingar bæði í rannsóknum og klíník.

- Afleiðingar í rannsóknum:
    - Rannsóknarniðurstöður verða óáreiðanlegar
    - Rannsóknarniðurstöður verða misvísandi
    - Rannsóknarniðurstöður verður erfitt eða ómögulegt að endurtaka – *ekki vegna breytileika í fyrirbærinu sem áhugi beinist að heldur vegna breytileika (vankanta) í því hvernig við metum þau*
    - Þekking staðnar

[1] Matstæki er hér notað sem samheiti yfir mat sem byggir á svörum fólks við atriðum (t.d. spurningum). Þetta eru m.ö.o. hverslags spurningalistar, s.s. sjálfsmatskvarðar (e. self-assessment scales) og einkennalistar (e. symptom check-lists).

- Afleiðingar í klíník:
    - Inngrip og meðferðir sem eru *raunverulega gagnlegar* fá ekki brautargengi, inngrip og meðferðir sem eru *ekki* gagnlegar fá brautargengi
    - Einstaklingar sem *þurfa* meðferð fá hana ekki, einstaklingar sem *þurfa hana ekki* fá hana
    - Skert gæði þjónustu, rangar ákvarðanir

Í stuttu máli: Niðurstöður rannsókna verða aldrei merkilegri eða marktækari en mælingarnar sem þær byggja á. Að sama skapi erum við engu nær um áhrif og árangur meðferða ef tækin sem við notum til að mæla áhrif og árangur eru gölluð.

Þess vegna skiptir þetta máli.

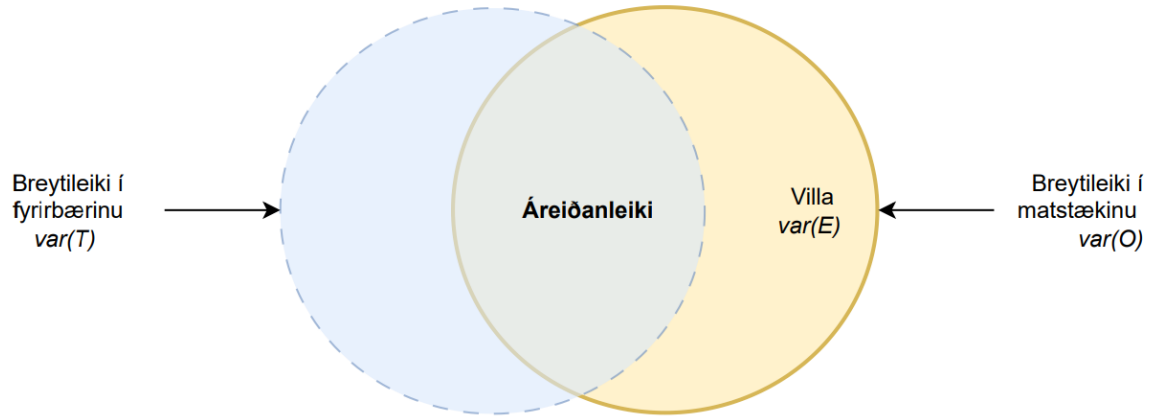
Að því sögðu skulum við hefja umfjöllunina. Fyrst beinum við sjónum okkar að áreiðanleika.

Áreiðanleiki

Áreiðanleiki (e. reliability) segir til um stöðugleika eða nákvæmni mælinga. Matstæki veitir áreiðanlega niðurstöðu ef það metur sama fyrirbærið með sömu nákvæmni yfir tíma, á milli einstaklinga og aðstæðna.

Hefðbundin skilgreining á áreiðanleika er fengin úr kenningu innan próffræði sem nefnist klassíska raungildiskenningin (e. classical test theory). Kenningin kveður á um að mælt skor á matstæki (e. observed score) sé summa raungildis (e. true score) og villu (e. error). Mælt skor á matstæki getur t.a.m. verið heildarskor á kvarða sem metur þunglyndiseinkenni. Fræðileg skilgreining á raungildi er meðaltal óendanlegra marga fyrirlagna á sama einstaklingi með sama matstæki (þetta er stærð sem við gefum okkur að sé til – hið „sanna“ þunglyndisskor). Villa er frávikið í mati okkar frá hinu „sanna“ raungildi.

Áreiðanleiki matstækis er í klassískri skilgreiningu það *að hve miklu leyti breytileiki í skorum á matstæki endurspeglar raunverulegan breytileika þess fyrirbæris sem meta á* (sjá mynd 1).



Mynd 1. Myndræn framsetning á áreiðanleika í klassískri skilgreiningu

Mikilvægt er að hafa í huga að áreiðanleiki vísar til niðurstaðna úr matstæki í tilteknum aðstæðum. Við getum því ekki talað um að matstæki sé áreiðanlegt, heldur frekar að niðurstöður þess séu áreiðanlegar fyrir ákveðinn hóp í ákveðnum aðstæðum. Þetta þýðir að meta þarf áreiðanleika matstækis í hver skipti sem það er notað. Það er ekki nægjanlegt að vísa í áreiðanleika sem fengist hefur í fyrri rannsóknum til þess að færa rök fyrir notkun þess – *áreiðanleiki er breytilegur frá einu úrtaki til annars og hann þarf að meta í hverju tilviki fyrir sig.*

Aðferðir við að meta áreiðanleika

Áreiðanleika matstækja má meta með ýmsum hætti. Yfirleitt eru reiknaðir áreiðanleikastuðlar sem taka gildi á bilinu 0 til 1 þar sem hærra gildi vitnar um hærri áreiðanleika (stöðugri niðurstöðu yfir tíma, einstaklinga, aðstæður).

Ýmis viðmið fyrir „ásættanlegan áreiðanleika" hafa verið gefin út fyrir klínískar aðstæður og rannsóknir. Slíkum viðmiðum ber þó að taka með fyrirvara því hvað telst ásættanlegt fer mjög eftir eðli þess sem meta á og aðstæðum sem það er metið í. Er fyrirbærið t.d. vítt eða þröngt, einsleitt eða margbrotið, stöðugt eða óstöðugt? Hversu nákvæmt þarf mat að vera? Hverjar eru afleiðingar rangra niðurstaðna? Þegar rannsakendur túlka tölulegar niðurstöður áreiðanleikamats er mikilvægt að líta til þessara spurninga.

Innri áreiðanleiki

Innri áreiðanleiki (e. internal consistency reliability) er reiknaður þegar safn atriða er notað til að meta eitthvert undirliggjandi fyrirbæri. Innri áreiðanleiki vísar til samræmis í niðurstöðum atriða á sama matstækinu, þ.e. hvort atriði sem eiga að meta sömu hugsmíðina veita okkur svipaðar niðurstöður um þann sem svarar. Tökum dæmi: Ef svarandi samþykkir staðhæfingarnar „Mér finnst tölfræði skemmtileg" og „Ég hef gaman af því að leysa tölfræðileg vandamál" en er ósammála staðhæfingunni „Ég hata tölfræði" myndum við segja að það væri samræmi í svörum frá einu atriði til annars. Ef svarandi væri hins vegar sammála því að hata tölfræði væri það á skjön við það sem fyrri svörin bentu til. Í slíku tilviki myndu vakna spurningar um innri áreiðanleika atriðasafnsins.

Cronbach's alpha er algengasta aðferðin til þess að meta innri áreiðanleika matstækja. Innri áreiðanleiki mældur með alfa er stundum kallaður *internal consistency*, og vísar þá til samræmis í niðurstöðum atriða á sama matstækinu (samræmi í svarmynstri, sjá að ofan).[2]

Þegar alfastuðullinn er notaður við mat á áreiðanleika þarf að hafa forsendur hans í huga.

