Op weg naar een vernieuwde rapportage van examenprestaties aan scholen

De cognitieve complexiteit van examenvragen onderzocht met Bloom en RTTI

■ **Hans Kuhlemeier, Lody Smeets en Arjen Galema**
De heer dr. H. Kuhlemeier is als onderzoeker werkzaam bij Cito; de heer drs. L. Smeets is clustermanager exacte vakken bij Cito. De heer drs. A. Galema is toetsdeskundige biologie bij Cito. E-mail: hans.kuhlemeier@cito.nl.

Als service aan de scholen die centraal schriftelijke examens afnemen, verzorgt Cito een terugrapportage. Die bevat onder meer informatie over hoe de kandidaten presteerden op de inhoudelijke domeinen van het examenprogramma en op de vragen die een beroep doen op reproductie of toepassing van het geleerde. Het is wenselijk dat de feedback aan scholen gebaseerd is op een betrouwbare en valide indeling van de geëxamineerde vaardigheden. Dit artikel doet verslag van de betrouwbaarheid en validiteit van de taxonomieën van Bloom en RTTI voor het vaststellen van de cognitieve complexiteit van examenvragen.

De aanleiding tot het onderzoek

De belangrijkste aanleiding tot het onderzoek was de vraag van sommige docenten of Cito de terugrapportage zou willen aanvullen met RTTI (Drost & Verra, 2016). Net als Bloom (Anderson & Krathwohl, 2001) is RTTI een classificatiesysteem voor het coderen van het cognitieve niveau van leerdoelen, leerstof en toets- en examenvragen. Docenten gebruiken taxonomieën onder meer om meer greep te krijgen op hun leerdoelen, leerstof en toetsen zodat zij het leren beter kunnen sturen en verantwoorden. Doel van het onderzoek is na te gaan in hoeverre de cognitieve complexiteit van examenvragen betrouwbaar en valide met Bloom en RTTI geclassificeerd kan worden.

De taxonomieën van Bloom en RTTI

Bloom kent zes hoofdcategorieën ofwel beheersingsniveaus: Memoriseren, Begrijpen, Toepassen, Analyseren, Evalueren en Creëren. RTTI kent vier categorieën: Reproductie (R), Toepassing in bekende situaties (T1), Toepassing in nieuwe situaties (T2) en Inzicht (I). R-vragen beantwoordt de leerling op basis van reproductie die in de kennisbasis aanwezig is, T1-vragen vereisen het toepassen van de leerstof volgens een getrainde methode in vergelijkbare situaties als de geoefende situatie, T2-vragen doen een beroep op het toepassen van de leerstof in nieuwe situaties (die dus niet in de les behandeld zijn) en bij I-vragen moet de leerling zelf de context en methode construeren om tot een antwoord te komen. Zie voor meer informatie het RTTI-handboek (Drost & Verra, 2016) en www.docentplus.nl. Een belangrijk taxonomisch verschil is dat de RTTI-categorie Inzicht bij Bloom is uitgesplitst in vaardigheden van Analyseren, Evalueren en Creëren. Naast verschillen zijn er ook overeenkomsten. Zo vertonen T1 en T2 uit RTTI grote

gelijkenis met de beide vaardigheden uit de hoofdcategorie Toepassen uit Bloom: uitvoeren van een procedure bij een routinematige taak / in een bekende context (categorie 3.1) versus uitvoeren in een nieuwe taaksituatie / in een onbekende context (categorie 3.2).

Onderzoeksvragen

In dit artikel over de cognitieve complexiteit van examenvragen onderscheiden we de volgende vier onderzoeksvragen:

- In hoeverre kunnen getrainde codeurs de cognitieve complexiteit van examenvragen betrouwbaar met Bloom en RTTI vaststellen?
- Hoe vaak komen de onderscheiden cognitieve niveaus in de examens voor?
- In hoeverre hangt de cognitieve complexiteit van een examenvraag samen met de moeilijkheidsgraad (p-waarde)?
- In hoeverre zijn de opeenvolgende cognitieve beheersingsniveaus hiërarchische geordend?

