Die bepunting van meerkeusige vrae

Author:
John J. Barnard[1]

Affiliation:
[1]Excel Psychological and Educational Consultancy (EPEC) Pty. Ltd.

Correspondence to:
John Barnard

Email:
John@EPECat.com

Postal address:
PO Box 3147, Doncaster East, VIC 3109, Australia

Dates:
Received: 18 Apr. 2013
Accepted: 09 May 2013
Published: 27 June 2013

How to cite this article:
Barnard, J.J., 2013, 'Die bepunting van meerkeusige vrae', *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 32(1), Art #402, 7 pages. http://dx.doi.org/10.4102/satnt.v32i1.402

In hierdie artikel word verskillende metingsteorieë kortliks bespreek en vergelyk wat betref die bepunting van meerkeusige vrae. Daar word gedemonstreer hoe die hantering van ontbrekende response 'n beduidende impak kan hê op 'n toetsling se prestasie en hoe dit ten beste gehanteer word volgens die verskillende moderne teorieë. Die feit dat toetslinge meerkeusige vrae korrek kan beantwoord deur te raai is jare lank al die onderwerp van bespreking. Daar word aangeneem dat toetslinge meestal nie géén kennis hoegenaamd van die inhoud van 'n toepaslike toets het nie, en dus geneig is om ingeligte raaiskote te waag eerder as wilde raaiskote. Probleme met die klassieke korreksie vir raai, die gebruik van passingsindekse in Rasch en die raai-itemparameter in die itemresponsteorie (IRT) word beklemtoon. Die slotsom waartoe gekom word, is dat die keusewaarskynlikheidsteorie moontlik die beste oplossing bied. Die drieparameter- logistiese itemresponsteorie (IRT)-model sluit 'n 'raai-itemparameter' in om die kanse aan te dui dat 'n toetsling die regte antwoord geraai het. Daar moet egter onthou word dat dit 'n persoon is wat raai, nie 'n item nie, en dat die raaiparameter dus eintlik 'n persoonsparameter moet wees. Die doel van die keusewaarskynlikheidsteorie is om hierdie probleem te oorkom deur te probeer vasstel wat die graad van sekerheid is wat die toetsling het dat 'n bepaalde opsie die regte een is. Relastiese allokasies van hierdie waarskynlikhede dui die graad van raai aan en bied dus meer noukeurige maniere om vermoëns te meet.

Scoring multiple choice questions. This article briefly touches on how different measurement theories can be used to score responses on multiple choice questions (MCQs). How missing data is treated may have a profound effect on a person's score and is dealt with most elegantly in modern theories. The issue of guessing a correct answer has been a topic of discussion for many years. It is asserted that test takers almost never have no knowledge whatsoever of the content in an appropriate test and therefore tend to make educated guesses rather than random guesses. Problems related to the classical correction for guessing is highlighted and the Rasch approach to use fit statistics to identify possible guessing, is briefly discussed. The three-parameter 'logistic' item response theory (IRT) model includes a 'guessing item parameter' to indicate the chances that a test taker guessed the correct answer to an item. However, it is pointed out that it is a person that guesses, not an item, and therefore a guessing parameter should be a person parameter. Option probability theory (OPT) purports to overcome this problem through requiring an indication of the degree of certainty the test taker has that a particular option is the correct one. Realistic allocations of these probabilities indicate the degree of guessing and hence more precise measures of ability.

Inleiding

Toetse is deel van ons daaglikse lewens. Studente word deurlopend getoets om te bepaal of sekere inhoude bemeester is, leerprobleme te identifiseer en so meer. Ons moet toetse doen om rybewyse te kry, wanneer ons aansoek doen vir werk of vir verdere studies, ens. Maar hoe geldig is hierdie toetse? Wat kan afgelei word van die punte wat toegeken is? Sekere toetse het natuurlik meer impak as ander, maar as 'n toets gegee word, word die punte wat behaal is, altyd geïnterpreteer en gebruik.

Die bepunting van toetse is eenvoudig – 'n getal vrae word gevra, die getal regte antwoorde word bymekaar getel en 'n telling word verkry. Indien meer as een persoon die toets skryf, kan die tellings met mekaar vergelyk word en gevolgtrekkings oor die relatiewe prestasies gemaak word. Maar presies wat is getoets en kan die individuele punte sonder meer bymekaar getel word om 'n totaal te gee?

Veronderstel 'n toets in wiskundige vermoë moet opgestel word. Eerstens moet die konstruk 'wiskundige vermoë' gedefinieer word. Die definisie kan gebaseer word op gekristalliseerde

kennis en 'n persoon kan 'n hoë telling behaal vanweë harde werk en memorisering van inhoud. Aan die ander kant kan die definisie gebaseer word op laterale denke in die sin dat probleme wat nie voorheen teëgekom is nie, opgelos moet word. Dit is duidelik dat die bloudruk van die toetse, gebaseer op hierdie uiteenlopende definisies, sal verskil met die gevolg dat die punte nie op dieselfde wyse geïnterpreteer kan word nie. Ongeag watter definisie gebruik word, kan daar ander faktore wees wat die punte beïnvloed. Indien daar byvoorbeeld 'storieprobleme' ingesluit word, kan taal 'n rol speel in die sin dat die persoon nie die vraag verstaan nie. Taal sou dan 'n tweede dimensie wees en die toetsresultate kan nie gebruik word vir die meting van bevoegdheid in suiwer wiskundige vermoë nie. So 'n toets sou ongeldig as 'n toets vir wiskundige vermoë alleenlik wees.

