# Hoe beoordeel je een evaluatie-instrument?

Het juiste evaluatie-instrument kiezen is niet makkelijk. Er zijn immers heel wat instrumenten op de markt. Breed evalueren houdt namelijk in dat je de evaluatie niet laat afhangen van de resultaten die voortkomen uit één evaluatie-instrument en -moment, maar dat je op geregelde tijdstippen en op verschillende manieren (lees: met verschillende instrumenten) evalueert, om zo een volledig beeld te krijgen van het kennen en kunnen van de leerlingen. De keuze voor een specifiek instrument kadert best binnen een breder taal- en evaluatiebeleid.

## 1. Evaluatiebeleid

Een instrument kiezen en beoordelen is eigenlijk niet de opdracht van een individuele leerkracht maar gebeurt best op schoolniveau, in samenspraak met het ganse team en in overeenstemming met het evaluatiebeleid van de school. Zicht hebben op de taalontwikkeling van leerlingen en op basis van de noden van de leerlingen een taalbeleid uitbouwen veronderstelt ook dat je nadenkt hoe je die taalontwikkeling zult opvolgen (=evaluatiebeleid). Hoe een evaluatiebeleid past binnen een taalbeleid lees je hier. Voor je met de school kiest voor bepaalde evaluatie-instrumenten of volgsystemen, dien je te weten waarom, wat, hoe en wanneer je wilt evalueren en ook wie de evaluatie zal uitvoeren. Het scenario 'Wat is breed evalueren?' kan het team op weg zetten om op elk van deze vragen een antwoord te vinden aangepast aan de realiteit van de school en in functie van de volledige schoolloopbaan van de leerlingen. Pas dan heb je zicht op een volledige evaluatieaanpak: waarom je evalueert, hoe je evalueert, wie je erbij betrekt en hoe je aan de slag gaat met de resultaten en kan je bestaande evaluatie-instrumenten selecteren en aan een kritisch onderzoek onderwerpen. Kies je ervoor om zelf een aantal instrumenten te ontwikkelen dan vind je in dit scenario tips en aandachtspunten. Binnen de Toolkit zijn identificatiefiches van bestaande instrumenten opgenomen, maar hoe je kunt beoordelen of het instrument ook een *geschikt* instrument is, of het in kaart brengt wat jij in kaart *wilt* brengen, leggen we hier uit.



## 2. Ruis

Geen enkel instrument is zaligmakend. Er is geen instrument voorhanden waarvan we zeker kunnen zeggen dat het alles meet wat we willen meten, dat het een betrouwbare score geeft en dat het de competenties van de leerling volledig en waarheidsgetrouw kan weergeven. Wat we wél kunnen is een kritische bril opzetten bij het bekijken en het selecteren van evaluatie-instrumenten, zodat we met de uiteindelijke selectie van een mix aan instrumenten een zo volledig en accuraat mogelijk beeld krijgen van de competenties van de leerlingen.

Toetsen hebben onvermijdelijk tekortkomingen. Zowel bij het ontwikkelen zelf, het afnemen, het scoren en het interpreteren van de score kan er immers een bepaalde vorm van 'ruis' optreden. Ruis

beïnvloedt de waarde, de kwaliteit (zie: punt 3 VRIP-parameters) van de toets. Hieronder overlopen we de vier fasen van het opstellen over het afleggen tot het verbeteren en interpreteren van toetsen en geven daarbij telkens aan welke ruis daarbij kan optreden. In punt 3. VRIP parameters voor een 'goed' evaluatie-instrument, gaan we verder in op de vier parameters die je helpen een waardeoordeel te vellen over een toets.

Als we toetsen opstellen, hebben we een specifiek leerdoel voor ogen. We stellen bepaalde vragen op om na te gaan of de leerlingen het **doel** bereikt hebben of in welke fase van het leerproces ze zich bevinden. Maar stellen we de juiste vragen? Wat verstaan we onder taalvaardigheid, hoe vullen we dat begrip in? Stellen we voldoende vragen om die vaardigheid na te gaan? Zien we er over het hoofd of schort er iets aan de vragen, waardoor het voor de leerling niet helemaal duidelijk is wat er bedoeld wordt. Een voorbeeld kan verduidelijken wat we bedoelen met ruis tijdens de toetsontwikkeling.