Toelichting op de vraagstelling

De onderzoeksvragen zijn van praktisch belang. Als de codering van vakinhoudelijke deskundigen bijvoorbeeld sterk van elkaar zouden verschillen, zijn de gegevens niet bruikbaar voor het doel van terugrapportage.

De derde en vierde onderzoeksvraag behoeven een wat langere toelichting. Een taxonomie kan gevalideerd worden door aan te tonen dat de ermee verkregen resultaten in overeenstemming zijn met gangbare theoretische noties in het desbetreffende onderzoeksgebied (Bloom e.a., 1956; De Block & Velema, 1971). In Bloom zijn de zes beheersingsniveaus hiërarchisch geordend. De kenmerken van opeenvolgende niveaus beschrijven een progressie van eenvoudig naar complex (Bloom e.a., 1956; De Block & Velema, 1971; Krathwohl, 2002). De lagere niveaus zijn daarbij voorwaardelijk voor de hogere niveaus in de zin dat leerlingen de hogere cognitieve niveaus doorgaans niet zonder de lagere zullen bereiken. Om een inzichtvraag te kunnen beantwoorden, zal de kandidaat bijvoorbeeld de belangrijkste begrippen uit de vraagstelling moeten kennen. Onder verder gelijkblijvende omstandigheden zouden examenvragen die hogere-ordevaardigheden vereisen daarom gemiddeld moeilijker moeten zijn dan vragen die een beroep doen op lagere-ordevaardigheden. In lijn hiermee is de bevinding dat leerlingen doorgaans lagere scores behalen op vragen die hogere-ordevaardigheden vereisen dan op vragen die een beroep doen op lagere-orde vaardigheden (De Block & Velema, 1971). Dunham, Yapa en Yu (2015) vonden een sterk negatieve correlatie ($r = -0{,}75$) tussen het percentage hogere-ordevragen in examens en de prestaties die de studenten op die examens behaalden. Overigens gaat Bloom er niet vanuit dat de samenhang tussen de cognitieve complexiteit en de moeilijkheid van examenvragen perfect zal zijn. Zo kan een inzichtvraag bij een deskundige een beroep doen op lagere-ordevaardigheden (bijvoorbeeld het ophalen van direct toegankelijke informatie uit het geheugen) terwijl een leek diezelfde vraag alleen kan oplossen door middel van hogere-ordevaardigheden (bijvoorbeeld analyseren en evalueren). Ook is het incidenteel mogelijk dat een examenvraag op Bloom-niveau 1 (Memoriseren) moeilijker is (een lagere p-waarde heeft) dan een vraag op niveau 6 (Creëren). Evenzo kunnen sommige leerlingen wel punten behalen op de vragen over hogere-ordevaardigheden en niet of nauwelijks op de reproductievragen. Al met al verwachten we geen perfecte maar wel een middelmatige samenhang tussen de moeilijkheidsgraad van examenopgaven en de cognitieve complexiteit volgens Bloom en RTTI.

Anderson en Krathwohl (2001) wijzen erop dat bestudering van de samenhang tussen beheersingsniveau en itemmoeilijkheid hooguit zwak bewijs kan leveren voor de hiërarchische ordening van de categorieën. Ervaren ontwikkelaars kunnen immers vrijwel elk item moeilijk of makkelijk maken ongeacht

de cognitieve complexiteit van de beoogde vaardigheid. Anderson en Krathwohl (2001) breken daarom een lans voor sterker valideringsonderzoek aan de hand van het bestuderen van de samenhang tussen opeenvolgende beheersingsniveaus. Voor elk niveau wordt er dan een schaal geconstrueerd, bijvoorbeeld door de scores op de items die dat niveau meten te sommeren. Als de veronderstelde hiërarchische structuur geldt, zouden niet alle leerlingen die de vragen bij niveau 1 goed maken de vragen bij niveau 2 goed mogen maken, nog minder van hen de vragen bij niveau 3, et cetera. In dat geval zou het patroon van correlaties tussen de schalen een zogenoemde simplex-structuur te zien moeten geven. De samenhang tussen aangrenzende niveaus is dan hoger dan die tussen verder van elkaar afliggende niveaus; in de correlatietabel nemen de correlaties dan af naarmate ze verder van de diagonaal verwijderd zijn. In het onderzoek zijn we nagegaan in hoeverre de correlaties inderdaad de veronderstelde simplex-structuur vertonen.

