Universiteit Utrecht | Onderwijsadvies & Training

Lindy Wijsman

Toetsanalyse in Remindo: *Learning Analytics*

Toetsanalyse in Remindo

Remindo heeft in de afnameomgeving onder de naam *Learning Analytics* de functionaliteit die inzicht geeft in de kwaliteit van de toets als geheel en de kwaliteit van de afzonderlijke vragen. In deze handout geven we uitleg over de betekenis van de analyseresultaten en hoe deze te gebruiken zijn om inzicht te krijgen in de kwaliteit van de toets, de vraagkwaliteit waar nodig te verbeteren, en om zo nodig inhoudelijk gefundeerde aanpassingen door te voeren in de toets en/of de uitslag.

Het belang van toetsanalyse

De analyseresultaten geven de docent allereerst inhoudelijke feedback. Op welke onderdelen presteerden de studenten goed, en op welke minder? En is dat wel volledig toe te schrijven aan de beheersing van de stof of ook aan de wijze waarop de vraag was geformuleerd?
De analyse kan aanwijzingen bevatten om afwijkende antwoordpatronen te begrijpen en zonodig het antwoordmodel aan te passen. Deze informatie is in die gevallen ook van belang om te bepalen of een vraag geschikt is voor opname in een vragenbank of dat er nog aan geschaafd moet worden.
Omdat geen enkele toets feilloos meet, en er dus altijd sprake zal zijn van een aandeel van de studenten dat op basis van deze toets onterecht geslaagd dan wel gezakt is, is zorgvuldigheid nodig bij het doorvoeren van aanpassingen achteraf. De analyseresultaten kunnen daarvoor aanwijzingen geven.
Wat te doen met de analyse en of, en welke consequenties uit de conclusie over de toetskwaliteit naar voren komen, wordt belicht in het laatste deel van deze handout. Want, de analyse is vooral van belang als informatiebron wanneer het slaagpercentage en/of de betrouwbaarheid van de toets als geheel tegenvalt en aanpassing van de uitslag gewenst is. Er zijn dan verschillende mogelijkheden om wijzigingen aan te brengen. Zo kan bijvoorbeeld besloten worden om bij bepaalde vragen meerdere alternatieven goed te rekenen, een vraag te verwijderen uit de toets en/of de voldoende-onvoldoende grens (cesuur) van de toets aan te passen.
Deze beslissingen baseert de docent op inhoudelijke gronden en de informatie die uit de analyse naar voren komt. Omdat de betekenis van de resultaten afhangt van verschillende factoren is het goed om deze vooraf in beeld te hebben.

Vragen vóóraf

Gaat het om een eerste afname of een herkans? En is het aantal deelnemende studenten > 60?

De richtlijnen die verderop worden gegeven voor een goede en onderscheidende toets en voor goede toetsvragen zijn afhankelijk van de grootte en de samenstelling van de groep deelnemers aan de toets.
De taak van de toets is onder andere beslissingen te nemen over zakken of slagen, en om verschillen tussen studenten in beheersing van de leerstof tot uitdrukking te brengen. Hoe groter en heterogener de groep des te groter de kans op verschillen en hoe betrouwbaarder de analyseresultaten naar verwachting zullen zijn. Vuistregel voor het minimum aantal deelnemers aan een toets is 60, om waarde te kunnen hechten aan de analyses.
Wanneer er minder onderling verschil is tussen de studenten in stofbeheersing, dan zal het onderscheid dat in de resultaten van de toetsanalyse wordt uitgedrukt in de diverse parameters naar verwachting ook niet groot zijn, waardoor de betrouwbaarheid van de toets tegenvalt. Bij een herkansing heb je waarschijnlijk een kleinere, en ook homogenere studentenpopulatie dan bij de eerste afname. Hierdoor is er minder onderscheid tussen de studenten en zal de betrouwbaarheid naar verwachting lager uitvallen dan bij een eerste kans.

Is er bij studenten al iets bekend over de norm?

Zijn er al verwachtingen gewekt over de norm (cesuur)? Met het gebruiken van analyseresultaten om achteraf aanpassingen aan te brengen in de toets of cesuur is het van belang dat de aanpassing strookt met de informatie die vooraf aan studenten is beloofd, dan wel dat de aanpassing een gelijke of verbetering van de individuele studentprestatie tot gevolg heeft.

Is dit de enige toets waarop het eindcijfer wordt gebaseerd, of zijn er meer toetsonderdelen?

Bij de uitleg over de interpretatie van de analyseresultaten die verderop in deze handout te vinden is, is ervan uitgegaan dat de beslissing over zakken en slagen geheel of voornamelijk op de uitslag van deze toets is gebaseerd. Meestal zijn er echter meer prestaties geleverd door de studenten, op een gevarieerde mix aan toetsvormen. In die gevallen kunnen de gehanteerde normen naar rato worden bijgesteld.

Stappenplan voor het gebruik van de analyseresultaten *(zoals gegenereerd door Remindo)*

Stap 1: *De eerste, algemene indruk*
Is het slaagpercentage conform verwachtingen? En, is de betrouwbaarheid van de toets voldoende?
NB: Is het de enige toets waarop het cursusoordeel is gebaseerd? Zo ja, dan volstaat een Quick Scan (1A)! Zo niet, ga verder met Stap 2.

Stap 1A: *Quick-scan**: Kwaliteit van de vragen (aanpassing voor opname in de vragenbank)***
Gebruik de item-analyse om mogelijk dubieuze vragen te markeren (P' en Rir < 0.1). Kijk naar die afleiders die aantrekkelijk blijken voor veel studenten (a-waarde) en de goed presterende studenten (Rar-waarde). Zoek naar een verklaring voor het afwijkende scorepatroon door naar de inhoud van de vraag te kijken. Is de formulering van de vraag of zijn de alternatieven te verbeteren?