Die meting van konstrukte word gebaseer op teorieë. 'n Metingsteorie kan gedefinieer word as 'n versameling aannames, definisies en stellings wat, wanneer aan die aannames voldoen word, kan lei tot die bepaling van die psigometriese eienskappe van die metings. In so 'n teorie kan punte toegeken word op 'n gefundeerde wyse deur die administrasie van 'n metingsinstrument (toets). Toetse word saamgestel uit vrae (items) en items is dus die boustene van toetse. (In die lig daarvan dat verskillende tipes item verskillende analitiese tegnieke vereis, fokus hierdie artikel slegs op meerkeusige tipe vrae (MCQs) met slegs een korrekte antwoord.) Die metingsteorie moet dit moontlik maak om getalle (metings) op 'n kontinue skaal te genereer. Deur die metings te vergelyk moet betekenis gegee kan word aan dit wat gemeet is.

In hierdie artikel word verskillende metingsteorieë kortliks toegelig en vergelyk. Dit is belangrik om in ag te hou dat verskillende teorieë verskillende vlakke van data (response op items) vereis en dat sommige teorieë oorwegend toetsgerig en ander weer itemgerig is. Jare gelede is al besef dat alle data nie op dieselfde vlak is nie. Stevens (1946) het byvoorbeeld vier vlakke (skale) van data onderskei, naamlik nominale, ordinale, interval en verhoudings. Een van die opmerklikste uitkomste van hierdie onderskeid is dat sekere vlakke van data vereis word voordat bepaalde statistiese tegnieke gebruik kan word. Byvoorbeeld, 'n *t*-toets kan nie uitgevoer word met twee stelle nominale data nie.

Klassieke toetsteorie (ook verwys na as tradisionele toetsteorie) is 'n toetsgebaseerde teorie en sal eerste kortliks toegelig word – kyk byvoorbeeld Crocker en Algina (1986). Itemresponsteorie (IRT) – kyk, byvoorbeeld, Hambleton en Swaminathan (1985) en Rasch se metingsteorie – kyk, byvoorbeeld, Bond en Fox (2007) is voorbeelde van itemgebaseerde teorieë. Die finale teorie, keusewaarskynlikheidsteorie Barnard (2012), word laastens aan die orde gestel.

Klassieke toetsteorie

Klassieke toetsteorie (KTT) dien al meer as 'n eeu as 'n hoeksteen in item- en toetsontleding en word steeds algemeen gebruik. KTT is 'n metingsteorie gebaseer op die grondbeginsel dat 'n waargenome telling behaal in 'n toets 'n funksie van die persoon se 'ware' telling en 'n fouttelling is. Die waargenome telling is die totale telling gebaseer op die itemtellings soos toegeken tydens die nasien van die toets. Indien daar byvoorbeeld tien items in 'n toets is en een punt word toegeken aan 'n korrekte antwoord, sal 'n persoon wat ses vrae korrek beantwoord het 'n waargenome telling van ses (uit 10) ofte wel 60% behaal. Daar word dikwels 'n onderskeid gemaak tussen items wat nie beantwoord is nie en oorgeslaan is en items wat nie bereik is nie. Eersgenoemde word gewoonlik as verkeerd gemerk op grond van die redenasie dat die toetsling die item moes gesien het omdat 'n volgende item beantwoord is. Items aan die einde van 'n toets wat nie beantwoord is nie, kan verskillend geïnterpreteer word. Indien gereken word dat genoeg tyd toegestaan is, kan sulke ontbrekende response verkeerd gemerk word. Indien sulke ontbrekende response egter as ontbrekend beskou word, sal 'n persoon wat slegs die eerste agt items van die 10-itemtoets beantwoord het en ses daarvan korrek beantwoord het, 'n waargenome telling van ses uit agt oftewel 75% behaal. Die regverdigheid, al dan nie, teenoor 'n ander persoon wat al tien items beantwoord het en ses daarvan korrek en dus 'n waargenome telling van 60% het, spreek vanself.

Die metingsfout dui aan dat die waargenome telling nie die absolute telling van die persoon se werklike vermoë verteenwoordig nie. Indien ander items byvoorbeeld in die toets ingesluit was, kon die persoon beter of slegter gepresteer het. Indien die toets meer items bevat het, sou 'n beter skatting van die persoon se werklike vermoeë verkry kon word omdat meer items meer inligting verstrek. Dit sou dus beteken dat die metingsfout kleiner is. Al hierdie aspekte hou verband met die basiese begrippe van KTT, naamlik geldigheid, betroubaarheid en die standaardmetingsfout.