- **Atriði skulu aðeins meta eina vídd[3]:** Það er algengur misskilningur að alfastuðullinn sé mælikvarði á það hvort atriði á matstæki meti öll eitt og sama fyrirbærið – hvort matstæki sé það sem kallað er einvítt (e. unidimensional). Rétt er hins vegar að alfastuðullinn *gerir ráð fyrir* því að atriði meti sama fyrirbærið. Sýnt hefur verið fram á að hár alfastuðull getur fengist fyrir safn atriða sem metur eitthvað fleira en eitt (sem hefur fjölvíða þáttabyggingu, e. multidimensional) og lágur alfastuðull getur fengist fyrir safn atriða sem metur í raun eitt fyrirbæri. Af þessu leiðir að alfastuðullinn upplýsir okkur ekki um þáttabyggingu matstækja. Stuðullinn gerir ráð fyrir einni vídd, og þá forsendu ætti með réttu að meta sérstaklega með viðeigandi aðferð (s.s. þáttagreiningu, sjá umfjöllun síðar) *áður en stuðullinn er reiknaður.*

[2] Það geta komið upp aðstæður þar sem alfastuðullinn endurspeglar samræmi í svarmynstri ekki sérlega vel. Að öðru jöfnu er stuðullinn hærri því hærri sem meðalfylgni svara er – en sú meðalfylgni getur verið há þrátt fyrir að stakir fylgnistuðlar séu lágir eða jafnvel 0 (Hayes & Coutts, 2020).

[3] Vídd er einfaldlega fyrirbærið sem á að meta. Við notum þetta orð til jafns við fyrirbæri, hugsmíð og þátt.

Afleiðingar forsendubrests: Alfastuðullinn veitir ekki endilega skakkt mat á innri áreiðanleika þegar atriði að baki honum meta eitthvað fleira en eitt: Hann getur eftir sem áður gefið rétt mat á innbyrðis fylgni þeirra (McNeish, 2018). Fjölvíð þáttabygging getur hins vegar líka leitt til ofmats á áreiðanleika. Þar sem alfa gerir ráð fyrir að eitt fyrirbæri búi að baki atriðum gerir stuðullinn líka ráð fyrir því að öll innbyrðis fylgni atriða sé tilkomin vegna eins og sama fyrirbærisins. Stuðullinn gerir m.ö.o. ekki greinarmun á fylgni sem tilkomin er vegna þess undirliggjandi fyrirbæris sem áhugi beinist að og þeirri sem skýrist af öðrum áhrifaþáttum.

- **Atriði skulu vera jafngild (e. tau-equivalent):** Hvert atriði í matstækinu á að hafa jafn sterk tengsl við fyrirbærið sem ætlunin er að meta. Þetta er ströng forsenda sem heldur sjaldan – yfirleitt tengjast sum atriði fyrirbærinu sterkt (t.d. kjarnaeinkenni þunglyndis) á meðan önnur gera það minna (jaðareinkenni þunglyndis).

  Afleiðingar forsendubrests: Þegar forsendan um jafngildi atriða heldur ekki (og fylgni milli villuliða er í lágmarki, sjá að neðan) veitir alfastuðullinn mat á lægri mörkum áreiðanleika og skilar því oft niðurstöðum sem fela í sér vanmat. Hversu mikið það er fer eftir samhengi. Rannsóknir hafa sýnt að vanmat getur numið 10% (þ.e. stuðull getur reiknast 10% lægri en hann ætti með réttu að vera) ef um ræðir einvíðan kvarða með fáum atriðum, eða 20% ef um ræðir fjölvíðan kvarða með fáum atriðum (Green & Yang, 2009a). Fyrir safn 5 atriða þar sem 1 þáttahleðsla sker sig verulega frá hinum getur alfastuðull reiknast 0,76 eða jafnvel 0,56 (Graham, 2006). Stuðull sem krefst ekki jafngildis atriða getur aftur á móti reiknast 0,97 eða 0,99 fyrir sama atriðasafn. Rétt er að taka fram að aðrir rannsakendur hafa bent á að forsendubrestur hvað varðar jafngildi þurfi ekki endilega að hafa afdrifaríkar afleiðingar – sjá t.d. Savalei & Reise, 2019.

- **Atriði skulu vera normaldreifð og mæld á samfelldum kvarða:** Við útreikning á alfastuðlinum er notast við Pearson samdreifnifylki (e. covariance matrix). Forsenda Pearson fylgnistuðulsins er að allar breytur séu samfelldar (e. continuous) og nokkurn veginn normaldreifðar.

  Afleiðingar forsendubrests: Rannsóknir á áhrifum skekktra dreifinga á alfastuðulinn hafa ekki gefið einhlítar niðurstöður. Sumar benda til þess að stuðullinn sé fremur traustur gagnvart skekkju (Zimmerman, 1993; Xiao & Hau, 2023). Aðrar hafa bent til þess að ólíkar tegundir skekkju geti bæði kallað fram vanmat (þegar dreifing er háreist, e. leptokurtic) og ofmat (þegar dreifing er lágreist, e. platykurtic) (Sheng & Sheng, 2012). Áhrif þess þegar alfastuðull er reiknaður fyrir matstæki sem inniheldur atriði sem hafa aðeins fáa svarkosti, eins og t.d. tvíkosta atriði (e. binary) eða atriði með þriggja punkta raðkvarða, eru að samdreifnin milli þeirra verður lægri en ella. Tengsl atriða virðast minni en þau raunverulega eru og alfastuðullinn skilar vanmati á áreiðanleika. Ef skekkja og fáir svarkostir fara saman getur alfa leitt til verulegs vanmats (Gadermann, Guhn & Zumbo, 2012).

- **Engin fylgni skal vera milli villu atriða:** Fylgni milli villu atriða verður til þegar eitthvað annað en fyrirbærið sem ætlunin er að meta býr til umfram tengsl á milli þeirra. Þessi umfram tengsl milli atriða geta skýrst af aðferðafræðilegum þáttum (t.d. orðalagi atriða – þegar jákvætt orðuð atriði eða atriði með svipað orðalag hafa umfram tengsl sín á milli) eða því að atriði meti annað fyrirbæri samhliða því sem áhugi beinist að. Hvort sem um ræðir aðferðafræðilega aukaafurð eða fjölvíða þáttabyggingu er fylgni villuliða atriða oftar jákvæð, sem skilar sér í hærri alfastuðli og þar með ofmati á áreiðanleika (Bentler, 2009; Green & Hershberger, 2000; Green & Yang, 2009b). Ofmatið getur numið allt að 20% (Gessaroli & Folske, 2002).

Fyrir utan að byggja á forsendum sem sumar eru líklegar til að bresta hefur alfastuðullinn þann ókost að hann ræðst ekki aðeins af innbyrðis fylgni atriða heldur einnig *fjölda* atriða. Almennt gildir að stuðulinn hækkar eftir því sem atriðum fjölgar og lækkar eftir því sem atriðum fækkar. Hann getur m.ö.o. verið lágur aðallega vegna þess að atriði eru fá, eða hár aðallega vegna þess að atriði eru mörg. Þetta er ekki eftirsóknarverður eiginleiki fyrir áreiðanleikastuðul.

Ofangreindir eiginleikar alfastuðulsins virðast fáum kunnir og fátítt er að sjá umfjöllun rannsakenda um þá. Það ætti hins vegar að vera prinsipp atriði að leggja mat á forsendur stuðulsins áður en hann er reiknaður og túlkaður. Sem dæmi, er ljóst að matstækinu er ætlað að meta eitthvað eitt fyrirbæri? Hefur sú þáttabygging fengist í endurteknum rannsóknum? Ef já – er hægt að líta á atriði matstækisins sem nokkurn veginn „útskiptanleg" – jafn þung eða alvarleg? Fræðilega? Klínískt? Að svo miklu leyti sem forsendur standast ekki getur stuðulinn gefið skekkt mat á innri áreiðanleika með þeim hætti sem greinir frá að ofan. Ef rannsakendur velja að halda sig við alfastuðulinn þrátt fyrir augljósa forsendubresti þurfa þeir að átta sig á og greina frá þeirri skekkju sem sennilega er fyrir hendi í áreiðanleikamati þeirra.

