Notitie Richtlijn Handelingsadvies bij Betrouwbaarheids- en Itemanalyse v. EC PW mei 2017

Deze notitie Richtlijn Handelingsadvies bij Betrouwbaarheids- en Itemanalyse beoogt docenten/examinatoren beter in staat te stellen om op basis van een betrouwbaarheids- en itemanalyse de kwaliteit van toetsing (bij voorbeeld bij mc-toetsen en open-vraag-toetsen) en bevorderen en verbeterslagen te maken voor het volgende studiejaar. In deze notitie wordt ingegaan op de indices die naar voren komen in de betrouwbaarheids- en itemanalyse en getracht te komen tot richtlijnen voor actie op basis van deze kwantitatieve analyse.

1. Toelichting bij de indices in de betrouwbaarheids- en itemanalyse

- Cronbach's alpha (of KR-20 bij meerkeuzetoetsen): betrouwbaarheid van de toets; de mate waarin de scores op de toets naar verwachting overeenkomen met de scores op dezelfde toets, wanneer deze nogmaals zou worden afgenomen (onder dezelfde condities).
- KR-20 (75): de KR-20 genormeerd naar 75 vragen, oftewel de verwachte betrouwbaarheid op een meerkeuzetoets wanneer deze uit 75 vragen zou bestaan; geeft de mate van samenhang van de toetsvragen aan (die bijdraagt aan de betrouwbaarheid van de toets), waarbij voor het effect van de toetslengte is gecorrigeerd[1].
- p-waarde: proportie studenten dat het desbetreffende item juist heeft beantwoord; de moeilijkheidsgraad van een item.
- gemiddelde p-waarde: het gemiddelde over p-waarden van alle items; de moeilijkheidsgraad van de toets.
- p': de p-waarde gecorrigeerd voor kans; geeft het gemiddelde kennisniveau van de studenten op het desbetreffende item weer. p' = p - (1-p)/(a-1), waarbij a is het aantal alternatieven van de meerkeuzevraag. Bij meerkeuzevragen hebben studenten altijd een kans het juiste alternatief te gokken. Bij een 4-keuze item is die kans 25%; een p-waarde van .25 betekent dus dat het gemiddelde kennisniveau 0 is; een p-waarde van .7 op een 4-keuze vraag komt overeen met een p' van .6.
- rir-waarde: item-restcorrelatie van een item; de mate waarin de score op een item overeenkomt met de scores van de andere items in de toets; onderscheidend vermogen van een item tussen studenten met kennis van de stof en studenten zonder kennis van de stof. Items met een lage rir hebben een gering discriminerend vermogen; items met een negatieve rir kunnen duiden op een sleutelfout of een misleidende vraagstelling.
- a-waarde: proportie studenten dat het desbetreffende alternatief heeft gekozen

2. Interpretatie van de betrouwbaarheids- en itemanalyse

Een verantwoorde betrouwbaarheids- en itemanalyse veronderstelt dat de groep studenten dat aan de toets heeft deelgenomen representatief is voor de doelpopulatie, d.w.z. dat er zich zowel goede studenten en minder goede studenten in de afgenomen groep bevinden, wat leidt tot een gebruikelijke spreiding van scores onder de studenten, zowel hoge als lage scores. Dit betekent dat met name de eerste gelegenheid geschikt is voor en betrouwbaarheidsanalyse; een herkansing is in de regel niet geschikt voor een dergelijke analyse. Ook het studentaantal is van invloed op de kwantitatieve waarden; in de praktijk hechten wij minder waarde aan een analyse van een toets aan minder dan 20 studenten.

[1] N.B. Dit is niet hetzelfde als de betrouwbaarheid; in de betrouwbaarheid van een toets is de lengte van de toets een invloedrijke factor.

Bij de interpretatie kijken we in de eerste plaats naar de betrouwbaarheid (Cronbach's alpha / KR-20) van de toets. Boven de .80 noemen we deze 'hoog', d.w.z. de scores van de studenten op de toets geven een goed beeld van de kennis van de student, de toets is geschikt voor summatieve toetsing; onder de .65 noemen we de betrouwbaarheid 'laag', d.w.z. er is een reëel risico dat studenten een score krijgen op de toets die te hoog of te laag is in vergelijking met het niveau van hun kennis van de stof; tussen de .65 en .80 noemende we de betrouwbaarheid 'redelijk'.

Naarmate de betrouwbaarheid hoger is, zijn rir-waarden meer indicatief, en wegen deze zwaarder mee in de interpretatie. Bij lagere betrouwbaarheid zijn rir-waarden minder veelzeggend. De p-waarden zijn in het algemeen minder vatbaar voor hoge of lage betrouwbaarheid.