Stap 2: *Bij een tegenvallend slaagpercentage*
Vraag vooraf: Is het slaag% te verklaren (kleine groep, slechter presterend cohort)? Of is het tegenvallende resultaat naar waarschijnlijkheid toe te schrijven aan de toets zelf? Voor je meteen de diepte in gaat (inspectie vraagkwaliteit) is het verstandig om te kijken of een marginale aanpassing al zou kunnen leiden tot een grote stijging van het percentage geslaagden. Is dat het geval dan haalt dat veel druk van de ketel. Daarvoor zou je de volgende vragen kunnen stellen: Bij welke cesuur zou het slaagpercentage wel volgens verwachting (aanvaardbaar) zijn? Hoeveel scheelt dit met de gehanteerde cesuur, is met een kleine wijziging een groot effect te bewerkstelligen?

Stap 3: *Deep Scan**: Maak gebruik van de uitleg, de interpretatie van analyseresultaten***

- Bij een laag slaagpercentage vooral richten op (te) moeilijke vragen: P'-waarde < 0.15
- Bij een laag slaagpercentage en een lage betrouwbaarheid: richten op vragen met een (te) lage Rir-waarde < 0.10 èn P'-waarde die laag is (< 0.10).

Stap 4: *Zoek naar een afwijkend scorepatroon* (altijd op basis van inhoudelijke argumenten!):

- Bij negatieve Rir-waarde: zoek naar het alternatief met een positieve Rar-waarde;
- Bij een lage P'-waarde: zoek naar de aantrekkelijke afleider (veel gekozen en dus een relatief hoge a-waarde).

In beide gevallen is het de vraag of er niet iets te zeggen is voor het ook goed rekenen van dit alternatief, of zelfs alle alternatieven.

Op de volgende pagina's is uitleg gegeven van de parameters die op toetsniveau en op vraagniveau in Remindo zijn weergegeven. Eerst samengevat en vervolgens meer uitvoerig aan de hand van voorbeelden - screenshots- zoals gepresenteerd in Remindo.

Samenvatting *Learning Analytics* in Remindo

Volledig overzicht

Cirkeldiagram slaag%:	1 Het percentage gezakt/geslaagd is het aandeel van de studenten dat voldoet aan de gestelde norm voor deze toets, en dus ook een resultante van de gestelde zak-slaaggrens (cesuur) en de wijze van normering. De normering kan op drie punten worden (bij)gesteld: de score voor het hoogste cijfer (10); voor het laagste cijfer een 1 of 0; en de score waarmee nét een voldoende resultaat is behaald (cijfer 5,5; of 6).
Lijndiagram	2 De verdeling van de scores is in deze diagram af te lezen. Is het scorepatroon normaal verdeeld? Zijn er pieken rond de zak-, slaaggrens? Wat zijn de hoogste en laagste scores?
Cronbachs α:	3 *Betrouwbaarheid en consistentie van de toets.* Alpha is een maat voor de betrouwbaarheid. Dat wil zeggen de interne consistentie van de toets, en of de vragen hetzelfde (aan kennis begrip) meten. Is de betrouwbaarheid (α) laag dan zegt dat ook iets over de geringe precisie van de toetsuitslag en de mate van toeval. Het opnemen van meer vragen helpt de betrouwbaarheid omhoog te krijgen.
SEM:	3 *Standard Error of Measurement* (*standaardmeetfout).* De standaardmeetfout is een maat voor de gemiddelde (on-)nauwkeurigheid van de gemeten toetsscores. In de praktijk is een standaardmeetfout lager dan 10% van de maximale score gebruikelijk, en aanvaardbaar.

Analyseer vragen

Max. score:	4 Is het aantal maximaal te verkrijgen punten voor de vraag.
Resultaten:	*Respons.* Is het feitelijk aantal studenten dat de betreffende vraag heeft beantwoord. Is dat aantal laag dan zegt dat mogelijk iets over de moeilijkheidsgraad van de vraag (zie P'-waarde), of mogelijk tijdgebrek.
P':	5 *Moeilijkheidsgraad*. In Remindo is de P'-waarde het gemiddelde score%: het totaal aantal behaalde punten van alle studenten op de vraag, gedeeld door het maximum aantal mogelijke punten op de vraag van alle studenten. De P'-waarde is een indicatie van de moeilijkheidsgraad van de vraag. Hoe hoger de P'-waarde des te beter hebben de studenten gepresteerd op de vraag. Is de P'-waarde laag dan was de vraag moeilijk.

Rir (Rar):	❻ *Onderscheidend vermogen.* Rir-, en Rar-waarden zijn indicaties voor het onderscheidend vermogen van het goede antwoord (Rir) en de afleiders (Rar) van een vraag. Zij geven aan in welke mate het alternatief hoog-, en laagscorende studenten onderscheidt. De Rir is de item-rest correlatie: de correlatie tussen de itemscore en de totaalscore van alle resterende items van dezelfde toets. De Rar is de correlatie tussen de score op de betreffende afleider en de resterende items. Een goed discriminerend item geeft een hoge positieve Rir-waarde en negatieve Rar-waarden.
STD:	❼ *De standaarddeviatie* geeft de spreiding van de toegekende scores tov de gemiddelde score (P'-waarde).
Tijdsduur:	❽ De gemiddelde tijd nodig voor het beantwoorden van een vraag, zegt iets over de moeilijkheid/complexiteit van de vraag en de tijd nodig voor lezen, oplossen en geven van het antwoord. Bij energievreters is het dus relevant te kijken naar de P'-, en de Rir-waarde van de vraag, maar ook om na te gaan of de benodigde tijd conform verwachtingen is. Mogelijk dat de vraagstelling onnodig complex is.
Interactie-analyse	
A-waarde:	❾ *Aantrekkelijkheid van de afleider.* De A-waarde of wel de A(fleider)-waarde is de proportie kandidaten die, of het percentage kandidaten dat, bij een meerkeuzevraag de desbetreffende afleider(foutief alternatief) als antwoord koos.
(Rit) en Rat:	❿ *Onderscheidend vermogen.* Rit geeft de correlatie tussen de toetsscores van de studenten voor het goede alternatief (Rit) en de scores op de hele toets. Eenzelfde correlatieberekening wordt gemaakt voor de afleiders: Rat-waarde. Deze indices vallen hoger uit dan de Rir- en Rar-waarden omdat hier de itemscore in de totaalscore van de toets wordt meegeteld, waardoor de correlatie geflatteerd is.