De p-waarde (moeilijkheid) zegt weinig over cognitieve complexiteit

De uitvoering van het onderzoek

In het Bloom-onderzoek hebben 22 codeurs in totaal 178 examenvragen van vijf examenvakken gecodeerd. Het betrof biologie havo 2015, economie havo 2014, management & organisatie vwo 2016, Engels havo 2016 en aardrijkskunde havo 2016; alle eerste tijdvak.
In het RTTI-onderzoek hebben 22 codeurs in totaal 177 vragen van vijf examenvakken gecodeerd volgens de RTTI-systematiek. De vijf examens waren biologie havo 2014, economie havo 2014, management & organisatie vwo 2014, Engels havo 2014 en aardrijkskunde vwo 2014; alle eerste tijdvak. Alle codeurs waren vakinhoudelijke deskundigen die professioneel bij de constructie van de examens betrokken waren.

De codeurs codeerden de examenvragen eerst onafhankelijk van elkaar op een zelfgekozen locatie. Vervolgens werden de toegekende codes vraag-voor-vraag besproken in een plenair overleg onder leiding van de toetsdeskundige van het betreffende vak. Overeenkomstig de aanwijzingen kregen de codeurs ruimschoots de gelegenheid met elkaar over de juistheid van de toegekende codes van gedachten te wisselen. De panelleider zorgde ervoor dat ieders stem gehoord werd. De discussie gehoord hebbende, kenden de codeurs hun definitieve codes toe, wederom onafhankelijk van elkaar. Voor elke examenvraag zijn er derhalve twee codes beschikbaar: één vóór en één na plenaire discussie. De vergelijking van de codering vóór en na discussie maakt het mogelijk de invloed van de plenaire discussies op de betrouwbaarheid van de codering te kwantificeren.

Resultaten

Betrouwbaarheid van de coderingen

De overeenstemming tussen de codeurs is bepaald met Cohen's multi-rater Kappa

(Cohen, 1960). Deze maat is 1 als de codeurs perfect overeenstemmen en 0 als de overeenstemming niet groter is dan men op basis van toeval mag verwachten. Landis en Koch (1977) geven de volgende vuistregel voor het interpreteren van de hoogte van Kappa: 0,00 tot 0,20 is slight, 0,21 tot 0,40 is fair, 0,41 tot 0,60 is moderate, 0,61 tot 0,80 is substantial en 0,81 tot 1,00 is almost perfect. Tabel 1 toont Cohen's multi rater Kappa vóór en na het plenaire overleg, uitgesplitst naar taxonomie en examen. Omdat de codeurs Engels hun codes na discussie niet onafhankelijk van elkaar gegeven hebben, zijn voor dat vak alleen de gegevens vóór discussie beschikbaar.

Voorafgaand aan het plenair overleg is de betrouwbaarheid bij biologie, economie, Engels en aardrijkskunde laag en bij m&o middelmatig. Dit geldt zowel voor de codering met Bloom als RTTI. Kennelijk is er voor een goede betrouwbaarheid meer nodig dan een training van vier à zes uur in het voortraject.
Na het overleg is de betrouwbaarheid bij drie vakken significant hoger dan ervoor: biologie (zowel Bloom als RTTI), economie (alleen RTTI) en aardrijkskunde (alleen RTTI). De betrouwbaarheid varieert hier van aanzienlijk tot nagenoeg perfect. Daarentegen treedt er bij Engels en m&o geen noemenswaardige betrouwbaarheidswinst op en blijft de betrouwbaarheid laag of middelmatig.

De cognitieve complexiteit van examenvragen met Bloom en RTTI

De verdeling van de toegekende Bloom-codes is per examen weergegeven in tabel 2. Bij de berekening van de percentages is er rekening mee gehouden dat een examenvraag een beroep kan doen op meer vaardigheden tegelijkertijd. Met uitzondering van Engels betreft het de verdeling van de scores na afloop van de plenaire discussie.