Ordinal alpha: Til er tilbrigði við hefðbundinn alfastuðul sem kallað hefur verið *ordinal alpha* (Zumbo, Gadermann & Zeisser, 2007). Sá stuðull er algerlega hliðstæður hefðbundnum alfastuðli nema að því leyti að hann styðst við *polychoric* samdreifnifylki í stað hefðbundins Pearson samdreifnifylkis. Polychoric fylgni er notuð við mat á fylgni milli tveggja undirliggjandi breyta (sbr. e. latent breytur – ekki mælanlegar með beinum hætti) með tveimur mældum breytum (sbr. e. manifest breytur – t.d. atriði á spurningalista). Gengið er út frá því að hinar undirliggjandi breytur séu í raun samfelldar og normallaga, en mældar á raðkvarða. Þetta er ólíkt forsendum hefðbundins Pearson fylgnistuðuls sem lúta að dreifingu hinna mældu breyta, og hún á að vera samfelld og normal. Ástæða þess að ordinal alfastuðull ætti að gefa réttara mat á áreiðanleika matstækis sem inniheldur atriði á raðkvarða er sú að þegar atriði eru mæld á raðkvarða en meðhöndluð líkt og þau væru mæld á samfelldum kvarða leiðir það yfirleitt til vanmats á fylgni milli þeirra. Tökum (ýkt) dæmi: Segjum að við höfum aðeins 3 tvíkosta atriði (Gadermann, Guhn & Zumbo, 2012). Dreifing þeirra strjál (e. discrete) og getur verið mjög skökk. Í slíku tilviki getur hefðbundinn alfastuðull fyrir atriðin 3 reiknast 0,46. Sú niðurstaða myndi sennilega þýða að matstækið væri afskrifað. Alfastuðull byggður á polychoric fylgni getur hins vegar reiknast 0,85. Sú niðurstaða myndi leiða okkur að allt annarri ályktun um gæði matstækisins.

Kuder-Richardson (KR-20): Þegar um ræðir matstæki sem samanstendur af tvíkosta atriðum er auk ordinal alfastuðuls hægt að notast við Kuder-Richardson (KR-20) áreiðanleikastuðulinn. Hann er túlkaður á sama hátt og alfa, nema hvað hann tekur tillit til þess að atriðin séu tvíkosta. Aðrar forsendur KR-20 eru þær sömu og eiga við um alfastuðulinn – að atriðin meti aðeins eina vídd, þau séu jafngild og að engin fylgni sé á milli villuliða þeirra. Þegar þetta er skrifað virðist sem stuðullinn sé ekki mikið notaður.

Omega er áreiðanleikastuðull sem hefur í auknu mæli verið notaður til þess meta innri áreiðanleika matstækja. Rétt eins og alfastuðullinn metur hann samræmi í niðurstöðum atriða, en með öðrum hætti. Stuðullinn á uppruna sinn í þáttagreiningu (sjá umfjöllun síðar). Mikilvægasti kostur omega umfram alfa er að stuðullinn var hannaður fyrir svokölluð samstofna (e. congeneric) atriði, þ.e. atriði sem tengjast fyrirbærinu sem ætlunin er að meta missterkt, og getur því verið réttmætari mælikvarði á áreiðanleika þegar forsenda um jafngildi atriða stenst ekki. Þar sem omega stuðullinn byggir á þáttagreiningu snúa megin forsendur hans að því sem um hana gildir. Í stuttu máli – forsenda fyrir réttmætum omegastuðli er að þáttagreiningarlíkanið sem hann er reiknaður upp úr sé nægilega góð nálgun að veruleikanum, þ.e. endurspegli hugsmíðina, passi gögnunum.

Til eru nokkrar útgáfur af omega en hér verður fjallað stuttlega um tvær: Omega total og omega hierarchical.

Omega total (ω_t) endurspeglar að hve miklu leyti breytileiki í skori á matstæki skýrist af því fyrirbæri (þætti) sem þau tengjast öll. Stuðullinn gerir, eins og alfa, ráð fyrir því að aðeins sé um einn þátt að ræða – þ.e. að öll samdreifing atriða sé tilkomin vegna sama fyrirbærisins. Áður en ω_t er reiknaður eru atriðin þáttagreind til að kanna tengsl þeirra við fyrirbærið sem meta á (meta þáttahleðslur atriða (e. factor loadings), sem svo eru notaðar við útreikning stuðulsins). Ef atriði eru jafngild mun ω_t gefa sömu niðurstöðu og alfastuðullinn. Ef atriðin eru hins vegar samstofna (þ.e. tengjast sama fyrirbæri en missterkt) getur ω_t hins vegar gefið réttara mat á áreiðanleika en alfa, sem í slíku tilviki myndi skila vanmati. Fyrir umræðu um ω_t og skilgreiningu viðeigandi þáttalíkans, sjá t.d. Savalei & Reise, 2019.

Omega hierarchical (ω_h) lýsir því að hve miklu leyti skor á matstæki endurspeglar það fyrirbæri sem atriði tengjast öll, jafnvel þótt þau geti metið önnur fyrirbæri samhliða. Stuðullinn gerir m.o.ö. ekki ráð fyrir því að aðeins einn þáttur skýri samdreifni atriða, heldur aðeins að fyrir hendi sé einn *almennur þáttur* sem atriði meta öll að einhverju marki. Þegar ω_h er reiknaður er fyrst gerð þáttagreining þar sem þáttabygging er skilgreind með viðeigandi hætti (hver tengsl atriða við almennan þátt eru, og hvort atriði hafi tengsl sín á milli umfram þau sem almennur þáttur skýrir). Fyrir umræðu um ω_h og viðeigandi útfærslur á þáttalíkani, sjá t.d. Savalei & Reise, 2019. ω_h getur dregið úr ofmati á áreiðanleika sem annars fengist með alfastuðlinum í sömu aðstæðum – á meðan alfa myndi gera ráð fyrir því að öll fylgni milli atriða væri tilkomin vegna eins og sama þáttarins getur ω_h aðgreint fylgni milli atriða í þá sem tilkomin er vegna víddarinnar sem meta á (hins almenna þáttar) og vegna annarra smærri vídda (sértækra þátta). Sumir vilja reyndar meina að sama gildi í tilviki ω_t, sem gæti ofmetið áreiðanleika með svipuðum hætti og alfa vegna þess að hann gerir líka ráð fyrir einvíðri þáttabyggingu. Fyrir umræðu um túlkun ω_h, sjá t.d. Zinbarg o.fl., 2006.

Aðrir stuðlar

Greatest lower bound stuðlar byggja líkt og alfastuðullinn á klassísku raungildiskenningunni. Stuðlarnir miða að því að finna það hámarks hlutfall villu í skori sem samræmist gögnunum, þ.e. sennilega hámarks villu í mati okkar. Af gildi hámarks villu má leiða lágmarks áreiðanleika (samanber *lower bound*). GBL stuðla er tiltölulega erfitt að reikna. Annar ókostur er að vísbendingar eru um að þeir ofmeti áreiðanleika þegar úrtak er smátt og skili ónákvæmu mati (Malkewitz o.fl., 2023).

H stuðullinn (athuga, ekki eini stuðullinn sem ber það heiti!) byggir ólíkt alfastuðlinum en líkt og omega á þáttagreiningu (sjá síðar). Hann gefur hugmynd um *hámarks* áreiðanleika þegar um ræðir vigtuð atriði. Öfugt við það sem gildir um alfa og omega hafa neikvæðar þáttahleðslur ekki óæskileg áhrif á stuðulinn. Að sama skapi eru atriði með hlutfallslega veikar þáttahleðslur ekki vandamál – þau eru bara látin vega minna í heildarskori en hin. Stuðullinn er þannig viðeigandi þegar ástæða eða vilji er til að mynda vigtuð skor úr matstæki.