Een toetsontwikkelaar ontwikkelt een toets om de leescompetentie van leerlingen uit het vijfde leerjaar in kaart te brengen. De toets omvat echter enkel leeskaarten. De maker van de toets kan dan wel zeggen dat hij de leesvaardigheid wil weten, maar in realiteit geeft de toets enkel de vaardigheid in technisch lezen weer. Dat is een mogelijke vorm van 'ruis' die kan optreden bij het operationaliseren, namelijk het vertalen van een evaluatiedoel in een concreet instrument.

Er kan ook ruis optreden bij het **afleggen van een toets**. Misschien is de leerling niet goed uitgeslapen? Misschien zit hij met zijn hoofd even ergens anders? Misschien kan de leerling het goed mondeling uitleggen, maar wanneer je hem vraagt om het antwoord op te schrijven, lukt het minder goed, etc.



In een volgende stap verbetert de leerkracht de toetsen. Als een leerkracht 20 schrijfopdrachten na elkaar leest, worden ze misschien niet telkens op dezelfde manier **gescoord**, ook al doet een leerkracht zijn uiterste best om dat zo objectief en eerlijk mogelijk te doen en gebruikt hij verbetersleutels en criterialijsten. Een goede toets geeft duidelijke richtlijnen en criteria voor het verbeteren, maar toch kan ook in deze fase ruis optreden. Het voordeel van digitale toetsen is dat ze onmiddellijk gescoord worden. De keerzijde van de medaille is dat, als er enkel een cijfer wordt weergegeven, de leerkracht soms geen of minder zicht heeft op waar de moeilijkheden zich precies situeren.

Tot slot krijgt de leerling een cijfer. Die score wordt vaak veralgemeend tot een resultaat op het rapport. Maar **wat zegt dat cijfer**? Zeven op tien is voor de ene leerling) een hoog cijfer, terwijl dat voor de andere eerder laag is. Bovendien kan je je afvragen wat het cijfer zegt over de soort ondersteuning die een leerling nog nodig heeft om zijn doelen te behalen op het einde van het schooljaar of op het einde van het basisonderwijs? Als taalvaardigheid op het rapport samengevat wordt in een algemeen cijfer, bv. 7/10 (het gemiddelde van 9/10 voor leesvaardigheid, 7/10 voor spreekvaardigheid, 5/10 voor luistervaardigheid en 7/10 voor schrijfvaardigheid), kan een leerling onmogelijk achterhalen dat hij een tandje moet bijsteken voor luisteren. Bovendien is het ook niet altijd duidelijk wat die cijfers voor de verschillende vaardigheden apart betekenen. Waarom rapporteren we dan zo algemeen? Wat leren die cijfers ons echt? Het schoolteam kan daarom best nadenken over andere manieren van resultaten communiceren naar ouders en leerlingen toe. Inspiratie over hoe dat kan, vind je in de scenario's 'Hoe ga je aan de slag met de resultaten van breed evalueren?' en 'Breed evalueren leidt naar breed rapporteren'.

## 3. VRIP-parameters voor een 'goed' evaluatie-instrument

In de identificatiefiches in de Toolkit worden bij elk instrument vier parameters besproken als criteria voor een goed evaluatie-instrument. Die parameters worden door experts wel eens 'VRIP-parameters' genoemd. VRIP staat voor 'Validitity' (validiteit of wat meet het instrument), 'Reliability' (betrouwbaarheid),' Impact' (in de praktijk) en 'Practicality' (praktische haalbaarheid). De informatie die de parameters je geven, kan je op weg helpen om een instrument te beoordelen.