Items met een hoge rir hebben een hoog onderscheidend vermogen tussen studenten met en zonder kennis van de stof, wat een kwaliteitskenmerk is van deze items. Items met een lage rir[2] hebben een gering tot geen discriminerend vermogen, wat erop kan duiden dat de meeste studenten (zowel studenten zonder als met kennis van de stof) de vraag hebben gegokt. Dit kan erop duiden dat de stof niet voldoende bij de studenten bekend was (gemaakt). Items met een negatieve R-waarde worden erdoor gekenmerkt dat studenten met kennis over de stof (die de 'rest' van de items juist beantwoorden) op dit item juist minder vaak het juiste antwoord geven dan studenten met minder kennis. Dit kan erop duiden dat een verkeerd alternatief als het juiste was aangegeven, of dat er een ander probleem is met de (formulering van de) vraag, waardoor studenten met kennis kiezen voor een onjuist alternatief.

Voor de p-waarde is minder gemakkelijk een kwaliteitsoordeel te geven; er bestaat niet zoiets als een 'optimale p-waarde', vanuit kwaliteitsoogmerk. Vragen met een zeer hoge p-waarde worden door veel studenten juist beantwoord, en kunnen dus gemakkelijk worden genoemd. Dit zijn in het algemeen geschikte items voor parate kennistoetsen (waarin wordt verwacht dat alle studenten van een bepaald niveau de veronderstelde kennis hebben). Items met een zeer hoge p-waarde hebben in het algemeen niet een sterk onderscheidend vermogen (rir). Bij items met een zeer lage p-waarde kan men zich afvragen of de studenten de desbetreffende stof wel in voldoende mate beheersen. Dit kan het gevolg zijn van voorbereiding van de student (bijv. tijdgebrek), uitleg over het onderwerp (onderwijs) of onwetendheid over het belang van het onderwerp (bijv. detailvraag).

De verschillende waarden uit de kwantitatieve analyse van een toets duiden we in samenhang. Lage rir-waarden bij zeer hoge p-waarden betekenen weinig; hoge rir-waarden bij een lage betrouwbaarheid zijn niet erg indicatief. Uit de waarden trachten we een beeld te krijgen over wat er met een item of met de toets als geheel aan de hand kan zijn: zijn er andere alternatieven ook (deels) juist? Is de formulering van de vraag éénduidig? Is de stof wel behandeld of voldoende duidelijk gemaakt? Konden de studenten zich goed voorbereiden op de toets? Heeft het onderwijs onder goede omstandigheden plaats kunnen vinden?

Bij beoordelen van een vraag dient voorts de inhoud van de vraag leidend te zijn, niet de kwantitatieve analyse; deze laatste kan slechts een signaal-functie vervullen, door bovenstaande vuistregels. Na signalering van een mogelijk problematisch item op grond van de kwantitatieve analyse, geeft de inhoudelijke analyse de doorslag over de kwaliteit van de vraag en eventuele verbetermaatregelen.

[2] In theorie wordt rir van groter dan .25 'goed' genoemd; in de praktijk komt dit niet veelvuldig voor.

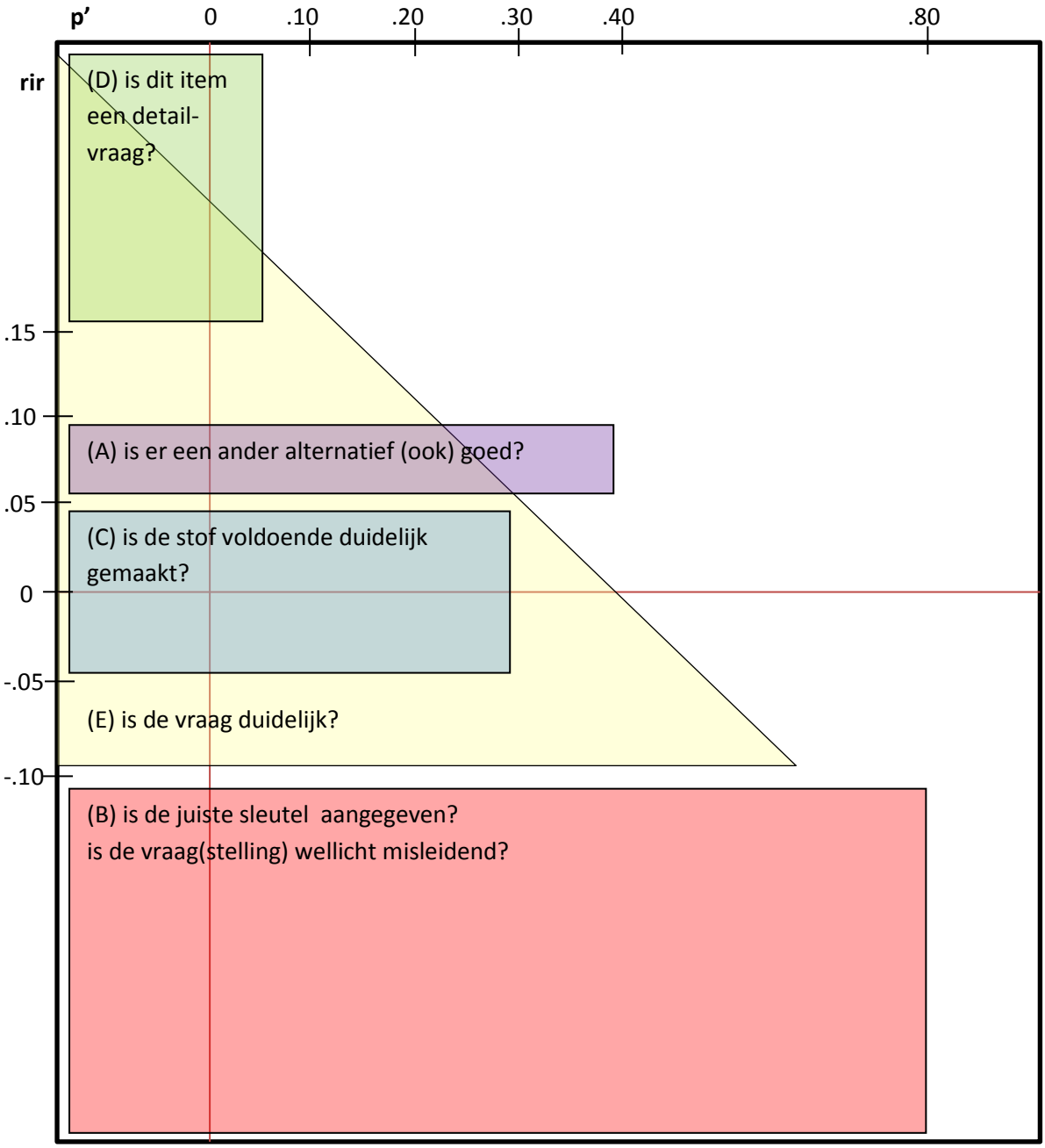
3. Schema voor handelingsadvies bij bepaalde gegeven indices