Endurprófunaráreiðanleiki

Endurprófunaráreiðanleiki (e. test-retest reliability) byggir á því að leggja sama matstækið fyrir sama hóp einstaklinga á tveimur mismunandi tímapunktum og reikna fylgni á milli skora í fyrirlögnunum tveimur. Við mat á endurprófunaráreiðanleika þarf að vera ljóst að fyrirbæri sem meta á sé stöðugt yfir tíma og að aðstæður við mat séu sambærilegar (hið síðara tengist forsendu um að dreifing villu í mati á að vera sú sama).

Endurprófunaráreiðanleika má meta með tvennum hætti – annars vegar með hefðbundnum fylgnistuðlum (Pearson, Spearman, Kendall) og hins vegar með svokölluðum *innanflokka* fylgnistuðlum (e. intraclass correlation coefficients, ICC – sjá einnig umfjöllun að neðan undir áreiðanleika

matstmanna). Þegar kemur að mati á endurprófunaráreiðanleika hafa ICC stuðlar þann kost umfram einfalda fylgnistuðla að með ICC má reikna fylgni milli fleiri en tveggja prófana. Að sama skapi lýsa ICC stuðlar ekki aðeins fylgni milli mælinga heldur einnig sammæli eða samræmi (e. agreement / consistency). ICC stuðlar eru þannig fágaðri stuðlar sem geta gefið fyllri mynd af áreiðanleika endurprófunar heldur en einfaldir fylgnistuðlar. Nánar verður farið í ólíkar gerðir þeirra að neðan undir áreiðanleika matsmanna, en fyrir umfjöllun um þær gerðir sem viðeigandi eru við mat á endurprófunaráreiðanleika, sjá bók Portney (2020) eða umfjöllun í grein Koo o.fl. (2016).

Áreiðanleiki matsmanna

Áreiðanleiki matsmanna (e. interrater reliability) er viðeigandi þegar gögn byggja á mati / úrskurði matsaðila (en ekki t.d. þegar gögn eru fengin með sjálfsmati). Áreiðanleiki matsmanna segir okkur að hvaða marki matsmenn gefa sömu athugunum (e. observations – t.d. sjúklingum) sama gildi eða sömu flokkun (t.d. „með sjúkdóm" eða „ekki með sjúkdóm"). Í slíku tilviki vitnar áreiðanleiki um hversu mikið sammæli er í mati / ályktunum matsmanna. Mismunandi aðferðir eru til við að meta áreiðanleika matsmanna – hér fyrir neðan er fjallað um þær helstu.

Hlutfall sammælis (e. percentage agreement) byggist á því að telja hversu oft matsmenn eru sammála í mati sínu og deila þeirri tölu með heildarfjölda athugana. Hlutfall sammælis er eins og nafnið bendir til tala á bilinu 0 til 1 (eða 0 til 100%) þar sem hærri tala vitnar um aukið sammæli. Gallinn við þessa aðferð er að eftir því sem matsmönnum fjölgar verður útreikningurinn viðameiri. Þar að auki er ekki tekið tillit til þess að sammæli geti komið til fyrir tilviljun (t.d. ef matsmenn eru óvissir og giska) og því getur stuðullinn ofmetið sammæli (og þar með áreiðanleika).

Kappa stuðullinn (e. Cohen's kappa) var hannaður til þess að bregðast við vanköntum þess að reikna aðeins hlutfall sammælis, en hann tekur tillit til ágiskunar í mati. Kappa tekur gildi á bilinu –1 til 1 þar sem 0 endurspeglar það sammæli sem búast mætti við fyrir tilviljun og 1 vitnar um fullkomið sammæli meðal matsmanna. Gildi undir 0 eru möguleg en þó sjaldséð í praxís og gefa til kynna að sammæli meðal matsmanna sé verra en búast mætti við fyrir tilviljun. Forsendur Kappa stuðulsins eru eftirfarandi:

- Aðeins tveir matsmenn
- Sömu matsmenn framkvæma matið. Ef matsmenn fyrir hverja athugun eru ekki þeir sömu ætti frekar að notast við t.d. Kappa stuðul Fleiss sem leyfir matsmenn sem valdir eru með tilviljunaraðferð fyrir hverja athugun
- Mælingar eru paraðar sem þýðir að báðir matsmenn meta sömu athuganir
- Fylgibreytan (matið) þarf að vera flokkabreyta (röðuð eða óröðuð) með ósamrýmanlegum flokkum. Sem dæmi, ef flokkarnir eru tveir (t.d. já / nei) þá ætti hver athugun að tilheyra öðrum hvorum flokki en aldrei báðum

Intraclass fylgnistuðla má einnig nota til þess að meta áreiðanleika matsmanna þegar fylgibreytan er samfelld (t.d. talning en ekki tvíkosta flokkun eins og „með sjúkdóm" – „ekki með sjúkdóm"). Eins og áður sagði koma stuðlarnir í ýmsum gerðum. Markmið áreiðanleikamats (hvort meiningin er að álykta út fyrir gögnin og þá hvernig) ræður því hvaða líkan (e. model) er rétt að miða útreikninga við. Gögnin (hvort um ræðir stakar mælingar eða meðaltal nokkurra mælinga) skilgreina týpu stuðulsins (e. type). Skilgreining á sambandi mælinga ræður því hvort réttara er að reikna samræmi (e. consistency) eða sammæli (e. agreement). Fyrir nánari útlistun á ólíkum gerðum ICC stuðla, sjá bók Portney (2020) eða umfjöllun í grein Koo o.fl. (2016).

Helmingunaráreiðanleiki

Helmingunaráreiðanleiki (e. split-half reliability) byggir á því að atriðum matstækis er skipt í tvennt eftir ákveðinni reglu (t.d. atriði númer 1, 3, 5 og númer 2, 4, 6) og fylgni svo reiknuð á milli skora sama hóps einstaklinga úr hvorum hluta. Almennt gildir að lengri próf (með fleiri atriðum) mælast áreiðanlegri. Formúlan sem notuð er við útreikning helmingunaráreiðanleika tekur tillit til þess að aðeins hluti atriða matstækis eru metin í hvorum helmingi, og leiðréttir þannig fyrir vanmat á áreiðanleika sem annars hefði fengist.

Meðal vankanta þessa áreiðaleikastuðuls er að hann getur tekið breytingum eftir því hvaða atriði veljast í helminga. Alfastuðullinn (sem má líta á sem „meðaltal allra mögulegra helmingunaráreiðanleikastuðla") var m.a. hugsaður til þess að leysa þann vanda. Í praxís sést þessi stuðull sjaldan eða aldrei.

Áreiðanleiki hliðstæðra matstækja

Áreiðanleiki hliðstæðra matstækja (e. alternate forms reliability) er tegund áreiðanleikastuðla sem ýmist eru notaðir til að meta samræmi milli tveggja matstækja sem eiga að vera hliðstæð með tilliti til innihalds (t.d. tveir spurningalistar með sambærilegum spurningum sem eiga að meta sama fyrirbærið) eða milli ólíkra fyrirlagna á sama matstæki (t.d. samræmi niðurstaðna í fyrirlögn á pappír og á netinu). Í fyrra tilvikinu eru hliðstæð matstæki lögð fyrir sama hóp einstaklinga á tveimur tímapunktum og fylgni á milli niðurstaðna reiknuð, oft með Pearson fylgnistuðli. Í seinna tilvikinu er fylgni milli niðurstaðna einstaklinga í hinum ólíku fyrirlögnum reiknuð. Í báðum tilvikum jafngildir fylgnin áreiðanleikastuðli þar sem hærri gildi vitna um aukinn áreiðanleika.