Uitleg en interpretatie van analyseresultaten in Remindo: *Learning Analytics*

In de bespreking van de toetsanalyse (Learning analytics) en hoe deze te interpreteren gaan we de drie tabbladen langs: **Tabblad A:** *Volledig overzicht*; **Tabblad B:** *De details van het resultaat*, en **Tabblad C:** *Analyseer vragen*, waar je door kunt klikken op **D:** *Interactie analyse* (analyse details van een vraag).

Tabblad (A). Volledig overzicht

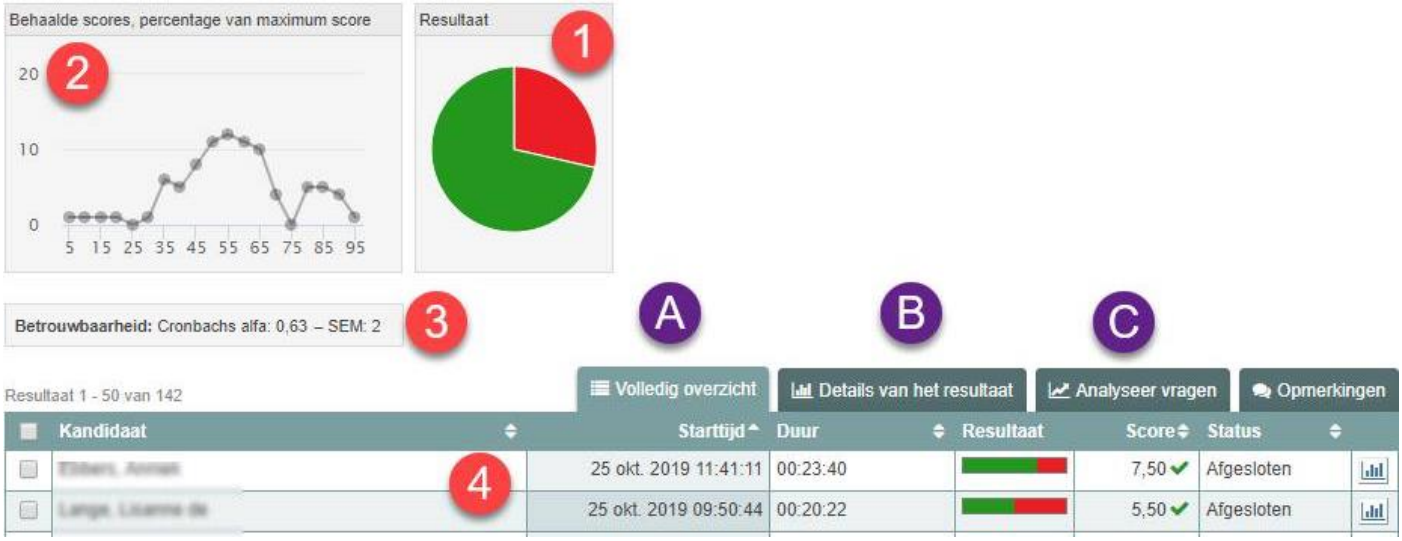
Het tabblad 'volledig overzicht' bestaat uit vier onderdelen, waarvan de eerste drie de toets als geheel betreffen: de slaagpercentages, de scoreverdeling en de betrouwbaarheid van de toets.

❶ Cirkeldiagram met percentages voldoendes en onvoldoendes

❷ Lijndiagram / frequentieverdeling van scorepercentages ten opzichte van het maximaal aantal te behalen punten.

❸ Betrouwbaarheid van de toets: de Cronbach's α, en de daarvan afgeleide SEM (standaardmeetfout), en

❹ een tabel met resultaten per kandidaat.



Figuur 1. Startscherm 'Volledig overzicht' in Remindo

❶ Cirkeldiagram met slaag/zak-percentage

Door met de cursor op het rode/groene vlak te gaan staan van de cirkeldiagram is precies af te lezen wat het zak/slaagpercentage is van deze toets. In het voorbeeld (figuur 1) is 17% gezakt en 83% geslaagd.

❷ De frequentieverdeling

De lijndiagram met de frequentieverdeling geeft informatie hoe de scores zijn verdeeld. Let op want in het diagram zijn de scores weergegeven als percentage van de maximum te behalen score (100%). Als je met de cursor op één van de lijnpunten staat krijg je het percentage te zien van de deelnemers met de betreffende score ten opzichte van de maximale score (in %). Bijvoorbeeld zoals afgebeeld in figuur 1 heeft 10% van de deelnemers een score behaald van 65% t.o.v. van de maximum haalbare score. NB: in dit voorbeeld is ook te zien dat de hoogste score de 95%-score is. De 100% score wordt door geen van de studenten gehaald.

Veel van de statistiek en de interpretatie van de parameters is gebaseerd op een 'normaal verdeeld' scorepatroon, en berust erop dat ook de te toetsen kennis van de studenten de normale verdeling volgt: hoogst en laagst gemeten scores komen weinig voor, en het merendeel van de scores ligt rond het gemiddelde.