Indien meerkeusige vrae gevra word, kan daar gereken word dat een of meer van die items korrek beantwoord is deur te raai. Verskillende variasies van formules om waargenome tellings te korrigeer deur moontlike raai te akkommodeer, kan in die literatuur gevind word. Een hiervan is om te reken dat indien 'n persoon V items verkeerd beantwoord het, dit aanvaar kan word dat die persoon 'n aantal van die ander items korrek geraai het en dat dit verband hou met hoeveel items verkeerd beantwoord is en hoeveel keuses daar in die items was (V/(k-1)) en dat hierdie faktor afgetrek moet word van die waargenome telling:

$$G = R - \frac{V}{k-1} \qquad \text{[Vergelyking 1]}$$

Waar G = gekorrigeerde telling, R = getal items reg beantwoord, V = getal items verkeerd beantwoord en k = getal keuses in die items is. Indien hierdie korreksie byvoorbeeld toegepas word op 'n persoon wat ses uit tien vierkeusige items korrek beantwoord het, is die gekorrigeerde telling 4.7 oftewel 47%. Die toepassing van sulke korreksies en besluite oor ontbrekende response kan dus 'n betekenisvolle invloed op 'n persoon se waargenome telling hê.

doi:10.4102/satnt.v32i1.402

Die korreksie-vir-raaiformule veronderstel dat die toetsling 'n item korrek kan beantwoord deur blindweg te raai. Dit kan wel die geval wees by sommige items, maar by baie items kan van die keuses uitgeskakel word as die moontlike korrekte antwoord. Gedeeltelike kennis kan dus die kanse verhoog om die item korrek te raai deur van die keuses uit te skakel. Die blindelingse raai word dus eerder 'n ingeligte raai. Hier in is 'n paradoks: 'n persoon wat byvoorbeeld 25 uit 'n honderd vierkeusige items korrek beantwoord het, sal 'n gekorrigeerde waargenome telling van nul kry. Dit is hoogs onwaarskynlik dat 'n persoon wat deel van 'n groep mense is vir wie die toets bedoel was, geen kennis sal hê nie. Die persoon sal sommige antwoorde weet en ander gedeeltelik ken, maar nie glad geen kennis hê nie. Dit is duidelik dat die korreksie-vir-raaiformule nie die probleem van moontlike raai in toetse oplos nie.

Twee sentrale begrippe in KTT om 'goeie' items te identifiseer is die item se moeilikheidswaarde en die diskriminasiewaarde. Eersgenoemde gee 'n aanduiding van hoe moeilik of maklik 'n item is en word bereken as die verhouding van die getal persone wat die item korrek beantwoord het tot die getal persone wat in die toetsgroep was. Itemdiskriminasie word gebruik om 'n aanduiding van die kwaliteit van die item te verkry. Hierdie indeks word tradisioneel bereken as die punttwee-reekskorrelasie:

$$r_{pbis} = \frac{\overline{X_r} - \overline{X_t}}{SD_t}\sqrt{\frac{p_i}{q_i}} \qquad \text{[Vergelyking 2]}$$

Waar $\overline{X_r}$ = gemiddeld is van die toetslinge wat die item korrek beantwoord het, $\overline{X_t}$ = gemiddeld van al die toetslinge se tellings, SD_t = die standaardafwyling van al die toetstellings, p_i = die moeilikheidswaarde van die item en $q_i = 1 - p_i$

In KTT word die betroubaarheid van 'n toets gedefinieer as die verhouding van die variansie van die ware tellings tot die variansie van die waargenome tellings:

$$Betroubaarheid = r_{tt} = \frac{Variansie_{Ware}}{Variansie_{Waargenome}} = \frac{SD_T^2}{SD_O^2} \qquad \text{[Vergelyking 3]}$$

'n Betroubaarheidsindeks word bereken as 'n korrelasie tussen twee stelle tellings. Weens probleme geassosieer met die administrasie van dieselfde toets by twee geleenthede of die gebruik van parallelle vorms van dieselfde toets, het psigometriste die begrip 'interne konsistense' ontwikkel sodat 'n toets net een keer toegepas hoef te word om 'n betroubaarheidsindeks te bereken. Cronbach se koëffisiënt alpha gee 'n algemene indeks van die betroubaarheid van 'n toets:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} SD_i^2}{SD_t^2}\right) \qquad \text{[Vergelyking 4]}$$

Waar k = getal items in die toets is. In die spesiale geval waar meerkeusige vrae digotomies bepunt word, kan die noemer in die breuk vereenvoudig word sodat die formule die algemeen toegepaste Kuder-Richardson-formule 20 word:

$$\alpha = KR\,20 = \frac{k}{k-1}\left(1 - \frac{\sum pq}{SD_t^2}\right) \qquad \text{[Vergelyking 5]}$$

'n Minimumbetroubaarheidsindeks van 0.8 word algemeen as 'n minimum aanvaar omdat minstens 60% van die variansie verklaar kan word. Indien hierdie minimum nie met die toets bereik word nie, word die Spearman Brown-formule dikwels gebruik om te bepaal hoeveel items in die toets moet wees om 'n vereiste betroubaarheid te bereik. Die betroubaarheidsindeks kan ook gebruik word om 'n aanduiding van die akkuraatheid van waargenome tellings te verkry. Dit word gedoen deur die standaardmetingsfout (SMF) te bereken:

$$SMF = SD_{tt}\sqrt{1 - r_{tt}} \qquad \text{[Vergelyking 6]}$$

Waar SD_{tt} = standaardafwyking van die waargenome tellings en r_{tt} = betroubaarheidsindeks van die toets.