Vooropgesteld dat, zoals hierboven al aangegeven, de verschillende waarden uit de kwantitatieve analyse in samenhang geduid dienen te worden en bij beoordelen van de kwaliteit van vragen de inhoud leidend dient te zijn, wordt hieronder in Tabel 1. een schema gegeven, dat dient als handreiking voor docenten/examinatoren om signalen in de kwantitatieve indices te herkennen en een aanknopingspunt te beiden voor nadere inhoudelijke analyse van de toets en mogelijke acties ter verbetering van de kwaliteit ervan, hetzij ter reparatie van de huidige toets, hetzij ter verbetering van de toetsen bij volgende gelegenheden.

Tabel 1. Stappenplan met gecategoriseerd handelingsadvies bij kwantitatieve analyse

Toets als geheel				
stap	**In geval van**	**Actie**	**Toelichting**	
1	als herkansing of aantal studenten < 20	voer alleen kwalitatieve, inhoudelijke analyse uit op de toets	Alleen bij de 1e gelegenheid en bij een voldoende groot aantal studenten wordt een kwantitatieve betrouwbaarheidsanalyse voldoende indicatief geacht	
2	Cronbach's alpha of KR-20 > .65	betrek rir- en p-waarden in de analyse; anders: kijk alleen naar p-waarden	Bij een voldoende hoge betrouwbaarheid zijn rir-waarden voldoende indicatief; p-waarden zijn robuuster voor lage betrouwbaarheden.	
Per vraag				
stap	**In geval van**	**Actie**	**Toelichting**	*****
3	rir < -.10 en p' < .80	kijk of de juiste sleutel was aangegeven of de vraag(stelling) misleidend was	Bij negatieve rir-waarden (en de p niet zeer hoog) scoren de betere studenten slechter dan anderen op een bepaalde vraag	B
4	-.05 < rir < .05 en p' < .30 en a' < .20	ga na of de stof voldoende duidelijk was (gemaakt)	Als rir rond de 0 is en de p-en a-waarden zijn laag, lijken de studenten te hebben gegokt	C
5	rir < .10 en p' < .40 en a' > .30	kijk of er een ander alternatief (ook) goed is	Als rir laag i, de p niet hoog, en een a-waarde ten opzichte van de p hoog is wellicht een ander alternatief ook (bijna) goed	A
6	rir ≥ .15 en p' < .05	ga na of dit item een detailvraag betreft	als p rond de gokkans ligt, maar de betere studenten kiezen wel het juiste antwoord, kan het wellicht om een detail in de stof gaan	D
7	rir + p' < .40	ga na of de vraag(formulering) voldoende duidelijk was	Als p en rir beide laag zijn, is noch het onderscheidend vermogen van de vraag goed, noch konden voldoende studenten het juiste antwoord herkennen tussen de afleiders	E

* De letters A – E verwijzen naar de categorieën zoals die worden gerapporteerd in de meerkeuze-tentamenanalyse van de tentamenservice VU.



Figuur 1. Grafische weergave van gecategoriseerd handelingsadvies bij kwantitatieve analyse, twee dimensionaal weergegeven naar p' en rir