Om onderscheid te kunnen maken tussen hoog-, en laagpresteerders, maar ook verfijnder tussen net voldoendes en net onvoldoendes, hoop je dat de frequentieverdeling een verscheidenheid aan scores laat zien. Die waarschijnlijkheid is groter naarmate de studentgroep groter is, en ook meer divers is in stofbeheersing.

Is de toets afgenomen onder een grote groep studenten dan mag je verwachten dat de toets onderlinge kennisverschillen blootlegt. Maar ook geldt dat hoe meer vragen de toets bevat des te nauwkeuriger en betrouwbaarder de prestaties en onderlinge verschillen kunnen worden vastgesteld.

Als het hoogst behaalde scorepercentage een stuk lager is dan de maximale 100%-score, terwijl het aantal deelnemende studenten groot is, dan is er mogelijk wat mis met de moeilijkheidsgraad van de toets als geheel en de doenlijkheid van specifieke vragen afzonderlijk.

❸ De betrouwbaarheid van de toets: Coëfficiënt alfa (α)

De coëfficiënt α is een maat voor de betrouwbaarheid van de toets. De coëfficiënt α kan maximaal 1 aannemen (volledig betrouwbaar) en minimaal de waarde 0 (volkomen onbetrouwbaar: scores zijn toevallig tot stand gekomen). Hoe betrouwbaarder de toets, des te nauwkeuriger de scores geïnterpreteerd kunnen worden. Voor een toets waarvan de consequenties van zakken of slagen groot zijn (*high stakes* test) wordt een α nagestreefd van 0.8 om een uitspraak te kunnen doen over het prestatieniveau van de student. Is het de enige toets waarop het oordeel is gebaseerd, dan is de 0.8 norm het streven. Is de beslissing (geslaagd-gezakt voor de cursus) mede gebaseerd op het resultaat van andere toetsen (open-vragen, tussentoetsen) dan is een lagere betrouwbaarheid dan de gewenste 0.8 verdedigbaar.

> **Een voorbeeld:** in een cursus zijn twee mc-toetsen afgenomen: een toets halverwege en een aan het eind van de cursus. Beide toetsen bestaan uit 40 MC-vragen en hebben elk een betrouwbaarheid van 0.60 (Cronbachs α). Voor de beslissing gezakt of geslaagd *voor de cursus* kunnen beide toetsresultaten 'samen worden genomen': Uit tabel 1 is af te lezen dat de verwachte betrouwbaarheid voor de zak-slaag beslissingen (d.w.z. voor beide toetsen samen) is gestegen naar α = 0.75.

Regel is dat hoe langer de toets is, hoe hoger de betrouwbaarheid en hoe beter de differentiatiegraad. De differentiatiegraad zegt iets over hoe 'verfijnd' de toets is. Maakt deze toets bijvoorbeeld ook nog voldoende onderscheid tussen een net voldoende score en een net onvoldoende score? Is de scorerange voldoende groot, en zijn zowel de minimale als de maximale te behalen scores in het scorepatroon terug te vinden ? De relatie tussen toetslengte en een schatting van de betrouwbaarheid is in Tabel 1 weergegeven.

Voorbeeld: Stel een toets bestaat uit 30 vragen (K) en de betrouwbaarheid (α) is 0.40. In dat geval zou de toets verlengd moeten worden tot 90 vragen (3K) om een betrouwbaarheid van 0.60 (α''')te bereiken.

Tabel 1: Betrouwbaarheidsschatting (α) bij toetsverlenging/inkorting, bij K-vragen

K	1.5 K	2 K	3 K
α	α'	α''	α'''
0.20	0.27	0.33	0.40
0.40	0.50	0.57	0.60
0.60	0.69	0.75	0.77
0.80	0.86	0.89	0.91

De analyse van een herkansing is een geval apart. Daarbij ligt het in de rede dat de betrouwbaarheid lager uitvalt omdat de onderlinge verschillen in stofbeheersing in de populatie (spreiding) geringer zijn dan bij een eerste afname.
De gemeten betrouwbaarheid zegt ook iets over het aantal inconsistente beslissingen. Een toets is nooit 100% betrouwbaar; er zijn altijd onterecht gezakten of onterecht geslaagden. Hoe minder betrouwbaar de toets des te groter de kans op foutieve beslissingen. In Tabel 2 is weergegeven wat het percentage niet-consistente beslissingen is als functie van het percentage gezakten en de betrouwbaarheid (α). In het afgebeelde voorbeeld was ca. 15% van de studenten gezakt, met een α van bijna 0.70. In dat geval is er dus in ca. 14% van de gevallen een inconsistente beslissing genomen: dat wil zeggen 7% is onterecht geslaagd en 7% is onterecht gezakt. De vraag is wat nog aanvaardbaar is. De consequentie van de beslissing dat 7% van de studenten wel een voldoende krijgt, maar mogelijk niet aan de kwalificaties voldoet, hangt af van veel en andere factoren die daarbij een rol spelen. Zoals, de plaats en het belang van de cursus in de opleiding, of soortgelijke doelen terugkomen in het vervolg van de opleiding.

Tabel 2: Percentages niet-consistente beslissingen als functie van afwijzingspercentage en toetsbetrouwbaarheid (α). Bron: Dousma, Horsten, Brants, Tentamineren (1997).