Hoewel die KTT algemeen gebruik is (en steeds gebruik word) deur metingsteoretici, het sekere belangrike probleme aan die lig gekom. Hierdie probleme is hoofsaaklik in die toepassing van die teorie. Eerstens is die item statisties steekproefafhanklik. Indien dieselfde item ingesluit word in 'n toets wat by twee uiteenlopende groepe mense toegepas word, sal hierdie waarde verskillend wees. Die waarde bereken uit 'n groep hoë presteerders sal aandui dat die item maklik is en die waarde bereken uit 'n groep onderpresteerders sal die item moeilik laat voorkom. In dieselfde lig, indien 'n item op 'n homogene groep mense toegepas word, sou die diskriminasiewaarde laer wees as wanneer dieselfde item op 'n heterogene groep toegepas word. Dit is een van die redes waarom groot steekproewe nodig is sodat sulke ekstreme geminimaliseer kan word. Tweedens is die waargenome tellings direk afhanklik van die moeilikheidsgraad van die toets. Laer prestasie sal uiteraard volg uit 'n toets wat uit moeilike items bestaan. 'n Derde probleem is dat die waargenome tellingskaal nie 'n intervalskaal is nie. Indien 'n toetsling byvoorbeeld wil verbeter van 'n telling van 23 na 24, word minder vermoë vereis as wanneer die verbetering van 48 na 49 in 'n 50-itemtoets is omdat die toetsling in laasgenoemde 'n moeiliker vraag moet reg beantwoord. Maar op die waargenome tellingskaal is dit net een punt meer – ongeag van waar op die skaal die verbetering is. Vierdens word, verkeerdelik, veronderstel dat die metingspresisie konstant is vir al die toetslinge. Dit is intuïtief duidelik dat toetslinge met lae tellings geneig is om meer te raai as toetslinge wat hoog presteer. Die SMF word egter as 'n konstante bereken vir alle toetslinge. Ten slotte is die moeilikheidswaardes van die items en die waargenome tellings van die toetslinge nie op dieselfde skaal nie. 'n Item met 'n moeilikheidswaarde van 0.6 kan byvoorbeeld nie sonder meer geassosieer word met 'n waargenome telling van 60% nie – eersgenoemde is 'n psigometriese begrip wat aandui watter verhouding toetslinge 'n bepaalde item korrek beantwoord het terwyl

laasgenoemde 'n statistiese begrip is wat aandui hoeveel items 'n sekere toetsling korrek beantwoord het. Oplossing vir hierdie en ander verwante probleme het die soeke na 'n sterker teorie aangespoor.

Rasch se metingsteorie

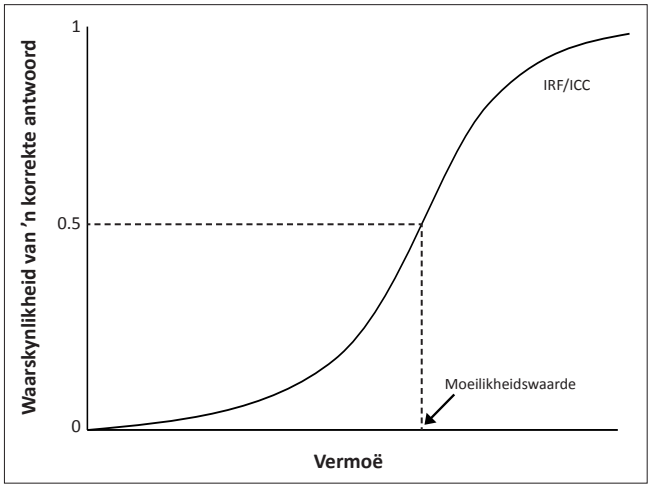
Rasch se metingsteorie is 'n filsosofie wat die belofte inhou om die KTT-probleme te oorkom. In wese is dit gegrond op die idee dat daar 'n bepaalde ordelikheid is wat verder strek as 'rou' tellings. Die waargenome telling word gesien as 'n voldoende statistiek in die lig daarvan dat dit 'n bepaalde responspatroon verteenwoordig. Indien 'n waargenome telling gegee word, kan daar bepaal word watter items korrek beantwoord behoort te geword het. Dit veronderstel natuurlik dat die moeilikheidswaardes van die items en die toetslinge se vermoeëns op dieselfde skaal geplaas kan word. Dit word gedoen deur die itemresponsfunksie (IRF) wat die verwantskap tussen die toetslinge se vermoëns en die waarskynlikheid vir 'n korrekte respons beskryf in terme van 'n wiskundige vergelyking (Figuur 1).