Þegar reikna á áreiðanleika með þessum hætti er mikilvægt að fyrirbærið sem ætlunin er að meta sé stöðugt yfir tíma. Fylgni hliðstæðra matstækja hentar einna best ef kanna á stöðug fyrirbæri, eins og t.d. greind, en ekki eiginleika sem geta breyst frá degi til dags. Annað skilyrði fyrir því að meta áreiðanleika með hliðstæðum matstækjum er að dreifing villu í fyrri fyrirlögn sé sú sama og dreifing villu í seinni fyrirlögn. Þetta þýðir að mikilvægt er að tryggja að aðstæður fyrirlagna séu sem líkastar og fyrirlagnirnar sjálfar sömuleiðis.

Helsti vandinn við að meta áreiðanleika með hliðstæðum prófunum (sjá fyrra dæmi að ofan – tveir hliðstæðir spurningalistar) er að það getur reynst erfitt að hanna tvö hliðstæð matstæki. Sem dæmi má nefna að hliðstæðir spurningalistar eru þeim eiginleikum gæddir að sérhvert atriði í einum á sér fullkomna hliðstæðu í öðrum. Tölfræðilega myndi þetta þýða að atriði beggja lista hefðu sama vænta meðaltal og villu. Slík atriði hafa sama innihald – án þess þó að hafa sama orðalag – og tengjast þeirri vídd sem ætlunin er að meta með nákvæmlega sama hætti. Fátítt er að sjá áreiðanleika af þessu tagi reiknaðan.

Réttmæti

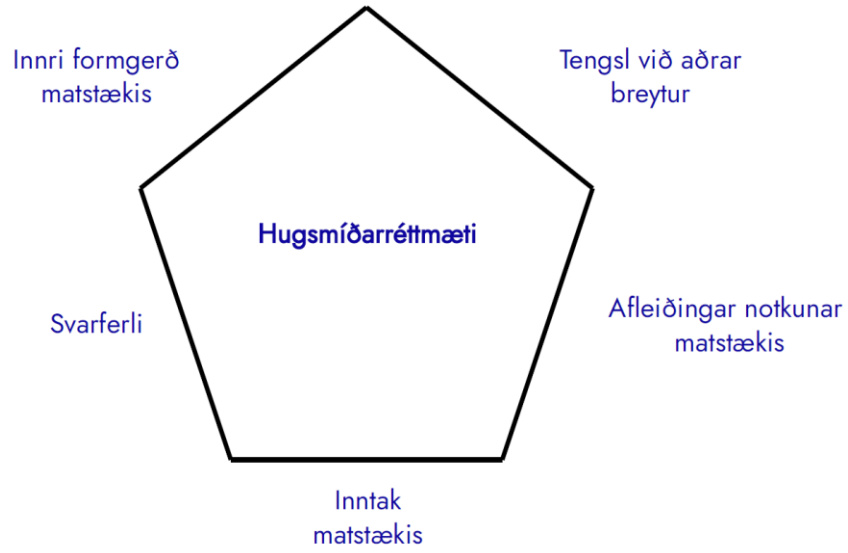
Réttmæti (e. validity) hefur verið skilgreint á ýmsa vegu. Talað hefur verið um svona og hinsegin réttmæti – hugsmíðar, inntaks, samleitni, sundurgreiningar, samtíma, forspár og viðmiðunarréttmæti. Í seinni tíð hafa vísindastofnanir lagt til að horfið sé frá því að tala um ólíkar tegundir réttmætis og hvatt til þess að litið sé á það sem einn og sama hlutinn sem má afla margskonar mismunandi *vísbendinga um*. Í grunninn hlýtur réttmæti matstækis að ráðast af því að a) fyrirbæri sem meta á sé til og að b) breytileiki í fyrirbærinu orsaki breytileika í útkomu á matstækinu – sjá umfjöllun í grein Borsboom (2004) um nauðsynlegar (og nægjanlegar!) forsendur réttmætis**. Það gefur augaleið að ef það sem við teljum okkur vera að meta er ekki til er tómt mál að tala um réttmæti! Að sama skapi, ef breytileiki í skorum á matstæki endurspeglar ekki raunverulegan breytileika í fyrirbærinu sem við teljum okkur meta er ekki hægt að tala um neinskonar réttmæti!**

Einföld og algeng skilgreining er að réttmæti vísi til þess að hvaða marki matstæki metur það sem því er ætlað að meta. Þessi skilgreining er einföldun, en ekki alslæm. Yfirgripsmeiri skilgreiningu má finna í Standards for Educational and Psychologocal Testing þar sem segir að réttmæti vísi til þess að hvaða marki kenningarlegur bakgrunnur (e. theoretical background) og raunvís gögn (e. empirical data) styðja við túlkun á niðurstöðu matstækis. Sú skilgreining felur í sér að réttmæti vísar ekki til matstækisins sjálfs – við getum ekki fullyrt að matstæki eða skor sem fást með þeim séu réttmæt eða óréttmæt í eðli sínu. Réttmæti vísar þess í stað til túlkunar og notkunar, og það er því *túlkun og notkun sem getur verið réttmæt eða óréttmæt.* Þegar skilgreiningin er ígrunduð ætti að vera ljóst að við mat á réttmæti þarf að taka tillit til margra ólíkra þátta: Kenningalegs bakgrunns, aðstæðna við notkun, tilgangs notkunar – og raunvísra gagna. Hið síðastnefnda (t.d. reiknuð fylgni skora á matstæki við annað matstæki) er stundum sett fram sem tæmandi mat, en rétt er að tölulegar niðurstöður eru einungis eitt af mörgu sem veita vísbendingar um réttmæti. Vísbendingar um réttmæti (fræðilegar, raunvísar) geta verið „sterkar“ eða „veikar” – þær „benda til” eða „styðja við” en sýna ekki fram á. Engin ein greining veitir tæmandi upplýsingar um réttmæti heldur notum við margskonar vísbendingar til að ályktа um það.

Aðferðir við að meta réttmæti

Eins og nefnt var að ofan hafa hugmyndir um réttmæti verið margskonar. Áður fyrr var til að mynda litið svo á sem að skipta mætti hugtakinu í þrjár ólíkar „tegundir“: Inntaksréttmæti (e. content validity), viðmiðsréttmæti (e. criterion validity) og hugsmíðarréttmæti (e. construct validity). Líkt og áður segir er réttara að tala um *mismunandi vísbendingar* fyrir *mismunandi hliðar* réttmætis.

Kjarninn í hugmyndinni um réttmæti má segja að sé hugsmíðarréttmæti, en það vísar til þess að hvaða marki hægt er að túlka skor á matstæki sem endurspeglun á því fyrirbæri sem meta á. Út frá hugsmíðarréttmæti má skilgreina fimm gerðir vísbendinga um réttmæti sem verður fjallað nánar um hér fyrir neðan.[4]



[4] Hér verður m.a. stuðst við texta úr *Standards for Educational and Psychological Testing* sem var gefin út af American Educational Research Association, American Psychological Association og National Council on Measurement in Education árið 2014.

Vísbendingar byggðar á inntaki

Hægt er að afla upplýsinga um réttmæti með því að gaumgæfa tengsl á milli inntaks atriða á matstæki (e. test content) og þeirra hugsmíðar sem ætlunin er að meta. Inntak vísar til þema, orðalags og uppsetningar atriða, verkefna eða spurninga á matstækinu. Mat á réttmæti inntaks hefst í raun strax við samningu nýs matstækis. Þá er mikilvægt að skýr skilgreining á efnissviði (e. content domain) og tilgangi matstækisins liggi fyrir. Tryggja þarf að atriði matstækisins nái yfir allt efnissviðið og endurspegli það með merkingarbærum hætti, þ.e., að atriði nái utan um allar hliðar þess fyrirbæris sem ætlunin er að meta.