	Betrouwbaarheid (α)						
Afwijzings% (gezakt)	**0,50**	**0,60**	**0,70**	**0,80**	**0,90**	**0,95**	**1,00**
5	8	7	6	5	4	3	0
10	14	12	11	9	6	4	0
15	18	17	14	12	8	6	0
20	23	20	17	14	10	7	0
25	26	23	20	16	11	8	0
30	29	25	22	18	12	9	0
35	31	27	23	19	13	9	0
40	32	29	24	20	14	10	0
45	33	**29**	**25**	20	14	10	0
50	33	30	25	20	14	10	0

❸ De standaardmeetfout (SEM)

De standaardmeetfout geeft de gemiddelde (on-)nauwkeurigheid van de gemeten toetsscores weer. Het is een maat om te kunnen schatten in hoeverre een student dezelfde score zou behalen bij het maken van een denkbeeldige, andere vergelijkbare toets over dezelfde leerstof.
In de standaardmeetfout (Sm) is de betrouwbaarheid van de toets (α) verdisconteerd volgens de formule: $Sm=Sa\sqrt{1-\alpha}$ (waarbij Sa de gemeten standaardafwijking is van de toetsscores).
De formule laat zien dat hoe betrouwbaarder de toets is des te kleiner de meetfout zal zijn en hoe waarschijnlijker het is dat de gemeten scores van de studenten overeenkomen met de 'ware scores'.
Is de toets onbetrouwbaar (grote mate van toevalligheid) dan kan aan de gemeten scores geen betekenis worden toegekend.
Is de standaardmeetfout 2, zoals in figuur 1, dan betekent dit voor een student met een gemeten score van 13, dat hij/zij met 67% waarschijnlijkheid kennis heeft overeenkomend met een score tussen de 11 en 15 (13 ± 2). Met een waarschijnlijkheid van 95% zal de ware score liggen tussen de gemetenscore ± 2xSm = 13 ± 4. Aanname voor de interpretatie van de standaardmeetfout is dat de toetsscores normaal verdeeld zijn.

❹ Tabel met de resultaten per kandidaat

In de tabel is per kandidaat het volgende terug te vinden. Hoelang hij/zij met de toets bezig was: *de tijdsduur*, *het resultaat* in een kleurenbalk weergegeven (van groen =aandeel goed, tot rood =aandeel fout) en *de score* weergegeven in een percentage, het aantal behaalde punten of het cijfer.
Bij deze tabel staan we verder niet stil omdat de parameters per student zijn uitgezet en daar weinig belang aan is toe te kennen. De tijdsduur heeft meer betekenis uitgezet als gemiddelde tijd die de studenten met de betreffende vraag bezig waren. Dit komt terug in de bespreking van Tabblad C: *Analyseer vragen.*

Tabblad (B). Details van het resultaat

Het tabblad *Details van het resultaat* is één tabel waarin per vraag de volgende kenmerken zijn terug te vinden: Het aantal studenten dat de vraag heeft beantwoord en verder:

5 De P'-waarde van de vraag;
6 De Rit/Rir-waarde van de vraag;
7 STD: de standaardafwijking;

Uitleg over de interpretatie van deze parameters is te vinden onder het volgende tabblad C: Analyseer vragen.

Tabblad (C). Analyseer vragen

5 P'-Waarde
6 Rir – Waarde
7 STD (standaardeviatie)
8 Gemiddelde tijdsduur

#	Vraagcode/kenmerk	Max. score	Resultaten	P'	R_{ir}	STD	Gemiddelde tijdsduur
1	[redacted]	1 pt.	621	0,67	-0,06 ⚠	0,47	00:01:03
2	[redacted]	1 pt.	621	0,80	0,05	0,40	00:01:19
3	[redacted]	1 pt.	621	0,24 ⚠	0,05	0,43	00:00:51
4	[redacted]	1 pt.	621	0,32 ⚠	0,07	0,47	00:01:54
5	[redacted]	1 pt.	621	0,71	0,08	0,45	00:01:14
6	[redacted]	1 pt.	621	0,75	0,09	0,43	00:01:16

Figuur 2. Schermafbeelding 'Analyseer vragen' in Remindo

5 De P'-waarde: De moeilijkheidsgraad (*Proportie goed*)

P': Max 1; *Min*: 0
Streefwaarde: is afhankelijk van een eventuele raadkans
⚠ Remindo plaatst een vlag bij de P' van een vraag als de waarde hoger is dan 0,9 *('mogelijk een te gemakkelijke vraag?')* of als de P'-waarde lager is dan 0,4 *('mogelijk een te moeilijke vraag?')*.

De P'-waarde geeft de moeilijkheidsgraad aan van een vraag. De P'-waarde in Remindo is een relatieve maat. Bij een 4-keuzevraag betekent een P'-waarde van 0.5 dat de vraag pittig was, maar naar verwachting scoort. Vraag 3 in figuur 2 is eveneens een 4-keuzevraag en heeft een P'-waarde van 0.24 wat overeenkomt met de raadkans (de kans dat iemand een vraag goed beantwoord door blind te raden).
Is de P' lager dan de raadkans dan kiezen de studenten in dat geval bewust voor een afleider, en niet door blind raden. Mogelijk dat er iets met de vraag aan de hand is. De bijbehorende rir-waarde kan dan aanwijzingen bevatten of het met name de minder goed of beter presterende studenten waren die de vraag goed hadden.
Bij de één-uit meervraag is het dus nodig om ook naar het aantal keuzemogelijkheden te kijken bij de interpretatie van de P'-waarde. Bij alle andere vraagvormen, waar geen sprake is van een raadscore is de P'-waarde een goede (absolute) indicatie voor de moeilijkheidsgraad.
De toets wordt meestal zo samengesteld dat er een goede mix is qua moeilijkheidsgraad van de vragen. De analyse zou dit beeld moeten bevestigen: p'-waarden geven een gevarieerd beeld. Er is een klein aantal heel moeilijke vragen (lage p'-waarden) om de 'negens' van de 'tienen' te kunnen scheiden, en ook een klein aantal zeer gemakkelijke vragen (hoge p'-waarden) om de 'vieren' van de 'vijven' te kunnen onderscheiden. Maar het merendeel van de vragen bestaat uit p'-waarden die in het midden liggen (tussen 0.3 en 0.7) en bijdragen in het onderscheid tussen zakken en/of slagen op de toets.