Hierdie monotone toenemende funksie met wiskundige vorm $y = \frac{1}{1+e^{-x}}$ is die basis van Rasch (en IRT) en wordalgemeen in die volgende vorm geskryf in die Rasch-literatuur:

$$P\{x_{ni} = 1 / \beta_n, \delta_i\} = \frac{e^{\beta_n - \delta_i}}{1 + e^{\beta_n - \delta_i}} \qquad \text{[Vergelyking 7]}$$

Hierdie vergelying verbind die waarskynlikheid vir 'n uitkoms om een te wees vir toetsling n met vermoë β en item i met moeilikheidswaarde δ. Elke item word beskryf deur so 'n IRF waar die moeilikheidswaarde van die item gedefinieer word as die plek waar die toetsling 'n 50% waarskynlikheid het om die item korrek te beantwoord. Die IRF's van al die items het dieselfde vorm en die kurwe skuif verder regs op die as soos wat die item moeiliker word.

Hierdie funksie kan gebruik word om 'n getal verskillende aspekte te verklaar. Veronderstel dieselfde item word aan twee verskillende groepe toetslinge gegee en die een groep se vermoëns is almal onder gemiddeld en die ander groep se vermoëns is almal bo gemiddeld. Vir die eerste groep sal al die data aan die linkerkant (kyk Figuur 1) onder die aangeduide moeilikheidswaarde wees. Vir dié data kan hierdie deel van die IRF gepas word en deur die vergelyking te gebruik, kan die funksie geëkstrapoleer word om die IRF te voltooi – selfs vir vermoëns waar geen toetslinge was nie. Dieselfde kan gedoen word vir die ander groep. In ideale omstandighede sal die twee funksies identies wees, maar in die werklikheid sal hulle binne metingsfout wees. Hierdie redenasie het ten gevolg dat die moeilikheidswaarde van die item bepaal kan word van enige steekproef in 'n populasie wat nie verskillend sal wees vanweë die vermoëns van toetslinge in 'n spesifieke steekproef nie.



FIGUUR 1: 'n Itemresponsfunksie wat die waarskynlikheid van 'n korrekte antwoord op 'n item as 'n funksie van die toetsling se vermoë aandui.

Metings (punte) volgens Rasch word bereken deur logaritmiese waarskynlikhede en uitgedruk op 'n vermoënskaal as 'logits'. Let wel dat 'n toetsling se vermoë nie die som van korrekte punte is nie, maar bereken word uit die IRF's. Die IRF's word vermenigvuldig om 'n aanneemlikheidsfunksie te genereer en die maksimum van hierdie funksie word beskou as die beste skatting van die toetsling se vermoë. Die toetslinge se vermoëns op die logitskaal, wat byna uitsluitlik strek van -3 tot 3, kan na enige gerieflike skaal getransformeer word – selfs na 'n skaal wat metings laat lyk soos waargenome tellings. Die verskil is dat die metings op 'n intervalskaal geplaas is en dus weerspieël dit dat die vermoë om van een punt op die skaal na die volgende te skuif, nie konstant is nie – die punte is tipies nader aan mekaar in die middel van die skaal en verder van mekaar namate weg beweeg word van die middel. Merk ook op dat die relatiewe posisies van twee toetslinge behoue bly, ongeag of 'n maklike of 'n moeilike toets toegepas word. Dit word bereik deur die logaritmiese transformasies. Dit kan byvoorbeeld nie in die klassieke sin gedoen word deur z-tellings of persentiele nie, omdat transformasies na standaardtellings (z) liniêr is en gevolglik sal dieselfde verskil tussen twee waargenome tellings behoue bly in z-tellings. Persentiele, aan die ander kant, sien af van die werklike verskille tussen toetslinge se waargenome tellings en oorkom die probleme van verskillende afstande tussen twee toetslinge, maar die waargenome tellings word getransformeer na range en is dus nie meer beskikbaar nie.

Een van die mees prominente gebruike van die toepassing van Rasch-metodes word moontlik gemaak deur die vermoëns van die toetslinge en die moeilikheidswaardes van die items op dieselfde skaal te kan plaas. Indien die moeilikheidswaardes van 'n getal items bekend is, kan die waarskynlikheid dat enige item korrek beantwoord sal word deur enige toetsling, bereken word. Dít is een van die kragtigste onderskeidings in Rasch se metingsteorie teenoor die KTT.

In KTT word die waargenome tellings gebruik as die basis van prestasie en gevolgtrekkings word daarop gebaseer. Sodanige tellings word verkry uit die response op die items in 'n toets en is dus beperk daardeur. In teenstelling hiermee word die response op items in Rasch gebruik om metings te

kry op die onderliggende konstruk wat gemeet word. Die verskil is baie subtiel, maar van kardinale belang. In KTT roer die items sekere aspekte aan en word die totaal bereken as die som van die korrekte antwoorde. In Rasch meet elke item 'n aspek van die latente eienskap en word die items deeglik ondersoek vir die mate waarin hieraan voldoen word. KTT kan dus as beskrywend beskou word, omdat geen aannames aangaande die latente eienskap gemaak word nie, terwyl Rasch eerder inferensieel is.