Allereerst valt op dat de hogere-ordevaardigheden Evalueren en Creëren niet of nauwelijks in de examens bevraagd worden. Het examen m&o kent veel toepassingsvragen terwijl dit type vragen bij biologie ondervertegenwoordigd zijn. Het examen Engels doet alleen een beroep op Begrijpen en Analyseren. Een mogelijke verklaring veronderstelt dat er bij Engels alleen maar tekstbegrip bevraagd wordt en niet bijvoorbeeld ook het stampen van tweetalige woordenlijstjes.

Tabel 3 toont de verdeling van de toegekende RTTI-codes per examen. Weergegeven zijn percentages na de plenaire discussie.

Tabel 1. Betrouwbaarheid (Kappa) van de codering per examenvak voor Bloom en RTTI

	Biologie	Economie	M&O	Engels	Aardrijkskunde
Bloom					
voor plenair overleg	0,35	0,13	0,43	0,19	0,13
na plenair overleg	0,82*	0,34	0,53	-.-	0,39
RTTI					
voor plenair overleg	0,20	0,07	0,44	0,14	0,22
na plenair overleg	0,70*	0,65*	0,31	0,26	0,77*

* significant verschil tussen betrouwbaarheid vóór en na plenair overleg; -.-: niet beschikbaar

Tabel 2. Verdeling van de toegekende codes volgens Bloom per examenvak (kolompercentages optellend tot 100%)

Bloom-niveaus	Biologie	Economie	M&O	Engels*	Aardrijks-kunde
1 Memoriseren	21	17	11	0	17
2 Begrijpen	56	38	20	76	44
3 Toepassen	6	16	53	0	15
4 Analyseren	17	25	16	24	22
5 Evalueren	0	1	0	0	1
6 Creëren	0	3	0	0	0
Totaal	100	100	100	100	100

* vóór plenaire discussie

Tabel 3. Verdeling van de toegekende codes volgens RTTI per examenvak (kolompercentages optellend tot 100%)

RTTI-niveaus	Biologie	Economie	M&O	Engels*	Aardrijks-kunde
1 R (Reproductie)	8	3	15	0	9
2 T1 (Toepassing in bekende situaties)	42	37	35	51	12
3 T2 (Toepassing in nieuwe situaties)	43	53	38	39	47
4 I (Inzicht)	7	6	12	10	32
Totaal	100	100	100	100	100

* vóór plenaire discussie

Allereerst valt op dat Reproductie en Inzicht over de hele linie zwak vertegenwoordigd zijn. Afhankelijk van het vak, gaat 59% tot 90% van de examenvragen over Toepassing in bekende of nieuwe situaties. Meer dan de andere vakken doet het examen aardrijkskunde een beroep op Inzicht, terwijl dat vak veel minder Toepassing in bekende situaties vereist. Engels is wederom een buitenbeentje, nu omdat vragen naar Reproductie geheel ontbreken (wat wederom logisch lijkt als men bedenkt dat dit examen alleen over tekstbegrip gaat).

Samenhang tussen cognitieve complexiteit en moeilijkheidsgraad

Eerder spraken we de verwachting uit dat examenvragen in de populatie van examenkandidaten gemiddeld moeilijker zijn (d.w.z. een lagere p-waarde bezitten) naarmate het cognitieve niveau zoals gecodeerd met Bloom of RTTI hoger is. De cognitieve complexiteit van een examenvraag is vastgesteld door het gemiddelde te berekenen over de door de codeurs toegekende codes (na afloop van het plenaire overleg). Het gemiddelde Bloom-niveau is hiermee uitgedrukt op een schaal van 1 t/m 6 en het gemiddelde RTTI-niveau op een schaal van 1 t/m 4. Hierbij staat de score 1 voor een lage complexiteit en de hoogste score voor een hoge complexiteit. Op de Bloom-schaal staat 1 voor zuiver Memoriseren en 6 voor zuiver Creëren. Op de RTTI-schaal

staat 1 voor zuiver Reproductie en 4 voor zuiver Inzicht. Tabel 4 toont de correlaties tussen cognitieve complexiteit en moeilijkheidsgraad (p-waarde) voor de codering met Bloom en RTTI.