6 Het onderscheidend vermogen: de Rir-, en/of Rit-waarde

Max: 1; *Min*: -1
Streefwaarde: positief, hoger dan 0.10
⚠ Remindo plaatst een vlag bij de Rir van een vraag als de waarde negatief is.

De p'-waarde geeft het aandeel weer van de studenten die de vraag goed hadden. De Rit (Item-totaalscore correlatie), en Rir (Item-restscore correlatie) geven aan in hoeverre de vraag de goede van de slechte studenten heeft gescheiden. Het verschil tussen Rit en de Rir-waarde is dat bij de Rit de correlatie berekend wordt tussen score op de vraag en de ***T***otaalscores van de studenten die de vraag goed hadden. De Rir is een zuiverder maat omdat de totaalscores minus de score op de vraag zelf (de ***R***est) in de correlatieberekening wordt gebruikt.
Is de Rir hoog (0.3-0.5) dan heeft de vraag zijn werk gedaan: de goede studenten hebben de vraag goed, en de slechte studenten kiezen voor een afleider. Is de Rir lager, dan is de vraag niet zeer onderscheidend geweest. Wordt de Rir negatief dan kan er iets aan de hand zijn. Juist de goede studenten kiezen voor een afleider, terwijl de slechter presterende studenten op de toets voor het juiste alternatief hebben gekozen. Als de Rir negatief is betekent dit per definitie dat één van de afleiders positief correleert met beter presterende studenten (positieve Rar-waarde). Gecontroleerd moet worden of er niet iets te zeggen valt voor die afleider, immers de beter presterende studenten kiezen daar tenslotte voor.

Interpretatie van combinaties van P'-, en Rir-waarden

Om een uitspraak te doen over de kwaliteit van een vraag is het dus van belang niet alleen naar de moeilijkheid of het onderscheidend vermogen te kijken, maar naar de samenhang tussen gemeten P'-, en de Rir-waarde. Zo is een onderscheidend vermogen (rir) van bijna 0.0 niet erg verontrustend als vrijwel alle studenten de vraag goed hadden (hoge p'-waarde). En, hebben alle studenten de vraag goed (p'=1.0), dan is per definitie het onderscheidend vermogen (rir) 0.0.
Maar is van een vraag de rir-waarde 0.0 en de p'-waarde is laag (< 0.3) dan is het een geheel ander verhaal. In tabel 3 zijn interpretaties gegeven van verschillende combinaties rir-, en p'-waarden, deze zijn iets verfijnder dan Remindo hanteert.

Tabel 3: Interpretaties van mogelijke combinaties van P'-, en Rir-waarden

Moeilijk-heidsgraad (P-waarde)	**Onderscheidend vermogen (Rir-waarde)**	
	Rir<0.1: Vraag maakt geen onderscheid, en is bij negatieve waarden zelfs tegenovergesteld aan de verwachting: goed presterende studenten op de toets scoren slecht op deze vraag.	**Rir> 0.1**:Redelijk tot goed (rir>0.3) onderscheidende vraag
Moeilijke vraag **P' < 0.3**	De vraag is slecht gemaakt, ook door de beter presterende studenten op de toets. *Vraag: Is de antwoordsleutel wel correct? Meer antwoorden goed rekenen? Tip: vergelijk de alternatieven met hoge A-, en positieve Rar-waarden.*	De vraag is over het algemeen slecht gemaakt, maar maakt nog wel onderscheid tussen de beter, en slecht presterende studenten op de toets. *Is de vraag te moeilijk, te complex? Instinker? Let op dat er niet te veel van dit soort moeilijke vragen in de toets voorkomen.*
Doenlijke vraag **0.3<P'<0.8**	De vraag is weliswaar redelijk goed gemaakt, maar onderscheid onvoldoende tussen slecht, en goed presterende studenten. *Misschien zijn er andere antwoorden ook -deels- goed te rekenen? Tip: vergelijk de alternatieven met hoge A-, en positieve Rar-waarden.*	De vraag is redelijk goed gemaakt en het onderscheidend vermogen is in orde. *Deze vraag behoeft verder geen actie.*
Gemakkelijk vraag **P'> 0.8**	De vraag was voor het overgrote deel van de studenten makkelijk te beantwoorden, ongeacht hun prestaties op de toets. *Is de vraag een onbedoelde weggever en op te lossen met boerenverstand?*	De vraag is goed gemaakt en heeft toch nog onderscheid gemaakt tussen goed-, en minder goed presterende studenten. *Behoeft verder geen actie.*

7 STD: de standaarddeviatie

De standaarddeviatie (STD) geeft aan hoeveel de scores op een vraag uiteenlopen ten opzichte van de gemiddelde score. De standaarddeviatie is niet zo interessant voor vragen die of goed (max. aantal punten) of fout (0 punten) worden toegekend, maar wel als het gaat om vragen waar geschaald meer of minder punten zijn te verdienen. Dat geldt bijvoorbeeld voor open vragen waar meer antwoordelementen zijn