Algeng leið til þess að meta réttmæti inntaks er að fá tvo eða fleiri sérfræðinga á sviði þess sem ætlunin er að meta til þess að leggja mat á innihald atriða. Sem dæmi, ef kanna á inntak matstækis sem metur þunglyndi meðal barna er leitað til aðila sem eru sérfræðingar á því sviði, s.s. klínískra barnasálfræðinga. Mat sérfræðinganna getur falist í því að gefa hverju atriði matstækisins einkunn eftir því hversu viðeigandi það er, eða hversu vel það nær utan um fyrirbærið. Góð atriði ættu þannig að hafa hátt meðaltal (metin viðeigandi) og lágt staðalfrávik (með nægilegu sammæli). Einkunnagjöf af þessu tagi ásamt útreikningum á meðaltali og staðalfráviki er viðeigandi þegar fjöldi sérfræðinga er nægilegur. Mat sérfræðinga getur einnig verið með öðrum hætti eða óformlegra. Stundum felst það t.a.m. í yfirlestri atriða og tillögum að breytingum á inntaki eða orðalagi án þess að einkunnagjöf komi til.

Önnur leið til þess að kanna réttmæti inntaks eru svokölluð ítarviðtöl (e. cognitive inteviews). Í slíkum viðtölum, sem ýmist eru stöðluð eða hálfstöðluð, má afla upplýsinga um hvort atriðin séu auðskilin, og hvort svarendur skilji og svari með þeim hætti sem lagt er upp með. Ítarviðtöl eru einnig algeng til þess að kanna réttmæti svarferils (e. response process – sjá neðar).

Til að fá sem fyllsta mynd af réttmæti inntaks er ákjósanlegt að samþætta niðurstöður nokkurra aðferða, þ.e. byggja á niðurstöðum úr mati sérfræðinga, ítarviðtölum og tölulegum niðurstöðum þegar við á.

Vísbendingar byggðar á svarferli

Hægt er að upplýsa um réttmæti með því kanna að hve miklu leyti svarferli (e. response processes) eða útkoma einstaklings á matstæki samræmist kenningum að baki þeirri hugsmíð sem matstækinu er ætlað að fanga. Svarferli er það sem fólk gerir, hugsar eða skynjar þegar það bregst við og svarar atriðum matstækis. Sem dæmi þá viljum við að fólk *sé fært* um og *hafi kost* á að veita svar sem samræmist þeirra upplifun, og að það *sé viljugt* til að veita heiðarlegt svar. Ýmsir eiginleikar matstækis eða einstakra atriða geta komið í veg fyrir að þetta gangi eftir. Þar má nefna flækjustig atriðis (sbr. e. cognitive load), og orðalag atriðis og svarkosta.

Hægt er að styðjast við nokkrar tegundir gagna til að álykta um réttmæti svarferlis. Meðal þeirra eru upplýsingar sem fást með ítarviðtölum við mögulega svarendur (þ.e. úr því þýði sem matstæki er ætlað). Fyrst svarar fólk matstækinu (eða leysir verkefni þess) og svo er það spurt, annað hvort jafn óðum eða í lok fyrirlagnar, hvernig það hafi farið að því að velja svörin sín: Hvað það hugsaði um þegar það svaraði, hvort það geti umorðað atriði o.s.frv. Með ítarviðtölum má einnig leggja mat á áhrifaþætti eins og félagslega æskilega svörun (e. social desirability) eða svarskekkjur á borð við jáhneigð (e. acquiescence). Annars konar gögn sem nota má til að álykta um réttmæti út frá svarferli er hlutlæg frammistaða svarenda, t.a.m. tíminn sem það tekur að svara eða hvort svaranda veitir rétt svar í þeim tilvikum sem það á við. Sem dæmi myndum við vilja að mynstur í svartíma í skynjunarrannsókn samræmdist kenningunni að baki því skynjunarferli sem verið er að rannsaka (flóknara sjónleitarverkefni = lengri svartími, aukin æfing = skemmri svartími, svo dæmi sé tekið).

Vísbendingar byggðar á innri formgerð

Hægt er að varpa ljósi á réttmæti með því að kanna hvernig svonefnd innri formgerð matstækis (e. internal structure) samræmist fræðilegri skilgreiningu á þeirri hugsmíð sem ætlunin er að meta. Innri formgerð vísar til fjölda vídda og innbyrðis tengsla atriða og undirkvarða. Segja má að innri formgerð matstækja hafi tvær megin hliðar: Fjölda vídda (e. dimensionality) og stöðugleika þáttabyggingar (stundum kallað *measurement invariance*, sem mætti þýða sem *óbreytni mælinga* en vísar í raun til þess að formgerð sé eins, og víddir / atriði virki eins óháð stað, stund og hópum svarenda).

Með athugun á innri formgerð matstækis viljum við meta hvort innbyrðis tengsl atriða samrýmist kenningarlegum bakgrunni og styðji þannig við ætlaða notkun á tækinu. Tökum dæmi. Niðurstaða matstækja er oft á formi einfaldra summuskora. Skor eru iðulega notuð til að flokka fólk í hópa („mikið þunglyndir“ vs. „lítið þunglyndir) eða til að upplýsa um klíníska ákvarðanatöku og mat (frekari meðferð, ekki frekari meðferð – í bata, í versnun). Þegar eitt heildarskor er reiknað og túlkað þarf matstæki að vera í megindráttum einvítt. Þetta er grundvallaratriði: Þegar atriði eru lögð saman í eitt heildarskor krefst það þess að þau a) meti sama fyrirbærið og b) séu u.þ.b. jafngildir vísar á því. Ef atriði meta ekki sama fyrirbærið verður heildarskorið ekki merkingarbær magnbinding á því – við vitum ekki að hvaða marki skor vitnar um fyrirbærið sem áhugi beinist að (t.d. þunglyndi) og að hvaða marki það vitnar um eitthvað allt annað (síþreytu? Vanvirkan skjaldkirtil?). Ef atriði eru lögð saman sem eru augljóslega ekki jafngild (fræðilega, klínískt) getur það leitt til þess að á bak við hvert heildarskor liggi sundurleitur hópur einstaklinga sem er jafnvel í grundvallaratriðum ólíkur með tilliti til þess sem á að meta. Það hefur afleiðingar bæði í rannsóknum og klíník. *Innri formgerð matstækis skiptir því ekki síst máli þegar kemur að notkun og túlkun heildarskora.*

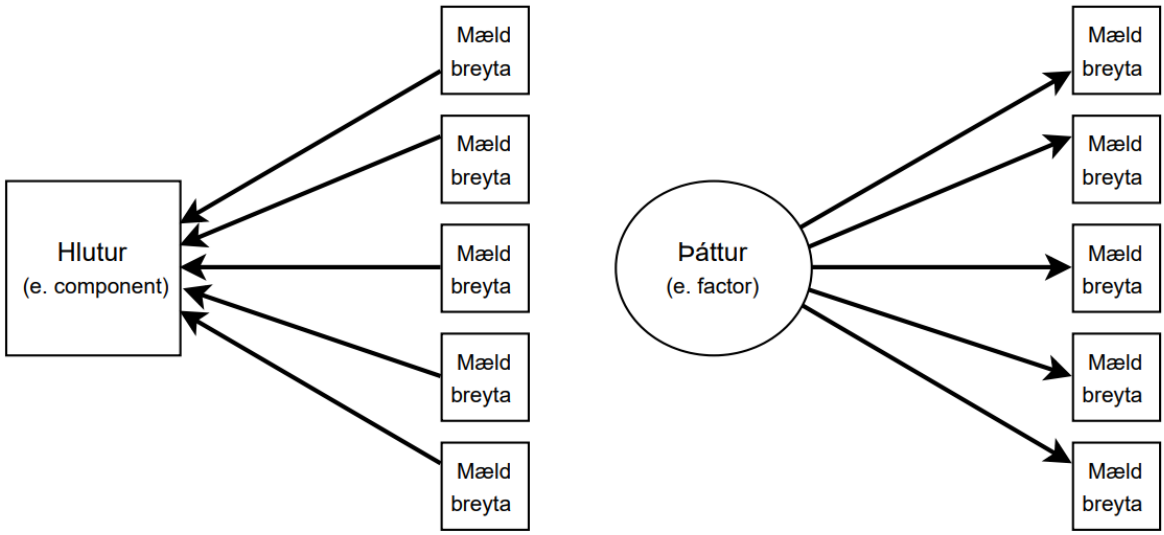
Athugun á óbreytni mælinga miðar að því að varpa ljósi á það hvort eiginleikar eða „virkni“ einstakra atriða sé sambærileg í ólíkum hópum einstaklinga (t.d. milli kynja eða ólíkra aldurshópa). Það felur m.a. í sér að matstæki meti sama fjölda vídda í ólíkum hópum og að tengsl atriða við tiltekna vídd séu svipuð milli hópa. Stöðugleiki í því tilliti er eftirsóknarverður því við viljum yfirleitt ekki að ólíkir hópar einstaklinga, sem hafa þó svipað „magn” af hugsmíð (færni, einkennum, skerðingu, viðhorfi), veiti kerfisbundið ólík svör við tilteknum atriðum vegna þess hvaða hópi þeir tilheyra frekar en vegna inntaks atriðis. Athugið einnig að ef matstæki metur hugsmíð ekki með sama hætti milli mismunandi hópa einstaklinga þá er ekki hægt að nota það til þess að bera hópana saman með tilliti til hugsmíðarinnar.