### 3.1 Validiteit: meet het instrument de competenties die het zegt te meten?

Een instrument is valide als het effectief de vaardigheid (of de kennis, de attitude) meet waarover men uitspraken wil doen. Het is dus belangrijk na te gaan wat het **doel** is van de toets. **Wat meet het instrument?** Het gebeurt bijvoorbeeld dat een instrument dat de leesvaardigheid van leerlingen wil nagaan, vraagt naar de betekenis van woorden. Zo een instrument meet wel de woordenschatkennis van leerlingen, maar zegt niets over hoe goed een leerling kan begrijpend lezen. Woordenschattoetsen die louter vragen naar een omschrijving van de betekenis van een woord kunnen bijvoorbeeld aangevuld worden of zelfs vervangen worden door opdrachten waarbij de leerlingen bepaalde woorden moeten gebruiken en door dat gebruik hun kennis aantonen. Op die manier krijg je een vollediger beeld van de taalvaardigheid van de leerling in plaats van te focussen op slechts één aspect ervan.

Door na te denken over wat een instrument meet en dat naast de doelstellingen van het instrument te leggen, vorm je je een beeld van de **validiteit** van het instrument: meet de toets effectief wat hij beoogt of pretendeert te meten? Beantwoordt het instrument aan het doel waarvoor het wordt gebruikt? De evaluatievorm moet de vaardigheid/competentie in al zijn aspecten bevatten. Alleen zo krijg je als leraar een werkelijkheidsgetrouw beeld van de vaardigheid/competentie die je wilt meten.



Dat betekent ook dat je als leerkracht goed moet weten wat je met je evaluatie in kaart wilt brengen. Bij het nadenken over validiteit kan je je ook afvragen hoe goed 'schoolse taalvaardigheid' zich eigenlijk laat meten. Taalvaardigheid meten is van een andere orde dan nagaan of een leerling een kwadraat kan berekenen. Taalvaardigheid is een complex concept en daardoor niet altijd eenduidig en/of eenvoudig te meten. Je kunt taalvaardigheid opvatten als een verzameling van competenties. Leerkrachten (of leerkrachtenteams) kunnen best vooraf vastleggen hoe zij die schoolse taalvaardigheid zien en wat zij eronder verstaan, om vervolgens een evaluatiestrategie op te stellen, waarbij ze gericht opteren voor evaluatie-instrumenten die dat precies nagaan. In een ideale situatie krijgt dit proces een plaats binnen het taal- en evaluatiebeleid van de school.

In de identificatiefiches van de evaluatie-instrumenten uit de Toolkit lees je informatie over:

1) de doelen waarop ze zich baseerden voor het opstellen van het evaluatie-instrument (bijvoorbeeld eindtermen, ERK[1], etc.). Aan de hand daarvan kan je aftoetsen of het instrument representatief is voor de leerstof.
2) de mate waarin het instrument aansluit bij de doelstellingen die vooropgesteld zijn (wat wil je evalueren versus hoe doe je dat?)

[1] Het ERK is het Europees Referentiekader voor Moderne Vreemde Talen. Dit kader werd opgesteld door de Raad van Europa. Het referentiekader geeft een uitgebreide en precieze omschrijving van het niveau waarop iemand zijn talen beheerst. Op de website van de Nederlandse Taalunie vind je een Nederlandse vertaling (www.taaluniversum.org)

### 3.2 Betrouwbaarheid: hoe betrouwbaar is de uitkomst van mijn evaluatie?

Betrouwbaarheid houdt in dat de uitkomst van een evaluatie niet door omgevingsfactoren of subjectiviteit wordt beïnvloed. Een betrouwbaar instrument zal bij herhaalde afname (min of meer) dezelfde uitkomsten opleveren.

Hoe weet je nu of een instrument betrouwbaar is? De factor die een belangrijke rol speelt bij het bepalen van de betrouwbaarheid van een evaluatie-instrument is de **afname**. Daarom geven sommige instrumenten in een handleiding duidelijk weer hoe de afname dient te gebeuren:

- wat je tegen de leerlingen mag zeggen vooraf,
- wat ze mogen gebruiken,
- hoe je de instructie moet geven,
- hoe het klaslokaal best ingericht wordt, etc.