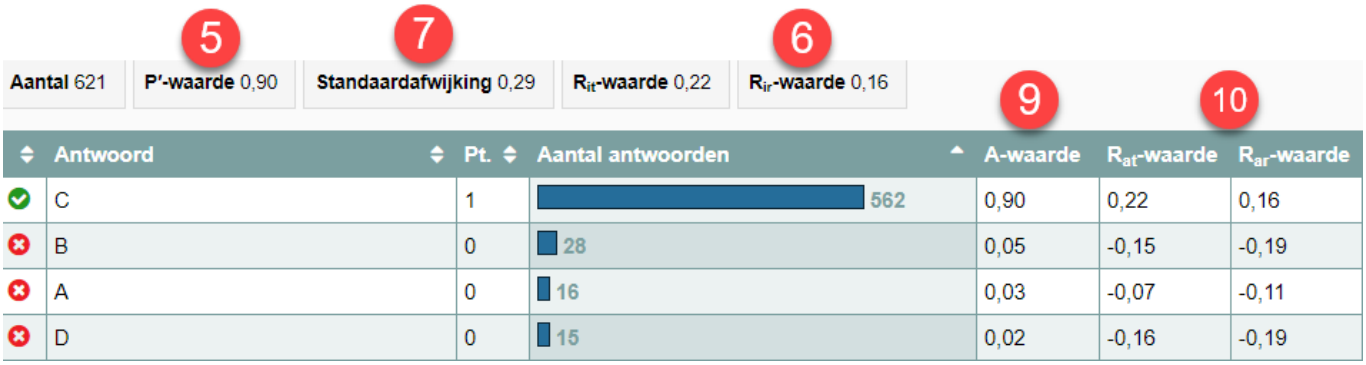
onderscheiden en voor gesloten vragen met een ‘gedeelde scoring’ waar toegekende punten en eventueel aftrekpunten tot een gedifferentieerd scorepatroon leiden.
Is bij dit type vragen de STD 0 dan is de conclusie dat de vraag niet heeft gedifferentieerd. Hoe hoger de STD des te groter de spreiding in scores. Wat je hoopt is dat de vraag, naast een redelijke spreiding in scores, ook voldoende differentieert tussen goed en slecht-presterende studenten op de toets. Daarvoor kan beter gekeken worden naar de Rir-waarde.

8 De tijdsduur

De tijd dat studenten gemiddeld met een vraag bezig waren is natuurlijk ook interessant, en de bedenk- of schrijftijd die de vraag kostte zegt mogelijk iets over de moeilijkheidsgraad/complexiteit van de vraag of de cognitieve vermogens waar de vraag een beroep op doet. Het geeft je als docent dus inhoudelijke feedback over wat studenten lastige vraagstukken vinden, en waar ze hun hand niet voor omdraaien.
Zijn studenten veel langer bezig met een vraag dan gedacht dan zou er mogelijk iets mis kunnen zijn met de vraag of vraaginstructie. De P’-, en Rir-waarde voor deze vraag geven dan misschien uitsluitsel.
Overall zouden studenten voldoende tijd moeten krijgen om de toets te kunnen maken. Zijn er onbedoeld ‘tijd-slurpers’ dan kan dat aanleiding zijn soortgelijke vragen voor een volgende keer te vermijden.

Tabblad (D). Interactie-analyse: Analysedetails voor een specifieke vraag

Bij doorklikken op de afbeelding rechts van het tabblad ‘Analyseer vragen’ kom je in het scherm van de interactie-analyse. Hierin staan parameters die al zijn besproken. Zoals de P’-, en Rir-waarde, en de Standaarddeviatie.
Wat nieuw is in dit scherm zijn de volgende parameters:
9 A-waarde
10 Rat-, en Rar-waarde
De betekenis ervan leggen we hieronder uit aan de hand van het schermvoorbeeld. Wat lastig is in Remindo is dat A en P-waarden door elkaar heen worden gebruikt, en dat zelfde geldt voor de Rir-, en Rar-waarden (Rit-, en Rat).
In figuur 3 staan op de eerste regel onder 5, 6 en 7, de parameters voor de goed gegeven antwoorden op de vraag. Dat zijn de P’-, Rir/Rit-, waarde en de standaardafwijking. In de tabel daaronder is onder 9 en 10 alleen de A-, en de Rar-, en Rat-waarde terug te vinden. Hiervoor is gekozen vanwege vormtechnische redenen. In de figuur zijn de berekende Rar/Rat van het goede antwoord C, dus feitelijk de P’,- Rir-, en Rit-waarde van deze vraag.



Antwoord	Pt.	Aantal antwoorden	A-waarde	R_{at}-waarde	R_{ar}-waarde
C	1	562	0,90	0,22	0,16
B	0	28	0,05	-0,15	-0,19
A	0	16	0,03	-0,07	-0,11
D	0	15	0,02	-0,16	-0,19

Figuur 3: Schermafbeelding ‘Interactie-analyse’ in Remindo

9 De A-waarde: aantrekkelijkheid van de afleiders (alleen relevant voor gesloten vragen):

Waarde: Max 1; *Min* 0
⚠ Streefwaarde: de som van de a-waarden (proportie fout), is lager dan de p’-waarde (proportie goed) van de vraag.
De A-waarde geeft de proportie (van 0 tot max. 1.0) weer van de studenten die voor de betreffende afleider heeft gekozen. De A-waarde is wat de p’-waarde is voor het goede antwoord, maar dan voor de afleiders. Is de A-waarde vergelijkbaar of hoger dan de p’-waarde, dan is deze afleider zeer aantrekkelijk geweest (misschien een instinker?). Is de A-waarde laag, dan trekt de afleider weinig studenten en is de afleider niet effectief gebleken. In zo’n geval (bijvoorbeeld geen of slechts één enkele student heeft de afleider gekozen) is er iets voor te zeggen om de afleider voor hergebruik te schrappen (4-keuzevraag wordt een 3-keuzevraag), óf om een betere afleider te formuleren.

In het voorbeeld van figuur 3 is de A-waarde van het goede antwoord (lees de P'-waarde) 0.90, dat wil zeggen dat 90% van de studenten deze vraag goed had. De 10% fout gegeven antwoorden zijn redelijk gelijk verdeeld over de drie afleiders. Conclusie lijkt aannemelijk dat als de studenten het antwoord niet wisten, ze willekeurig een alternatief hebben aangestreept

⑩De Rar-waarde: het discriminerend vermogen van de afleiders

Waarderange tussen de -1 en 1.
Streefwaarde: Som van alle Rar-waarden is lager dan de Rir.
De Rar-waarde heeft een vergelijkbare betekenis als de Rir-waarde, maar nu voor de afleider/alternatief.
. De Rar-waarde geeft extra informatie, namelijk welk deel van de studentgroep voor het juiste alternatief (Rir) of de afleiders (Rar) heeft gekozen. Waren het de beter presterende studenten op de toets dan zal de Rir positief zijn. Maar als met name de goed presterende studenten voor een afleider kiezen, en de slecht-presterenden voor het juiste kiezen, dan is er met deze vraag mogelijk wat aan de hand.