Meginhlutagreining (e. principle components analysis, PCA) og þáttagreining (e. factor analysis) eru gjarnan notaðar til þess að álykta um innri formgerð matstækja. Meginhlutagreining dregur sameiginlega dreifingu mældra breyta (atriða) saman í færri og stærri hluta eða víddir (e. components), en hlutar eru nokkurs konar samantektarbreytur *lýsa* (e. descripe) innbyrðis tengslum atriða. Meginhlutagreining gerir ekki strangar kröfur til gagna og getur gefið hugmynd um fjölda vídda sem atriði mynda, þ.e.a.s. innri formgerð.

Þáttagreining byggir á annars konar hugmyndafræði. Hún miðar að því að *skýra* (e. explain) sameiginlega dreifingu mældra breyta með færri undirliggjandi breytum, svonefndum hugsmíðum eða þáttum. Þáttagreining gerir ríkari kröfur til gagna heldur en meginhlutagreining en gerir okkur jafnframt kleift að vinna sérstaklega með villudreifingu atriða (ólíkt meginhlutagreiningu).

Helsti munurinn á meginhlutagreiningu og þáttagreiningu er að „hlutur” í meginhlutagreiningu er samantektarbreyta sem lýsir tengslum hinna mældu breyta. Þáttur í þáttagreiningu er aftur á móti skýrandi breyta sem litið er á sem orsakavald í þeim breytileika sem sést í mældu breytunum. M.ö.o, í meginhlutagreiningu lítum við á hlut sem fall af mældum breytum (t.d. er „meðaleinkunn” (hlutur) fall af einkunnum áfanga (mældum breytum)) en í þáttagreiningu lítum við á mældar breytur sem fall af

þætti eða hugsmíð (t.d. göngum við út frá því að „hugræn geta" (hugsmíð) skýri útkomu á atriðum taugasálfræðilegs prófs – eða þunglyndi skýri svarmynstur á einkennakvarða) – sjá mynd 2. Rétt er að leggja nokkra áherslu á þennan mun: Þótt stærðfræðin að baki aðferðunum sé svipuð er hugmyndafræðin það ekki. Meginhlutagreining er ekki eiginleg þáttagreining, heldur dæmi um það sem kallað er *data reduction* tækni. Hlutar / víddir sem aðferðin skilar eru einungis samantektarbreytur – hentug lýsing – og breyturnar sem raðast saman í hlut / á vídd *þurfa ekki að tengjast fræðilega*. Þáttur í þáttagreiningu er hins vegar skilgreindur fræðilega sem undirliggjandi hugsmíð sem skýrir tengsl breyta. Breytur sem raðast á sama þátt ættu að eiga fræðilega saman.



Mynd 2. Myndræn framsetning á muninum á hlut og þætti: Hlutur er fall af breytum – þáttur orsakar gildin sem breytur taka (samanber stefna örvanna)

Tvær gerðir þáttagreiningar eru leitandi og staðfestandi þáttagreining. Leitandi þáttagreining (e. exploratory factor analysis, EFA) getur gefið vísbendingu um innri formgerð að baki atriðasafni þegar það er ekki skýr kenning að baki (pre-liminary). Staðfestandi þáttagreining (e. confirmatory factor analysis, CFA) gefur vísbendingu um að hvaða marki þáttabygging samræmist því sem kenningar kveða á um (empirískur stuðningur fyrir því að matstæki endurspegli hugsmíð eins og hún er skilgreind fræðilega).

Fyrir ýtarlegri umfjöllun um leitandi og staðfestandi þáttagreiningu, sjá t.d. bók Fabrigar (2012) eða bók Brown (2015).

Vísbendingar byggðar á tengslum

Hægt er að varpa ljósi á réttmæti með því að bera niðurstöður á matstæki saman við önnur skyld og óskyld matstæki. Athuganir á tengslum skora veita upplýsingar um samleitni, sundurgreiningu, sammæli við ytri viðmiðsmælingar og forspárgetu m.t.t. annarra mælinga.

Tengsl skora á matstækjum sem meta sömu eða svipuð fyrirbæri veita okkur vísbendingar um samleitni (e. convergent evidence – þess sem sum hafa kallað „samleitniréttmæti") á meðan tengsl skora á matstækjum sem meta ólík fyrirbæri veita vísbendingar um sundurgreiningu (e. discriminant evidence – „sundurgreiningarréttmæti"). Sem dæmi myndi sterk jákvæð fylgni skora á Lífsgæðakvarðanum (e. Quality of Life Scale) við skor á Lífsánægjukvarðanum (e. Satisfaction with Life Scale) veita vísbendingu um samleitni þess fyrri við hinn síðari (þ.e. ef áhugi okkar beinist að réttmæti

Lífsgæðakvarðans – annars á hinn veginn). Þegar niðurstaða á einu matstæki samræmist niðurstöðu á öðru viðurkenndu matstæki á sama eða skyldu fyrirbæri eykur það traust okkar á því að það nái sannarlega utan um þá hugsmíð sem því er ætlað. Að sama skapi myndi hlutfallslega lág fylgni skora á kvíðakvarða Becks (e. Beck´s Anxiety Inventory) við skor á sjálfsmatskvarða Rosenbergs veita vísbendingu um sundurgreiningu. Skortur á tengslum (eða neikvæð tengsl) matstækja sem meta aðskilin / ólík / andstæð fyrirbæri auka traust okkar á því tæki sem áhugi beinist að. Athugun á sundurgreiningu matstækja er einkum mikilvæg í tilfellum þar sem fyrirbæri sem meta á hefur umtalsverðan samslátt við önnur fyrirbæri (dæmi um slíkt eru geðraskanir).