Duidelijke en strikte richtlijnen voor de afname van een toets of het gebruik van een ander evaluatie-instrument verhogen de betrouwbaarheid, omdat ze ervoor zorgen dat verschillende beoordelaars tot eenzelfde resultaat komen. Het mag met andere woorden niet uitmaken welke leraar op welk moment de evaluatie uitvoert; het resultaat zou altijd min of meer hetzelfde moeten zijn.

Een andere manier om naar de betrouwbaarheid te kijken, is door te letten op de gebruikte evaluatie**criteria**. Zijn ze vooraf bepaald? Zijn ze transparant? Is er een duidelijke richtlijn voor de correctie van het product en het bepalen van een score? Een positief antwoord op voorgaande vragen draagt bij aan de betrouwbaarheid van een evaluatie-instrument.



Betrouwbaarheid heeft vele facetten. Naast betrouwbaarheid van de toetsafname en de scoring speelt ook de **kwaliteit van de toets** een rol. Onduidelijke opgaven, te moeilijke of te gemakkelijke vragen, een evenwichtige spreiding van inhouden, aandacht voor het aantal opgaven etc. bepalen de kwaliteit en daarmee ook de betrouwbaarheid van een instrument.

Auteurs van instrumenten kiezen er soms voor om bepaalde aspecten van validiteit of betrouwbaarheid sterker te belichten dan andere. Sommige handleidingen maken ook melding van een statistische waarde die de mate van betrouwbaarheid van het instrument weergeeft. In de identificatiefiches van de evaluatie-instrumenten van de Toolkit is steeds uitgegaan van de informatie die de auteurs zelf gaven. Die informatie is misschien niet altijd volledig, vandaar dat je als leerkracht ook alert dient te zijn voor wat niét in de handleiding vermeld staat.

### 3.3 Hoe wordt het instrument in de praktijk ingezet?

Sommige instrumenten helpen leerkrachten in hun handleiding een eind op weg met de interpretatie van de resultaten. Wat kunnen leerkrachten na de testafname doen? Hoe kunnen leerlingen, leerkrachten en het schoolteam aan de slag met het instrument? Hoe kunnen de evaluatieresultaten de lespraktijk verbeteren. Hoe kunnen de resultaten aangewend worden om de competenties van zwakkere leerlingen te vergroten? Wanneer hier concrete aanwijzingen voor gegeven worden, helpt je dat als leerkracht al een eind vooruit.

Als je een evaluatie-instrument wilt inzetten om de prestaties van je **leerlingen** te **vergelijken** met die van andere leerlingen (een norm- of referentiegroep) dien je bij de beoordeling en selectie van een instrument ook rekening te houden met het al dan niet beschikbaar zijn van zo een norm en met

het feit of die norm uit Vlaanderen komt of niet. Sommige instrumenten die ontwikkeld zijn in Nederland geven enkel normen voor Nederland aan. Hun toetsen zijn dan niet alleen gebaseerd op wat leerlingen in Nederland moeten kennen en kunnen (bijvoorbeeld toetsen leesvaardigheid waarvan de teksten aansluiten bij de Nederlandse kerndoelen), maar de resultaten kunnen dan ook enkel vergeleken worden met wat leerlingen uit Nederland presteren. De resultaten van jouw klas of school kunnen aan de hand van een Nederlandse norm niet vergeleken worden met een Vlaams gemiddelde.

Wat doe je als school met een grote instroom van **anderstalige nieuwkomers**? Moet je die leerlingen vergelijken met het 'regulier' Vlaamse gemiddelde? Loop je dan niet het risico dat je de vooruitgang van je leerlingen niet ziet? Een bijkomend criterium kan zijn dat het instrument ook in aangepaste normen voor specifieke **doelgroepen** voorziet die beter geschikt zijn om de evaluatie van je leerlingen mee te volgen. LVS VCLB Lezen, bijvoorbeeld gebruikt aparte normen voor anderstalige nieuwkomers. In dit scenario leggen we uit hoe je de taalontwikkeling van anderstalige nieuwkomers kan opvolgen.

Bij het kiezen van een goed evaluatie-instrument reflecteert degene die de keuze maakt ook over de **gevolgen** van de evaluatie.