Volgens Rasch se metingsteorie verskil items slegs op grond van hul moeilikheidswaardes. Itemdiskriminasie en raai word geakkommodeer in passingsindekse, gewoonlik bereken as 'n tipe X^2 -statistiek. Waardes buite 'n bepaalde interval word gewoonlik geïnterpreteer as 'statistiese geraas' waarna verwys kan word as die gevolg van raai of ander faktore. Dit is belangrik om daarop te let dat, anders as in KTT, 'n SMF vir elke item en elke toetsling bereken word. Die SMF word gedefinieer as die omgekeerde van die vierkantswortel van statistiese inligting wat bereken word deur die itemvariansie. Die SMF van 'n toetsling wat meer geraai het (moeilike items korrek beantwoord het) sal groter wees as vir 'n toetsling vir wie dit nie die geval was nie. Soortgelyk sal 'n item waarvan die antwoord deur baie toetslinge geraai is, 'n groter SMF hê as vir 'n item waarin dit nie die geval was nie.

Itemresponsteorie

Itemresponsteorie (IRT) is 'n statistiese benadering wat verskillende wiskundige modelle gebruik om eienskappe van items te akkommodeer en skattings van die vermoëns van toetslinge te lewer. Die eenparameter- logistiese model skyn dieselfde te wees as die Rasch-model, terwyl die tweeparameter- logistiese model IRF's toelaat om verskillende hellings te hê wat beteken dat verskillende diskriminasiewaardes vir items toegelaat word. 'n Item sal optimaal diskrimineer by die moeilikheidswaarde van die item ('n raaklyn aan die IRF sal hier die grootste gradiënt hê) en minder effektief diskrimineer namate verder weg beweeg word van hierdie punt. Dit kan soos volg verduidelik word: 'n Item met gemiddelde moeilikheidswaarde sal optimaal diskrimineer vir gemiddelde toetslinge. Die item sal hoogs waarskynlik korrek beantwoord word deur toetslinge met hoë vermoëns, maar sal nie werklik tussen twee toetslinge met hoë vermoëns kan onderskei nie, omdat beide die item waarskynlik korrek sal beantwoord. Dieselfde redenasie geld vir toetslinge met lae vermoëns wat die item hoogs waarskynlik verkeerd sal beantwoord (tensy korrek geraai word). In teenstelling met KTT wat een diskriminasiewaarde vir 'n item bereken, is hierdie indeks eerder 'n funksie van vermoë in IRT. Dit is dus belangrik dat items korrek geteiken moet word.

Tot dusver is aanvaar dat daar geen waarskynlikheid vir 'n toetsling is om 'n item korrek te beantwoord deur te raai nie. In KTT is korreksie vir raai kortliks bespreek en in Rasch en die een- en tweeparameter- IRT-modelle word moontlike raai geïdentifiseer deur passingsindekse. Die feit bly staan dat dit moontlik is om 'n meerkeusige item korrek te beantwoord deur te raai. Blindelingse raai op 'n vierkeusige item beteken dat daar 'n 25% kans is dat die item korrek beantwoord kan word. Sommige navorsing het bevind dat toetslinge met ondergemiddelde vermoëns beter kan presteer deur blindelings te raai eerder as om die korrekte antwoord te probeer identifiseer omdat afleiers sulke toetslinge soms maklik kan aflei. Die drieparameter- IRT-model sluit 'n 'raai'-parameter (pseudokans) in wat 'n aanduiding gee van die waarskynlikheid vir enige toetsling om die korrekte antwoord te raai. Let op dat hierdie 'n itemparameter is.

Dit sal nooit bekend wees of 'n toetsling 'n item korrek beantwoord het deur te raai of omdat die toetsling die antwoord geken het nie tensy die toetsling gevra word. Somige studies het 'n vraelys ingesluit waarin toetslinge gevra is of hulle die antwoord geweet of geraai het. Die vraag is of dit nie in 'n toets ingesluit kan word nie. Let op dat dit mense is wat raai en nie items nie en dus moet so 'n parameter eerder 'n toetslingparameter wees as 'n itemparameter.

Keusewaarskynlikheidsteorie

Tegnologiese vooruitgang het die afneem van toetse op rekenaars moontlik gemaak. Dit is ook moontlik om toetslinge te vra of antwoorde geraai is of nie. Daar kan selfs gevra word tot watter mate die toetsling die antwoord geweet het, of seker is van die antwoord. Dit is presies die beweegrede agter die keusewaarskynlikheidsteorie – Barnard (2012). Ongeag of 'n toetsling blindelings of ingelig geraai het, is daar 'geraas' in tellings – met ander woorde 'n metingsfout. Die metingsfout kan geminimaliseer word deur die toetsling te vra na die waarskynlikheid dat enigeen van die opsies die korrekte antwoord kan wees. Indien die toetsling die korrekte antwoord weet (of so dink), kan 100% toegeken word aan die opsie, andersins kan verskillende waarskynlikhede toegeken word aan die verskillende opsies. Dit is ook moontlik om 'n realisme-indeks te bereken wat 'n aanduiding gee van hoe realisties die toetsling oor sy of haar kennis is. 'n Toetsling kan byvoorbeeld hoë waarskynlikhede toeken aan verkeerde response wat oorskatting van vermoë beteken.