Wat te doen met de gegevens uit de analyse?

Als uit de toetsanalyse achteraf blijkt dat een vraag niet voldoet, dan moeten inhoudelijke motieven de doorslag geven om al dan niet wijzigingen in de toets aan te brengen. Het gaat er tenslotte om dat de docent het goed of fout aan een student kan uitleggen en aannemelijk kan maken, en dat gebeurt op inhoudelijke gronden.
Wil je de toets en/of de uitslag aanpassen dan kun je verschillende dingen doen op vraagniveau, maar ook op toetsniveau (zak-/slaaggrens, en de score-cijfer omzetting).

1) Je verandert niks
De vraagstelling is correct en het correcte alternatief is eenduidig juist. De vraag blijft dus in de toets in zijn oorspronkelijke vorm en er wordt niets aangepast. Omdat de vraag om onduidelijke redenen zijn werk niet goed heeft gedaan kan de docent besluiten de vraag uit het vragenbestand te verwijderen of de vraag in een verbeterde versie op te nemen.

2) Je verwijdert de vraag uit de toets
Deze actie is alleen aan te raden als de vraagstelling zo dubieus is dat een keuze uit de alternatieven niet goed mogelijk was. Belangrijk probleem bij het laten vervallen van vragen is, ook al zijn de argumenten steekhoudend, dat studenten die de vraag wel goed hadden zich gedupeerd voelen.

3) Je rekent meer antwoorden van de vraag goed
Op basis van de analyse wordt duidelijk voor welk alternatief de meeste studenten (P'-, en A-waarden) en de best presterende studenten hebben gekozen (Rir-, en Rar-waarden). Als de docent van mening is dat voor een afleider (oorspronkelijk fout antwoord) ook iets te zeggen is, dan ligt het in de rede dit alternatief ook goed te rekenen. Bij een ondeugdelijke vraagstelling kan de docent er zelfs voor kiezen alle antwoorden goed te rekenen. In vergelijking met methode 2, waar de studenten die de vraag wel goed hadden zich benadeeld zullen voelen, gaan de gemiddelde scores bij methode 3 omhoog en krijgt de student het voordeel van de twijfel.
NB: Bij het goed rekenen van meer alternatieven gaat de raadkans voor de vraag omhoog en daarmee dus ook de raadscore voor de toets, en die raadscore is mogelijk van invloed op het cijfer en toetsresultaat.
Bij een 4-keuzevraag neemt de raadkans met 0.25 toe per extra goed gerekend alternatief. Voor de 3-keuzevraag is dat 0.33, etc.. Omdat de raadscore vaak gebruikt wordt als referentiepunt voor het minimum cijfer (1.0 of 0.0), heeft de aanpassing ook gevolgen voor de score-cijfer transformatie, tenzij de docent besluit wel meer antwoorden goed rekenen, maar niet de cijferbepaling aanpast.

4) Je verlegt de zak-, slaaggrens (cesuur)
Er zijn twee redenen om de norm (cesuur) voor een voldoende aan te passen (te verlagen).

- De toets was gewoonweg te moeilijk. Is de toets slechter gemaakt dan voorgaande jaren, terwijl er geen reden is om aan te nemen dat de studenten zich minder goed hebben voorbereid? Dan is het waarschijnlijk dat de toets te moeilijk was.
- De toets was onvoldoende betrouwbaar. Is de betrouwbaarheid laag dan is het aandeel onterecht gezakten mogelijk onaanvaardbaar hoog. Door de norm te verlagen kan het aandeel onterecht gezakten tot

aanvaardbare proporties worden teruggebracht. Je neemt daarmee wel voor lief dat het aantal onterecht geslaagden groter is!

5) Een combinatie van methode 2, 3 en 4 *Aanpassingen van de cesuur*
Als de samenstelling van de toets achteraf is veranderd (meerdere alternatieven goed, vervallen vragen) dan is het goed om te bekijken of de zak-slaaggrens (cesuur) niet ook moet worden aangepast. Een tentamen waarvan een aantal vragen zijn komen te vervallen is korter geworden en behoeft een aanpassing in de cesuur en score-cijfertransformatie. Maar ook als er meerdere antwoorden goed worden gerekend kan dit gevolgen hebben voor de cesuur, doordat de raadkans-score is toegenomen en deze score vaak gebruikt wordt als referentie voor het minimale cijfer (1, of 0). De cesuur kan niet verhoogd worden als deze aan studenten al bekend was gemaakt. Verlagen mag daarentegen altijd (dit zal immers niet tot protesten leiden).

Verdieping / Literatuur

Dousma, T., Hortsen, A., & Brants, J. (1997). Tentamineren.Groningen: Wolters-Noordhoff.
De Groot, A.D.,&Van Naerssen,R.F. (1969, 1973). Studietoetsen. Den Haag: Mouton.
De Gruijter, D.N.M. (2008). Toetsing en toetsanalyse. Leiden: ICLON
Milius, J.J. (2007). Schriftelijk tentamineren: een draaiboek voor docenten in het hoger onderwijs. Utrecht: COLUU
Milius, J.J. (2016). Handleiding Remindo toets versie 1.0. Utrecht: Educate-it.

Workshops Toetsanalyse in Remindo van Onderwijsadvies & Training (O&T)

Met enige regelmaat organiseert O&T de workshop Toetsanalyse in Remindo (1 dagdeel).
Meer informatie is te vinden op de website van O&T.