Sommige evaluatie-instrumenten geven, op basis van de resultaten, suggesties en adviezen om de lespraktijk te verbeteren. Andere instrumenten zijn eerder op doorverwijzing en remediëring gericht. Nog andere op een combinatie van beide.

### 3.4 Praktische haalbaarheid?



Tot slot is er nog het criterium van de praktische haalbaarheid. Sommige instrumenten zijn misschien heel valide, maar daardoor misschien minder haalbaar in de praktijk, omdat ze teveel **tijd** in beslag nemen om ze af te nemen en/of te scoren. Evaluatie-instrumenten die veel vragen stellen aan de leerling zullen wellicht veel aspecten van taalvaardigheid in kaart kunnen brengen, maar ze kosten veel toetstijd. De tijd die leerlingen aan een toets spenderen moet realistisch zijn, niet alleen omdat leerlingen vaak een korte aandacht spanne hebben, maar ook omdat het extra inspanningen van de leerkracht vraagt. MILOS, bijvoorbeeld, zal voor leerlingen minder een probleem vormen op vlak van aandacht, omdat het uit verschillende deeltaken bestaat die uitdagend zijn voor de leerling en omdat er voldoende tijd is tussen de taken. Voor de leerkracht vraagt het echter een behoorlijke tijdsinvestering. Het schoolteam of de leerkracht zal de **kosten** (tijdsinvestering) en de **baten** (opgeleverde informatie) **afwegen**. Een 'goed' instrument vindt vaak een evenwicht tussen validiteit en praktische haalbaarheid.

## 4. Beoordelen en afwegen

De vier genoemde criteria voor een goed evaluatie-instrument staan vaak op gespannen voet met elkaar. Er zijn niet veel instrumenten die op alle vier de paramaters goed scoren. Een leerkracht zal dus een weloverwogen keuze moeten maken en alle pro's en contra's tegen elkaar moeten afwegen.

June Eyckmans, professor aan de vakgroep Taalkunde van de UGent geeft het volgende mee: Leerkrachten zijn vaak zelf het best geplaatst om een toets te beoordelen. Het is in eerste instantie de leerkracht die kan nagaan of de inspanning en de tijd die een toetsafname heeft gekost in een goede verhouding staat tot de hoeveelheid en de kwaliteit van de informatie die met de toets werd

ingewonnen. Als geen ander is de leerkracht immers in staat de ervaringen van de leerlingen tijdens de toetsafname in te schatten en te beoordelen of de toets de beoogde leerdoelen evalueert (validiteit). Daarnaast kan hij uitmaken aan de hand van de eigen verbeterervaring of het toetsformaat een objectieve beoordeling mogelijk maakte (betrouwbaarheid). Tot slot is hij vaak het eerste en enige aanspreekpunt van de leerlingen wat maakt dat hij informatie kan vergaren over de transparantie van de toetsinstructies.

Ook het schoolteam speelt een rol bij het beoordelen van evaluatie-instrumenten en hier komen we terug bij evaluatiebeleid. Binnen het evaluatiebeleid van je school kan je afspraken maken over evaluatiecriteria. Naast de VRIP-parameters kunnen ook andere criteria een rol spelen. Denk bijvoorbeeld aan aanschouwelijkheid; hoe zien onze toetsen eruit, spreken ze aan, hoe zijn ze vormelijk opgebouwd, enzovoort. Maar ook de rol en de ervaring met specifieke toetsen kan meespelen, als een school al vijftien jaar hetzelfde toetst, bijvoorbeeld spelling, en uit de resultaten blijkt dat de school wat spelling betreft goede resultaten neerzet kan er nagedacht worden over de inzet van andere evaluatie-instrumenten of toetsen die andere aspecten in kaart brengen. Het is aan het schoolteam en het beleid om samen te bepalen hoe het evaluatiebeleid eruit ziet, wat en waarom er geëvalueerd wordt en wie in welke fase betrokken wordt.