Van de twaalf correlaties zijn er twee significant op 1%-niveau. De significantie doet zich voor bij biologie zoals gecodeerd met Bloom en bij aardrijkskunde zoals gecodeerd met RTTI. Overeenkomstig onze verwachtingen is de samenhang in beide gevallen negatief: hoe hoger de cognitieve complexiteit, hoe lager de p-waarde (en dus hoe moeilijker de examenvraag). Bij m&o is er met Bloom sprake van een significante samenhang op 5%-niveau, maar hier is de richting positief en dus in strijd met onze verwachtingen. De correlaties zijn in alle tien gevallen echter laag. De conclusie is dat de cognitieve complexiteit van examenvragen niet of nauwelijks met de moeilijkheidsgraad samenhangt.

De hiërarchische ordening van de beheersingsniveaus

Voor de vijf examens die relatief betrouwbaar gecodeerd konden worden zijn we nagegaan in hoeverre de correlaties inderdaad de veronderstelde simplex-structuur vertoonden. De vijf onderzochte examens waren biologie havo 2015 en m&o havo 2016 zoals gecodeerd met Bloom en biologie havo 2014, economie havo 2014 en aardrijkskunde vwo 2014 met RTTI. Zoals eerder uiteengezet is het cognitieve niveau van de opgaven voor Bloom uitgedrukt op een schaal van 1 t/m 6 en voor RTTI op een schaal van 1 t/m 4. Omdat het beperkte aantal items geen indeling in de zes Bloom-categorieën toelaat, zijn de gemiddelde coderingen op de Bloom-schaal ingedikt tot drie categorieën. Van items in het bereik van 1 t/m 2 is de complexiteit als laag beschouwd, van 2 t/m 3 als middelmatig en van 3 t/m 6 als hoog. Per categorie zijn schalen geconstrueerd door de scores van de kandidaten over de desbetreffende items te sommeren. Tabel 5 toont de correlaties tussen de aldus geconstrueerde schalen voor de opeenvolgende niveaus voor de vijf onderzochte examens.

Bij het examen Biologie havo 2015 biedt het patroon van correlaties tussen de schalen voor de opeenvolgende Bloom-niveaus geen enkele ondersteuning voor de aanname dat de opeenvolgende beheersingsniveaus een hiërarchische ordening kennen.

Bij Aardrijkskunde vwo 2014 correleert het hoogste beheersingsniveau I relatief hoog met het direct daaronder liggende niveau T2 (r = 0,55). Zoals verondersteld is deze correlatie duidelijk hoger dan die tussen I en T1 (r = 0,26). In strijd met de veronderstelling van een hiërarchische ordening is dat de overige correlaties weinig van elkaar verschillen (en bovendien erg laag zijn).

Bij Biologie havo 2014 is er strikt genomen sprake van een simplex-structuur aangezien de correlaties tussen aangrenzende niveaus

Tabel 4: Correlaties tussen cognitieve complexiteit en moeilijkheidsgraad (p-waarde)

	Biologie	Economie	M&O	Engels[1]	Aardrijkskunde
Bloom	-0,28***	-0,13^{ns}	0,18**	0,03^{ns}	0,03^{ns}
RTTI	-0,01^{ns}	-0,02^{ns}	-0,07^{ns}	-0,05^{ns}	-0,26***

[1] vóór discussie; ***: p < 0,001; **: p = 0,014; ns = niet significant op 5%-niveau

Tabel 5. Correlaties tussen de schalen voor opeenvolgende niveaus van Bloom en RTTI per examen (N: aantal examenkandidaten)

Biologie havo 2015 gecodeerd met Bloom (N = 19335)

Beheersingsniveau	Laag	Midden
Midden	0,37	
Hoog	0,40	0,32

M&O havo 2016 gecodeerd met Bloom (N = 14188)

Beheersingsniveau	Laag	Midden
Midden	0,42	
Hoog	0,30	0,61

Aardrijkskunde vwo 2014 gecodeerd met RTTI (N = 10695)