Vísbendingar um samleitni og sundurgreiningu eru yfirleitt fengnar með því að reikna fylgni milli skora matstækja. Í birtum greinum er algengt að sjá fylgnistuðla reiknaða í þeim tilgangi – en þó ber að hafa í huga að allt það sem hefur áhrif á fylgnistuðla almennt hefur áhrif á túlkunina.[5]

Önnur leið við réttmætisathuganir af þessu tagi er að skoða tengsl skora á matstæki við eitthvert ytra viðmið (e. test-criterion relationships – áður kallað „viðmiðsréttmæti“). Þá er spurt hversu vel skor á matstækinu sem um er rætt spá fyrir um þetta tiltekna viðmið. Breytan sem notuð er sem viðmið er eiginleiki eða útkoma sem er aðskilin frá matstækinu (sjá dæmi að neðan). Aðferðir byggja yfirleitt á fylgniútreikningum (t.d. aðhvarfsgreiningu) og má skipta þeim í tvennt eftir því hvort þær gefi vísbendingar um forspárhæfni (e. predictive evidence) matstækis eða vísbendingar um samtíma sammæli (e. concurrent evidence). Mat á forspárhæfni felur í sér að kanna tengsl á milli skora á matstækinu á einum tímapunkti og viðmiðsmælingunni á einhverjum öðrum tímapunkti. Dæmi um þetta væri t.d. að kanna hversu vel frammistaða á samræmdu könnunarprófi í 10. bekk (matstækið) spái fyrir um það hvort einstaklingarnir sem þreyttu prófið fari í háskóla seinna á lífsleiðinni (viðmiðið). Vísbendingar um samtíma sammæli fást aftur á móti með því að kanna tengsl á milli skora á matstækinu á tilteknum tímapunkti og viðmiðinu á u.þ.b. sama tímapunkti. Hér mætti t.d. nefna rannsókn þar sem þátttakendur byrja á því að svara spurningalista um kvíðaeinkenni (matstækið) og fara svo strax í kjölfarið í greiningarviðtal hjá geðlækni sem sker úr um hvort það sé með kvíðaröskun eða ekki (viðmiðið).

Rétt er að ítreka að fylgni eins matstækis við annað getur aldrei vitnað um „réttmæti þess” ein og sér. Munum orð Borsboom að matstæki sé réttmætt ef það metur fyrirbæri sem er til og sem orsakar breytileika í því sem við mælum. Fylgni matstækis við skyldar og óskyldar hugsmíðar, forspárhæfni og sammæli við ytri viðmið eru allt vísbendingar um réttmæti, en einar og sér geta þær aldrei gefið tæmandi mat á því.

Vísbendingar byggðar á afleiðingum notkunar

Réttmæti notkunar matstækis ræðst af ýmsu, meðal annars af því hversu viðeigandi notkun er í samhengi / aðstæðum mats. Þar skiptir máli að matstæki sé notað í þeim tilgangi sem það var hannað til – til að *magnbinda* einkenni eða *greina* sjúkdóm í ákveðnum *hópi* (ungmenna eða aldraðra, fatlaðra eða ófatlaðra, einstaklinga með verki, einstaklinga með sögu um áföll, o.s.frv.). Notkun er réttmæt að því marki sem aðstæður notkunar og einkenni svarendahóps eru í samræmi við hugmyndafræði að baki matstæki. Réttmæti notkunar ræðst sömuleiðis af því að hvaða marki skor eru túlkuð með hætti sem matstækið (próffræðilegir eiginleikar þess og ætlað notagildi) heimila. Til dæmis myndi slakur áreiðanleiki eða skortur á vísbendingum um réttmæti matstækis draga mjög úr trúverðugleika ályktana sem af skorum þess væru dregnar. Síðast en ekki síst geta afleiðingar notkunar matstækis

[5] Pearson fylgnistuðullinn gerir t.a.m. ráð fyrir tveimur samfelldum normallaga breytum. Þegar breytur eru ekki normallaga er rétt að nota Spearman eða Kendall stuðlana. Útreikningur fylgnistuðla gerir þar að auki ráð fyrir því að það sé einhvers konar línulegt samband fyrir hendi. Útlagar geta skekkt fylgnistuðla töluvert mikið, og skert dreifisvið breyta einnig.

upplýst um réttmæti hennar. Slíkar afleiðingar geta verið jákvæðar (aukin þjónusta) og neikvæðar („stimplun“), fyrirséðar (uppáskrift lyfja) og ófyrirséðar (aukaverkanir lyfja).

Samantekt

Meiningin með þessari umfjöllun um próffræðilega eiginleika var að skýra hugtökin áreiðanleiki og réttmæti með sem einföldustum hætti, fjalla um þær aðferðir sem helst eru notaðar við mat á þeim og takmarkanir þeirra aðferða eftir atvitvikum. Okkar von er að efnið hafi skýrt sýn lesenda á flækjustig og blæbrigði við mat á próffræðilegum eiginleikum. Munum að niðurstöður rannsókna okkar verða aldrei merkilegri eða marktækari en mælingarnar sem þær byggja á.

Heimildir

American Educational Research Association, American Psychological Association, & National Council on Measurment in Education. (2014). *The Standards for Educational and Psychological Testing*. American Educational Research Association. https://www.apa.org/science/programs/testing/standards

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137-143. https://doi.org/10.1007/s11336-008-9100-1

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Einar Guðmundsson. (2017). [Fyrirlestrar um áreiðanleika og réttmæti]. Sálfræðideild, Háskóli Íslands.

Fabrigar, L. R., & Wegener, D. T. (2012*). Exploratory factor analysis*. Oxford University Press.

Flake, J. K. & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*(4), 484–501. https://doi.org/10.1177/2515245920951747

Foster, R. C. (2021). KR20 and KR21 for Some Nondichotomous Data (It's Not Just Cronbach's Alpha). *Educational and Psychological Measurement, 81*(6), 1172-1202. https://doi.org/10.1177/0013164421992535

Gadermann A. M., Guhn M., Zumbo B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. Practical Assessment, Research & Evaluation, 17(1), 1–13. https://doi.org/10.7275/n560-j767

Gessaroli, M. E., & Folske, J. C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing, 2*(3-4), 277-295. https://doi.org/10.1080/15305058.2002.9669496

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and psychological measurement, 66*(6), 930-944. https://doi.org/10.1177/0013164406288165

Green, S. B., & Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural equation modeling, 7*(2), 251-270. https://doi.org/10.1207/S15328007SEM0702_

Green, S. B., & Yang, Y. (2009a). Commentary on coefficient alpha: A cautionary tale.

Psychometrika, 74(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Green, S. B., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika, 74*, 155-167. https://doi.org/10.1007/s11336-008-9099-3

Hayes, A. F. & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley

Malkewitz, C. P., Schwall, P., Meesters, C., & Hardt, J. (2023). Estimating reliability: A comparison of Cronbach's α, McDonald's ωt and the greatest lower bound. *Social Sciences & Humanities Open, 7*(1), 100368.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. British Journal of Psychology, 88(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. Frontiers in Psychology, 3. https://doi.org/10.3389/fpsyg.2012.00261

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Portney, L. G. (2020). *Foundations of clinical research: applications to evidence-based practice.* FA Davis.

Savalei, V., & Reise, S. P. (2019). Don't Forget the Model in Your Model-based Reliability Coefficients: A Reply to McNeish (2018). *Collabra: Psychology, 5(*1), 36. https://doi.org/10.1525/collabra.247

Sheng Y., Sheng Z. (2012). Is coefficient alpha robust to non-normal data? *Frontiers in Psychology, 3*, 1–13. https://doi.org/10.3389/fpsyg.2012.00034

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements*. Frontiers in psychology, 7*, 769. https://doi.org/10.3389/fpsyg.2016.00769

Zimmerman D. W., Zumbo B. D., Lalonde C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement, 53*, 33–49. https://doi.org/10.1177/07399863870092005

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating Generalizability to a

Latent Variable Common to All of a Scale's Indicators: A Comparison of Estimators for ωh. Applied Psychological Measurement, 30(2), 121-144. https://doi.org/10.1177/0146621605278814

Zumbo, B.D., Gadermann, A.M. & Zeisser, C. (2007). Ordinal Versions of Coefficients Alpha and Theta for Likert Rating Scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29. https://doi.org/10.22237/jmasm/1177992180

Xiao, L., & Hau, K.-T. (2023). Performance of Coefficient Alpha and Its Alternatives: Effects of Different Types of Non-Normality. *Educational and Psychological Measurement, 83*(1), 5-27. https://doi.org/10.1177/00131644221088240

Yang, Y., & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling, 17*(1), 66–81. https://doi.org/10.1080/10705510903438963