Navorsing oor die toekennings van waarskynlikhede aan opsies dateer terug na die sestigerjare en verskeie benaderings is oorweeg en gekritiseer – kyk byvoorbeeld De Finetti (1970). Een van die kernprobleme in die implementering van die modelle was die praktiese uitvoerbaarheid daarvan, net soos rekenaaraanpassingstoetse wat toenemend veld wen. Die implementering van die modelle het rekenaars vereis vir die administrasie en bepunting van die toetse en in die 60s en 70s was rekenaars nie geredelik toeganklik nie.

Hierdie modelle het dus, anders as tipiese digotomiese bepunting van veelkeusige vrae, 'n bepuntingsreël gebruik as die basis van die waarskynlikhede wat 'n toetsling aan die verskillende opsies toeken. Een hiervan is 'n sferiese reël.

Punt = $\frac{P_k}{\sqrt{\sum P_o^2}}$ waar P_k die waarskynlikheid toegeken aan die korrekte antwoord is en P_o die waarskynlikhede toegeken aan die opsies. Hierdie bepuntingsreël het verlangde eienskappe soos 'n minimum telling indien 'n waarskynlikheid van nul aan die korrekte antwoord toegeken word, ongeag

doi:10.4102/satnt.v32i1.402

die waarskynlikhede toegeken aan die ander opsies en 'n maksimum telling indien 'n 100% waarskynlikheid aan die korrekte antwoord toegeken word. Dit kan nagegaan word dat die reël 'n toetsling wat waarskynlikhede meer realisties toeken, bevoordeel en dat 'n ervare toetsling die telling kan verhoog deur sekere opsies uit te skakel as moontlike korrekte antwoorde. Maar, die verwagte telling is in die algemeen laer as die waargenome telling.

'n Ander benadering was om die waarskynlikhede toegeken aan al die opsies te oorweeg en 'n kwadratiese bepuntingsreël te gebruik: telling $= 2P_k - P_o^2$. 'n Meer gesofistikeerde kwadratiese reël is soos volg:

Telling $= 1 + P_k^2 - (1 - P_k)^2 - \sum P_o^2$. [Vergelyking 8]

Die wesenlike probleem met hierdie tipe bepuntingsreël is dat die itemtelling beïnvloed word deur hoe die waarskynlikhede toegeken word aan die verskillende opsies. Dit kan byvoorbeeld nagegaan word dat 'n toetsling wat sekere opsies uitskakel, 'n laer telling behaal as 'n toetsling wat dit nie doen nie. Dit is teenstrydig met die logika dat 'n toetsling wat sekere opsies kan elimineer, gedeeltelike kennis het. Dus, alhoewel hierdie bepuntingsreëls sekere verlangde eienskappe het, is sulke reëls afhanklik van hoe die waarskynlikhede toegeken word aan die opsies in die item.

In teenstelling met oorweging van die waarskynlikhede toegeken aan al die opsies om die item te bepunt, is dit duidelik dat die bepuntingsreël gebaseeer moet word op die waarskynlikheid toegeken aan die korrekte opsie. Dit kan wiskundig bevestig word dat 'n logaritmiese bepuntingsreël die enigste reël is met hierdie eienskap vir items met drie of meer opsies. Ongeag van hoe die item bepunt word, word waarskynlikhede $p(i)$ toegeken aan die opsies en die verwagte telling word gemaksimeer as 'n produk van die waarskynlikhede en die telling:

$E = \sum p(i) \times s(r(i))$ waar $s(r(i)) = F(r(i))$ vir die itemtelling as i die korrekte opsie is. [Vergelyking 9]

Die funksie F definieer die bepuntingsreël. Indien die toetsling realisties is oor sy of haar kennis, sal hoër waarskynlikhede in die algemeen toegeken word aan korrekte antwoorde en laer waarskynlikhede aan items wat verkeerd beantwoord word. Dit sal lei tot die meeste items met dieselfde verspreiding van $p(i)$, as die relatiewe frekwensie korrek van al die opsies toegeken $r(i)$ namate $p(i)$ benader word. Dit sal lei tot akkurate skattings en 'n meer realistiese resultaat as byvoorbeeld hoë waarskynlikhede toegeken aan verkeerde opsies. 'n Bepuntingsreël F moet dus realistiese toekennings van waarskynlikhede aanmoedig, veral aan die korrekte opsie. Die telling $s(r(i))$ is 'n funksie van F en moet maksimaal wees slegs as $r(i) = p(i)$ vir alle i. Sodanige funksie kan afgelei word deur parsiële differensiasie op voorwaarde dat $\sum r(i) = 1$ deur die gebruik van die Lagrange-vermenigvuldiger λ.