Beheersingsniveau	R	T1	T2
T1	0,20		
T2	0,26	0,33	
I	0,21	0,26	0,55

Biologie havo 2014 gecodeerd met RTTI (N = 16806)

Beheersingsniveau	R	T1	T2
T1	0,25		
T2	0,13	0,35	
I	0,08	0,20	0,27

Economie havo 2014 gecodeerd met RTTI (N = 24993)

Beheersingsniveau	R	T1	T2
R	-.-		
T1	-.-		
T2	-.-	0,37	
I	-.-	0,07	0,20

telkens net iets hoger zijn dan die tussen verder van elkaar verwijderde niveaus. Maar ook bij dit examen zijn de verschillen tussen de correlaties klein.

Bij de overige twee examens zien we wel enige evidentie voor een hiërarchische ordening. Bij M&O havo 2016 correleert het hoogste beheersingsniveau aanzienlijk hoger met het middelste niveau (r = 0,61) dan met het laagste niveau (r = 0,30). En bij Economie havo 2014 correleert T1 hoger met het aangrenzende T2 (r = 0,37) dan met het verder verwijderde I (r = 0,07).Daar staat echter tegenover dat de overige correlaties weinig van elkaar verschillen.

Discussie en aanbevelingen

Coderen van de examenvragen blijkt niet eenvoudig

De codes voor de cognitieve complexiteit kenden de codeurs toe op basis van een nauwgezette bestudering van de inhoud van de examenvraag en het bijbehorende antwoordmodel. Zowel bij Bloom als bij RTTI bleken de codeurs vooral moeite te hebben met het verschil tussen toepassing in bekende en nieuwe situaties. Zo kan een examenvraag een beroep doen op zeer complexe denkprocessen als het een toepassing in een volledig nieuwe situatie betreft, terwijl leerlingen diezelfde examenvraag grotendeels vanuit hun geheugen kunnen beantwoorden als de toepassingssituatie uitgebreid in de lessen geïnstrueerd en geoefend is. Om toepassingsvragen eenduidig te kunnen coderen, is het dus noodzakelijk de voorafgaande schoolervaring van de leerlingen te kennen of op zijn minst als bekend te veronderstellen (Bloom e.a., 1956; De Block & Velema, 1971). Voor de codeurs was het vaak onduidelijk in hoeverre de kandidaten in de landelijke populatie de geëxamineerde leerstof geleerd en geoefend hadden. De codeurs is er uitdrukkelijk op gewezen dat zij bij twijfel niet mochten uitgaan van hun eigen lessen en hun eigen kandidaten. Een procedure of een toepassingscontext kan immers voor de ene klas bekend zijn, maar voor de andere klas totaal nieuw. Als de codeurs bijvoorbeeld twijfelden of een toepassing aangeleerd of nieuw was, moesten zij de codering mede baseren op hun kennis van het 'gemiddelde' landelijke leerstofaanbod (methoden!), de intentie van de toetsdeskundigen en de omschrijving van het toetsdoel in de syllabus. Deze richtlijn bleek in de praktijk weinig houvast te bieden.

Nadere vakspecifieke uitwerking

Bloom en RTTI zijn in essentie vakoverstijgende taxonomieën. In ons onderzoek onderschreven alle codeurs de noodzaak van een nadere vakspecifieke uitwerking. Allen hadden behoefte aan een gedegen vakspecifieke handleiding voor het coderen met heldere vakinhoudelijke definities van de categorieën, duidelijke regels voor het toewijzen van codes aan examenvragen en illustratieve voorbeelden van juist en onjuist gecodeerde examenvragen. Het verdient aanbeveling de aanwijzingen voor het coderen verder uit te werken alvorens de betrouwbaarheid met andere codeurs en examens opnieuw te onderzoeken. Daarbij lijkt het raadzaam meer codeurs en examens in het onderzoek te betrekken dan in het huidige onderzoek gebeurd is.