## 5. Doel

Niet elk instrument is geschikt voor elk doel. Bart Deygers (2013) maakt de vergelijking met het openen van een fles met een aansteker. De aansteker symboliseert hier het instrument en de fles de te evalueren competentie of (deel)vaardigheid. Wanneer je taalvaardigheid (de fles) wilt evalueren aan de hand van een spellingstoets (de aansteker) zal je wel een resultaat verkrijgen, je zult de leerling een cijfer kunnen geven, maar dan geef je een cijfer op de spellingvaardigheid van de leerling en niet op zijn taalvaardigheid. Je kunt een flesje openen met een aansteker, maar dat lukt natuurlijk minder goed dan met een flessenopener. Zo is het ook met toetsen. Toetsen die gebruikt worden om het doel na te gaan waarvoor ze ontwikkeld werden, doen hun werk beter. Bart Deygers adviseert daarom om elke toets te gebruiken waarvoor hij bedoeld is.




**Bron: Studieavond 'Taalinstaptoetsen in het hoger onderwijs', 25 april 2013, Bart Deygers.**

In de identificatiefiches van de Toolkit staat telkens aangegeven welke vaardigheden en eindtermen geëvalueerd kunnen worden met het instrument. De fiches helpen je al een eind op weg bij het selecteren van geschikt evaluatie-instrumenten.

## 6. Tot slot

Kiezen voor geschikte evaluatie-instrumenten is niet gemakkelijk. Daarom vatten we het voorgaande nog eens samen in een aantal tips:

1. Wees je bewust van de **ruis** die kan optreden. Houd rekening met die ruis bij het interpreteren van een toetsscore en laat je beslissingen bijgevolg niet afhangen van één toetsresultaat.

2. Wees kritisch ten opzichte van evaluatie-instrumenten. Je kunt een instrument beoordelen aan de hand van deze checklist bij de **VRIP**-criteria.

| | |
|---|---|
| **Validiteit** | Wat willen we meten? : Hoe benaderen wij taalvaardigheid?<br>- Welke (deel)competenties vallen onder onze definitie van taalvaardigheid?<br><br>Welke (deel)competenties meet het evaluatie-instrument?<br>- Welke (deel)competenties zegt het te meten?<br>- Welke (deel)competenties worden effectief gemeten?<br>o Wat zegt de handleiding?<br>o Wat kan ik afleiden uit de opgaven? |
| **Betrouwbaarheid** | Zijn er duidelijke instructies voor de afname van het instrument?<br>Zijn er duidelijke criteria voor de evaluatie van de prestatie van de leerlingen?<br>Zijn er duidelijke richtlijnen voor de correctie en het bepalen van een score?<br>Zijn er voldoende vragen voorzien? |
| **Impact** | Zijn er duidelijke richtlijnen voor de interpretatie van de score?<br>Hoe kan je met de resultaten aan de slag?<br>- Laat het instrument toe om leerlingen te vergelijken?<br>- Geeft het instrument aanwijzingen voor het aanpassen van de lespraktijk of richt het zich enkel op diagnose en remediëring? |
| **Praktische haalbaarheid** | Hoeveel tijd vraagt de afname van het instrument?<br>Is onze infrastructuur aangepast aan de afname van dit instrument?<br>Welke informatie levert dit instrument ons op?<br>Zijn afnametijd (kost) en uitkomst (winst) in verhouding? (efficiëntie) |

3. Gebruik evaluatie-instrumenten enkel voor het **doel** waarvoor ze ontwikkeld zijn.

Wanneer je besluit om eigen evaluatie-instrumenten te ontwikkelen, vind je hier meer informatie.

## 7. Bronnen

Deygers, B. Studieavond 'Taalinstaptoetsen in het hoger onderwijs', 25 april 2013.

Van Avermaet, P., Mondt, K., & Pulinx, R. (2011). *Bruikbaarheidsstudie Screeningsinstrument Nederlands Aanvang Secundair Onderwijs. Eindrapport*. https://biblio.ugent.be/publication/4270360/file/4270365.pdf

Van den Branden, K. (2010). *Handboek taalbeleid basisonderwijs*. Leuven: Acco.