$$\frac{\partial[\sum(p(i)F(r(i)) + \lambda(1 - \sum r(i)]}{\partial r(i)} = 0 \text{ as } p(i) = r(i) \text{ vir alle } i$$

met gevolg dat $F(r(i)) = A1n(r(i)) + B$ vir konstantes A en B. Indien A en B sodanig gekies word dat die telling nul is vir 'n uniforme distribusie $r(i)$, sal 'n toetsling se telling nul wees indien gelyke waarskynlikhede aan al die opsies toegeken word. Hierdie waardes kan ook sodanig gekies word dat 'n telling van een verkry word indien 'n toetsling 'n waarskynlikheid van een aan die korrekte antwoord toeken. Met ander woorde $s(i) = 0$ as $r(i) = \frac{1}{k}$ en $s(i) = 1$ as $r(i) = 1$. In die ekstreme geval waar 'n waarskynlikheid van nul aan die korrekte opsie toegeken word, sal $s(r(i)) = -\infty$. Hiedie 'oneindige' telling kan reggestel word deur 'n korreksieparameter (toleransie) wat veranderlik kan wees as 'n funksie van die getal opsies in 'n item.

Sommige toetslinge oorskat hul kennis of neig om te 'dobbel'. Dit lei dikwels tot hoë waarskynlikhede toegeken aan verkeerde opsies. Die teenoorgestelde is ook waar, naamlik dat sommige toetslinge te lae waarskynlikhede toeken aan korrekte opsies. Wanneer 'n waarskynlikheid r toegeken word aan 'n getal f – opsies vir verskillende items waarvan ft korrek is, word verwag dat $\frac{f}{f(t)} = r$. 'n Toetsling is realisties gekalibreer indien $r(i) = p(i)$ vir waarskynlikhede p. Dit kan bereik word deur 'n kleinste kwadraat of 'n X^2 – skatting. Die hipotese dat die toetsling realisties in die toekenning van waarskynlikhede is, kan getoets word – die toetsling is perfek gekalibreer indien vir elke waarde van p, die verhouding korrek van die opsies waaraan hierdie p toegeken is, gelyk is aan p.

Die bepuntingsreël met 'n korreksieparameter en 'n indeks vir realisme sal 'n telling op 'n kontinue skaal genereer. Net soos in die geval van vermoëskattings in Rasch of IRT, kan hierdie tellings na 'n sinvolle skaal getransformeer word wat meer 'tradisioneel' is.

Samevatting

In hierdie artikel is die toekenning van punte in toetse en eksamens vanuit verskillende perspektiewe bespreek. Voordat enige ontledings gedoen of punte toegeken word, moet 'n verantwoordelike besluit oor die hantering van ontbrekende response gemaak word. Daar is aangetoon dat verskillende besluite tot betekenisvolle verskille in punte aanleiding kan gee.

Verskillende metingsteorieë is oorsigtelik aan die orde gestel. Enkele kernbegrippe van die teorieë is kortliks toegelig. Die beantwoording van veelkeusige vrae is gebruik om verskille uit te lig. Dit is 'n feit dat toetslinge die korrekte antwoorde van sulke vrae kan raai. Indien blindelings geraai word, het enige toetsling 'n kans om die regte antwoord te raai. Daar is egter van die vertrekpunt uitgegaan dat toetslinge meestal nie blindelings raai nie, maar sommige opsies probeer uitskakel wat natuurlik die kanse om die regte antwoord te raai uit die oorblywende opsies verhoog. Die klassieke toetsteorie beskik oor 'n korreksie vir raai wat die punte van alle toetslinge, behalwe diegene wat al die items korrrek beantwoord het,

sal verlaag. Die Rasch-model gebruik passingsindekse om moontlike raai te identifiseer en die drieparameter- IRT-model sluit 'n itemparameter in wat die kanse aandui vir korrek raai. Die keusewaarskynlikheidsteorie hou stellig die beste moontlikheid in om die raaifaktor te betrek. In hierdie teorie word 'raai' as 'n toetslingparameter gehanteer.

Dié oorsig roer vir seker nie alle aspekte van die bepunting van toetse aan nie, maar lewer hopelik 'n bydrae tot die insig dat dit nie 'n eenvoudige proses is nie.

Erkenning

Mededingende belange

Die outeur verklaar hiermee dat hy geen finansiële of persoonlike verbintenis het met enige party wat hom nadelig kon beïnvloed het in die skryf van hierdie artikel nie.

Literatuurverwysings

Barnard, J.J., 2012, *A Primer on Measurement Theory*, Excel Psychological and Educational Consultancy, Melbourne.

Bond, T.G. & Fox, C.M., 2007, *Applying the Rasch Model*, 2 edn., Lawrence Erlbaum, London.

Crocker, L.M., en Algina, J., 1986, *Introduction to Classical and Modern Test Theory*, Holt, Rinehart and Winston Inc, New York.

De Finetti, B., 1970, 'Logical foundations and measurement of subjective probability', *Acta Psychologica* 34, 129–145. http://dx.doi.org/10.1016/0001-6918(70)90012-0

Hambleton, R.K., en Swaminathan, H., 1985, *Item Response Theory: Principles and Applications*, Kluwer-Nijhoff, Boston.

Stevens, S.S., 1946, 'On the theory of scales of measurement', *Science* 103, 677–680. http://dx.doi.org/10.1126/science.103.2684.677