Naast beheersingsniveau ook inhoudelijke complexiteit coderen

Op de moeilijkheid van een examenvraag zijn vele factoren van invloed. Behalve het vereiste beheersingsniveau zoals bepaald met een taxonomie speelt ook de complexiteit van de vakinhoud een belangrijke rol. Een vraag waarbij een leerling een oplossing moet bedenken waarbij slechts inzicht in één inhoudelijk begrip nodig is, zal in de regel eenvoudiger zijn dan een vraag waarbij de leerling bij de oplossing meer begrippen in onderlinge samenhang zal moeten gebruiken. Voor eventueel vervolgonderzoek verdient het aanbeveling om examenvragen zowel te coderen op vereist beheersingsniveau als vakinhoud.

Spreiding over de categorieën van Bloom en RTTI

Het onderzoek geeft inzicht in de spreiding van de examenvragen over de verschillende catego-

rieën van Bloom en RTTI. Dit geeft examenmakers en docenten informatie die waardevol kan zijn voor een toekomstige aanpassing van examenprogramma's, syllabi en de daarvan af te leiden examens. Het kan ook bijdragen aan een goede verdeling van inhouden en vaardigheden over het centrale examen en het schoolexamen.

De hiërarchische ordening van de beheersingsniveaus

Gezien de niet eensluidende en elkaar soms tegensprekende resultaten kan er op de vraag naar de hiërarchische ordening van de beheersingsniveaus geen eenduidig antwoord worden gegeven. De samenhang tussen de schalen voor opeenvolgende beheersingsniveaus zijn vrijwel zonder uitzondering lager dan uit de meta-analyse van Anderson & Krathwohl (2001) naar voren komt. Hierbij past de kanttekening dat de correlaties in onze verkenning gedrukt worden door de lage betrouwbaarheid van sommige schalen als gevolg van het kleine aantal items. Voor vervolgonderzoek verdient het aanbeveling om ervoor te zorgen dat alle beheersingsniveaus met voldoende items gemeten worden.

Hoe nu verder met de terugrapportage van examenresultaten?

Na de afname van het examen geven scholen de scores door aan Cito. Cito verzorgt een terugrapportage met informatie over de beheersing van examenonderdelen en vaardigheden die kandidaten goed en minder goed beheersen. Aanleiding tot het onderhavige onderzoek was de vraag van sommige docenten om deze terugrapportage uit te breiden met een rapportage overeenkomstig RTTI. Uit het onderzoek komt naar voren dat het voor een betrouwbare codering nodig is om na een onafhankelijke codering van examenvragen intensief plenair te overleggen. Het betrouwbaar classificeren van examenvragen kost tijd en geld. Cito zal moeten afwegen of deze extra inspanning genoeg oplevert voor betrouwbare rapportages en of het voor de scholen ook voldoende toevoegt aan wat ze zelf al weten over hun leerlingen.

Literatuur

- Alberts, R. (2017). Waardering per examen 2017. Verkregen op 19 september 2017, van http://www.cito.nl/onderwijs/voortgezet%20onderwijs/centrale_examens/examenverslagen/waardering_per_examen_2017
- Anderson, L.W., & Krathwohl, D.R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (Complete edition). New York: Longman.
- Bloom B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R (1956). Taxonomy of educational objectives: The classification of educational goals, Handbook I: Cognitive domain. New York: David McKay Co Inc.
- Block, A. de, & Velema, E. (Red) (1971). Taxonomie van een aantal in het onderwijs en de vorming gestelde doelen. Een systematische classificatie van expliciet gewenste leerresultaten. I Het cognitieve gebied (Reeks algemene onderwijskunde). Rotterdam: Universitaire Pers.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological
- Measurement, 20, 37- 46.
- Drost, M., & Verra, P. (2014, 2016). Handboek RTTI. Bodegrave: Uitgeverijplus.
- Dunham, B., Yapa, G., & Yu, E. (2015). Calibrating the difficulty of an assessment tool: The Blooming of a statistics examination. Journal of Statistics Education, 23 , 3, 1-33.
- Krathwohl, D.R. (2002). Revising Bloom's taxonomy: An overview. Theory into Practice, 4, 213-218.
- Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 1, 159-174.